aws

用户指南

AWS Clean Rooms

AWS Clean Rooms: 用户指南

Table of Contents

什么是 AWS Clean Rooms？

AWS Clean Rooms 帮助您和您的合作伙伴对您的集体数据集进行分析和协作，以获得新的见解，而无需互相透露基础数据。 AWS Clean Rooms 是一个安全的协作工作空间，您可以在几分钟内创建自己的干净室，只需几个步骤即可分析您的集体数据集。您可以选择要与之合作的合作伙伴，选择他们的数据集，然后为这些合作伙伴配置隐私增强控件。

借助 AWS Clean Rooms，您可以与成千上万家已经在使用的公司进行协作 AWS。协作不需要将数据移出其他云服务提供商 AWS 或将其加载到其他云服务提供商。当您运行查询或作业时，会从该数据的原始位置 AWS Clean Rooms 读取数据，并应用内置的分析规则来帮助您保持对这些数据的控制。

AWS Clean Rooms 提供您可以配置的内置数据访问控制和审计支持控件。这些控制包括：

- 用于限制 SQL 查询和提供输出约束的@@ 分析规则。
- 加密计算，Clean Rooms用于保持数据加密，即使在处理查询时也是如此，以遵守严格的数据处理政策。
- 分析日志，用于查看查询和任务 AWS Clean Rooms 并帮助支持审计。
- 差异隐私，可防止用户识别尝试。 AWS Clean Rooms 差异隐私是一项完全托管的功能，可通过数学支持的技术和直观的控件来保护用户的隐私，只需几个步骤即可应用这些技术和直观的控件。
- AWS Clean Rooms ML 允许双方识别其数据中的相似用户，而无需彼此共享数据。第一方通过其训练数据创建并配置一个相似模型。然后，会将种子数据引入到协作中，以便创建与训练数据类似的相似细分。

以下视频解释了更多相关信息 AWS Clean Rooms。

AWS Clean Rooms

你是首次 AWS Clean Rooms 使用吗？

如果您是首次使用 AWS Clean Rooms，我们建议您先阅读以下章节：

- 如何 AWS Clean Rooms 运作
- 正在访问 AWS Clean Rooms
- 设置 AWS Clean Rooms
- AWS Clean Rooms 词汇表

如何 AWS Clean Rooms 运作

在中 AWS Clean Rooms，您可以创建协作并添加要邀请的人，或者创建成员资格以加入您已被邀请参加的协作。 AWS 账户 然后，您可以链接您的使用案例所需的数据资源：针对事件数据的配置表、针对机器学习建模的配置模型，或针对实体解析的 ID 命名空间。您可以选择创建或批准分析模板，以便事先就协作中要允许的确切查询和任务达成一致。最后，您可以通过在配置的表上运行 SQL 查询或 PySpark 作业、在 ID 映射表中执行实体解析或使用 ML 建模生成相似的受众细分来分析联合数据。

下图显示了 AWS Clean Rooms 工作原理。

相关服务

AWS 服务

以下内容与 AWS 服务 以下内容有关 AWS Clean Rooms：

- Amazon Athena

  协作成员可以将他们 AWS Clean Rooms 作为 AWS Glue Data Catalog 视图带入的数据存储在 Amazon Athena 中。有关更多信息，请参阅以下主题：

  有关更多信息，请参阅以下主题：

  在中为查询准备数据表 AWS Clean Rooms

  创建已配置表-Amazon Athena 数据源

  什么是 Amazon Athena？ 在亚马逊 Athena 用户指南中
- CloudFormation

  在中创建以下资源 CloudFormation：协作、已配置的表、配置的表关联和成员资格

  有关更多信息，请参阅 使用创建 AWS Clean Rooms 资源 AWS CloudFormation。
- AWS CloudTrail

  AWS Clean Rooms 与 CloudTrail 日志配合使用可增强对 AWS 服务 活动的分析。

  有关更多信息，请参阅 使用记录 AWS Clean Rooms API 调用 AWS CloudTrail。

- AWS Entity Resolution 数据匹配服务

  AWS Clean Rooms 与一起使用 AWS Entity Resolution 数据匹配服务 可执行实体解析。

  有关更多信息，请参阅 AWS Entity Resolution 数据匹配服务 in AWS Clean Rooms。

- AWS Glue

  协作成员可以根据自己在 Amazon S3 中的数据创建 AWS Glue 表以供在中使用 AWS Clean Rooms。

  有关更多信息，请参阅以下主题：

  在中为查询准备数据表 AWS Clean Rooms

  《AWS Glue 开发人员指南》中的 What is AWS Glue?

- Amazon Simple Storage Service（Amazon S3）

  协作成员可以将他们带入的 Amazon S3 AWS Clean Rooms 中的数据存储。

  有关更多信息，请参阅以下主题：

  在中为查询准备数据表 AWS Clean Rooms

  创建已配置的表-Amazon S3 数据源

  《Amazon Simple Storage Service 用户指南》中的什么是 Amazon S3？

- AWS Secrets Manager

  协作成员可以创建密钥来访问和读取 Snowflake 中存储的数据。

  有关更多信息，请参阅以下主题：

  创建服务角色以从 Snowflake 读取数据

  在中为查询准备数据表 AWS Clean Rooms

  AWS Secrets Manager 用户指南 中的什么是 AWS Secrets Manager？

第三方服务

以下第三方服务与以下内容有关 AWS Clean Rooms：

- Snowflake

  协作成员可以将他们带入 Snowflake 仓库中的数据存储 AWS Clean Rooms 在 Snowflake 仓库中。

  有关更多信息，请参阅以下主题：

  在中为查询准备数据表 AWS Clean Rooms

  创建已配置的表 — Snowflake 数据源

正在访问 AWS Clean Rooms

您可以 AWS Clean Rooms 通过以下选项进行访问：

- 直接通过 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
- 通过 AWS Clean Rooms API 以编程方式进行。有关更多信息，请参阅 AWS Clean Rooms API 参考。

的定价 AWS Clean Rooms

有关定价信息，请参阅 AWS Clean Rooms 定价。

> **Note**
>
> 对于关联存储在 Snowflake 中的数据的协作成员，每次运行使用存储在这些位置的数据的查询时，他们各自的数据仓库提供商或云提供商都将向您收取数据流出和计算费用。

的账单 AWS Clean Rooms

AWS Clean Rooms 使协作创建者能够指定哪个成员为协作中的查询或作业计算费用付费。

大多数情况下，可以查询的成员和为查询计算费用付费的成员是相同的人。但是，如果可以查询的成员和为查询计算费用付费的成员不同，则当可以查询的成员针对自己的成员身份资源运行查询时，将按支付查询计算费用的成员的成员身份资源计费。

支付查询计算费用的成员在其 CloudTrail 事件历史记录中看不到任何正在运行的查询的事件，因为付款人既不是运行查询的人，也不是运行查询所针对的资源的所有者。但是，对于可以在协作中运行查询的成员运行的所有查询，付款人确实会看到其会员资源上产生的费用。

有关如何创建协作和配置支付查询计算费用的成员的更多信息，请参阅创建协作。

中的分析规则 AWS Clean Rooms

作为启用表格 AWS Clean Rooms 用于协作分析的一部分，协作成员必须配置分析规则。

分析规则是每个数据所有者在配置表上设置的隐私增强控制。分析规则决定了如何分析配置表。

分析规则是对配置表（账户级资源）的账户级控制，并且在关联了配置表的任何协作中强制执行。如果未配置分析规则，则可以将配置表关联到协作，但无法对其进行查询。查询只能引用具有相同分析规则类型的配置表。

要配置分析规则，首先要选择分析类型，然后指定分析规则。在这两个步骤中，您都应考虑要启用的使用案例以及如何保护底层数据。

AWS Clean Rooms 对查询中引用的所有已配置表强制执行更严格的控制。

以下示例演示了限制性控制。

Example限制性控制：输出约束

- 协作者 A 对标识符列的输出限制为 100。
- 协作者 B 对标识符列的输出限制为 150。

  引用两个配置表的聚合查询要求输出行中至少有 150 个不同的标识符值，才能在查询输出中显示。查询输出并没有显示由于输出约束而删除了结果。

Example限制性控制：分析模板未获得批准

- 协作者 A 允许在其自定义分析规则中使用带有引用协作者 A 和协作者 B 配置表的查询的分析模板。
- 协作者 B 不允许使用分析模板。

  由于协作者 B 不允许使用分析模板，因此可以查询的成员无法运行该分析模板。

分析规则类型

有三种类型的分析规则：聚合、列表和自定义。下表对这些分析规则类型进行了比较。每种类型具有单独的部分，以描述如何指定分析规则。

> **Note**
>
> 有一种名为 ID 映射表分析规则的分析规则。但是，此分析规则由管理 AWS Clean Rooms 且无法修改。有关更多信息，请参阅 ID 映射表分析规则。

以下部分描述了每种分析规则类型支持的使用案例和控制。

支持的使用案例

下表显示了每种分析规则类型支持的使用案例的比较总结。

应用场景	聚合	列表	自定义
支持的分析	使用 COUNT、SUM 和 AVG 函数按可选维度聚合统计数据的查询	输出多个表之间重叠部分的行级列表的查询	经过分析模板或分析创建者审核并允许的任何自定义分析
常见使用案例	细分分析、衡量、归因	扩充、细分构建	首次接触归因、增量分析、受众发现
SQL 构造	• JOIN 语句：INNER JOIN • 聚合函数：DISTCOUNT/COUNT DISTINCT, SUM/SUM INCT 和 AVG • 标量函数：有限子集	• JOIN 语句：INNER JOIN • 标量函数：无	通过 SELECT 命令可使用的大多数 SQL 函数和 SQL 构造
子查询和公用表表达式 () CTEs	否	否	是
分析模板	否	否	是

支持的控制

下表显示了每种分析规则类型如何保护底层数据的比较总结。

控件	聚合	列表	自定义
控制机制	控制如何在查询中使用表中的数据 （例如，允许对列 hashed_email 进行 COUNT 和 SUM。）	控制如何在查询中使用表中的数据 （例如，仅允许使用 hashed_email 列进行联接。）	控制允许在表上运行哪些查询 （例如，仅允许在分析模板“自定义查询 1”中定义的查询。）
内置隐私增强技术	• 给匹配项设盲 • 需要聚合 • 最小聚合阈值 >= • 2 预定义的查询结构	• 给匹配项设盲 • 需要重叠 • 预定义的查询结构 • 允许的其他分析	• 差别隐私 • 不允许的输出列
在运行查询之前对其进行审核	否	否	是，正在使用分析模板

有关中提供的分析规则的更多信息 AWS Clean Rooms，请参阅以下主题。

- 聚合分析规则
- 列表分析规则
- 中的自定义分析规则 AWS Clean Rooms

聚合分析规则

在中 AWS Clean Rooms，聚合分析规则使用可选维度的 COUNT、SUM、 and/or AVG 函数生成聚合统计数据。将聚合分析规则添加到配置表后，可以查询的成员就能在配置表上运行查询。

聚合分析规则支持活动规划、媒体覆盖率、频率测量和归因等使用案例。

支持的查询结构和语法在 聚合查询结构和语法 中定义。

聚合分析规则 — 查询控制 中定义的分析规则的参数包括查询控制和查询结果控制。其查询控制包括要求一个配置表至少联接到一个可直接或临时查询的成员所拥有的配置表。此要求可使您确保在您的表及其他人的表的交叉点 (INNER JOIN) 上运行查询。

聚合查询结构和语法

对具有聚合分析规则的表的查询必须遵循以下语法。

```
--select_aggregate_function_expression
SELECT
aggregation_function(column_name) [[AS] column_alias ] [, ...]

 --select_grouping_column_expression
  [, {column_name|scalar_function(arguments)} [[AS] column_alias ]][, ...]

--table_expression
FROM table_name [[AS] table_alias ]
  [[INNER] JOIN table_name [[AS] table_alias] ON join_condition] [...]

--where_expression
[WHERE where_condition]

--group_by_expression
[GROUP BY {column_name|scalar_function(arguments)}, ...]]

--having_expression
[HAVING having_condition]

--order_by_expression
[ORDER BY {column_name|scalar_function(arguments)} [{ASC|DESC}]] [,...]]
```

下表解释前面语法中列出的每个表达式。

Expression	定义	示例
select_aggregate_function_expression	包含以下表达式的逗号分隔列表： • select_aggregation_function_expression	SELECT SUM(PRICE), user_segment

Expression	定义	示例
	• `select_aggregate_expression` **Note** `select_aggregate_expression` 中必须至少有一个 `select_aggregation_function_expression` 。	

`select_aggregation_function_expression`	应用于一个或多个列的一个或多个支持的聚合函数。只允许将列作为聚合函数的参数。
	
	Note
	`select_aggregate_expression` 中必须至少有一个 `select_aggregation_function_expression` 。
		`AVG(PRICE)`
		
		`COUNT(DISTINCT user_id)`

Expression	定义	示例

select_grouping_column_expression	可以包含任何使用以下元素的表达式的表达式：
	• 表列名称
	• 支持的标量函数
	• 字符串文本
	• 数值文本
	
	Note
	`select_aggregate_expression` 可以带或不带 AS 参数对列设置别名。有关更多信息，请参阅 AWS Clean Rooms SQL 参考。
		`TRUNC(timestampColumn)`
		
		`UPPER(campaignName)`

table_expression	使用 `join_condition` 连接联接条件表达式的表或表的联接。 `join_condition` 返回布尔值。 `table_expression` 支持： • 特定的 JOIN 类型 (INNER JOIN) • `join_condition` 中的相等比较条件 (=) • 逻辑运算符（AND、OR）。	`FROM consumer_table INNER JOIN provider_table ON consumer_table.identifier1 = provider_table.identifier1 AND consumer_table.identifier2 = provider_table.identifier2`

Expression	定义	示例
where_expression	返回布尔值的条件表达式。它可能包括以下内容： • 表列名称 • 支持的标量函数 • 数学运算符 • 字符串文本 • 数值文本 支持的比较条件是 (=, >, <, <=, >=, <>, !=, NOT, IN, NOT IN, LIKE, IS NULL, IS NOT NULL)。 支持的逻辑运算符是 (AND, OR)。 where_expression 是可选项。	WHERE where_condition WHERE price > 100 WHERE TRUNC(timestampColumn) = '1/1/2022' WHERE timestampColumn = timestampColumn2 - 14
group_by_expression	满足 select_grouping_column_expression 要求的表达式的逗号分隔列表。	GROUP BY TRUNC(timestampColumn), UPPER(campaignName), segment

Expression	定义	示例
having_expression	返回布尔值的条件表达式。它们具有应用于单列（例如 SUM(price) ）的支持聚合函数，并与数值文字进行比较。 支持的比较条件是 (=, >, <, <=, >=, <>, !=)。 支持的逻辑运算符是 (AND, OR)。 having_expression 是可选项。	HAVING SUM(SALES) > 500
order_by_expression	以逗号分隔的表达式列表，与前面定义的 select_aggregate_expression 中定义的要求一致。 order_by_expression 是可选项。 **Note** order_by_expression 允许 ASC 和 DESC 参数。有关更多信息，请参阅 AWS Clean Rooms SQL 参考中的 ASC DESC 参数。	ORDER BY SUM(SALES), UPPER(campaignName)

关于聚合查询的结构和语法，请注意以下几点：

- 不支持除 SELECT 之外的 SQL 命令。

- 不支持子查询和通用表格表达式（例如 WITH）。
- 不支持组合多个查询的运算符（例如 UNION）。
- TOP、LIMIT 和 OFFSET 参数不受支持。

聚合分析规则 — 查询控制

使用聚合查询控制，您可以控制如何使用表中的列来查询表。例如，您可以控制哪一列用于联接，哪一列可以计数，或者 WHERE 语句中可以使用哪一列。

下面几节解释每种控制。

主题

聚合控制

通过使用聚合控制，您可以定义允许哪些聚合函数以及必须将其应用于哪些列。聚合函数可以在 SELECT、HAVING 和 ORDER BY 表达式中使用。

控件	定义	用法
`aggregateColumns`	允许在聚合函数中使用的已配置表列的列。	`aggregateColumns` 可以在 SELECT、HAVING 和 ORDER BY 表达式中的聚合函数中使用。 有些 `aggregateColumns` 也可以归类为 `joinColumn`（稍后定义）。 给定的 `aggregateColumn` 也不能归类为 `dimensionColumn`（稍后定义）。

控件	定义	用法
function	允许在 aggregateColumns 上使用的 COUNT、SUM 和 AVG 函数。	function 可以应用于与之关联的 aggregateColumns 。

联接控制

JOIN 子句用于根据两个或多个表中的相关列合并两个或多个表中的行。

您可以使用联接控件来控制如何将您的表连接到中的其他表table_expression。 AWS Clean Rooms 仅支持INNERJOIN。 INNERJOIN语句只能使用在分析规则joinColumn中明确归类为 a 的列，但要遵守您定义的控件。

INNER JOIN 必须对您的已配置表中的 joinColumn 和协作中另一个已配置表中的 joinColumn 进行操作。您可以决定表中的哪些列可以用作 joinColumn。

ON 子句中的每个匹配条件都要求在两列之间使用相等比较条件 (=)。

ON 子句中的多个匹配条件可以是：

- 使用 AND 逻辑运算符组合
- 使用 OR 逻辑运算符分隔

> **Note**
>
> 所有 JOIN 匹配条件都必须与 JOIN 两侧各一条记录匹配。所有由 OR 或 AND 逻辑运算符连接的条件也必须遵守此要求。

以下是使用 AND 逻辑运算符的查询示例。

```
SELECT some_col, other_col
FROM table1
    JOIN table2
    ON table1.id = table2.id AND table1.name = table2.name
```

以下是使用 OR 逻辑运算符的查询示例。

```
SELECT some_col, other_col
FROM table1
    JOIN table2
    ON table1.id = table2.id OR table1.name = table2.name
```

控件	定义	用法
joinColumns	您希望允许可以查询的成员在 INNER JOIN 语句中使用的列（如果有）。	特定的 joinColumn 也可以归类为 aggregate Column （参阅聚合控制）。 同一列不能同时用作 joinColumn 和 dimension Columns （见下文）。 除非它也被归类为 aggregate Column ，否则除了 INNER JOIN 之外，查询的其他部分都不能使用 joinColumn 。
joinRequired	控制您是否要求与可以查询的成员的已配置表进行 INNER JOIN。	如果启用了此参数，则要求 INNER JOIN。如果未启用此参数，则 INNER JOIN 是可选的。 假设您启用了此参数，则可以查询的成员需要在 INNER JOIN 中包含他们拥有的表。他们必须将您的表与他们的表 JOIN，可以是直接，也可以是传递（也就是说，将他们的表联接到另一个表，而另一个表又联接到您的表）。

以下是传递联接的示例。

```
ON
```

```
my_table.identifer = third_party_table.identifier
....
ON
third_party_table.identifier = member_who_can_query_table.id
```

> **Note**
>
> 可以查询的成员也可以使用 joinRequired 参数。在这种情况下，查询必须将其表与至少一个其他表联接。

维度控制

维度控制控制可以对聚合列进行筛选、分组或聚合的列。

控件	定义	用法
dimensionColumns	您允许可以查询的成员在 SELECT、WHERE、GROUP、BY 和 ORDER BY 中使用的列（如果有）。	dimensionColumn 可以在 SELECT (select_grouping_column_expression)、WHERE、GROUP BY 和 ORDER BY 中使用。 同一列不能同时是 a dimensionColumn 、a joinColumn 、 and/or a aggregateColumn 。

标量函数

标量函数控制哪些标量函数可以在维度列上使用。

控件	定义	用法
scalarFunctions	可在查询的 dimension Columns 上使用的标量函数。	指定允许在 dimension Columns 上应用的标量函数（如果有）（例如 CAST）。

控件	定义	用法
		标量函数不能在其他函数之上使用，也不能在其他函数中使用。标量函数的参数可以是列、字符串文本或数字文本。

支持以下标量函数：

- 数学函数 - ABS、CEILING、FLOOR、LOG、LN、ROUND、SQRT
- 数据类型格式设置函数 — CAST, CONVERT, TO_CHAR, TO_DATE, TO_NUMBER, TO_TIMESTAMP
- 字符串函数 - LOWER、UPPER、RTRIM、SUBSTRING
  - 对于 RTRIM，不允许使用自定义字符集进行修剪。
- 条件表达式 — COALESCE
- 日期函数 - EXTRACT、GETDATE、CURRENT_DATE、DATEADD
- 其他函数 — TRUNC

有关详细信息，请参阅 AWS Clean Rooms SQL 参考。

聚合分析规则 — 查询结果控制

使用聚合查询结果控制，可以通过指定每个输出行必须满足的一个或多个条件来控制返回哪些结果。AWS Clean Rooms 支持 `COUNT (DISTINCT column) >= X` 形式的聚合约束。此形式要求每行至少聚合从配置表中选择的 X 个不同值（例如，不同 `user_id` 值的最少个数）。即使提交的查询本身不使用指定的列，也会自动强制执行此最低阈值。它们是在来自协作中每个成员的已配置表的查询中的每个已配置表中共同强制执行的。

每个配置表的分析规则中必须有至少一个聚合约束。配置表所有者可以添加多个 `columnName` 和关联的 `minimum`，这些表将共同强制执行。

聚合约束

聚合约束 控制返回查询结果中的哪些行。要返回，行必须满足聚合约束中指定的每列中指定的最小不同值数。即使在查询或分析规则的其他部分中未明确提及该列，此要求也适用。

控件	定义	用法
columnName	在每个输出行必须满足的条件中使用的 aggregateColumn 。	可以是已配置表中的任何列。
minimum	要在查询结果中返回输出行（例如 COUNT DISTINCT），关联 aggregateColumn 必须具有的最小不同值个数。	minimum 的值必须至少为 2。

聚合分析规则结构

以下示例显示了聚合分析规则的预定义结构。

在以下示例中，*MyTable* 指您的数据表。你可以用自己的信息替换每个*user input placeholder*信息。

```
{
  "aggregateColumns": [
    {
      "columnNames": [MyTable column names], "function": [Allowed Agg Functions]
    },
  ],
  "joinRequired": ["QUERY_RUNNER"],
  "joinColumns": [MyTable column names],
  "dimensionColumns": [MyTable column names],
  "scalarFunctions": [Allowed Scalar functions],
  "outputConstraints": [
    {
      "columnName": [MyTable column names], "minimum": [Numeric value]
    },
  ]
}
```

聚合分析规则 — 示例

以下示例演示了两家公司如何合作 AWS Clean Rooms 使用聚合分析。

A 公司有客户和销售数据。A 公司有兴趣了解产品退货活动。B 公司是 A 公司的零售商之一，有退货数据。B 公司也有对 A 公司有用的客户细分属性（例如，购买过相关产品、使用过零售商的客户服务）。B 公司不想提供行级客户退货数据和属性信息。B 公司只想为 A 公司启用一组查询，以最小聚合阈值获取重叠客户的聚合统计数据。

A 公司和 B 公司决定协作，以便 A 公司能够了解产品退货活动，并在 B 公司和其他渠道提供更好的产品。

为了创建协作并进行聚合分析，两家公司执行以下操作：

1. A 公司创建协作并创建成员身份。协作中的另一个成员是 B 公司。A 公司在协作中启用查询日志记录，并在其账户中启用查询日志记录。
2. B 公司在协作中创建成员身份。它在其账户中启用查询日志记录。
3. A 公司创建销售配置表。
4. A 公司将以下聚合分析规则添加到销售配置表中。

```
{
  "aggregateColumns": [
    {
      "columnNames": [
        "identifier"
      ],
      "function": "COUNT_DISTINCT"
    },
    {
      "columnNames": [
        "purchases"
      ],
      "function": "AVG"
    },
    {
      "columnNames": [
        "purchases"
      ],
      "function": "SUM"
    }
  ],
  "joinColumns": [
    "hashedemail"
  ],
  "dimensionColumns": [
```

```
    "demoseg",
    "purchasedate",
    "productline"
  ],
  "scalarFunctions": [
    "CAST",
    "COALESCE",
    "TRUNC"
  ],
  "outputConstraints": [
    {
      "columnName": "hashedemail",
      "minimum": 2,
      "type": "COUNT_DISTINCT"
    },
  ]
}
```

aggregateColumns — A 公司想要计算销售数据和退货数据之间重叠的唯一客户数量。A 公司还想汇总 purchases 数量，以便与 returns 数量进行比较。

joinColumns — A 公司想要使用 identifier 来匹配销售数据中的客户和退货数据中的客户。这将有助于 A 公司将退货与正确的采购相匹配。它还可以帮助 A 公司细分重叠的客户。

dimensionColumns — A 公司使用 dimensionColumns 来筛选特定产品，比较一段时期内的购买和退货情况，确保退货日期在产品日期之后，并帮助细分重叠客户。

scalarFunctions — A 公司选择 CAST 标量函数，以便在需要时根据 A 公司关联到协作的配置表更新数据类型格式。它还添加了标量函数，以便在需要时帮助格式化列。

outputConstraints — A 公司设定了最低输出约束。它不需要限制结果，因为允许分析师从销售表中查看行级数据

> **Note**
>
> A 公司没有在分析规则中加入 joinRequired。它为他们的分析师提供了单独查询销售表的灵活性。

5. B 公司创建退货配置表。
6. B 公司将以下聚合分析规则添加到退货配置表中。

```
{
  "aggregateColumns": [
    {
      "columnNames": [
        "identifier"
      ],
      "function": "COUNT_DISTINCT"
    },
    {
      "columnNames": [
        "returns"
      ],
      "function": "AVG"
    },
    {
      "columnNames": [
        "returns"
      ],
      "function": "SUM"
    }
  ],
  "joinColumns": [
    "hashedemail"
  ],
  "joinRequired": [
    "QUERY_RUNNER"
  ],
  "dimensionColumns": [
    "state",
    "popularpurchases",
    "customerserviceuser",
    "productline",
    "returndate"
  ],
  "scalarFunctions": [
    "CAST",
    "LOWER",
    "UPPER",
    "TRUNC"
  ],
  "outputConstraints": [
    {
      "columnName": "hashedemail",
```

```
      "minimum": 100,
      "type": "COUNT_DISTINCT"
    },
    {
      "columnName": "producttype",
      "minimum": 2,
      "type": "COUNT_DISTINCT"
    }
  ]
}
```

aggregateColumns — B 公司让 A 公司汇总 returns 以与购买数量进行比较。它们至少有一个聚合列，因为它们启用了聚合查询。

joinColumns — B 公司让 A 公司在 identifier 上进行联接，以将退货数据中的客户与销售数据中的客户进行匹配。identifier 数据特别敏感，将其作为 joinColumn 可确保数据永远不会在查询中输出。

joinRequired — B 公司要求对退货数据的查询必须与销售数据重叠。他们不想让 A 公司查询其数据集中的所有个人。他们还在协作协议中商定了这一限制。

dimensionColumns — B 公司让 A 公司按 state、popularpurchases 和 customerserviceuser 进行筛选和分组，这些独特的属性有助于 A 公司进行分析。B 公司让 A 公司使用 returndate 筛选在 purchasedate 之后发生的 returndate 的输出。通过这种筛选，输出可以更准确地评估产品变更的影响。

scalarFunctions — B 公司启用以下函数：

- TRUNNC（表示日期）
- LOWER 和 UPPER（如果 producttype 在数据中以不同的格式输入）
- CAST（如果 A 公司需要将销售表中的数据类型转换为与退货中表的数据类型相同）

A 公司不启用其他标量函数，因为他们认为查询不需要这些函数。

outputConstraints — B 公司对 hashedemail 设定了最低输出约束，以帮助降低重新识别客户身份的能力。它还对 producttype 增加了最低输出约束，以降低重新识别退回的特定产品的能力。根据输出的维度（例如 state），某些产品类型可能更占优势。无论 A 公司在其数据中添加了什么输出约束，他们的输出约束都将始终得到执行。

7. A 公司创建了与协作的销售表关联。
8. B 公司创建了与协作的退货表关联。

9. A 公司运行查询（如以下示例），以更好地了解 B 公司的退货数量与 2022 年各地采购总量的对比情况。

```
SELECT
  companyB.state,
  SUM(companyB.returns),
  COUNT(DISTINCT companyA.hashedemail)
FROM
  sales companyA
  INNER JOIN returns companyB ON companyA.identifier = companyB.identifier
WHERE
  companyA.purchasedate BETWEEN '2022-01-01' AND '2022-12-31' AND
  TRUNC(companyB.returndate) > companyA.purchasedate
GROUP BY
  companyB.state;
```

10A 公司和 B 公司查看查询日志。B 公司验证查询是否符合协作协议中上商定的内容。

聚合分析规则问题疑难解答

使用此处的信息可帮助您诊断和修复在使用聚合分析规则时出现的常见问题。

问题

- 我的查询没有返回任何结果

我的查询没有返回任何结果

当没有匹配结果或匹配结果不符合一个或多个最低聚合阈值时，就会发生这种情况。

有关最低聚合阈值的更多信息，请参阅聚合分析规则 — 示例。

列表分析规则

在中 AWS Clean Rooms，列表分析规则输出行级列表，列出其添加到的已配置表与可以查询的成员的配置表之间的重叠部分。可以查询的成员运行包含列表分析规则的查询。

列表分析规则类型支持扩充和受众构建等使用案例。

有关此分析规则的预定义查询结构和语法的更多信息，请参阅列表分析规则预定义结构。

列表分析规则的参数（在列表分析规则 — 查询控制中定义）具有查询控制。它的查询控制包括选择可以在输出中列出的列的功能。查询要求至少有一次与可以查询成员的配置表联接，可以是直接联接，也可以是传递联接。

不存在像聚合分析规则那样的查询结果控制。

列表查询只能使用数学运算符。它们不能使用其他函数（例如聚合或标量）。

主题

列表查询结构和语法

对具有列表分析规则的表的查询必须遵循以下语法。

```
--select_list_expression
SELECT DISTINCT column_name [[AS] column_alias ] [, ...]

--table_expression
FROM table_name [[AS] table_alias ]
  [[INNER] JOIN table_name [[AS] table_alias] ON join_condition] [...]

--where_expression
[WHERE where_condition]

--limit_expression
[LIMIT number]
```

下表解释前面语法中列出的每个表达式。

Expression	定义	示例
select_list_expression	包含至少一个表列名的逗号分隔列表。 DISTINCT 参数是必需的。	SELECT DISTINCT segment

Expression	定义	示例
	Note `select_list_expression` 可以带或不带 AS 参数对列设置别名。有关更多信息，请参阅 AWS Clean Rooms SQL 参考。	
table_expression	使用 `join_condition` 连接到 `join_condition` 的表或表的联接。 `join_condition` 返回布尔值。 `table_expression` 支持： • 特定的 JOIN 类型 (INNER JOIN) • `join_condition` 中的相等比较条件 (=) • 逻辑运算符（AND、OR）。	`FROM consumer_table INNER JOIN provider_table ON consumer_table.identifier1 = provider_table.identifier1 AND consumer_table.identifier2 = provider_table.identifier2`

Expression	定义	示例
where_expression	返回布尔值的条件表达式。它可能包括以下内容： • 表列名称 • 数学运算符 • 字符串文本 • 数值文本 支持的比较条件是 (=, >, <, <=, >=, <>, !=, NOT, IN, NOT IN, LIKE, IS NULL, IS NOT NULL)。 支持的逻辑运算符是 (AND, OR)。 where_expression 是可选项。	WHERE state + '_' + city = 'NY_NYC' WHERE timestampColumn = timestampColumn2 - 14
limit_expression	此表达式必须采用正整数。 limit_expression 是可选项。	LIMIT 100

关于列表查询的结构和语法，请注意以下几点：

- 不支持除 SELECT 之外的 SQL 命令。
- 不支持子查询和通用表格表达式（例如 WITH）。
- 不支持 HAVING、GROUP BY 和 ORDER BY 子句
- 不支持 OFFSET 参数

列表分析规则 — 查询控制

使用列表查询控制，您可以控制如何使用表中的列来查询表。例如，您可以控制哪一列用于联接，或者 SELECT 语句和 WHERE 子句中可以使用哪一列。

下面几节解释每种控制。

主题

联接控制

使用联接控制，您可以控制如何将您的表连接到 table_expression 中的其他表。 AWS Clean Rooms 仅支持 INNER JOIN。在列表分析规则中，至少需要一个 INNER JOIN，并且可以查询的成员必须在 INNER JOIN 中包含自己拥有的表。这意味着他们必须直接或通过传递方式将您的表与他们的表联接起来。

以下是传递联接的示例。

```
ON
my_table.identifer = third_party_table.identifier
....
ON
third_party_table.identifier = member_who_can_query_table.id
```

INNER JOIN 语句只能使用在分析规则中明确归类为 `joinColumn` 的列。

INNER JOIN 必须对您的已配置表中的 `joinColumn` 和协作中另一个已配置表中的 `joinColumn` 进行操作。您可以决定表中的哪些列可以用作 `joinColumn`。

ON 子句中的每个匹配条件都要求在两列之间使用相等比较条件 (=)。

ON 子句中的多个匹配条件可以是：

- 使用 AND 逻辑运算符组合
- 使用 OR 逻辑运算符分隔

> **Note**
>
> 所有 JOIN 匹配条件都必须与 JOIN 两侧各一条记录匹配。所有由 OR 或 AND 逻辑运算符连接的条件也必须遵守此要求。

以下是使用 AND 逻辑运算符的查询示例。

```
SELECT some_col, other_col
FROM table1
    JOIN table2
    ON table1.id = table2.id AND table1.name = table2.name
```

以下是使用 OR 逻辑运算符的查询示例。

```
SELECT some_col, other_col
FROM table1
    JOIN table2
    ON table1.id = table2.id OR table1.name = table2.name
```

控件	定义	用法
joinColumns	您希望允许可以查询的成员在 INNER JOIN 语句中使用的列。	同一列不能同时归类为 joinColumn 和 listColumn （参阅列表控制）。 除了 INNER JOIN 之外，不能在查询的任何其他部分中使用 joinColumn 。

列表控制

列表控制用于控制可在查询输出中列出（即在 SELECT 语句中使用）或用于筛选结果（即在 WHERE 语句中使用）的列。

控件	定义	用法
listColumns	您允许可以查询的成员在 SELECT 和 WHERE 中使用的列。	listColumn 可以在 SELECT 和 WHERE 中使用。 同一列不能同时用作 listColumn 和 joinColumn 。

列表分析规则预定义结构

以下示例包括一个预定义的结构，该结构向您展示了如何完成列表分析规则。

在以下示例中，*MyTable* 指您的数据表。你可以用自己的信息替换每个*user input placeholder*信息。

```
{
  "joinColumns": [MyTable column name(s)],
  "listColumns": [MyTable column name(s)],
}
```

列表分析规则 — 示例

以下示例演示了两家公司如何合作 AWS Clean Rooms 使用列表分析。

A 公司有客户关系管理 (CRM) 数据。A 公司希望获得有关其客户的更多细分数据，以进一步了解他们的客户，并有可能使用属性作为其他分析的输入。B 公司的细分数据由他们根据第一方数据创建的独特细分属性组成。B 公司只想向 A 公司提供其数据与 A 公司数据重叠的客户的唯一细分属性。

两家公司决定进行协作，以便 A 公司能够扩充重叠的数据。A 公司是可以查询的成员，B 公司是贡献者。

为创建协作并在协作中运行列表分析，两家公司执行以下操作：

1. A 公司创建协作并创建成员身份。协作中的另一个成员是 B 公司。A 公司在协作中启用查询日志记录，并在其账户中启用查询日志记录。
2. B 公司在协作中创建成员身份。它在其账户中启用查询日志记录。
3. A 公司创建 CRM 配置表。

4. A 公司将分析规则添加到客户配置表中，如以下示例所示。

```
{
  "joinColumns": [
    "identifier1",
    "identifier2"
  ],
  "listColumns": [
    "internalid",
    "segment1",
    "segment2",
    "customercategory"
  ]
}
```

joinColumns— A 公司希望使用 hashedemail and/or thirdpartyid（从身份供应商处获得）将 CRM 数据中的客户与来自细分市场的客户进行匹配。这将有助于确保 A 公司为合适的客户匹配扩充的数据。他们有两个 JoinColumns，可以提高分析的匹配率。

listColumns — A 公司使用 listColumns 来获取他们在自己系统中使用的 internalid 旁边的扩充列。他们添加了 segment1、segment2 和 customercategory，以便通过在筛选器中使用它们，将扩充限制到特定的细分。

5. B 公司创建细分配置表。

6. B 公司将分析规则添加到细分配置表中。

```
{
  "joinColumns": [
    "identifier2"
  ],
  "listColumns": [
    "segment3",
    "segment4"
  ]
}
```

joinColumns— B 公司让 A 公司在 identifier2 上进行联接，以便将细分数据中的客户与 CRM 数据相匹配。A 公司和 B 公司与身份供应商合作，以获得与此协作相匹配的 identifier2。他们之所以没有添加其他 joinColumns，是因为他们认为 identifier2 可以提供最高和最准确的匹配率，而且查询不需要其他标识符。

`listColumns` — B 公司让 A 公司使用 `segment3` 和 `segment4` 属性来扩充其数据，这些属性是他们（与客户 A）一起创建、收集和调整的独特属性，是数据扩充的一部分。他们希望 A 公司在行级获取这些重叠的细分，因为这是一项数据扩充协作。

7. A 公司创建了与协作的 CRM 表关联。
8. B 公司创建了与协作的细分表关联。
9. A 公司运行查询（例如以下查询）以扩充重叠的客户数据。

```
SELECT companyA.internalid, companyB.segment3, companyB.segment4
INNER JOIN returns companyB
 ON companyA.identifier2 = companyB.identifier2
WHERE companyA.customercategory > 'xxx'
```

10A 公司和 B 公司查看查询日志。B 公司验证查询是否符合协作协议中上商定的内容。

中的自定义分析规则 AWS Clean Rooms

在中 AWS Clean Rooms，自定义分析规则是一种新型的分析规则，它允许在配置的表上运行自定义查询。自定义 SQL 查询仍然仅限于只有SELECT命令，但与聚合和列表查询相比，可以使用更多的 SQL 结构（例如，窗口函数、OUTER JOIN 或子查询；有关完整列表 CTEs，请参阅 AWS Clean Rooms SQL 参考）。自定义 SQL 查询不必遵循聚合和列表查询之类的查询结构。

与聚合和列表分析规则支持的使用案例相比，自定义分析规则支持更高级的使用案例，例如自定义归因分析、基准测试、增量分析和受众发现。这是对聚合和列表分析规则支持的使用案例的超集的补充。

自定义分析规则还支持差别隐私。差别隐私是一种在数学上非常严格的数据隐私保护框架。有关更多信息，请参阅 AWS Clean Rooms 差异隐私。创建分析模板时， AWS Clean Rooms 差异隐私会检查该模板以确定其是否与 AWS Clean Rooms 差异隐私的通用查询结构兼容。此验证可确保您不会创建对于差别隐私保护表不允许的分析模板。

要配置自定义分析规则，数据所有者可以选择允许存储在分析模板中的特定自定义查询在其配置表上运行。数据所有者在将分析模板添加到自定义分析规则中允许的分析控制之前应先审核这些模板。分析模板仅在创建这些模板的协作中可用和可见（即使该表与其他协作关联），并且只能由可以在该协作中进行查询的成员运行。

或者，成员可以选择允许其他成员（查询提供者）无需审核即可创建查询。成员在自定义分析规则添加允许的查询提供者控制的查询提供者账户。如果查询提供者是可以查询的成员，则他们可以直接在配置表上运行任何查询。查询提供者还可以通过创建分析模板来创建查询。在存在查询提供程序和关联表的所有协作中，自动允许查询提供者创建的任何查询在表上运行。 AWS 账户

数据所有者只能允许分析模板或账户创建查询，不能同时允许两者创建。如果数据所有者将其留空，则可以查询的成员将无法对配置表运行查询。

主题

自定义分析规则预定义结构

以下示例包含一个预定义的结构，说明了如何完成开启差别隐私的自定义分析规则。`userIdentifier` 值是唯一地标识您的用户的列，例如 user_id。如果在协作中具有两个或更多开启了差别隐私的表， AWS Clean Rooms 要求您在两个分析规则中配置与用户标识符列相同的列，以在表之间保持一致的用户定义。

```
{
  "allowedAnalyses": ["ANY_QUERY"] | string[],
  "allowedAnalysisProviders": [],
  "differentialPrivacy": {
    "columns": [
      {
        "name": "userIdentifier"
      }
    ]
  }
}
```

您可以：

- 将分析模板 ARNs 添加到允许的分析控件。在这种情况下，不包括 `allowedAnalysisProviders` 控制。

  ```
  {
    allowedAnalyses: string[]
  }
  ```

- AWS 账户 IDs 向`allowedAnalysisProviders`控件添加成员。在这种情况下，您可以将 `ANY_QUERY` 添加到 `allowedAnalyses` 控制。

```
{
  allowedAnalyses: ["ANY_QUERY"],
  allowedAnalysisProviders: string[]
}
```

自定义分析规则示例

以下示例演示了两家公司如何合作 AWS Clean Rooms 使用自定义分析规则。

A 公司有客户和销售数据。A 公司有兴趣了解 B 公司网站上广告活动的销售增量。B 公司拥有对 A 公司有用的观众数据和细分属性（例如，他们观看广告时使用的设备）。

A 公司想在协作中运行一个特定的增量查询。

为创建协作并在协作中运行自定义分析，两家公司执行以下操作：

1. A 公司创建协作并创建成员身份。协作中的另一个成员是 B 公司。A 公司在协作中启用查询日志记录，并在其账户中启用查询日志记录。
2. B 公司在协作中创建成员身份。它在其账户中启用查询日志记录。
3. A 公司创建 CRM 配置表。
4. A 公司向销售配置表添加空的自定义分析规则。
5. A 公司将销售配置表与协作关联起来。
6. B 公司创建观众配置表。
7. B 公司在观众配置表中添加一个空的自定义分析规则。
8. B 公司将观众配置表与协作关联起来。
9. A 公司查看与协作关联的销售表和观众表，并创建分析模板，为活动月份添加增量查询和参数。

```
{
    "analysisParameters": [
    {
        "defaultValue": ""
        "type": "DATE"
        "name": "campaign_month"
    }
    ],
    "description": "Monthly incrementality query using sales and viewership data"
    "format": "SQL"
    "name": "Incrementality analysis"
```

```
    "source":
        "WITH labeleddata AS
        (
        SELECT hashedemail, deviceid, purchases, unitprice, purchasedate,
        CASE
            WHEN testvalue IN ('value1', 'value2', 'value3') THEN 0
            ELSE 1
        END AS testgroup
        FROM viewershipdata
        )
        SELECT labeleddata.purchases, provider.impressions
        FROM labeleddata
        INNER JOIN salesdata
          ON labeleddata.hashedemail = provider.hashedemail
        WHERE MONTH(labeleddata.purchasedate) > :campaignmonth
        AND testgroup = :group
       "
}
```

10A 公司将其账户（例如 444455556666）添加到自定义分析规则允许的分析提供者控制中。他们之所以使用允许的分析提供者控制，是因为他们希望允许在销售配置表上运行他们创建的任何查询。

```
{
  "allowedAnalyses": [
    "ANY_QUERY"
  ],
  "allowedAnalysisProviders": [
    "444455556666"
  ]
}
```

11B 公司在协作中看到创建的分析模板并查看其内容，包括查询字符串和参数。

12B 公司确定分析模板实现了增量使用案例，并满足如何查询其观众配置表的隐私要求。

13B 公司将分析模板 ARN 添加到观众表的自定义分析规则允许的分析控制中。他们之所以使用允许的分析控制，是因为他们只想允许在观众配置表上运行增量查询。

```
{
  "allowedAnalyses": [
    "arn:aws:cleanrooms:us-east-1:111122223333:membership/41327cc4-bbf0-43f1-b70c-
a160dddceb08/analysistemplate/1ff1bf9d-781c-418d-a6ac-2b80c09d6292"
  ]
}
```

14A 公司运行分析模板并使用参数值 `05-01-2023`。

具有差别隐私的自定义分析规则

在中 AWS Clean Rooms，自定义分析规则支持差异隐私。差别隐私是一种在数学上非常严格的数据隐私保护框架，可以帮助保护您的数据以防范重新识别尝试。

差异隐私支持综合分析，例如广告活动规划、 post-ad-campaign衡量、金融机构联盟中的基准测试以及医疗保健研究的 A/B 测试。

支持的查询结构和语法在 查询结构和语法 中定义。

具有差别隐私的自定义分析规则示例

> **Note**
>
> AWS Clean Rooms 差异隐私仅适用于数据存储在 Amazon S3 中的协作。

考虑上一节中介绍的自定义分析规则示例。该示例说明了如何使用差别隐私保护您的数据以防范重新识别尝试，同时允许您的合作伙伴从您的数据中了解业务关键型见解。假设 B 公司具有观众数据，并希望使用差别隐私保护其数据。为了完成差别隐私设置，B 公司完成以下步骤：

1. B 公司开启差别隐私，同时在观众配置表中添加自定义分析规则。B 公司选择 `viewershipdata.hashedemail` 以作为用户标识符列。
2. B 公司在协作中添加差别隐私策略，以使其观众数据表可供查询。B 公司选择默认策略以快速完成设置。

A 公司希望了解 B 公司网站上的广告活动的销售增量，并运行分析模板。由于该查询与 Diferation Privacy 的 AWS Clean Rooms 通用查询结构兼容，因此查询可以成功运行。

查询结构和语法

包含至少一个开启了差别隐私的表的查询必须遵循以下语法。

```
query_statement:
    [cte, ...] final_select
```

```
cte:
   WITH sub_query AS (
      inner_select
      [ UNION | INTERSECT | UNION_ALL | EXCEPT/MINUS ]
      [ inner_select ]
   )

inner_select:
    SELECT [user_id_column, ] expression [, ...]
    FROM table_reference [, ...]
    [ WHERE condition ]
    [ GROUP BY user_id_column[, expression] [, ...] ]
    [ HAVING condition ]

final_select:
    SELECT [expression, ...] | COUNT | COUNT_DISTINCT | SUM | AVG | STDDEV
    FROM table_reference [, ...]
    [ WHERE condition ]
    [ GROUP BY expression [, ...] ]
    [ HAVING COUNT | COUNT_DISTINCT | SUM | AVG | STDDEV | condition ]
    [ ORDER BY column_list ASC | DESC ]
    [ OFFSET literal ]
    [ LIMIT literal ]

expression:
   column_name [, ...] | expression AS alias | aggregation_functions |
window_functions_on_user_id | scalar_function | CASE | column_name math_expression [,
expression]

window_functions_on_user_id:
   function () OVER (PARTITION BY user_id_column, [column_name] [ORDER BY column_list
ASC|DESC])
```

Note

对于差别隐私查询结构和语法，请注意以下事项：

- 不支持子查询。
- 如果表或 CTE 涉及受差异隐私保护的数据，则公用表表达式 (CTEs) 应发出用户标识符列。应在用户级别完成筛选、分组和聚合。
- final_select 允许使用 COUNT DISTINCT、COUNT、SUM、AVG 和 STDDEV 聚合函数。

有关差别隐私支持哪些 SQL 关键字的更多详细信息，请参阅 AWS Clean Rooms 差异隐私的 SQL 功能。

ID 映射表分析规则

在中 AWS Clean Rooms，ID 映射表分析规则不是独立的分析规则。这种类型的分析规则由管理 AWS Clean Rooms 并用于连接不同的身份数据以方便查询。它会自动添加到 ID 映射表中，并且无法编辑。它会继承协作中其他分析规则的行为，前提是这些分析规则是同构分析规则。

ID 映射表分析规则对 ID 映射表强制执行安全措施。它限制协作成员使用 ID 映射表直接选择或检查两个成员数据集之间的非重叠人群。当在查询中与其他分析规则一起隐式使用时，ID 映射表分析规则用于保护 ID 映射表中的敏感数据。

使用 ID 映射表分析规则，在展开的 SQL 中 AWS Clean Rooms 强制执行 ID 映射表两侧的重叠。这样做可让您执行以下任务：

- 在 JOIN 语句中使用 ID 映射表的重叠部分。

  AWS Clean Rooms 允许在 ID 映射表上使用INNERLEFT、或RIGHT联接，前提是它尊重重叠之处。为了保护敏感的映射信息，ID 映射表必须始终位于任何JOIN操作的 inner “” 端。例如，以下JOIN操作是有效的：

  - table LEFT JOIN id_mapping_table
  - id_mapping_table RIGHT JOIN table
  - table INNER JOIN id_mapping_table

  以下JOIN操作无效：

  - id_mapping_table LEFT JOIN table
  - table RIGHT JOIN id_mapping_table

  这样可以防止数据集中没有相应匹配项的映射记录被泄露。允许此类操作可能会泄露有关其他协作成员数据映射的敏感信息。
- 在 JOIN 语句中使用映射表列。

  不能在以下语句中使用映射表列：SELECT、WHERE、HAVING、GROUP BY、或 ORDER BY（除非修改了源 ID 命名空间关联或目标 ID 命名空间关联的保护）。
- 在扩展的 SQL 中， AWS Clean Rooms 还支持OUTERJOINJOIN、隐式和 CROSS JOIN。这些联接无法满足重叠要求。而是 AWS Clean Rooms 使用`requireOverlap`来指定必须连接哪些列。

支持的查询结构和语法在 ID 映射表查询结构和语法 中定义。

ID 映射表分析规则查询控制 中定义的分析规则的参数包括查询控制和查询结果控制。其查询控制包括要求在 JOIN 语句中使用 ID 映射表重叠部分的功能（即 `requireOverlap`）。

主题

ID 映射表查询结构和语法

对具有 ID 映射表分析规则的表的查询必须遵循以下语法。

```
--select_list_expression
SELECT
provider.data_col, consumer.data_col

--table_expression
FROM provider

JOIN idMappingTable idmt ON provider.id = idmt.sourceId

JOIN consumer ON consumer.id = idmt.targetId
```

协作表

下表表示 AWS Clean Rooms 协作中存在的已配置表。cr_drivers_license 和 cr_insurance 表中的 id 列都表示与 ID 映射表匹配的列。

cr_drivers_license

id	司机姓名	注册状态
1	爱德华	TX

2	达纳	MA
3	Gweneth	IL

cr_insurance

id	保单持有人_电子邮件	保单编号
a	eduardo@internal.company.com	17f9d04e-f5be-4426-bdc4-250ed59c6529
b	gwen@internal.company.com	3f0092db-2316-48a8-8d44-09cf8f6e6c64
c	rosa@internal.company.com	d7692e84-3d3c-47b8-b46d-a0d5345f0601

ID 映射表

下表显示了在 cr_drivers_license 和 cr_insurance 表上匹配的现有 ID 映射表。并非所有条目都 IDs 适用于两个协作表。

cr_drivers_license_id	cr_insurance_id
1	a
2	null
3	b
null	c

ID 映射表分析规则仅允许对一组重叠数据运行查询，重叠数据如下所示：

cr_drivers_license_id	cr_insurance_id	司机姓名	注册状态	保单持有人_电子邮件	保单编号

1	a	爱德华	TX	eduardo@internal.company.com	17f9d04e-f5be-4426-bdc4-250ed59c6529
3	b	Gweneth	IL	gwen@internal.company.com	3f0092db-2316-48a8-8d44-09cf8f6e6c64

示例查询

以下示例显示了 ID 映射表联接的有效位置：

```
-- Single ID mapping table
SELECT
    [ select_items ]FROM
    cr_drivers_license cr_dl
    [ INNER | LEFT ] JOIN cr_identity_mapping_table idmt ON idmt.cr_drivers_license_id
 = cr_dl.id
    [ INNER | RIGHT ] JOIN cr_insurance cr_in            ON idmt.cr_insurance_id
 = cr_in.id
;
-- Single ID mapping table (Subquery)
SELECT
    [ select_items ]FROM (
    SELECT
        [ select_items ]
    FROM
        cr_drivers_license cr_dl
        [ INNER | LEFT ] JOIN cr_identity_mapping_table idmt ON
 idmt.cr_drivers_license_id = cr_dl.id
        [ INNER | RIGHT ] JOIN cr_insurance cr_in            ON idmt.cr_insurance_id
    = cr_in.id
)
;
-- Single ID mapping table (CTE)
WITH
    matched_ids AS (
        SELECT
            [ select_items ]
```

```
        FROM
            cr_drivers_license cr_dl
            [ INNER | LEFT ] JOIN cr_identity_mapping_table idmt ON
 idmt.cr_drivers_license_id = cr_dl.id
            [ INNER | RIGHT ] JOIN cr_insurance cr_in            ON
 idmt.cr_insurance_id       = cr_in.id
    )SELECT
    [ select_items ]FROM
    matched_ids
;
```

注意事项

关于 ID 映射表查询的结构和语法，请注意以下几点：

- 您不能对其进行编辑。
- 默认情况下，它会应用于 ID 映射表。
- 它在协作内部使用源和目标 ID 命名空间关联。
- 默认情况下，ID 映射表配置为向来自 ID 命名空间的列提供默认保护。您可以修改此配置，以便来自 ID 命名空间（`sourceID` 或 `targetID`）的列可以出现在查询中的任何位置。有关更多信息，请参阅 ID 中的命名空间 AWS Clean Rooms。
- ID 映射表分析规则将继承协作中其他分析规则的 SQL 限制。

ID 映射表分析规则查询控制

使用 ID 映射表查询 AWS Clean Rooms 控件，控制如何使用表中的列来查询表。例如，它可以控制哪些列用于联接，哪些列需要重叠。ID 映射表分析规则还包括允许在不需要 JOIN 的情况下投影 `sourceID` 和/或 `targetID` 的功能。

下表介绍了每种控制。

控件	定义	用法
`joinColumns`	可以查询的成员能在 INNER JOIN 语句中使用的列。	除了 INNER JOIN 之外，不能在查询的任何其他部分中使用 `joinColumns` 。 有关更多信息，请参阅 联接控制。

控件	定义	用法
dimensionColumns	可以查询的成员能在 SELECT 和 GROUP BY 语句中使用的列（如果有）。	可以在 SELECT 和 GROUP BY 中使用的 dimensionColumn 。 可以显示为 joinKeys 的 dimensionColumn 。 如果使用方括号指定 dimensionColumns ，则只能在 JOIN 子句中使用它。
queryContraints:RequireOverlap	ID 映射表中必须联接以便可以运行查询的列。	必须使用这些列对 ID 映射表和协作表执行 JOIN。

ID 映射表的分析规则预定义结构

ID 映射表分析规则的预定义结构对 sourceID 和 targetID 应用默认保护。这意味着在查询中必须使用应用了保护的列。

您可以通过以下方式配置 ID 映射表分析规则：

- sourceID 和 targetID 均受到保护

  在此配置中，不能同时投影 sourceID 和 targetID。引用 ID 映射表时，必须在 JOIN 中使用 sourceID 和 targetID。
- 仅保护 targetID

  在此配置中，不能投影 targetID。引用 ID 映射表时，必须在 JOIN 中使用 targetID。可以在查询中使用 sourceID。
- 仅保护 sourceID

  在此配置中，不能投影 sourceID。引用 ID 映射表时，必须在 JOIN 中使用 sourceID。可以在查询中使用 targetID。
- sourceID 或 targetID 均不受保护

  在此配置中，ID 映射表不受可在查询中使用的任何特定强制措施的约束。

以下示例显示了对 sourceID 和 targetID 应用默认保护的 ID 映射表分析规则的预定义结构。在此示例中，ID 映射表分析规则仅允许对 sourceID 列和 targetID 列执行 INNER JOIN。

```
{
  "joinColumns": [
    "source_id",
    "target_id"
  ],
  "queryConstraints": [
    {
      "requireOverlap": {
        "columns": [
          "source_id",
          "target_id"
        ]
      }
    }
  ],
  "dimensionColumns": [] // columns that can be used in SELECT and JOIN
}
```

以下示例显示了对 targetID 应用保护的 ID 映射表分析规则的预定义结构。在此示例中，ID 映射表分析规则仅允许对 sourceID 列执行 INNER JOIN。

```
{
  "joinColumns": [
    "source_id",
    "target_id"
  ],
  "queryConstraints": [
    {
      "requireOverlap": {
        "columns": [
          "target_id"
        ]
      }
    }
  ],
  "dimensionColumns": [
    "source_id"
  ]
}
```

以下示例显示了对 sourceID 应用保护的 ID 映射表分析规则的预定义结构。在此示例中，ID 映射表分析规则仅允许对 targetID 列执行 INNER JOIN。

```
{
  "joinColumns": [
    "source_id",
    "target_id"
  ],
  "queryConstraints": [
    {
      "requireOverlap": {
        "columns": [
          "source_id"
        ]
      }
    }
  ],
  "dimensionColumns": [
    "target_id"
  ]
}
```

以下示例显示了不对 sourceID 或 targetID 应用保护的 ID 映射表分析规则的预定义结构。在此示例中，ID 映射表分析规则支持对 sourceID 列和 targetID 列执行 INNER JOIN。

```
{
  "joinColumns": [
    "source_id",
    "target_id"
  ],
  "queryConstraints": [
    {
      "requireOverlap": {
        "columns": []
      }
    }
  ],
  "dimensionColumns": [
    "source_id",
    "target_id"
  ]
}
```

ID 映射表分析规则 - 示例

例如，公司可以使用 ID 映射表分析规则来使用多方 LiveRamp 转码，而不是编写引用个人身份信息 (PII) 的长瀑布语句。以下示例演示如何协作 AWS Clean Rooms 使用 ID 映射表分析规则。

A 公司是拥有客户和销售数据的广告商，这些数据将用作源。A 公司还代表合作各方进行转码，并提供 LiveRamp 证书。

B 公司是拥有事件数据的发布者，这些数据将被用作目标。

> **Note**
>
> A 公司或 B 公司均可提供 LiveRamp 转码凭证并执行转码。

为创建支持在协作中运行 ID 映射表分析的协作，两家公司执行以下操作：

1. A 公司创建协作并创建成员身份。添加公司 B ，该公司还在协作中创建成员身份。
2. 公司 A 要么关联现有 ID 命名空间来源，要么 AWS Entity Resolution 数据匹配服务 使用 AWS Clean Rooms 控制台创建新的 ID 命名空间源。

   公司 A 创建一个配置表，其中包含他们的销售数据，以及对应于 ID 映射表中的 `sourceId` 的列。

   ID 命名空间源提供要转码的数据。
3. B 公司要么关联现有 ID 命名空间目标，要么 AWS Entity Resolution 数据匹配服务 使用 AWS Clean Rooms 控制台创建一个新的 ID 命名空间目标。

   公司 B 创建一个配置表，其中包含他们的事件数据，以及对应于 ID 映射表中的 `targetId` 的列。

   ID 命名空间目标不提供要转码的数据，只提供有关 LiveRamp 配置的元数据。
4. 公司 A 发现与协作关联的两个 ID 命名空间，并创建且填充一个 ID 映射表。
5. 公司 A 通过联接 ID 映射表对这两个数据集运行查询。

```
--- this would be valid for Custom or List
SELECT provider.data_col, consumer.data_col
FROM provider
  JOIN idMappingTable-123123123123-myMappingWFName idmt
     ON provider.id = idmt.sourceId
  JOIN consumer
```

```
    ON consumer.id = idmt.targetId
```

AWS Clean Rooms 差异隐私

AWS Clean Rooms 差异隐私通过一种以数学为依据的技术帮助您保护用户的隐私，该技术只需单击几下即可通过直观的控件实现。作为一项完全托管的功能，无需事先体验差异化隐私即可帮助您防止重新识别用户。 AWS Clean Rooms 在运行时自动向查询结果添加经过精心校准的噪音量，以帮助保护您的个人级别数据。

AWS Clean Rooms Difersition Privacy 支持广泛的分析查询，非常适合各种用例，在这些用例中，查询结果中的少量错误不会影响分析的实用性。通过使用该功能，您的合作伙伴可以生成有关广告活动、投资决策、临床研究等的业务关键型见解，合作伙伴无需进行任何额外的设置。

AWS Clean Rooms 差异隐私可防止恶意使用标量函数或数学运算符符号的溢出或无效强制转换错误。

有关 AWS Clean Rooms 差分隐私的更多信息，请参阅以下主题。

主题

- 差异隐私
- 差分隐私是如何 AWS Clean Rooms 运作的
- 差别隐私策略
- AWS Clean Rooms 差异隐私的 SQL 功能
- Differential Privacy 查询技巧和示例
- AWS Clean Rooms 差异隐私的局限性

差异隐私

差别隐私仅允许聚合的见解，并掩盖任何个人数据在这些见解中的贡献。差别隐私保护协作数据，以防止可以接收结果的成员了解特定个人的数据。如果没有差别隐私，可以接收结果的成员可能会尝试添加或删除有关个人的记录，并观察查询结果差异以推断个人用户数据。

在开启差别隐私后，将在查询结果中添加指定数量的噪声以掩盖各个用户的贡献。如果能够接收结果的成员在从其数据集中删除有关个人的记录后试图观察查询结果的差异，则查询结果的可变性有助于阻止识别该个人的数据。 AWS Clean Rooms Difersial Privacy 使用SampCert采样器，这是由开发的经过验证的正确采样器实现。 AWS

差分隐私是如何 AWS Clean Rooms 运作的

在完成以下工作流程时，开启差异隐私的工作流程 AWS Clean Rooms需要执行以下额外步骤 AWS Clean Rooms：

1. 在添加自定义分析规则时，您可以开启差别隐私。
2. 您为协作配置差别隐私策略，以使受差别隐私保护的数据表可供查询。

完成这些步骤后，可以查询的成员可以开始对受差异隐私保护的数据进行查询。 AWS Clean Rooms 返回符合差异隐私政策的结果。 AWS Clean Rooms Differation Privacy 会跟踪您可以运行的剩余查询的估计数量，类似于显示汽车当前燃油水平的汽车中的汽油表。可以查询的成员可以运行的查询数量受差别隐私策略中设置的隐私预算和每个查询添加的噪声参数的限制。

注意事项

在中使用差分隐私时 AWS Clean Rooms，请考虑以下几点：

- 可以接收结果的成员无法使用差别隐私。他们将为配置的表配置自定义分析规则，并关闭差别隐私。
- 如果两个或更多数据提供者都开启了差别隐私，可以查询的成员无法联接来自这些数据提供者的表。

差别隐私策略

差别隐私策略控制允许可以查询的成员在协作中运行多少个聚合函数。隐私预算定义一种通用的有限资源，该资源应用于协作中的所有表。每个查询添加的噪声控制隐私预算的耗尽速率。

需要具有差别隐私策略，才能使受差别隐私保护的表可供查询。这是协作中的一次性步骤，其中包括两个输入：

- 隐私预算 - 以 epsilon 量化，隐私预算控制隐私保护级别。这是一种通用的有限资源，应用于协作中受差别隐私保护的所有表，因为目标是保护可能在多个表中包含信息的用户的隐私。

  每次对表运行查询时，都会使用隐私预算。在隐私预算用完时，可以查询的协作成员无法运行额外的查询，直到增加或刷新隐私预算。通过设置较大的隐私预算，可以接收结果的成员可以减少他们对数据中的个人的不确定性。在咨询业务决策者后，选择一个兼顾您的协作要求和隐私需求的隐私预算。

  如果您计划定期将新数据引入到一个协作中，您可以选择每月刷新隐私预算，以在每个日历月自动创建新的隐私预算。如果选择该选项，在两次刷新之间重复查询时，可能会泄露任意数量的数据行相关信息。如果在隐私预算刷新之间重复查询相同的行，请避免选择该选项。

- 每个查询添加的噪声是根据您希望掩盖其贡献的用户数量测量的。该值控制隐私预算的耗尽速率。较大的噪声值降低隐私预算的耗尽速率，因此，允许对数据运行更多查询。不过，这会导致发布的数据见解不太准确。在设置该值时，请考虑协作见解所需的准确性。

您可以使用默认的差异隐私策略来快速完成设置或根据您的用例自定义差异隐私政策。 AWS Clean Rooms 差异隐私提供了用于配置策略的直观控件。 AWS Clean Rooms Difersition Privacy 允许您根据数据的所有查询中可能的聚合数量来预览该实用程序，并估算在数据协作中可以运行多少查询。

您可以使用交互式示例，以了解隐私预算和每个查询添加的噪声的不同值如何影响不同类型的 SQL 查询的结果。一般来说，您需要兼顾隐私需求以及要允许的查询数量和这些查询的准确性。较小的隐私预算或较大的每个查询添加的噪声可以更好地保护用户隐私，但为协作合作伙伴提供不太有意义的见解。

如果您增加隐私预算，同时将每个查询添加的噪声参数保持不变，则可以查询的成员可以在协作中对您的表运行更多的聚合。您可以在协作期间随时增加隐私预算。如果您减少隐私预算，同时将每个查询添加的噪声参数保持不变，则可以查询的成员可以运行更少的聚合。在可以查询的成员开始分析您的数据后，您无法减少隐私预算。

如果您增加每个查询添加的噪声，同时将隐私预算输入保持不变，则可以查询的成员可以在协作中对您的表运行更多的聚合。如果您减少每个查询添加的噪声，同时将隐私预算输入保持不变，则可以查询的成员可以运行更少的聚合。您可以在协作期间随时增加或减少每个查询添加的噪声。

差别隐私策略是通过隐私预算模板 API 操作管理的。

AWS Clean Rooms 差异隐私的 SQL 功能

AWS Clean Rooms 差异隐私使用通用查询结构来支持复杂的 SQL 查询。根据此结构对自定义分析模板进行验证，以确保它们可以在受差别隐私保护的表上运行。下表指示支持哪些函数。请参阅查询结构和语法了解更多信息。

Spark
SELECT
(CTEs)
SQL
ANY_VALUE
APPROXIMATE
AVG
PERCENTILE_DISC
STDDEV
AVG
SUM。
COUNT
COUNT DISTINCT
MAX

Spark
SELECT
(CTEs)
SQL
结
构

- 有MEDIAN
用函
户数
- 级
MIN
记函
录数
的
- PERCENTIL
数E_CONT
据函
您数
- 应
STDDEV_SA
该MP
CTEs
使和
用STDDEV_PO
SELECT
P
userIdent
函
ifierColu
数
mn...' 格
- 式
SUM
编和
写SUM
SELECT DISTINCT
表函
数

ark
ELECT
TEs)
的
SQL
结
构

- 达VAR_SAMP
  式和
  VAR_POP
  函
  数

Spark
SELECT
TEs)
SQL
结
构

WITH
：

Spark
SELECT
CTEs)
SQL
有
用
户
级
记
录
的
数
据。
您
应
该
CTEs
使
用`SELECT
userIdent
ifierColu
mn...' 格
式
编
写
SELECT

Spark
SELECT
TEs)
的
SQL
结
构

达
式。

Spark
SELECT
(CTEs)
SQL
- SELECT
- HAVING
- JOIN
- JOIN
- FROM
- WHERE

Spark
SELECT
(CTEs)
SQL
结
构
询。
您
只
能
在
FROM
和
JOIN
子
句
中
使
用
任
何
引
用
差
异
隐
私
关
系
的
子

Spark
SELECT
TEs)
SQL
查
询。

Spark
SELECT
(CTEs)
的
SQL
结
构

INNER
JOIN
LEFT
JOIN

- 左
半
连
接
- 左
边
反
连
接

RIGHT
JOIN

FULL
JOIN

[JOIN]

OR

Spark
SELECT
TEs)
SQL
结
构

- CROSS JOIN

在查询两个或更多开启了差别隐私的表时，必须使用这

Spark
SELECT
(CTEs)
SQL
些
函
数。
确
保
必
需
的
等
值
联
接
条
件
是
正
确
的。
确
认
表
所
有
者
在

Spark
SELECT
(CTEs)
SQL
结
构

所
有
表
中
配
置
了
相
同
的
用
户
标
识
符
列，
以
便
用
户
的
定
义
在
表

Spark
SELECT
(CTEs)
的
SQL
结
构
之
间
保
持
一
致。

在
合
并
两
个
或
更
多
开
启
了
差
别
隐
私
的
关
系

Spark
SELECT
TEs)
SQL
结
构

时，
不
支
持
CROSS
JOIN
函
数。

NION、UNI
N
L、INTER
ECT、
除
外
‖
减
号
（这
些
是
同
义
词
）

Spark
SELECT
(CTEs)
的
SQL
结
构

- AVG
窗
口
函
数
- COUNT
窗
口
函
数
- CUME_DIST
开
窗
函
数
- DENSE_RANK
窗
口

Spark
SELECT
(CTEs)
的
SQL
结
构

用函
户数
标FIRST_VAL
识UE
符窗
列口
是函
必数
填LAG
的窗
条口
件函
是数
所LAST_VALU
有E
这窗
些口
都函
支数
持。

- LEAD
  窗
  口
  函
  数

Spark
SELECT
(CTEs)
的
SQL
结
构

- MAX
  窗
  口
  函
  数
- MEDIAN
  窗
  口
  函
  数
- MIN
  窗
  口
  函
  数
- NTH_VALUE
  窗
  口
  函
  数
- STDDEV_SA
  MP
  和
  STDDEV_PO

Spark
SELECT
TEs)
SQL
结
构

P
窗
口
函
数
（STDD
EV_SAMP
和
STDDEV
是
同
义
词）

- SUM
窗
口
函
数
- VAR_SAMP
和
VAR_POP
窗
口
函

Spark
SELECT
TEs)
的
SQL
结
构

数
（VAR_
SAMP
和
VARIANCE
是
同
义
词）

排
名
函
数

- DENSE_RAN
  K
  窗
  口
  函
  数
- NTILE
  窗
  口

Spark
SELECT
(CTEs)
的
SQL
结
构

函
数

- PERCENT_R
  ANK
  开
  窗
  函
  数
- RANK
  窗
  口
  函
  数
- ROW_NUMBE
  R
  窗
  口
  函
  数

Spark
SELECT
(CTEs)
SQL

CASE
表达式
- COALESCE 表达式
- GREATEST 和 LEAST 函数
- NVL 和 COALESCE 函数
- NVL2 函数

Spark
SELECT
TEs)
的
SQL
结
构

- NULLIF
  函
  数

[illegible] Spark [illegible] SELECT [illegible] (CTEs) 的 SQL 结构		
Conditions	比较条件 • 逻辑条件 • 模式匹配条件 • BETWEEN 范围条件 • Null 条件	EXISTS 并且 IN 无法使用，因为它们需要子查询支持所有其他内容。

Spark
SELECT
TEs)
SQL
结
构

事
务
中
的
日
期
和
时
间
函
数

- 串
  联
  运
  算
  符
- ADD_MONTH
  S
  函
  数
- CONVERT_T
  IMEZONE
  函
  数

Spark
SELECT
(CTEs)
的
SQL
结
构

- CURRENT_DATE 函数
- DATEADD 函数
- DATEDIFF 函数
- DATE_PART 函数
- DATE_TRUNC 函数
- EXTRACT 函数
- TO_TIMESTAMP 函数

Spark
SELECT
TEs)
SQL
结
构

- 日
期
或
时
间
戳
函
数
的
日
期
部
分

Spark
SELECT
TEs)
的
SQL
结
构

||
(串
联)
运
算
符

- BTRIM
  函
  数
- CHAR_LENG
  TH
  函
  数
- CHARACTER
  _LENGTH
  函
  数
- CONCAT
  函
  数
- LEFT
  和
  RIGHT

Spark
SELECT
公共表表达式 (CTEs) 的 SQL 结构

函数

- LEN 函数
- LENGTH 函数
- LOWER 函数
- LPAD 和 RPAD 函数
- LTRIM 函数
- POSITION 函数

Spark
SELECT
的
SQL
结
构

- REGEXP_COUNT 函数
- REGEXP_INSTR 函数
- REGEXP_REPLACE 函数
- REGEXP_SUBSTR 函数
- REPEAT 函数
- REPLACE 函数

Spark
SELECT
公共表表达式 (CTEs) 的 SQL 结构

- REVERSE 函数
- RTRIM 函数
- SPLIT_PART 函数
- SUBSTRING 函数
- TRANSLATE 函数
- TRIM 函数
- UPPER 函数

Spark
SELECT
(CTEs)
CAST
TO_CHAR
TO_DATE
TO_NUMBER

- 日期时间格式字符串
- 数字格式字

Spark
SELECT
(CTEs)
SQL
AES_
AES_DECRYPT

- ENCODE
- DECODE
- MD5
- SHA1
- SHA2
- XX_HASH64

Spark
SELECT
(CTEs)
的
SQL
结
构

-、*、/、%

ark
ELECT
TEs)
SQL
结
构

ABS
函
数
ACOS
函
数

- ASIN
  函
  数
- ATAN
  函
  数
- ATAN2
  函
  数
- CBRT
  函
  数
- CEILING（或
  CEIL）
  函
  数

Spark
SELECT
公共表表达式 (CTEs) 的 SQL 结构

- COS 函数
- COT 函数
- DEGREES 函数
- LTRIM 函数
- EXP 函数
- FLOOR 函数
- LN 函数

Spark
SELECT
(CTEs)
的
SQL
结
构

- LOG
  函
  数
- MOD
  函
  数
- PI
  函
  数
- POWER
  函
  数
- RADIANS
  函
  数
- RANDOM
  函
  数
- ROUND
  函
  数

Spark
SELECT
(CTEs)
SQL

- SIGN
- SIN
- SQRT
- TRUNC

Spark
SELECT
(CTEs)
SQL
UNBASE64
十六进制
HLL_SKETCH_AGG，
HLL_SKETCH_ESTIMATE
HLL_UNION
HLL_UNION_AGG
JSON_JSON
GET_JSON_OBJECT

Spark
SELECT
TEs)
的
SQL
结
构

数
数组
_
数包
含

- 数
  组
  _
  不
  同
- 数
  组
  _
  除
  外
- 数
  组
  _
  相
  交
- ARRAY_JOI
  N

Spark
SELECT
TEs)

- 数组_删除
- 数组_排序
- ARRAY_UNION

Spark
SELECT
(CTEs)
SQL
结
构

ORDER
ORDER
Y
，
条
件
是
只
有
在
开
启
差
分
隐
私
的
情
况
下
查
询

Spark
SELECT
TEs)
SQL
结
构

表
时，
窗
口
函
数
的
分
区
子
句
才
支
持
ORDER
BY
子
句。

[illegible]ark [illegible] [illegible]ELECT [illegible] [illegible] [illegible] [illegible](CTEs) 的 SQL 结构

[illegible]MIT、OFFSET [illegible] [illegible]TEs 使用受差异隐私保护的表

[illegible] [illegible] 列别名

Spark
SELECT
(CTEs)
的
SQL
结
构

函
数
上
的
数
学
函
数

函
数
中
的
标
量
函
数

不支持的 SQL 构造的常见替代方案

类别	SQL 构造	或者
窗口函数	• LISTAGG • PERCENTILE_CONT • PERCENTILE_DISC	您可以将等效的聚合函数与 GROUP BY 一起使用。
数学运算符符号	• $column ||/ 2 • $column |/ 2 • $column ^ 2	• CBRT • SQRT • POWER($column, 2)
标量函数	• SYSDATE • $column::integer • convert(type, $column)	• CURRENT_DATE • CAST $column AS integer • CAST $column AS type
文本	间隔 '1 秒'	间隔 '1' 秒
行限制	TOP n	限制 n
联接	• USING • NATURAL	ON 子句应明确包含连接标准。

Differential Privacy 查询技巧和示例

AWS Clean Rooms 差异隐私使用通用查询结构来支持各种 SQL 结构，例如用于数据准备的公用表表达式 (CTEs) 和常用的聚合函数COUNT，例如、或。SUM为了通过在运行时向聚合查询结果添加噪音来混淆任何可能的用户在数据中的贡献，Difersient Privacy 要求最终SELECT statement版本中的聚合函数在用户级数据上运行。 AWS Clean Rooms

以下示例使用来自一个媒体发布者的两个名为 socialco_impressions 和 socialco_users 的表，该发布者希望使用差别隐私保护数据，同时与一个具有 athletic_brand_sales 数据的运动品牌协作。该媒体发布者已将 user_id 列配置为用户标识符列，同时在 AWS Clean Rooms中启用差别隐私。广告商不需要差异隐私保护，而是希望使用组合数据 CTEs 进行查询。由于他们的 CTE 使用受差别隐私保护的表，因此，广告商将这些受保护的表中的用户标识符列包含在 CTE 列的列表中，并根据用户标识符列联接这些受保护的表。

```
WITH matches_table AS(
     SELECT si.user_id, si.campaign_id, s.sale_id, s.sale_price
     FROM socialco_impressions si
     JOIN socialco_users su
         ON su.user_id = si.user_id
     JOIN athletic_brand_sales s
         ON s.emailsha256 = su.emailsha256
     WHERE s.timestamp > si.timestamp

UNION ALL

     SELECT si.user_id, si.campaign_id, s.sale_id, s.sale_price
     FROM socialco_impressions si
     JOIN socialco_users su
         ON su.user_id = si.user_id
     JOIN athletic_brand_sales s
         ON s.phonesha256 = su.phonesha256
     WHERE s.timestamp > si.timestamp
)

SELECT COUNT (DISTINCT user_id) as unique_users
FROM matches_table
GROUP BY campaign_id
ORDER BY COUNT (DISTINCT user_id) DESC
LIMIT 5
```

同样，如果要对受差别隐私保护的数据表运行窗口函数，您必须在 PARTITION BY 子句中包含用户标识符列。

```
ROW_NUMBER() OVER (PARTITION BY conversion_id, user_id ORDER BY match_type, match_age)
 AS row
```

AWS Clean Rooms 差异隐私的局限性

AWS Clean Rooms 差异隐私不能解决以下情况：

1. AWS Clean Rooms 差异隐私仅支持使用 Amazon S3 支持的 AWS Glue 表进行查询。它不支持使用 Snowflake 或 Amazon Athena 表进行查询。
2. AWS Clean Rooms 差异隐私无法解决定时攻击。例如，如果单个用户贡献大量的行，并且添加或删除该用户显著改变查询计算时间，则可能会受到这些攻击。

3. AWS Clean Rooms 当 SQL 查询可能由于使用某些 SQL 结构而在运行时导致溢出或无效的强制转换错误时，差异隐私不能保证差异隐私。

   下表列出了一些（但不是全部）可能会产生运行时错误而应当在分析模板中进行验证的 SQL 构造。我们建议您批准能够最大限度地减少出现此类运行时错误次数的分析模板，并定期查看查询日志，确定查询是否符合协作协议。

   以下 SQL 构造容易出现溢出错误：

类别	SQL 构造容易出现 Spark SQL 分析引擎中的溢出错误
聚合函数	• AVG • 总和/总和不同
数据类型格式设置函数	• TO_TIMESTAMP • TO_DATE
日期和时间函数	• ADD_MONTHS • DATEADD • DATEDIFF
数学函数	• +, -, *, / • POWER
字符串函数	• || • CONCAT • 重复
窗口函数	• AVG • SUM

4. CAST 数据类型格式化函数容易出现无效的强制转换错误。

   您可以配置CloudWatch 为为日志组创建指标筛选器，然后在该指标筛选器上创建 CloudWatch 警报，以便在遇到潜在的溢出或投射错误时接收警报。

具体而言，您应该关注错误代码 `CastError`、`OverflowError`、`ConversionError`。存在这些错误代码表示可能存在侧信道攻击，但也可能表示存在错误的 SQL 查询。

有关更多信息，请参阅 分析登录 AWS Clean Rooms。

AWS 无尘室机器学习

AWS Clean Rooms ML 允许两个或多个参与方在其数据上运行机器学习模型，而无需彼此共享数据。该服务提供增强隐私的控件，使数据所有者能够安全地保护自己的数据和模型 IP。您可以使用 AWS 创作模型或自带自定义模型。

有关其工作方式的更详细说明，请参阅跨账户作业。

有关 Clean Rooms 机器学习模型功能的更多信息，请参阅以下主题。

主题

- AWS Clean Rooms 机器学习术语
- AWS Clean Rooms ML 如何与 AWS 模型配合使用
- AWS Clean Rooms ML 如何使用自定义模型
- AWS Clean Rooms ML 中的模型
- Clean Rooms ML 中的自定义模型

AWS Clean Rooms 机器学习术语

使用 Clean Rooms ML 时，了解以下术语非常重要：

- 训练数据提供者 - 贡献训练数据、创建和配置相似模型并将该相似模型与一个协作关联的一方。
- 种子数据提供者 - 贡献种子数据、生成相似细分并导出其相似细分的一方。
- 训练数据 - 训练数据提供者的数据，用于生成相似模型。训练数据用于测量用户行为的相似性。

  训练数据必须包含用户 ID、项目 ID 和时间戳列。（可选）训练数据可以包含其他交互作为数值或分类特征。举例而言，交互可以是观看的视频、购买的物品或阅读的文章列表。
- 种子数据 - 种子数据提供者的数据，用于创建相似细分。种子数据可以直接提供，也可以来自 AWS Clean Rooms 查询结果。相似细分输出是训练数据中与种子用户最相似的一组用户。
- 相似模型 - 训练数据的机器学习模型，用于在其他数据集中查找相似用户。

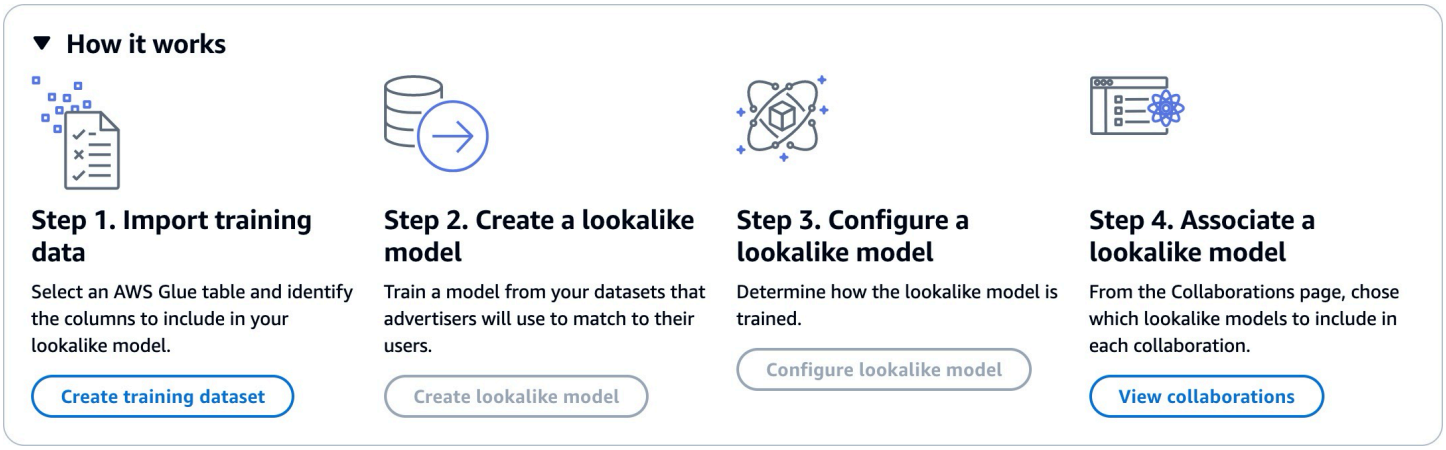
在使用 API 时，受众模型 术语等同于相似模型。例如，您可以使用 CreateAudienceModelAPI 创建外观相似的模型。

- 相似细分 - 是与种子数据最相似的训练数据子集。

  使用 API 时，您可以使用 API 创建外观相似的StartAudienceGenerationJob区段。

训练数据提供者的数据绝不会与种子数据提供者共享，并且种子数据提供者的数据绝不会与训练数据提供者共享。相似细分输出与训练数据提供者共享，但绝不会与种子数据提供者共享。

AWS Clean Rooms ML 如何与 AWS 模型配合使用



使用相似模型需要两方，即训练数据提供者和种子数据提供者，按顺序合作，将他们的数据整合到协作中 AWS Clean Rooms 。以下是训练数据提供者必须先完成的工作流程：

1. 训练数据提供者的数据必须存储在用户-项目交互 AWS Glue 的数据目录表中。训练数据必须至少包含用户 ID 列、交互 ID 列和时间戳列。
2. 训练数据提供者向注册训练数据 AWS Clean Rooms。
3. 训练数据提供者创建一个相似模型，可以将其与多个种子数据提供者共享。相似模型是一种深度神经网络，训练时间可能长达 24 小时。它不会自动重新训练，我们建议您每周重新训练一次。
4. 训练数据提供者配置相似模型，包括是否共享相关性指标以及输出细分的 Amazon S3 位置。训练数据提供者可以通过单个相似模型创建多个配置的相似模型。
5. 训练数据提供者将配置的受众模型关联到与某个种子数据提供者共享的协作。

以下是种子数据提供者接下来必须完成的工作流程：

1. 种子数据提供者的数据可以存储在 Amazon S3 存储桶中，也可以来自查询结果。

2. 种子数据提供者开启与训练数据提供者共享的协作。
3. 种子数据提供者从协作页面的“Clean Rooms ML”选项卡中创建一个相似细分。
4. 种子数据提供者可以评估相关性指标（如果已共享），并导出相似细分以在 AWS Clean Rooms外部使用。

AWS Clean Rooms ML 如何使用自定义模型

借助 Clean Rooms ML，协作成员可以使用存储在 Amazon ECR 中的 dockerized 自定义模型算法来共同分析他们的数据。为此，模型提供者必须创建图像并将其存储在 Amazon ECR 中。按照 Amazon Elastic Container Registry 用户指南中的步骤创建包含自定义 ML 模型的私有存储库。

协作中的任何成员都可以成为模型提供者，前提是他们拥有正确的权限。协作的所有成员都可以向模型贡献训练数据、推理数据或两者兼而有之。在本指南中，提供数据的成员被称为数据提供者。创建协作的成员是协作创建者，该成员可以是模型提供者，也可以是数据提供者之一，或者两者兼而有之。

在最高级别，以下是执行自定义 ML 建模必须完成的步骤：

1. 协作创建者创建协作并为每个成员分配适当的成员能力和付款配置。协作创建者必须在此步骤中将成员接收模型输出或接收推理结果的能力分配给相应的成员，因为协作创建后无法对其进行更新。有关更多信息，请参阅 在 AWS Clean Rooms ML 中创建和加入合作。
2. 模型提供者配置其容器化机器学习模型并将其与协作关联，并确保为导出的数据设置隐私约束。有关更多信息，请参阅 在 AWS Clean Rooms ML 中配置模型算法。
3. 数据提供者将其数据贡献给合作，并确保其隐私需求得到具体说明。数据提供者必须允许模型访问其数据。有关更多信息，请参阅在 AWS Clean Rooms ML 中贡献训练数据和在 AWS Clean Rooms ML 中关联配置的模型算法。
4. 协作成员创建 ML 配置，该配置定义了模型工件或推理结果的导出位置。
5. 协作成员创建一个 ML 输入通道，为训练容器或推理容器提供输入。机器学习输入通道是一个查询，用于定义要在模型算法的上下文中使用的数据。
6. 协作成员使用 ML 输入通道和配置的模型算法调用模型训练。有关更多信息，请参阅 在 AWS Clean Rooms ML 中创建经过训练的模型。
7. （可选）模型训练器调用模型导出作业，并将模型工件发送到模型结果接收器。只有具有有效 ML 配置且成员能够接收模型输出的成员才能接收模型工件。有关更多信息，请参阅 从 AWS Clean Rooms ML 中导出模型工件。
8. （可选）协作成员使用 ML 输入通道、经过训练的模型 ARN 和推理配置的模型算法调用模型推理。推理结果将发送到推理输出接收器。只有具有有效 ML 配置且成员能够接收推理输出的成员才能接收推理结果。

以下是模型提供者必须完成的步骤：

1. 创建与 A SageMaker I 兼容的 Amazon ECR docker 镜像。Clean Rooms ML 仅支持与 SageMaker AI 兼容的 docker 镜像。
2. 创建与 SageMaker AI 兼容的 docker 镜像后，将该镜像推送到 Amazon ECR。按照 Amazon 弹性容器注册表用户指南中的说明创建容器训练镜像。
3. 配置模型算法以在 Clean Rooms ML 中使用。
   a. 提供 Amazon ECR 存储库链接和配置模型算法所需的所有参数。
   b. 提供服务访问角色，允许 Clean Rooms ML 访问 Amazon ECR 存储库。
   c. 将配置的模型算法与协作关联。这包括提供隐私政策，该政策定义了对容器日志、故障日志、CloudWatch 指标的控制以及对可以从容器结果中导出多少数据的限制。

以下是数据提供者为与自定义 ML 模型协作而必须完成的步骤：

1. 使用自定义分析规则配置现有 AWS Glue 表。这允许一组特定的预先批准的查询或预先批准的账户使用您的数据。
2. 将您配置的表与协作关联，并提供可以访问您的 AWS Glue 表格的服务访问角色。
3. 向表中@@ 添加协作分析规则，允许配置的模型算法关联访问配置的表。
4. 在 Clean Rooms ML 中关联和配置模型和数据后，能够运行查询的成员提供 SQL 查询并选择要使用的模型算法。

模型训练完成后，该成员启动模型训练工件或推理结果的导出。这些工件或结果将发送给能够接收经过训练的模型输出的成员。结果接收器必须`MachineLearningConfiguration`先对其进行配置，然后才能接收模型输出。

AWS Clean Rooms ML 中的模型

AWS Clean Rooms ML 为双方提供了一种隐私保护方法，便于双方识别其数据中的相似用户，而无需彼此共享数据。第一方将训练数据带到， AWS Clean Rooms 这样他们就可以创建和配置外观相似的模型并将其与协作关联起来。然后，会将种子数据引入到协作中，以便创建与训练数据类似的相似细分。

有关其工作方式的更详细说明，请参阅跨账户作业。

以下主题提供有关如何在 Clean Rooms ML 中创建和配置 AWS 模型的信息。

主题

AWS Clean Rooms ML 的隐私保护

Clean Rooms ML 旨在降低成员身份推断攻击的风险；通过这种推断攻击，训练数据提供者可以了解哪些成员位于种子数据中，种子数据提供者可以了解哪些成员位于训练数据中。我们采取了一些措施以防范这种攻击。

首先，种子数据提供者不直接观察 Clean Rooms ML 输出，同时训练数据提供者也根本无法观察种子数据。种子数据提供者可以选择将种子数据包含在输出细分中。

接下来，通过训练数据的随机样本创建相似模型。该样本包含大量与种子受众不匹配的用户。此过程使得确定用户是否不在数据中变得更加困难，这是推断成员资格的另一种途径。

此外，可以在种子特定的相似模型训练的每个参数中使用多个种子客户。这限制了模型可以过度拟合的程度，从而限制了可以推断的用户相关数据量。因此，我们建议种子数据的最小大小为 500 个用户。

最后，一定不要向训练数据提供者提供用户级指标，这可以阻断成员身份推断攻击的另一种途径。

Clean Rooms ML 的训练数据要求

要成功创建相似模型，您的训练数据必须满足以下要求：

- 训练数据必须采用 Parquet、CSV 或 JSON 格式。

> **Note**
>
> 不支持 Zstandard (ZSTD) 压缩的 Parquet 数据。

- 您的训练数据必须编入 AWS Glue目录。有关更多信息，请参阅 AWS Glue 开发人员指南中的 AWS Glue 数据目录入门。我们建议使用 AWS Glue 爬虫来创建表，因为架构是自动推断出来的。
- 包含训练数据和种子数据的 Amazon S3 存储桶与您的其他 Clean Rooms 机器学习资源位于同一 AWS 区域。
- 训练数据必须包含至少 100,000 个独立用户 IDs ，每个用户至少有两个项目互动。
- 训练数据必须包含至少 100 万条记录。

- CreateTrainingDataset操作中指定的架构必须与创建 AWS Glue 表时定义的架构保持一致。
- 所提供的表中定义的必填字段是在 CreateTrainingDataset 操作中定义的。

字段类型	支持的数据类型	必需	说明
USER_ID	string、int、bigint	是	数据集中每个用户的唯一标识符。它应该是非个人身份信息 (PII)。这可能是经过哈希处理的标识符或客户 ID。
ITEM_ID	string、int、bigint	是	用户与之交互的每个商品的唯一标识符。
TIMESTAMP	bigint、int、timestamp	是	用户与商品交互的时间。值必须采用 Unix 纪元时间格式，以秒为单位。
CATEGORICAL_FEATURE	string、int、float、bigint、double、boolean、array	否	捕获与用户或商品相关的分类数据。这可能包括事件类型（例如点击或购

字段类型	支持的数据类型	必需	说明
			买）、用户人口统计信息（年龄组、性别 - 匿名）、用户位置（城市、国家 - 匿名）、商品类别（例如服装或电子产品）或商品品牌。
NUMERICAL_FEATURE	double、float、int、bigint	否	捕获与用户或商品相关的数值数据。这可能包括用户购买历史记录（总消费金额）、商品价格、访问某件商品的次数或用户对商品的评分。

- 或者，您最多可以提供 10 个分类或数值特征。

以下是 CSV 格式的有效训练数据集的示例

```
USER_ID,ITEM_ID,TIMESTAMP,EVENT_TYPE(CATEGORICAL FEATURE),EVENT_VALUE (NUMERICAL
 FEATURE)
196,242,881250949,click,15
186,302,891717742,click,13
```

```
22,377,878887116,click,10
244,51,880606923,click,20
166,346,886397596,click,10
```

Clean Rooms ML 的种子数据要求

相似模型的种子数据可以直接来自 Amazon S3 存储桶，也可以来自 SQL 查询结果。

直接提供的种子数据必须满足以下要求：

- 种子数据必须采用 JSON 行格式，并包含用户列表 IDs。
- 种子大小应介于 25 到 500,000 个唯一用户之间 IDs。
- 种子用户的最小数量必须与您在创建配置的受众模型时指定的最小匹配种子大小值相匹配。

以下是 CSV 格式的有效训练数据集的示例

```
{"user_id": "abc"}
{"user_id": "def"}
{"user_id": "ghijkl"}
{"user_id": "123"}
{"user_id": "456"}
{"user_id": "7890"}
```

AWS Clean Rooms 机器学习模型评估指标

Clean Rooms ML 计算召回率和相关性分数以确定模型的性能。召回率比较相似数据和训练数据之间的相似性。相关性分数用于确定受众规模应该有多大，而不是模型是否性能很好。

召回率是衡量相似细分与训练数据相似程度的公正标准。召回率是受众生成作业在种子受众中包含的训练数据样本中最相似用户的百分比（默认情况下，最相似百分比为 20%）。值范围为 0-1，值越大表示受众越好。召回值大致等于最大区间百分比就表示受众模型等同于随机选择。

我们认为这是比准确性、精度和 F1 分数更好的评估指标，因为 Clean Rooms ML 在构建模型时没有准确地标记真正的负面用户。

细分级相关性分数 是一个相似性指标，值范围从 -1（最不相似）到 1（最相似）。Clean Rooms ML 为不同的细分大小计算一组相关性分数，以帮助您确定数据的最佳细分大小。随着区段大小的增加，相

关性分数会单调降低，因此，随着区段大小的增加，它可能与种子数据不太相似。在细分级相关性分数达到 0 时，模型预测相似细分中的所有用户来自与种子数据相同的分布。增加输出大小可能会包括相似细分中来自与种子数据不同的分布的用户。

相关性分数是在单个活动中标准化的，不应用于比较不同的活动。不应将相关性分数用作任何业务结果的单一来源证据，因为除了相关性外，这些分数还会受到多个复杂因素的影响，例如库存质量、库存类型、广告投放时间等。

相关性分数不应用于判断种子质量，而应用于判断它是否可以增加或减少。考虑以下示例：

- 全部为正分 - 这表明预测为相似的输出用户比相似细分中包含的用户多。这对于属于大型市场的种子数据来说很常见，例如，过去一个月内购买过牙膏的每个人。我们建议查看较小的种子数据，例如，过去一个月内多次购买牙膏的每个人。
- 全部为负分或您所需的相似细分大小为负分 - 这表明 Clean Rooms ML 预测在所需的相似细分大小中没有足够的相似用户。这可能是因为，种子数据太具体或市场太小。我们建议为种子数据应用更少的筛选条件，或者扩大市场。例如，如果原始种子数据是购买婴儿车和汽车座椅的客户，您可以将市场扩大到购买多种婴儿产品的客户。

训练数据提供者确定是否公开相关性分数以及计算相关性分数的桶区间。

Clean Rooms ML 中的自定义模型

借助 Clean Rooms ML，协作成员可以使用存储在 Amazon ECR 中的 dockerized 自定义模型算法来共同分析他们的数据。为此，模型提供者必须创建图像并将其存储在 Amazon ECR 中。按照 Amazon Elastic Container Registry 用户指南中的步骤创建包含自定义 ML 模型的私有存储库。

协作中的任何成员都可以成为模型提供者，前提是他们拥有正确的权限。协作的所有成员都可以向模型贡献数据。在本指南中，提供数据的成员被称为数据提供者。创建协作的成员是协作创建者，该成员可以是模型提供者，也可以是数据提供者之一，或者两者兼而有之。

以下主题描述了创建自定义 ML 模型所需的信息

主题

- 自定义 ML 建模先决条件
- 训练容器的模型创作指南
- 推理容器的模型创作指南
- 接收模型日志和指标

自定义 ML 建模先决条件

在执行自定义 ML 建模之前，应考虑以下几点：

- 确定是否将在协作中同时对训练过的模型进行模型训练和推理。
- 确定每个协作成员将扮演的角色并为他们分配适当的能力。
  - 将该CAN_QUERY能力分配给将训练模型并对训练过的模型进行推理的成员。
  - 将分配CAN_RECEIVE_RESULTS给至少一名协作成员。
  - 为将分别接收训练模型导出或推理输出的成员分配CAN_RECEIVE_MODEL_OUTPUT或CAN_RECEIVE_INFERENCE_OUTPUT能力。如果您的用例需要这两种技能，则可以选择使用这两种技能。
- 确定允许导出的训练模型工件或推理结果的最大大小。
- 我们建议所有用户的角色都附加CleanrooomsFullAccess和CleanroomsMLFullAccess策略。使用自定义 ML 模型需要同时使用 AWS Clean Rooms 和 AWS Clean Rooms ML SDKs。
- 请考虑以下有关 IAM 角色的信息。
  - 所有数据提供者都必须具有服务访问角色， AWS Clean Rooms 允许从其 AWS Glue 目录和表以及底层 Amazon S3 位置读取数据。这些角色与 SQL 查询所需的角色类似。这允许您使用该CreateConfiguredTableAssociation操作。有关更多信息，请参阅 创建服务角色以创建已配置的表关联。
  - 所有想要接收指标的成员都必须具有服务访问角色，允许他们写入 CloudWatch 指标和日志。在模型训练和推理 AWS 账户 期间，Clean Rooms ML 使用此角色将所有模型指标和日志写入成员的指标。我们还提供隐私控制，以确定哪些成员有权访问指标和日志。这允许您使用该CreateMLConfiguration操作。有关更多信息，请参阅为自定义 ML 建模创建服务角色-机器学习配置。

    接收结果的成员必须为服务访问角色提供写入其 Amazon S3 存储桶的权限。此角色允许 Clean Rooms ML 将结果（经过训练的模型工件或推理结果）导出到 Amazon S3 存储桶。这允许您使用该CreateMLConfiguration操作。有关更多信息，请参阅 为自定义 ML 建模创建服务角色-机器学习配置。
  - 模型提供者必须为服务访问角色提供读取其 Amazon ECR 存储库和图像的权限。这允许您使用该CreateConfigureModelAlgorithm操作。有关更多信息，请参阅 创建服务角色以提供自定义 ML 模型。
  - 创建MLInputChannel以生成用于训练或推理的数据集的成员必须提供允许 Clean Rooms ML 在中 AWS Clean Rooms执行 SQL 查询的服务访问角色。这允许您使

用CreateTrainedModel和StartTrainedModelInferenceJob操作。有关更多信息，请参阅 创建用于查询数据集的服务角色。

- 模型作者应遵循训练容器的模型创作指南和推理容器的模型创作指南，以确保模型输入和输出按预期进行配置 AWS Clean Rooms。

训练容器的模型创作指南

本节详细介绍了模型提供者在为 Clean Rooms ML 创建自定义 ML 模型算法时应遵循的指南。

- 使用 SageMaker AI 训练支持的相应容器基础镜像，如 SageMaker AI 开发者指南中所述。以下代码允许您从公共 SageMaker AI 终端节点提取支持的容器基础镜像。

```
ecr_registry_endpoint='763104351884.dkr.ecr.$REGION.amazonaws.com'
base_image='pytorch-training:2.3.0-cpu-py311-ubuntu20.04-sagemaker'
aws ecr get-login-password --region $REGION | docker login --username AWS --password-
stdin $ecr_registry_endpoint
docker pull $ecr_registry_endpoint/$base_image
```

- 在本地创作模型时，请确保满足以下条件，以便可以在本地、开发实例、在自己的 SageMaker AI Training 和 Clean Roo AWS 账户 ms ML 上测试模型。
  - 我们建议编写一个训练脚本，通过各种环境变量访问有关训练环境的有用属性。Clean Rooms ML 使用以下参数来调用模型代码的训练：SM_MODEL_DIRSM_OUTPUT_DIRSM_CHANNEL_TRAIN、、和FILE_FORMAT。Clean Rooms ML 使用这些默认值在自己的执行环境中使用来自各方的数据训练机器学习模型。
  - Clean Rooms ML 通过 docker 容器中的/opt/ml/input/data/*channel-name*目录提供您的训练输入频道。每个 ML 输入通道均根据CreateTrainedModel请求中channel_name提供的相应通道进行映射。

```
parser = argparse.ArgumentParser()# Data, model, and output directories

parser.add_argument('--model_dir', type=str, default=os.environ.get('SM_MODEL_DIR',
 "/opt/ml/model"))
parser.add_argument('--output_dir', type=str,
 default=os.environ.get('SM_OUTPUT_DIR', "/opt/ml/output/data"))
parser.add_argument('--train_dir', type=str,
 default=os.environ.get('SM_CHANNEL_TRAIN', "/opt/ml/input/data/train"))
parser.add_argument('--train_file_format', type=str,
 default=os.environ.get('FILE_FORMAT', "csv"))
```

- 确保您能够根据模型代码中使用的协作者架构生成合成数据集或测试数据集。
- 在将模型算法与 AWS Clean Rooms 协作关联 AWS 账户 之前，请确保您可以自己运行 SageMaker AI 训练作业。

  以下代码包含与本地测试、 SageMaker AI 训练环境测试和 Clean Rooms ML 兼容的示例 Docker 文件

```
FROM  763104351884.dkr.ecr.us-west-2.amazonaws.com/pytorch-training:2.3.0-cpu-
py311-ubuntu20.04-sagemaker
MAINTAINER $author_name

ENV PYTHONDONTWRITEBYTECODE=1 \
    PYTHONUNBUFFERED=1 \
    LD_LIBRARY_PATH="${LD_LIBRARY_PATH}:/usr/local/lib"

ENV PATH="/opt/ml/code:${PATH}"

# this environment variable is used by the SageMaker PyTorch container to determine
 our user code directory
ENV SAGEMAKER_SUBMIT_DIRECTORY /opt/ml/code

# copy the training script inside the container
COPY train.py /opt/ml/code/train.py
# define train.py as the script entry point
ENV SAGEMAKER_PROGRAM train.py
ENTRYPOINT ["python", "/opt/ml/code/train.py"]
```

- 为了最好地监控容器故障，我们建议导出日志并出于故障原因进行调试。作为`GetTrainedModel`响应，Clean Rooms ML 返回了该文件中的前 1024 个字符`StatusDetails`。
- 完成所有模型更改并准备好在 SageMaker AI 环境中对其进行测试后，请按提供的顺序运行以下命令。

```
export ACCOUNT_ID=xxx
export REPO_NAME=xxx
export REPO_TAG=xxx
export REGION=xxx

docker build -t $ACCOUNT_ID.dkr.ecr.us-west-2.amazonaws.com/$REPO_NAME:$REPO_TAG

# Sign into AWS $ACCOUNT_ID/ Run aws configure
# Check the account and make sure it is the correct role/credentials
```

```
aws sts get-caller-identity
aws ecr create-repository --repository-name $REPO_NAME --region $REGION
aws ecr describe-repositories --repository-name $REPO_NAME --region $REGION

# Authenticate Doker
aws ecr get-login-password --region $REGION | docker login --username AWS --password-
stdin $ACCOUNT_ID.dkr.ecr.$REGION.amazonaws.com

# Push To ECR Images
docker push  $ACCOUNT_ID.dkr.ecr.$REGION.amazonaws.com$REPO_NAME:$REPO_TAG

# Create Sagemaker Training job
# Configure the training_job.json with
# 1. TrainingImage
# 2. Input DataConfig
# 3. Output DataConfig
aws sagemaker create-training-job --cli-input-json file://training_job.json --region
 $REGION
```

在 SageMaker AI 任务完成并且您对模型算法感到满意后，您可以使用 AWS Clean Rooms ML 注册 Amazon ECR 注册表。使用`CreateConfiguredModelAlgorithm`操作注册模型算法并将其`CreateConfiguredModelAlgorithmAssociation`与协作关联。

推理容器的模型创作指南

本节详细介绍了模型提供者在为 Clean Rooms ML 创建推理算法时应遵循的指南。

- 按照《 SageMaker AI 开发者指南》中所述，使用支持人工智能推理的相应容器基础镜像。SageMaker 以下代码允许您从公共 SageMaker AI 终端节点提取支持的容器基础镜像。

```
ecr_registry_endpoint='763104351884.dkr.ecr.$REGION.amazonaws.com'
base_image='pytorch-inference:2.3.0-cpu-py311-ubuntu20.04-sagemaker'
aws ecr get-login-password --region $REGION | docker login --username AWS --password-
stdin $ecr_registry_endpoint
docker pull $ecr_registry_endpoint/$base_image
```

- 在本地创作模型时，请确保满足以下条件，以便可以在本地、开发实例、您的 SageMaker AI Batch Transform 和 Clean Rooms ML 上测试模型。 AWS 账户
  - Clean Rooms ML 通过 docker 容器中的`/opt/ml/model`目录使推理中的模型工件可供推理代码使用。

- Clean Rooms ML 按行拆分输入，使用`MultiRecord`批处理策略，并在每条转换后的记录的末尾添加一个换行符。
- 确保您能够根据将在模型代码中使用的协作者的架构生成合成或测试推理数据集。
- 在将模型算法与 AWS Clean Rooms 协作关联 AWS 账户 之前，请确保您可以自己运行 SageMaker AI 批量转换作业。

  以下代码包含与本地测试、 SageMaker AI 转换环境测试和 Clean Rooms ML 兼容的示例 Docker 文件

```
FROM 763104351884.dkr.ecr.us-east-1.amazonaws.com/pytorch-inference:1.12.1-cpu-
py38-ubuntu20.04-sagemaker

ENV PYTHONUNBUFFERED=1

COPY serve.py /opt/ml/code/serve.py
COPY inference_handler.py /opt/ml/code/inference_handler.py
COPY handler_service.py /opt/ml/code/handler_service.py
COPY model.py /opt/ml/code/model.py

RUN chmod +x /opt/ml/code/serve.py

ENTRYPOINT ["/opt/ml/code/serve.py"]
```

- 完成所有模型更改并准备好在 SageMaker AI 环境中对其进行测试后，请按提供的顺序运行以下命令。

```
export ACCOUNT_ID=xxx
export REPO_NAME=xxx
export REPO_TAG=xxx
export REGION=xxx

docker build -t $ACCOUNT_ID.dkr.ecr.us-west-2.amazonaws.com/$REPO_NAME:$REPO_TAG

# Sign into AWS $ACCOUNT_ID/ Run aws configure
# Check the account and make sure it is the correct role/credentials
aws sts get-caller-identity
aws ecr create-repository --repository-name $REPO_NAME --region $REGION
aws ecr describe-repositories --repository-name $REPO_NAME --region $REGION

# Authenticate Docker
```

```
aws ecr get-login-password --region $REGION | docker login --username AWS --password-
stdin $ACCOUNT_ID.dkr.ecr.$REGION.amazonaws.com

# Push To ECR Repository
docker push $ACCOUNT_ID.dkr.ecr.$REGION.amazonaws.com$REPO_NAME:$REPO_TAG

# Create Sagemaker Model
# Configure the create_model.json with
# 1. Primary container -
    # a. ModelDataUrl - S3 Uri of the model.tar from your training job
aws sagemaker create-model --cli-input-json file://create_model.json --region $REGION

# Create Sagemaker Transform Job
# Configure the transform_job.json with
# 1. Model created in the step above
# 2. MultiRecord batch strategy
# 3. Line SplitType for TransformInput
# 4. AssembleWith Line for TransformOutput
aws sagemaker create-transform-job --cli-input-json file://transform_job.json --
region $REGION
```

在 SageMaker AI 任务完成并且您对批量转换感到满意后，您可以使用 AWS Clean Rooms ML 注册 Amazon ECR 注册表。使用`CreateConfiguredModelAlgorithm`操作注册模型算法并将其`CreateConfiguredModelAlgorithmAssociation`与协作关联。

接收模型日志和指标

要接收来自自定义模型训练或推理的日志和指标，成员必须创建具有提供必要 CloudWatch 权限的有效角色的 ML 配置（请参阅为自定义 ML 建模创建服务角色——机器学习配置）。

系统指标

训练和推理的系统指标（例如 CPU 和内存利用率）通过有效的机器学习配置发布给所有成员。随着任务的进展，可以分别通过`/aws/cleanroomsml/TrainedModels`或`/aws/cleanroomsml/TrainedModelInferenceJobs`命名空间中的 CloudWatch指标查看这些指标。

模型日志

模型日志的访问权限由每个已配置的模型算法的隐私配置策略提供。模型作者在将配置的模型算法（通过控制台或 `CreateConfiguredModelAlgorithmAssociation` API）关联到协作时设置隐私配置策略。设置隐私配置策略可控制哪些成员可以接收模型日志。

此外，模型作者可以在隐私配置策略中设置过滤器模式来过滤日志事件。模型容器发送到stdout或stderr且符合筛选模式（如果已设置）的所有 CloudWatch 日志都将发送到 Amazon Logs。模型日志分别在 CloudWatch 日志组/aws/cleanroomsml/TrainedModels或/aws/cleanroomsml/TrainedModelInferenceJobs中可用。

自定义指标

在配置模型算法（通过控制台或 CreateConfiguredModelAlgorithm API）时，模型作者可以在输出日志中提供要搜索的特定指标名称和正则表达式语句。这些可以在任务进行时通过/aws/cleanroomsml/TrainedModels命名空间中的 CloudWatch 指标进行查看。关联已配置的模型算法时，模型作者可以在指标隐私配置中设置可选的噪声级别，以避免输出原始数据，同时仍然提供对自定义指标趋势的可见性。如果设置了噪音水平，则指标将在作业结束时发布，而不是实时发布。

加密计算 Clean Rooms

加密计算 Clean Rooms (C3R) 是一种除了分析 AWS Clean Rooms规则之外还可以使用的功能。借助 C3R，组织可以将敏感数据整合在一起，从数据分析中获得新的见解，同时以加密方式限制任何一方在流程中可以了解到的信息。C3R 可供想要协作处理其敏感数据但只需要在云中使用加密数据的两方或多方使用。

C3R 加密客户端是一种客户端加密工具，您可以使用它来加密数据以供使用。 AWS Clean Rooms使用 C3R 加密客户端时，数据在协作中使用时仍会受到加密保护。 AWS Clean Rooms 与常规 AWS Clean Rooms 协作一样，输入数据是关系数据库表，计算以 SQL 查询表示。但是，C3R 仅支持对加密数据的有限 SQL 查询子集。

具体而言，C3R 支持 SQL JOIN 以及 SELECT 关于受加密保护的数据的声明。输入表中的每列只能用于以下 SQL 语句类型之一：

- 受加密保护的列，可用于 JOIN 语句被称为 fingerprint 列。
- 受加密保护的列，可用于 SELECT 语句被称为 sealed 列。
- 未受加密保护的列，无法用于 JOIN 或 SELECT 语句被称为 cleartext 列。

在某些情况下，GROUP BY 支持语句 fingerprint 列。有关更多信息，请参阅 Fingerprint 列。目前，C3R 不支持在加密数据上使用其他 SQL 结构，例如 WHERE 子句或聚合函数，比如 SUM 以及 AVERAGE, 即使相关分析规则允许这样做.

C3R 旨在保护表中各个单元格中的数据。使用 C3R 的默认配置，当内容在 AWS Clean Rooms中使用时，客户通过协作向第三方提供的底层数据将保持加密。C3R 对所有人使用行业标准 AES-GCM 加密

sealed 列和行业标准的伪随机函数，称为基于哈希的消息身份验证码 (HMAC)，用于保护 fingerprint 列。

尽管 C3R 会对表中的数据进行加密，但仍可以推断出以下信息：

- 有关表本身的信息，包括表中的列数、列名和行数。
- 与大多数标准加密形式一样，C3R 不会尝试隐藏加密值的长度。C3R 确实提供了填充加密值以隐藏明文确切长度的功能。但是，仍然可以向另一方揭示每列明文长度的上限。
- 日志级别的信息，例如何时将特定行添加到加密的 C3R 表中。

有关 C3R 的更多信息，请参阅以下主题。

主题

使用 Clean Rooms 加密计算时的注意事项

Clean Rooms 加密计算 (C3R) 旨在最大限度地保护数据。但是，某些使用案例可能会受益于较低级别的数据保护，以换取额外的功能。您可以通过修改 C3R 最安全的配置来做出这些特定的权衡。作为客户，您应该了解这些权衡，并确定它们是否适合您的使用案例。要考虑的权衡包括以下内容：

主题

有关如何为这些场景设置参数的更多信息，请参阅加密计算参数。

允许在表中混合 cleartext 和加密数据

对所有数据进行客户端加密可最大限度地保护数据。但是，这限制了某些类型的查询（例如，SUM 聚合函数）。允许 cleartext 数据的风险在于，任何有权访问加密表的人都可以推断出一些有关加密值的信息。这可以通过对 cleartext 和关联数据进行统计分析来实现。

例如，假设您的列为 `City` 和 `State`。`City` 列为 cleartext，`State` 列加密。当您看到 `City` 列中的 `Chicago` 值时，这有助于您确定 `State` 很有可能是 `Illinois`。相比之下，如果一列是 `City`，另一列是 `EmailAddress`，则 cleartext `City` 不太可能揭示加密 `EmailAddress` 的任何信息。

有关此场景的参数的更多信息，请参阅允许 cleartext 列参数。

允许 fingerprint 列中有重复值

对于最安全的方法，我们假设任何 fingerprint 列都只包含一个变量实例。fingerprint 列中的任何项目都不能重复。C3R 加密客户端将这些 cleartext 值映射为与随机值无法区分的唯一值。因此，不可能从这些随机值中推断出 cleartext 信息。

fingerprint 列中有重复值的风险在于，重复的值会导致重复的随机值。因此，从理论上讲，任何有权访问加密表的人都可以对可能揭示 cleartext 值信息的 fingerprint 列进行统计分析。

同样，假设 fingerprint 列是 `State`，并且表中的每一行都对应一个美国家庭。通过频率分析，人们很有可能推断出哪个州是 `California`，哪个州是 `Wyoming`。这种推断是可能的，因为 `California` 的居民人数远远超过 `Wyoming`。相比之下，假设 fingerprint 列位于家庭标识符上，在包含数百万个条目的数据库中，每个家庭出现 1 到 4 次。频率分析不太可能揭示任何有用的信息。

有关此场景的参数的更多信息，请参阅“允许重复”参数。

放宽对 fingerprint 列命名方式的限制

默认情况下，我们假设当使用加密 fingerprint 列联接两个表时，这些列在每个表中的名称相同。此结果的技术原因是，默认情况下，我们派生出不同的加密密钥来加密每个 fingerprint 列。该密钥源自协作共享密钥和列名的组合。如果我们尝试联接具有不同列名的两列，则会派生出不同的密钥，并且无法计算出有效的联接。

要解决这个问题，可以关闭从每个列名派生密钥的功能。然后，C3R 加密客户端对所有 fingerprint 列使用一个派生密钥。风险在于可以进行另一种可能揭示信息的频率分析。

让我们再次以 City 和 State 为例。如果我们为每个 fingerprint 列派生相同的随机值（不包含列名）。New York 在 City 和 State 列中的随机值相同。纽约是美国为数不多的 City 名称与 State 名称相同的城市之一。相比之下，如果数据集的每一列都有完全不同的值，则不会泄露任何信息。

有关此场景的参数的更多信息，请参阅"允许对具有不同名称的列进行 JOIN"参数。

确定 NULL 值的表示方式

您可以选择是否像处理其他值一样对 NULL 值进行加密处理（加密和 HMAC）。如果您不像处理其他值一样处理 NULL 值，可能会揭示信息。

例如，假设 cleartext 中 Middle Name 列中的 NULL 表示没有中间名的人。如果您不加密这些值，则会泄露加密表中哪些行用于没有中间名的人。对于某些人群中的某些人来说，这些信息可能是一个识别信号。但是，如果您对 NULL 值进行加密处理，某些 SQL 查询的行为就会有所不同。例如，GROUP BY 子句不会将 fingerprint 列中的 fingerprintNULL 值分组在一起。

有关此场景的参数的更多信息，请参阅"保留 NULL 值"参数。

Clean Rooms 加密计算中支持的文件和数据类型

C3R 加密客户端可识别以下文件类型：

- CSV 文件
- Parquet 文件

您可以在 C3R 加密客户端中使用 --fileFormat 标志来明确指定文件格式。如果明确指定，则文件格式不取决于文件扩展名。

主题

CSV 文件

假定扩展名为 .csv 的文件采用 CSV 格式并包含 UTF-8 编码的文本。C3R 加密客户端将所有值视为字符串。

.csv 文件中支持的属性

C3R 加密客户端要求 .csv 文件具有以下属性：

- 可能包含也可能不包含唯一命名每列的初始标题行。
- 逗号分隔。（目前，不支持自定义分隔符。）
- UTF-8 编码的文本。

从 .csv 条目中修剪空格

.csv 条目中的前导和尾部空格都会被修剪。

.csv 文件的自定义 NULL 编码

.csv 文件可以使用自定义 NULL 编码。

使用 C3R 加密客户端，您可以使用 `--csvInputNULLValue=<csv-input-null>` 标志为输入数据中的 NULL 条目指定自定义编码。通过使用 `--csvOutputNULLValue=<csv-output-null>` 标志，C3R 加密客户端可以在生成的输出文件中为 NULL 条目使用自定义编码。

> **Note**
>
> NULL 条目被认为缺少 内容，特别是在 SQL 表等更丰富的表格格式的上下文中。尽管由于历史原因 .csv 并不明确支持这种描述，但通常的惯例是将仅包含空格的空条目视为 NULL。因此，这是 C3R 加密客户端的默认行为，可以根据需要对其进行自定义。

C3R 如何解释 .csv 条目

下表举例说明了如何根据为 `--csvInputNULLValue=<csv-input-null>` 和 `--csvOutputNULLValue=<csv-output-null>` 标志提供的值（如果有）编组 .csv 条目（为清楚起见，cleartext 至 cleartext）。在 C3R 解释任何值的含义之前，将修剪引号之外的前导和尾部空格。

<csv-input-null>	**<csv-output-null>**	输入条目	输出条目
无	无	`,AnyProduct,`	`,AnyProduct,`
无	无	`, AnyProduct ,`	`,AnyProduct,`

<csv-input-null>	**<csv-output-null>**	输入条目	输出条目
无	无	,"AnyProduct",	,AnyProduct,
无	无	, "AnyProduct" ,	,AnyProduct,
无	无	,,	,,
无	无	, ,	,,
无	无	,"",	,,
无	无	," ",	," ",
无	无	, " " ,	," ",
"AnyProduct"	"NULL"	,AnyProduct,	,NULL,
"AnyProduct"	"NULL"	, AnyProduct ,	,NULL,
"AnyProduct"	"NULL"	,"AnyProduct",	,NULL,
"AnyProduct"	"NULL"	, "AnyProduct" ,	,NULL,
无	"NULL"	,,	,NULL,
无	"NULL"	, ,	,NULL,
无	"NULL"	,"",	,NULL,
无	"NULL"	," ",	," ",
无	"NULL"	, " " ,	," ",
""	"NULL"	,,	,NULL,
""	"NULL"	, ,	,NULL,
""	"NULL"	,"",	,"",

<csv-input-null>	**<csv-output-null>**	输入条目	输出条目
""	"NULL"	," ",	," ",
""	"NULL"	, " " ,	," ",
"\"\""	"NULL"	,,	,,
"\"\""	"NULL"	, ,	,,
"\"\""	"NULL"	,"",	,NULL,
"\"\""	"NULL"	," ",	," ",
"\"\""	"NULL"	, " " ,	," ",

不带标题的 CSV 文件

源 .csv 文件不必在第一行中包含标题来唯一命名每列。但是，没有标题行的 .csv 文件需要位置加密架构。需要的是位置加密架构，而不是带标题行的 .csv 文件和 Parquet 文件使用的典型映射架构。

位置加密架构按位置而不是按名称指定输出列。映射加密架构将源列名映射到目标列名。有关更多信息，包括两种架构格式的详细讨论和示例，请参阅映射和定位表架构。

Parquet 文件

假定扩展名为 .parquet 的文件采用 Apache Parquet 格式。

支持的 Parquet 数据类型

C3R 加密客户端可以在表示 AWS Clean Rooms支持的数据类型的 Parquet 文件中处理任何非复杂（即基本类型）数据。

但是，只能将字符串列用于 sealed 列。

支持以下 Parquet 数据类型：

- 带以下逻辑注释的 `Binary` 基本类型：
  - 如果已设置 `--parquetBinaryAsString`（STRING 数据类型），则为 None

  - Decimal(scale, precision)（DECIMAL 数据类型）
  - String（STRING 数据类型）
- 不带逻辑注释的 Boolean 基本数据类型（BOOLEAN 数据类型）
- 不带逻辑注释的 Double 基本数据类型（DOUBLE 数据类型）
- 带 Decimal(scale, precision) 逻辑注释的 Fixed_Len_Binary_Array 基本类型（DECIMAL 数据类型）
- 不带逻辑注释的 Float 基本数据类型（FLOAT 数据类型）
- 带以下逻辑注释的 Int32 基本类型：
  - 无（INT 数据类型）
  - Date（DATE 数据类型）
  - Decimal(scale, precision)（DECIMAL 数据类型）
  - Int(16, true)（SMALLINT 数据类型）
  - Int(32, true)（INT 数据类型）
- 带以下逻辑注释的 Int64 基本数据类型：
  - 无（BIGINT 数据类型）
  - Decimal(scale, precision)（DECIMAL 数据类型）
  - Int(64, true)（BIGINT 数据类型）
  - Timestamp(isUTCAdjusted, TimeUnit.MILLIS)（TIMESTAMP 数据类型）
  - Timestamp(isUTCAdjusted, TimeUnit.MICROS)（TIMESTAMP 数据类型）
  - Timestamp(isUTCAdjusted, TimeUnit.NANOS)（TIMESTAMP 数据类型）

加密非字符串值

当前，sealed 列仅支持字符串值。

对于 .csv 文件，C3R 加密客户端会将所有值视为 UTF-8 编码的文本，并且在加密之前不会尝试对其进行不同的解释。

对于指纹列，类型被分为等价类。等价类是一组数据类型，可以通过代表性数据类型明确比较其是否相等。

等价类允许将相同的指纹分配给相同的语义值，而不管原始表示形式如何。但是，两个等价类中的相同值不会生成相同的指纹列。

例如，无论 INTEGRAL 值 42 最初是 SMALLINT、INT 还是 BIGINT，都将为其分配相同的指纹。此外，INTEGRAL 值 0 永远不会与 BOOLEAN 值 FALSE 匹配（由值 0 表示）。

指纹列支持以下等价类和相应 AWS Clean Rooms 的数据类型：

等价类	支持的 AWS Clean Rooms 数据类型
BOOLEAN	BOOLEAN
DATE	DATE
INTEGRAL	BIGINT, INT, SMALLINT
STRING	CHAR, STRING, VARCHAR

Clean Rooms 加密计算中的列名

默认情况下，在 Clean Rooms 加密计算中，列的名称很重要。

如果允许对具有不同名称的列进行 JOIN 参数的值为 false，则在加密 fingerprint 列时将使用列名。因此，默认情况下，协作者必须事先进行协调，并对将在查询中使用 JOIN 语句的数据使用相同的目标列名称。默认情况下，为 JOIN 加密的列如果名称不同，就不能成功地对任何值进行 JOIN。

如果允许对具有不同名称的列进行 JOIN 参数的值为 true，则跨加密为 fingerprint 列的列的 JOIN 语句会成功。使用此参数加密数据可能允许对 cleartext 值进行一些推断。例如，如果某行的 `City` 列和 `State` 列中都具有相同的 HMAC 散列消息认证码值，则该值可能为 `New York`。

列标题名称的标准化

列标题名称由 C3R 加密客户端进行标准化。所有前导和尾随的空格都将被删除，转换后的输出将列名改为小写。

标准化应用于可能受列名影响的所有其他计算或其他运算。发出的输出文件仅包含标准化名称。

Clean Rooms 加密计算中的列类型

本主题提供有关 Clean Rooms 加密计算中列类型的信息。

主题

- Fingerprint 列
- 密封列
- Cleartext 列

Fingerprint 列

Fingerprint 列是在 JOIN 语句中使用的受加密保护的列。

fingerprint 列中的数据无法解密。只有密封列中的数据才能解密。

Fingerprint 列只能在以下 SQL 子句和函数中使用：

- JOIN (INNER, OUTER, LEFT, RIGHT, or FULL) 与其他 fingerprint 列对比：
  - 如果将 `allowJoinsOnColumnsWithDifferentNames` 参数的值设置为 `false`，则 JOIN 的两个 fingerprint 列的名称也必须相同。
- `SELECT COUNT()`
- `SELECT COUNT(DISTINCT )`
- `GROUP BY`（仅当协作将 `preserveNulls` 参数的值设置为 `true` 时才使用。）

违反这些限制条件的查询可能会生成不正确的结果。

密封列

密封列 是 SELECT 语句中使用的通过加密保护的列。

密封列只能在以下 SQL 子句和函数中使用：

- `SELECT`
- `SELECT ... AS`
- `SELECT COUNT()`

> **Note**
>
> 不支持 `SELECT COUNT(DISTINCT )`。

违反这些限制条件的查询可能会生成不正确的结果。

加密前为 sealed 列填充数据

当您指定列应该是 sealed 列时，C3R 会询问您要选择哪种填充。加密前填充数据是可选的。如果不使用填充（填充类型为 none），则加密数据的长度表示 cleartext 的大小。在某些情况下，cleartext 的大小可能会暴露明文。如果使用填充（填充类型为 `fixed` 或 `max`），则先将所有值填充到常见大小，然后再加密。使用填充时，加密数据的长度除了给出其大小的上限外，不提供有关原始 cleartext 长度的信息。

如果要为列填充并且已知该列中数据的最大字节长度，请使用 `fixed` 填充。使用至少与该列中最长值的字节长度一样大的 `length` 值。

> **Note**
>
> 如果值长于提供的 `length` 值，则会发生错误并导致加密失败。

如果要为列填充并且未知该列中数据的最大字节长度，请使用 `max` 填充。这种填充模式将所有数据填充到最长值的长度加上额外的 `length` 字节。

> **Note**
>
> 您可能需要批量加密数据，或者定期使用新数据更新表。请注意，`max` 填充会将条目填充到给定批次中最长的明文条目的长度（加 `length` 字节）。这意味着加密文字长度可能因批次而异。因此，如果您知道列的最大字节长度，则应使用 `fixed` 而不是 `max`。

Cleartext 列

Cleartext 列是在 JOIN 或 SELECT 语句中使用的未受加密保护的列。

Cleartext 列可以用于 SQL 查询的任何部分。

加密计算参数

在创建协作时，可使用 Clean Rooms 加密计算 (C3R) 为协作提供加密计算参数。您可以使用 AWS Clean Rooms 控制台或 `CreateCollaboration` API 操作创建协作。在控制台中，启用支持加密计算选项后，可以为加密计算参数中的参数设置值。有关更多信息，请参阅以下主题。

主题

- 允许 cleartext 列参数

- “允许重复”参数
- “允许对具有不同名称的列进行 JOIN”参数
- “保留 NULL 值”参数

允许 cleartext 列参数

在控制台中，您可以在创建协作时设置允许 cleartext 列参数，以指定是否允许在包含加密数据的表中包含 cleartext 数据。

下表描述了允许 cleartext 列参数的值。

参数值	说明
否	加密表中不允许有 Cleartext 列。所有数据都受到加密保护。
是	加密表中允许有 Cleartext 列。 Cleartext 列不受加密保护，包含为 cleartext。您应该注意行中的 cleartext 数据可能揭示了表中其他数据的哪些信息。 要在特定列上运行 SUM 或 AVG，这些列必须是 cleartext。

使用 `CreateCollaboration` API 操作，对于 `dataEncryptionMetadata` 参数，您可以将 `allowCleartext` 的值设置为 `true` 或 `false`。有关 API 操作的更多信息，请参阅 AWS Clean Rooms API 参考。

Cleartext 列对应于按表特定架构分类为 cleartext 的列。这些列中的数据未加密，可以以任何方式使用。 Cleartext如果需要比加密列或fingerprint列所允许的更大的灵活性， and/or 则如果数据不敏感，则sealed列可能很有用。

“允许重复”参数

在控制台中，您可以在创建协作时设置允许重复参数，以指定为 JOIN 查询加密的列是否可以包含重复的非 NULL 值。

> **Important**
>
> 允许重复、允许对不同名称列进行 JOIN 和保留 NULL 值参数具有单独但相关的效果。

下表描述了允许重复参数的值。

参数值	说明
否	fingerprint 列中不允许有重复的值。单个 fingerprint 列中的所有值都必须是唯一的。
是	fingerprint 列中允许有重复的值。 如果需要联接具有重复值的列，请将此值设置为是。如果设置为是，则 C3R 表或结果的 fingerprint 列中出现的频率模式可能意味着有关 cleartext 数据结构的一些其他信息。

使用 `CreateCollaboration` API 操作，对于 `dataEncryptionMetadata` 参数，您可以将 `allowDuplicates` 的值设置为 `true` 或 `false`。有关 API 操作的更多信息，请参阅 AWS Clean Rooms API 参考。

默认情况下，如果必须在 JOIN 查询中使用加密数据，则 C3R 加密客户端要求这些列没有重复值。此要求是为了加强数据保护。这种行为可以帮助确保无法观察到数据中的重复模式。但是，如果您想在 JOIN 查询中使用加密数据并且不担心重复值，则允许重复参数可以禁用此保守检查。

“允许对具有不同名称的列进行 JOIN”参数

在控制台中，您可以在创建协作时设置允许对具有不同名称的列进行 JOIN 参数，以指定是否支持具有不同名称的列之间的 JOIN 语句。

有关更多信息，请参阅 列标题名称的标准化。

下表描述了允许对具有不同名称的列进行 JOIN 参数的值。

参数值	说明
否	不支持联接具有不同名称的 fingerprint 列。JOIN 语句仅在具有相同名称的列上提供准确的结果。 **Important** 否值可提高信息安全性，但要求协作参与者事先就列名达成共识。如果两列在加密为 fingerprint 列时具有不同

参数值	说明
	的名称，并且允许对具有不同名称的列进行 JOIN 设置为否，则对这些列的 JOIN 语句不会生成任何结果。这是因为它们之间不共享加密后的值。
是	支持联接具有不同名称的 fingerprint 列。为了提高灵活性，用户可以将此值设置为是，这样无论列名如何，都允许对列执行 JOIN 语句。 如果设置为是，则 C3R 加密客户端在保护 fingerprint 列时将不会考虑列名。因此，在 C3R 表中可以观察到不同 fingerprint 列的共同值。 例如，如果一行在 `City` 列和 `State` 列中都具有相同的加密 JOIN 值，则可以合理地推断出该值为 `New York`。

使用 `CreateCollaboration` API 操作，对于 `dataEncryptionMetadata` 参数，您可以将 `allowJoinsOnColumnsWithDifferentNames` 的值设置为 `true` 或 `false`。有关 API 操作的更多信息，请参阅 AWS Clean Rooms API 参考。

默认情况下，fingerprint 列加密受该列的 `targetHeader` 所影响，它在步骤 4：为表格文件生成加密架构中设置。因此，同一个 cleartext 值在每个不同的 fingerprint 列中都有不同的加密表示。

在某些情况下，此参数可用于防止推断 cleartext 值。例如，在 fingerprint 列中看到相同的加密值时，可以通过 `City` 和 `State` 来合理地推断该值是 `New York`。但是，使用此参数需要事先进行额外的协调，以便查询中要联接的所有列都具有共享名称。

您可以使用允许对具有不同名称的列进行 JOIN 参数来放宽此限制。当参数值设置为 `Yes` 时，无论名称如何，都允许一起使用为 JOIN 加密的任何列。

“保留 NULL 值”参数

在控制台中，您可以在创建协作时设置保留 NULL 值参数，以指示该列不存在任何值。

下表描述了保留 NULL 值参数的值。

参数值	说明
否	NULL 值不会被保留。NULL 值在加密表中不会显示为 NULL。NULL 值在 C3R 表中显示为唯一的随机值。
是	NULL 值会被保留。NULL 值在加密表中显示为 NULL。如果您需要 NULL 值的 SQL 语义，则可以将此值设置为是。因此，无论列是否加密以及允许重复的参数设置如何，NULL 条目在 C3R 表中都会显示为 NULL。

使用 `CreateCollaboration` API 操作，对于 `dataEncryptionMetadata` 参数，您可以将 `preserveNulls` 的值设置为 `true` 或 `false`。有关 API 操作的更多信息，请参阅 AWS Clean Rooms API 参考。

当协作的保留 NULL 值参数设置为否时：

1. `cleartext` 列中的 NULL 条目保持不变。
2. 加密 `fingerprint` 列中的 NULL 条目被加密为随机值以隐藏其内容。在 cleartext 中联接带有 NULL 条目的加密列不会生成任何 NULL 条目的任何匹配项。不会进行任何匹配，因为它们各自会收到自己的唯一随机内容。
3. 加密 `sealed` 列中的 NULL 条目已加密。

当协作的保留 NULL 值参数的值设置为是时，无论列是否加密，所有列中的 NULL 条目都将保持为 NULL。

保留 NULL 值参数在数据扩充等情况下非常有用，在这些情况下，您需要共享以 NULL 表示的信息缺失。在 fingerprint 或 HMAC 格式中，如果要进行 JOIN 或 GROUP BY 的列中有 NULL 值，保留 NULL 值参数也很有用。

如果允许重复和保留 NULL 值参数的值设置为否，则在 fingerprint 列中包含多个 NULL 条目会产生错误并停止加密。如果任一参数的值设置为是，则不会发生此类错误。

Clean Rooms 加密计算中的可选标志

以下各节描述了在使用 C3R 加密客户端加密数据以进行表格文件自定义和测试时可以设置的可选标志。

主题

`--csvInputNULLValue` 标志

使用 C3R 加密客户端加密数据时，您可以使用 `--csvInputNULLValue` 标志为输入数据中的 NULL 条目指定自定义编码。

下表总结了此标志的用法和参数。

用法	参数
可选。用户可以为输入数据中的 NULL 条目指定自定义编码。	输入 CSV 文件中 NULL 值的用户指定编码

NULL 条目是被视为缺少内容的条目，特别是在 SQL 表等更丰富的表格格式的上下文中。尽管由于历史原因 .csv 并不明确支持这种描述，但通常的惯例是将仅包含空格的空条目视为 NULL。因此，这是 C3R 加密客户端的默认行为，可以根据需要对其进行自定义。

`--csvOutputNULLValue` 标志

使用 C3R 加密客户端加密数据时，您可以使用 `--csvOutputNULLValue` 标志为输出数据中的 NULL 条目指定自定义编码。

下表总结了此标志的用法和参数。

用法	参数
可选。用户可以在生成的输出文件中为 NULL 条目指定自定义编码。	输入 CSV 文件中 NULL 值的用户指定编码

NULL 条目是被视为缺少内容的条目，特别是在 SQL 表等更丰富的表格格式的上下文中。尽管由于历史原因 .csv 并不明确支持这种描述，但通常的惯例是将仅包含空格的空条目视为 NULL。因此，这是 C3R 加密客户端的默认行为，可以根据需要对其进行自定义。

`--enableStackTraces` 标志

使用 C3R 加密客户端加密数据时，可以使用 `--enableStackTraces` 标志提供其他上下文信息，以便在 C3R 遇到错误时报告错误。

AWS 不收集错误。如果遇到错误，请使用堆栈跟踪自行解决错误，或者将堆栈跟踪发送到 支持 寻求帮助。

下表总结了此标志的用法和参数。

用法	参数
可选。用于提供其他上下文信息，以便在 C3R 加密客户端遇到错误时报告错误。	无

`--dryRun` 标志

加密和解密 C3R 加密客户端命令包括一个可选的 `--dryRun` 标志。该标志采用用户提供的所有参数，并检查它们的有效性和一致性。

您可以使用 `--dryRun` 标志来检查您的架构文件是否有效且与其相应的输入文件一致。

下表总结了此标志的用法和参数。

用法	参数
可选。使 C3R 加密客户端解析参数和检查文件，但不执行加密或解密。	无

`--tempDir` 标志

您可能需要使用临时目录，因为加密文件有时可能比非加密文件大，具体取决于它们的设置。每个协作还必须对数据集进行加密才能正常工作。

使用 C3R 加密数据时，可以使用 `--tempDir` 标志来指定在处理输入时可以创建临时文件的位置。

下表总结了此标志的用法和参数。

用法	参数
用户可以指定在处理输入时可以创建临时文件的位置。	默认为系统临时目录。

使用 Clean Rooms 加密计算进行查询

本主题提供有关编写查询的信息，这些查询使用已使用 Clean Rooms 加密计算加密的数据表。

主题

- 在 NULL 上分支的查询
- 将一个源列映射到多个目标列
- 在 JOIN 和 SELECT 查询中使用相同的数据

在 NULL 上分支的查询

要在 NULL 语句上设置查询分支，需要使用 `IF x IS NULL THEN 0 ELSE 1` 这样的语法。

查询总是可以在 cleartext 列中的 NULL 语句上分支。

只有当保留 NULL 值参数 (`preserveNulls`) 的值设置为 `true` 时，查询才能在 sealed 列和 fingerprint 列中的 NULL 语句上分支。

违反这些限制条件的查询可能会生成不正确的结果。

将一个源列映射到多个目标列

将一个源列映射到多个目标列。例如，您可能希望在一列上同时进行 JOIN 和 SELECT。

有关更多信息，请参阅 在 JOIN 和 SELECT 查询中使用相同的数据。

在 JOIN 和 SELECT 查询中使用相同的数据

如果列中的数据不敏感，则它可以出现在 cleartext 目标列中，这样就可以用于任何目的。

如果列中的数据很敏感，必须同时用于 JOIN 和 SELECT 查询，则应将该源列映射到输出文件中的两个目标列。一列作为 fingerprint 列进行 `type` 加密，一列作为密封列进行 `type` 加密。C3R 加密客户端的交互式架构生成建议标题后缀为 `_fingerprint` 和 `_sealed`。这些标题后缀可以成为快速区分此类列的有用惯例。

C3R 加密客户端指南

C3R 加密客户端是一种工具，它使组织能够将敏感数据整合在一起，从数据分析中获得新的洞察。该工具以加密方式限制了任何一方和在此过程中可以学到 AWS 的内容。尽管这一点至关重要，但以加密方式保护数据的过程可能会在计算和存储资源方面增加大量开销。因此，了解使用每种设置的利弊得失以及如何在保持所需的加密保障的同时优化设置非常重要。本主题重点介绍 C3R 加密客户端和架构中不同设置对性能的影响。

所有 C3R 加密客户端加密设置都提供不同的加密保障。默认情况下，协作级别的设置是最安全的。在创建协作时启用其他功能会削弱隐私保障，从而允许对加密文字进行频率分析等活动。有关如何使用这些设置及其影响的更多信息，请参阅the section called “加密计算”。

主题

对列类型的性能影响

C3R 使用三种列类型：cleartext、fingerprint 和 sealed。每种列类型都提供不同的加密保障，并且具有不同的预期用途。在以下各节中，将讨论列类型的性能影响以及每种设置对性能的影响。

主题

Cleartext 列

Cleartext 列不会改变其原始格式，也不会以任何方式进行加密处理。此列类型无法配置，也不会影响存储或计算性能。

Fingerprint 列

Fingerprint 列旨在用于联接多个表中的数据。为此，生成的加密文字大小必须始终相同。但是，这些列受协作级别设置的影响。Fingerprint 列可能会对输出文件大小产生不同程度的影响，具体取决于输入中包含的 cleartext。

主题

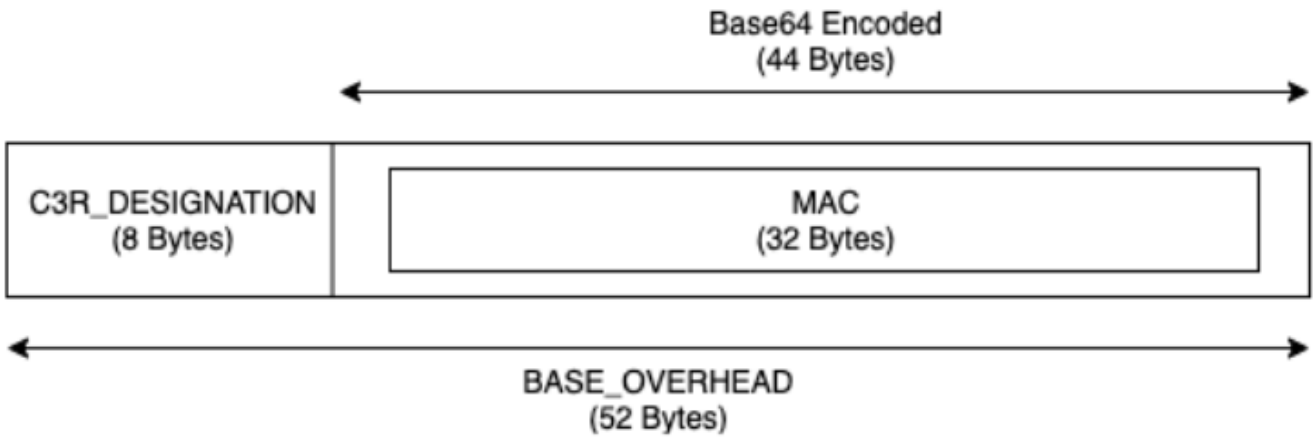
- fingerprint 列的基本开销
- fingerprint 列的协作设置
- fingerprint 列的示例数据
- fingerprint 列疑难解答

fingerprint 列的基本开销

fingerprint 列有基本开销。这种开销是恒定的，与 cleartext 字节的大小无关。

fingerprint 列中的数据通过 HMAC 散列消息认证码函数进行加密处理，该函数将数据转换为 32 字节的消息认证码 (MAC)。然后通过 base64 编码器对这些数据进行处理，使字节大小增加约 33%。它前面有一个 8 字节的 C3R 标识，用于指定数据所属的列类型以及生成数据的客户端版本。最终结果为 52 字节。然后将此结果乘以行数得出总基本开销（如果 `preserveNulls` 设置为 true，则使用非 `null` 值总数）。

下图显示了如何操作 *`BASE_OVERHEAD = C3R_DESIGNATION + (MAC * 1.33)`*



fingerprint 列中的输出加密文字将始终为 52 字节。如果输入 cleartext 数据的平均值超过 52 字节（例如，完整的街道地址），则存储空间可能会显著减少。如果输入 cleartext 数据的平均值小于 52 字节（例如，客户年龄），则存储空间可能会显著增加。

fingerprint 列的协作设置

preserveNulls 设置

当协作级别设置 preserveNulls 为 false（默认值）时，每个 null 值都将替换为一个唯一的随机 32 字节，并被当作非 null 处理。结果是，现在每个 null 值都是 52 字节。与此设置为 true 且 null 值作为 null 传递时相比，对于包含非常稀疏的数据的表，这可能会增加大量存储需求。

如果您不需要此设置的隐私保障，并且希望在数据集中保留 null 值，请在创建协作时启用 preserveNulls 设置。创建协作后将无法更改 preserveNulls 设置。

fingerprint 列的示例数据

以下是带有要重现设置的 fingerprint 列的输入和输出数据的示例集。其他协作级别的设置（例如 allowCleartext 和 allowDuplicates）不会影响结果，如果尝试在本地重现，则可以设置为 true 或 false。

共享密钥示例：wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY

协作 ID 示例：a1b2c3d4-5678-90ab-cdef-EXAMPLE11111

allowJoinsOnColumnsWithDifferentNames：True 此设置不会影响性能或存储要求。但是，在重现下表中显示的值时，此设置使列名的选择变得无关紧要。

示例 1

Input	null
preserveNulls	TRUE
Output	null
确定性	Yes
输入字节	0
输出字节	0

示例 2

Input	null

preserveNulls	FALSE
Output	01:hmac:3lkFjthvV3IUu6mMvFc1a+XAHwgw/ElmOq4p3Yg25kk=
确定性	No
输入字节	0
输出字节	52

示例 3

Input	empty string
preserveNulls	-
Output	01:hmac:oKTgi3Gba+eUb3JteSz2EMgXUkF1WgM77UP0Ydw5kPQ=
确定性	Yes
输入字节	0
输出字节	52

示例 4

Input	abcdefghijklmnopqrstuvwxyz
preserveNulls	-
Output	01:hmac:kU/IqwG7FMmzzshr0B9scomE0UJUEE7j9keTctplGww=
确定性	Yes
输入字节	26

输出字节	52

示例 5

Input	`abcdefghijklmnopqrstuvwxyzA BCDEFGHIJKLMNOPQRSTUVWXYZ01 23456789`
`preserveNulls`	-
Output	`01:hmac:ks3htnQbw2vdhCRFF6J NzW5LMndJaHG57uvE26mBtSs=`
确定性	`Yes`
输入字节	62
输出字节	52

fingerprint 列疑难解答

为什么我的 fingerprint 列中的加密文字比进入它的 cleartext 大几倍？

fingerprint 列中的加密文字的长度始终为 52 字节。如果您的输入数据很小（例如，客户的年龄），则其大小将显著增加。如果将 `preserveNulls` 设置设置为 `false`，也可能发生这种情况。

为什么我的 fingerprint 列中的加密文字比进入它的 cleartext 小几倍？

fingerprint 列中的加密文字的长度始终为 52 字节。如果您的输入数据很大（例如，客户的完整街道地址），则其大小将显著减小。

我怎么知道我是否需要 **`preserveNulls`** 提供的加密保障？

遗憾的是，答案是“看情况”。至少，应查看 the section called “参数” 了解 `preserveNulls` 设置如何保护您的数据。但是，我们建议您参考组织的数据处理要求以及适用于相应协作的任何合同。

为什么我必须承担 base64 的开销？

为了与 CSV 等表格文件格式兼容，必须进行 base64 编码。尽管某些文件格式（例如 Parquet）可能支持数据的二进制表示，但重要的是，协作中的所有参与者都必须以相同的方式表示数据，以确保查询结果正确。

Sealed 列

Sealed 列用于在协作成员之间传输数据。这些列中的加密文字是不确定的，并且会根据列的配置方式对性能和存储产生重大影响。这些列可以单独配置，通常对 C3R 加密客户端的性能和由此产生的输出文件大小影响最大。

主题

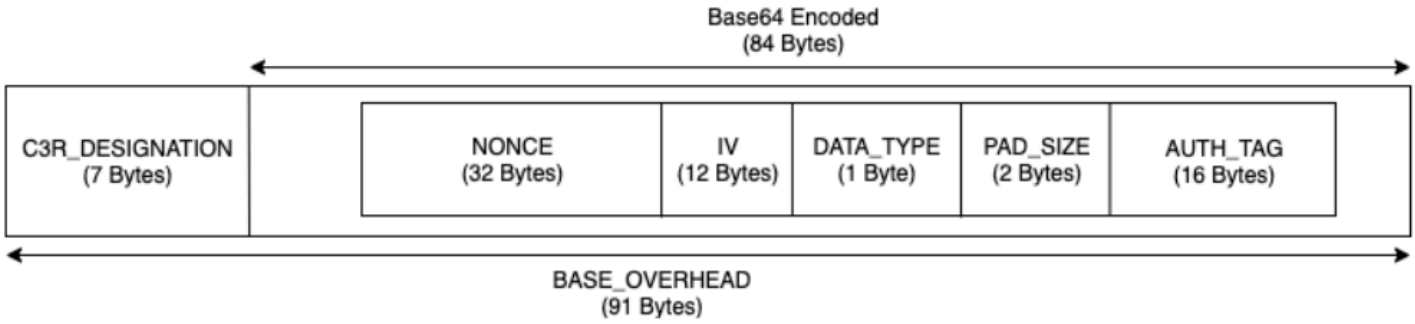
- sealed 列的基本开销
- sealed 列的协作设置
- 架构设置 sealed 列：填充类型
- sealed 列的示例数据
- sealed 列疑难解答

sealed 列的基本开销

sealed 列有基本开销。此开销是恒定的，是 cleartext 和填充（如有）字节大小之外的开销。

在进行任何加密之前，在 sealed 列中的数据前面加上一个 1 字节的字符，表示所包含的数据类型。如果选择了填充，则会对数据进行填充并追加 2 个字节，说明填充大小。添加这些字节后，使用 AES-GCM 对数据进行加密处理，并以 IV（12 字节）、nonce（32 字节）和 Auth Tag（16 字节）进行存储。然后通过 base64 编码器对这些数据进行处理，使字节大小增加约 33%。数据前面有一个 7 字节的 C3R 标识，用于指定数据属于哪种类型的列以及用于生成数据的客户端版本。结果是 91 字节的最终基本开销。然后将此结果乘以行数得出总基本开销（如果 `preserveNulls` 设置为 true，则使用非 null 值总数）。

下图显示了如何操作 *`BASE_OVERHEAD = C3R_DESIGNATION + ((NONCE + IV + DATA_TYPE + PAD_SIZE + AUTH_TAG) * 1.33)`*



sealed 列的协作设置

`preserveNulls` 设置

当协作级别设置 `preserveNulls` 为 `false`（默认值）时，每个 `null` 值都将是唯一的随机 32 字节，并被当作非 `null` 处理。结果是，现在每个 `null` 值都是 91 字节（如果填充，则更多）。与此设置为 `true` 且 `null` 值作为 `null` 传递时相比，对于包含非常稀疏的数据的表，这可能会增加大量存储需求。

如果您不需要此设置的隐私保障，并且希望在数据集中保留 `null` 值，请在创建协作时启用 `preserveNulls` 设置。创建协作后将无法更改 `preserveNulls` 设置。

架构设置 sealed 列：填充类型

主题

- none 的填充类型
- fixed 的填充类型
- max 的填充类型

`none` 的填充类型

选择 `none` 的填充类型不会向 cleartext 增加任何填充，也不会在前面描述的基本开销之外增加额外的开销。没有填充会产生最节省空间的输出大小。但是，它不提供与 `fixed` 和 `max` 填充类型相同的隐私保障。这是因为底层 cleartext 的大小可以从加密文字的大小中分辨出来。

`fixed` 的填充类型

选择 `fixed` 的填充类型是一种隐私保护措施，用于隐藏列中包含的数据的长度。这是通过在加密之前将所有 `pad_length` 填充到提供的 cleartext 来完成的。任何超过该大小的数据都会导致 C3R 加密客户端失败。

假设填充是在加密之前添加到 cleartext 的，因此 AES-GCM 具有 cleartext 到加密文字字节的 1:1 映射。base64 编码将增加 33%。填充的额外存储开销可以通过从 pad_length 的值中减去 cleartext 的平均长度，然后乘以 1.33 来计算。结果就是每条记录的平均填充开销。然后将此结果乘以行数得出总填充开销（如果 preserveNulls 设置为 true，则使用非 null 值总数）。

*PADDING_OVERHEAD = (PAD_LENGTH - AVG_CLEARTEXT_LENGTH) * 1.33 * ROW_COUNT*

我们建议您选择包含列中最大值的最小 pad_length。例如，如果最大值为 50 字节，则 pad_length 为 50 就足够了。大于该值的值只会增加额外的存储开销。

固定填充不会增加任何显著的计算开销。

max 的填充类型

选择 max 的填充类型是一种隐私保护措施，用于隐藏列中包含的数据的长度。这是通过将所有 cleartext 填充到列中的最大值再加上加密之前的额外 pad_length 来完成的。通常，max 填充提供的保障与单个数据集的 fixed 填充相同，同时允许不知道列中的最大 cleartext 值。但是，max 填充可能无法提供与跨更新的 fixed 填充相同的隐私保障，因为各个数据集中的最大值可能有所不同。

我们建议您在使用 max 填充时，额外选择 0 的 pad_length。此长度将所有值填充到与列中最大值的大小相同。大于该值的值只会增加额外的存储开销。

如果已知给定列的最大 cleartext 值，我们建议您改用 fixed 填充类型。使用 fixed 填充可在更新的数据集之间实现一致性。使用 max 填充会使每个数据子集填充到该子集中的最大值。

sealed 列的示例数据

以下是带有要重现设置的 sealed 列的输入和输出数据的示例集。其他协作级别的设置（例如 allowCleartext、allowJoinsOnColumnsWithDifferentNames 和 allowDuplicates）不会影响结果，如果尝试在本地重现，则可以设置为 true 或 false。尽管这些是要重现的基本设置，但 sealed 列是不确定的，值每次都会更改。目的是显示输入字节与输出字节的对比。示例 pad_length 值是故意选择的。它们表明，使用推荐的最低 pad_length 设置或需要额外的填充时，fixed 填充产生的值与 max 填充相同。

共享密钥示例：wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY

协作 ID 示例：a1b2c3d4-5678-90ab-cdef-EXAMPLE11111

主题

- none 的填充类型

none 的填充类型

示例 1

Input	`null`
`preserveNulls`	`TRUE`
Output	`null`
确定性	`Yes`
输入字节	0
输出字节	0

示例 2

Input	`null`
`preserveNulls`	`FALSE`
Output	`01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfssGSPbNIJfG3iXmu6cbCUrizuV`
确定性	No
输入字节	0
输出字节	91

示例 3

Input	empty string
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfstGSPEM6qR8DWC2PB2GMlX41YK
确定性	No
输入字节	0
输出字节	91

示例 4

Input	abcdefghijklmnopqrstuvwxyz
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfsteEE1GKEPiRzyh0h7t6OmWMLTWCvO2ckr6pkx9sGL5VLDQeHzh6DmPpyWNuI=
确定性	No
输入字节	26
输出字节	127

示例 5

Input	abcdefghijklmnopqrstuvwxyzA BCDEFGHIJKLMNOPQRSTUVWXYZ01 23456789
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfsteEE1GKEPiRzyh0 h7t6OmWMLTWCvO2ckr6plwtH/8t RFnn2rF91bcB9G4+n8GiRfJNmqdP4/ QOQ3cXb/pbvPcnnohrHIGSX54ua+1/ JfcVjc=
确定性	No
输入字节	62
输出字节	175

fixed 的填充类型（示例 1）

在此示例中，pad_length 为 62，最大输入为 62 字节。

示例 1

Input	null
preserveNulls	TRUE
Output	null
确定性	Yes
输入字节	0
输出字节	0

示例 2

Input	null
preserveNulls	FALSE
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfssGSNWfMRp7nSb7SMX2s3JKLOhK1+7r75Tk+Mx9jy48Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/hCz7oaIneVsrcoNpATs0GzbnLkor4L+/aSuA=
确定性	No
输入字节	0
输出字节	175

示例 3

Input	empty string
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfstGSNWfMRp7nSb7SMX2s3JKLOhK1+7r75Tk+Mx9jy48Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/hCz7oaIneVsrcoLB53l07VZpA6OwkuXu29CA=
确定性	No
输入字节	0
输出字节	175

示例 4

Input	abcdefghijklmnopqrstuvwxyz
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfsteEE1GKEPiRzyh0h7t6OmWMLTWCvO2ckr6pkx9jy48Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/hCz7oaIneVsrcutBAcO+Mb9tuU2KIHH31AWg=
确定性	No
输入字节	26
输出字节	175

示例 5

Input	abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfsteEE1GKEPiRzyh0h7t6OmWMLTWCvO2ckr6plwtH/8tRFnn2rF91bcB9G4+n8GiRfJNmqdP4/QOQ3cXb/pbvPcnnohrHIGSX54ua+1/JfcVjc=
确定性	No
输入字节	62

输出字节	175

fixed 的填充类型（示例 2）

在此示例中，pad_length 为 162，最大输入为 62 字节。

示例 1

Input	null
preserveNulls	TRUE
Output	null
确定性	Yes
输入字节	0
输出字节	0

示例 2

Input	null
preserveNulls	FALSE
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfssGSNWfMRp7nSb7S MX2s3JKLOhK1+7r75Tk+Mx9jy48 Fcg1yOPvBqRSZ7oqy1V3UKfYTLE Zb/hCz7oaIneVsrcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwv/xAySX+xcntotL703aBTBb

确定性	No
输入字节	0
输出字节	307

示例 3

Input	`empty string`
`preserveNulls`	-
Output	`01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfstGSNWfMRp7nSb7SMX2s3JKLOhK1+7r75Tk+Mx9jy48Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/hCz7oaIneVsrcnkB0xbLWD7zNdAqQGR0rXoSESdW0I0vpNoGcBfv4cJbG0A3h1DvtkSSVc2B80OOGppzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn+8o4WtG/ClipNcjDXvXVtK4vfCohcCA6uwrmwv84lVaT9Yd+6oQx65/+gdVT`
确定性	No
输入字节	0
输出字节	307

示例 4

Input	`abcdefghijklmnopqrstuvwxyz`
`preserveNulls`	-
Output	`01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRY`

	Z98t5KU6aWfsteEE1GKEPiRzyh0 h7t6OmWMLTWCvO2ckr6pkx9jy48 Fcg1yOPvBqRSZ7oqy1V3UKfYTLE Zb/hCz7oaIneVsrcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwtX5Hnl+WyfO6ks3QMaRDGSf
确定性	No
输入字节	26
输出字节	307

示例 5

Input	abcdefghijklmnopqrstuvwxyzA BCDEFGHIJKLMNOPQRSTUVWXYZ01 23456789
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfsteEE1GKEPiRzyh0 h7t6OmWMLTWCvO2ckr6plwtH/8t RFnn2rF91bcB9G4+n8GiRfJNmqd P4/QOQ3cXb/pbvPcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwjkJXQZOgPdeFX9Yr/8alV5i
确定性	No

输入字节	62
输出字节	307

max 的填充类型（示例 1）

在此示例中，pad_length 为 0，最大输入为 62 字节。

示例 1

Input	null
preserveNulls	TRUE
Output	null
确定性	Yes
输入字节	0
输出字节	0

示例 2

Input	null
preserveNulls	FALSE
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfssGSNWfMRp7nSb7S MX2s3JKLOhK1+7r75Tk+Mx9jy48 Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/ hCz7oaIneVsrcoNpATs0GzbnLkor4L+/ aSuA=
确定性	No
输入字节	0

输出字节	175

示例 3

Input	`empty string`
`preserveNulls`	-
Output	`01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfstGSNWfMRp7nSb7SMX2s3JKLOhK1+7r75Tk+Mx9jy48Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/hCz7oaIneVsrcoLB53l07VZpA6OwkuXu29CA=`
确定性	No
输入字节	0
输出字节	175

示例 4

Input	`abcdefghijklmnopqrstuvwxyz`
`preserveNulls`	-
Output	`01:enc:bm9uY2UwMTIzNDU2Nzg5MG5vbmNlMDEyMzQ1Njc4OTBqfRYZ98t5KU6aWfsteEE1GKEPiRzyh0h7t6OmWMLTWCvO2ckr6pkx9jy48Fcg1yOPvBqRSZ7oqy1V3UKfYTLEZb/hCz7oaIneVsrcutBAcO+Mb9tuU2KIHH31AWg=`
确定性	No

输入字节	26
输出字节	175

示例 5

Input	`abcdefghijklmnopqrstuvwxyzA BCDEFGHIJKLMNOPQRSTUVWXYZ01 23456789`
`preserveNulls`	-
Output	`01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfsteEE1GKEPiRzyh0 h7t6OmWMLTWCvO2ckr6plwtH/8t RFnn2rF91bcB9G4+n8GiRfJNmqdP4/ QOQ3cXb/pbvPcnnohrHIGSX54ua+1/ JfcVjc=`
确定性	No
输入字节	62
输出字节	175

`max` 的填充类型（示例 2）

在此示例中，`pad_length` 为 100，最大输入为 62 字节。

示例 1

Input	`null`
`preserveNulls`	`TRUE`
Output	`null`
确定性	`Yes`

输入字节	0
输出字节	0

示例 2

Input	null
preserveNulls	FALSE
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfssGSNWfMRp7nSb7S MX2s3JKLOhK1+7r75Tk+Mx9jy48 Fcg1yOPvBqRSZ7oqy1V3UKfYTLE Zb/hCz7oaIneVsrcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwv/xAySX+xcntotL703aBTBb
确定性	No
输入字节	0
输出字节	307

示例 3

Input	empty string
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfstGSNWfMRp7nSb7S MX2s3JKLOhK1+7r75Tk+Mx9jy48

	Fcg1yOPvBqRSZ7oqy1V3UKfYTLE Zb/hCz7oaIneVsrcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwv84lVaT9Yd+6oQx65/+gdVT
确定性	No
输入字节	0
输出字节	307

示例 4

Input	abcdefghijklmnopqrstuvwxyz
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfsteEE1GKEPiRzyh0 h7t6OmWMLTWCvO2ckr6pkx9jy48 Fcg1yOPvBqRSZ7oqy1V3UKfYTLE Zb/hCz7oaIneVsrcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwtX5Hnl+WyfO6ks3QMaRDGSf
确定性	No
输入字节	26
输出字节	307

示例 5

Input	abcdefghijklmnopqrstuvwxyzA BCDEFGHIJKLMNOPQRSTUVWXYZ01 23456789
preserveNulls	-
Output	01:enc:bm9uY2UwMTIzNDU2Nzg5 MG5vbmNlMDEyMzQ1Njc4OTBqfRY Z98t5KU6aWfsteEE1GKEPiRzyh0 h7t6OmWMLTWCvO2ckr6plwtH/8t RFnn2rF91bcB9G4+n8GiRfJNmqd P4/QOQ3cXb/pbvPcnkB0xbLWD7z NdAqQGR0rXoSESdW0I0vpNoGcBf v4cJbG0A3h1DvtkSSVc2B80OOGp pzdDqhrUVN5wFNyn8vgfPMqDaeJk5bn +8o4WtG/ClipNcjDXvXVtK4vfCohcCA6 uwrmwjkJXQZOgPdeFX9Yr/8alV5i
确定性	No
输入字节	62
输出字节	307

sealed 列疑难解答

为什么我的 sealed 列中的加密文字比进入它的 cleartext 大几倍？

这取决于多个因素。首先，Cleartext 列中的加密文字的长度始终至少为 91 字节。如果您的输入数据很小（例如，客户的年龄），则其大小将显著增加。其次，如果 preserveNulls 设置为 false，并且您的输入数据包含很多 null 值，则每个 null 值都将变成 91 字节的加密文字。最后，如果您使用填充，则根据定义，在加密 cleartext 数据之前会将字节添加到该数据中。

我的 sealed 列中的大部分数据都非常小，我需要使用填充。我可以删除大值并单独处理它们以节省空间吗？

我们不建议删除大值并单独处理。这样做会改变 C3R 加密客户端提供的隐私保证。作为威胁模型，假设观察者可以看到两个加密的数据集。如果观察者发现一个数据子集的列填充明显大于或小于另一个子集，则他们可以推断每个子集中数据的大小。例如，假设 `fullName` 列在一个文件中填充到总共 40 字节，而在另一个文件中填充到 800 字节。观察者可能会认为，其中一个数据集包含了世界上最长的名字（747 字节）。

使用 **`max`** 填充类型时，我需要提供额外的填充吗？

不需要。使用 `max` 填充时，我们建议将 `pad_length`（也称为列中最大值之外的额外填充）设置为 0。

我能否在使用 **`fixed`** 填充时选择大的 **`pad_length`** 来避免担心最大的值是否合适？

能，但是大的填充长度效率低下，并且占用的存储空间超出了必要的范围。我们建议您查看最大值有多大，并将 `pad_length` 设置为该值。

我怎么知道我是否需要 **`preserveNulls`** 提供的加密保障？

遗憾的是，答案是“看情况”。至少，应查看 加密计算 Clean Rooms 了解 `preserveNulls` 设置如何保护您的数据。但是，我们建议您参考组织的数据处理要求以及适用于相应协作的任何合同。

为什么我必须承担 base64 的开销？

为了与 CSV 等表格文件格式兼容，必须进行 base64 编码。尽管某些文件格式（例如 Parquet）可能支持数据的二进制表示，但重要的是，协作中的所有参与者都必须以相同的方式表示数据，以确保查询结果正确。

加密文字大小意外增加疑难解答

假设您加密了数据，结果数据的大小出人意料地大。以下步骤可以帮助您确定大小增加的位置以及可以采取的措施（如果有）。

确定大小增加的位置

在排查加密数据明显大于 cleartext 数据的原因之前，必须先确定大小的增加位置。可以放心地忽略 Cleartext 列，因为它们没有变化。查看其余的 fingerprint 和 sealed 列，然后选择一个看起来很重要的列。

确定大小增加的原因

fingerprint 列或 sealed 列可能会导致大小增加。

主题

大小增加是否来自 fingerprint 列？

如果对存储增加贡献最大的列是 fingerprint 列，则可能是因为 cleartext 数据很小（例如，客户年龄）。生成的每个 fingerprint 加密文字的长度为 52 字节。不幸的是，在这个问题上无能为 column-by-column力。有关更多信息，请参阅 fingerprint 列的基本开销 了解有关此列的详细信息，包括它如何影响存储要求。

导致 fingerprint 列中大小增加的另一个可能原因是协作设置 `preserveNulls`。如果禁用了 `preserveNulls` 的协作设置（默认设置），则 fingerprint 列中的所有 `null` 值都将变为 52 字节的加密文字。目前的协作对此无能为力。`preserveNulls` 设置是在创建协作时设置的，所有协作者必须使用相同的设置以确保查询结果正确。有关 `preserveNulls` 设置以及启用它会如何影响数据隐私保障的更多信息，请参阅 the section called “加密计算”。

大小增加是否来自 sealed 列？

如果对存储增加贡献最大的列是 sealed 列，那么有一些细节可能会导致大小增加。

如果 cleartext 数据很小（例如，客户年龄），则生成的每个 sealed 加密文字的长度至少为 91 字节。遗憾的是，我们对这个问题无能为力。有关更多信息，请参阅 sealed 列的基本开销 了解有关此列的详细信息，包括它如何影响存储要求。

sealed 列存储增加的第二个主要原因是填充。填充会在加密 cleartext 之前向其添加额外的字节，以隐藏数据集中各个值的大小。我们建议您将数据集的填充设置为可能的最小值。至少必须将 `fixed` 填充的 `pad_length` 设置为包含列中可能的最大值。任何高于此值的设置都不会增加额外的隐私保障。例如，如果您知道列中可能的最大值可能为 50 字节，我们建议您将 `pad_length` 设置为 50 字节。但是，如果 sealed 列使用 `max` 填充，我们建议您将 `pad_length` 设置为 0 字节。这是因为 `max` 填充是指列中最大值之外的额外填充。

导致 sealed 列中大小增加的最后一个可能原因是协作设置 `preserveNulls`。如果禁用了 `preserveNulls` 的协作设置（默认设置），则 sealed 列中的所有 `null` 值都将变为 91 字节的加密文字。目前的协作对此无能为力。`preserveNulls` 设置是在创建协作时设置的，所有协作者必须使用相同的设置以确保查询结果正确。有关此设置以及启用它会如何影响数据隐私保障的更多信息，请参阅 the section called “加密计算”。

设置 AWS Clean Rooms

以下主题说明了如何设置 AWS Clean Rooms。

主题

报名参加 AWS

在使用或使用 AWS Clean Rooms任何东西之前 AWS 服务，您必须先注册 AWS 账户。 AWS

如果您没有 AWS 账户，请完成以下步骤来创建一个。

报名参加 AWS 账户

1. 打开https://portal.aws.amazon.com/billing/注册。
2. 按照屏幕上的说明操作。

   在注册过程中，您将接到一个带有验证码的电话，您将在电话键盘上输入该验证码。
3. 当您注册时 AWS 账户，将创建一个 AWS 账户 root 用户。根用户有权访问该账户中的所有 AWS 服务 和资源。作为安全最佳实践，请为管理用户分配管理访问权限，并且只使用根用户执行需要根用户访问权限的任务。

为设置服务角色 AWS Clean Rooms

以下各节描述了执行每项任务所需的角色。

主题

- 创建服务角色以从 Snowflake 读取数据
- 创建用于从 S3 存储桶读取代码的服务角色（PySpark 分析模板角色）
- 创建服务角色以写入 PySpark 作业结果
- 创建服务角色来接收结果

创建管理员用户

要使用 AWS Clean Rooms，您需要为自己创建一个管理员用户，并将该管理员用户添加到管理员组中。

要创建管理员用户，请选择以下选项之一。

选择一种方法来管理您的管理员	目标	方式	您也可以
在 IAM Identity Center 中（推荐）	使用短期凭证访问 AWS。 这符合安全最佳实操。有关最佳实践的信息，请参阅《IAM 用户指南》中的 IAM 中的安全最佳实践。	有关说明，请参阅《AWS IAM Identity Center 用户指南》中的入门。	通过在《AWS Command Line Interface 用户指南》 AWS IAM Identity Center中配置 AWS CLI 要使用的来配置编程访问权限。
在 IAM 中（不推荐使用）	使用长期凭证访问 AWS。	按照《IAM 用户指南》中的创建用于紧急访问的 IAM 用户中的说明进行操作。	按照《IAM 用户指南》中的管理 IAM 用户的访问密钥，配置编程式访问。

为协作成员创建 IAM 角色

成员是 AWS 指参与协作的客户。

为协作成员创建 IAM 角色

1. 请按照《AWS Identity and Access Management 用户指南》的创建向 IAM 用户委派权限的角色步骤中的说明进行操作。
2. 在创建策略步骤中，选择策略编辑器中的 JSON 选项卡，然后根据授予协作成员的能力添加策略。

   AWS Clean Rooms 根据常见用例提供以下托管策略。

如果要...	然后使用...
查看资源和元数据	AWS 托管策略：AWSCleanRoomsReadOnlyAccess
Query	AWS 托管策略：AWSCleanRoomsFullAccess
查询和运行作业	AWS 托管策略：AWSCleanRoomsFullAccess
查询和接收结果	AWS 托管策略：AWSCleanRoomsFullAccess
管理协作资源但不查询	AWS 托管策略：AWSCleanRoomsFullAccessNoQuerying

   有关提供的不同托管策略的信息 AWS Clean RoomsAWS 的托管策略 AWS Clean Rooms，请参阅

创建服务角色以从 Amazon S3 读取数据

AWS Clean Rooms 使用服务角色从 Amazon S3 读取数据。

有两种方法可以创建此服务角色。

- 如果您拥有创建服务角色所必需的 IAM 权限，请使用 AWS Clean Rooms 控制台创建服务角色。
- 如果您没有`iam:CreateRole`、`iam:CreatePolicy`、`iam:AttachRolePolicy`权限或想要手动创建 IAM 角色，请执行以下任一操作：

- 使用以下步骤使用自定义信任策略创建服务角色。
- 要求管理员使用以下步骤创建服务角色。

> **Note**
>
> 只有在您没有使用 AWS Clean Rooms 控制台创建服务角色的必要权限时，您或您的 IAM 管理员才应遵循此过程。

使用自定义信任策略创建服务角色以从 Amazon S3 读取数据

1. 使用自定义信任策略创建角色。有关更多信息，请参阅《AWS Identity and Access Management 用户指南》中的使用自定义信任策略创建角色（控制台）过程。
2. 根据使用自定义信任策略创建角色（控制台）步骤使用以下自定义信任策略。

> **Note**
>
> 如果您想帮助确保该角色仅在特定的协作成员资格环境中使用，则可以进一步缩小信任策略的范围。有关更多信息，请参阅 防止跨服务混淆代理。

JSON

```
{

    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "RoleTrustPolicyForCleanRoomsService",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
     ]
}
```

3. 根据使用自定义信任策略创建角色（控制台）步骤使用以下权限策略。

> **Note**
>
> 以下示例策略支持读取 AWS Glue 元数据及其相应的 Amazon S3 数据所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。例如，如果您为 Amazon S3 数据设置了自定义 KMS 密钥，则可能需要使用额外 AWS Key Management Service (AWS KMS) 权限修改此政策。
> 您的 AWS Glue 资源和底层 Amazon S3 资源必须与 AWS Clean Rooms 协作 AWS 区域相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "NecessaryGluePermissions",
            "Effect": "Allow",
            "Action": [
                "glue:GetDatabase",
                "glue:GetDatabases",
                "glue:GetTable",
                "glue:GetTables",
                "glue:GetPartition",
                "glue:GetPartitions",
                "glue:BatchGetPartition"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:database/databaseName",
                "arn:aws:glue:us-
east-1:111122223333:table/databaseName/tableName",
                "arn:aws:glue:us-east-1:111122223333:catalog"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "glue:GetSchema",
                "glue:GetSchemaVersion"
```

```
            ],
            "Resource": [
                "*"
            ]
        },
        {
            "Sid": "NecessaryS3BucketPermissions",
            "Effect": "Allow",
            "Action": [
                "s3:GetBucketLocation",
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::bucket"
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "444455556666"
                    ]
                }
            }
        },
        {
            "Sid": "NecessaryS3ObjectPermissions",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject"
            ],
            "Resource": [
                "arn:aws:s3:::bucket/prefix/*"
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "444455556666"
                    ]
                }
            }
        }
    ]
}
```

> **Note**
>
> 该策略引用了两种不同的策略 AWS 账户 IDs 来支持由不同各方管理数据目录元数据和实际数据存储的 AWS Clean Rooms 协作：
>
> - 111122223333-这是拥有 AWS Glue 数据目录资源（数据库、表和目录）的帐户。第一条语句授予访问该账户 AWS Glue 目录中的表架构、分区信息和元数据的权限。
> - 444455556666- 这个账户拥有包含实际数据文件的 Amazon S3 存储桶。根据`s3:ResourceAccount`条件，Amazon S3 权限（声明 3 和声明 4）仅限于该账户拥有的存储桶。
>
> 此配置支持常见的企业数据架构，其中一个团队管理数据目录和架构定义，而另一个团队则拥有底层的数据存储基础架构。该`s3:ResourceAccount`条件通过确保 Amazon S3 操作仅适用于指定账户拥有的存储桶，从而提供了额外的安全层。

4. 将每个 *placeholder* 替换为您自己的信息。
5. 继续按照使用自定义信任策略创建角色（控制台）步骤创建角色。

创建服务角色以读取来自亚马逊 Athena 的数据

AWS Clean Rooms 使用服务角色从 Amazon Athena 读取数据。

使用自定义信任策略创建服务角色以从 Athena 读取数据

1. 使用自定义信任策略创建角色。有关更多信息，请参阅《AWS Identity and Access Management 用户指南》中的使用自定义信任策略创建角色（控制台）过程。
2. 根据使用自定义信任策略创建角色（控制台）步骤使用以下自定义信任策略。

> **Note**
>
> 如果您想帮助确保该角色仅在特定的协作成员资格环境中使用，则可以进一步缩小信任策略的范围。有关更多信息，请参阅 防止跨服务混淆代理。

JSON

```
{

    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "RoleTrustPolicyForCleanRoomsService",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
     ]
}
```

3. 根据使用自定义信任策略创建角色（控制台）步骤使用以下权限策略。

> **Note**
>
> 以下示例策略支持读取 AWS Glue 元数据及其相应的 Athena 数据所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。例如，如果您已经为自己的 Amazon S3 数据设置了自定义 KMS 密钥，则可能需要修改此政策，使其具有额外的 AWS KMS 权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "athena:GetWorkGroup",
                "athena:GetTableMetadata",
                "athena:GetDataCatalog",
                "athena:StartQueryExecution",
                "athena:GetQueryExecution",
                "athena:GetQueryResults"
```

```
            ],
            "Resource": [
                "arn:aws:athena:region:accountId:workgroup/workgroup",
                "arn:aws:athena:region:accountId:datacatalog/federatedCatalogName"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "glue:GetDatabase",
                "glue:GetTable",
                "glue:GetCatalog"
            ],
            "Resource": [
                "arn:aws:glue:region:accountId:catalog",
                "arn:aws:glue:region:accountId:catalog/federatedCatalogName",

"arn:aws:glue:region:accountId:database/federatedCatalogName/databaseName",

"arn:aws:glue:region:accountId:table/federatedCatalogName/databaseName/tableName"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetBucketLocation",
                "s3:AbortMultipartUpload",
                "s3:ListBucket",
                "s3:PutObject",
                "s3:ListMultipartUploadParts"
            ],
            "Resource": [
                "arn:aws:s3:::athenaResultsBucket",
                "arn:aws:s3:::athenaResultsBucket/*"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceAccount": "accountId"
                }
            }
        },
        {
            "Effect": "Allow",
```

```
            "Action": "lakeformation:GetDataAccess",
            "Resource": "*"
        }
    ]
}
```

4. 将每个 *placeholder* 替换为您自己的信息。
5. 继续按照使用自定义信任策略创建角色（控制台）步骤创建角色。

设置 Lake Formation 权限

如果您查询受 Lake Formation 权限保护的资源，则服务角色必须对 AWS Glue 数据库具有选择table/view/catalog和描述访问权限以及描述权限。

有关更多信息，请参阅:

- 使用 Athena 查询在亚马逊 A thena 用户 AWS Lake Formation指南中注册的数据
- AWS Lake Formation 开发者指南中的@@ 入门使用 Lake Formation 权限

创建服务角色以从 Snowflake 读取数据

AWS Clean Rooms 使用服务角色检索您的凭据，让 Snowflake 从该来源读取您的数据。

可以使用两种方法创建此服务角色：

- 如果您拥有创建服务角色所必需的 IAM 权限，请使用 AWS Clean Rooms 控制台创建服务角色。
- 如果您没有`iam:CreateRole`、`iam:CreatePolicy`、`iam:AttachRolePolicy`权限或想要手动创建 IAM 角色，请执行以下任一操作：
  - 使用以下步骤使用自定义信任策略创建服务角色。
  - 要求管理员使用以下步骤创建服务角色。

> **Note**
>
> 只有在您没有使用 AWS Clean Rooms 控制台创建服务角色的必要权限时，您或您的 IAM 管理员才应遵循此过程。

使用自定义信任策略创建服务角色以从 Snowflake 读取数据

1. 使用自定义信任策略创建角色。有关更多信息，请参阅《AWS Identity and Access Management 用户指南》中的使用自定义信任策略创建角色（控制台）过程。
2. 根据使用自定义信任策略创建角色（控制台）步骤使用以下自定义信任策略。

> **Note**
>
> 如果您想帮助确保该角色仅在特定的协作成员资格环境中使用，则可以进一步缩小信任策略的范围。有关更多信息，请参阅 防止跨服务混淆代理。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowIfSourceArnMatches",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "ForAnyValue:ArnEquals": {
                    "aws:SourceArn": [
                        "arn:aws:cleanrooms:us-
east-1:111122223333:membership/membershipId",
                        "arn:aws:cleanrooms:us-
east-1:444455556666:membership/queryRunnerMembershipId"
                    ]
                }
            }
        }
    ]
}
```

> **Note**
>
> 此信任策略引用了两种不同的策略 AWS 账户 IDs 来支持将查询执行责任分配给多方的 AWS Clean Rooms 协作：
>
> - 111122223333-该账户包含参与协作的成员资格。该成员可能拥有数据表、分析规则或其他需要角色访问权限的协作资源。
> - 444455556666-此帐户包含负责运行查询的成员资格（“查询运行器”）。此成员资格执行受保护的查询，并且需要担任此角色才能访问必要的计算和数据资源。
>
> 此配置支持一方提供数据或分析模板而另一方运行实际查询的场景。通过相同的执行角色，这两个角色都需要不同但互补的权限。该`aws:SourceArn`条件可确保只有源自这两个特定成员资格的 AWS Clean Rooms 操作才能担任该角色，从而在支持分布式作业执行和结果管理工作流程的同时维护安全性。

3. 根据使用自定义信任策略创建角色（控制台）过程使用以下权限策略之一。

使用客户拥有的 KMS 密钥加密的机密的权限策略

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": "secretsmanager:GetSecretValue",
            "Resource": "arn:aws:secretsmanager:us-
east-1:111122223333:secret:secretIdentifier",
            "Effect": "Allow"
        },
        {
            "Sid": "AllowDecryptViaSecretsManagerForKey",
            "Action": "kms:Decrypt",
            "Resource": "arn:aws:kms:us-
east-1:444455556666:key/keyIdentifier",
            "Effect": "Allow",
            "Condition": {
                "StringEquals": {
```

```
                    "kms:ViaService": "secretsmanager.us-
east-1.amazonaws.com",
                    "kms:EncryptionContext:SecretARN":
 "arn:aws:secretsmanager:us-east-1:111122223333:secret:secretIdentifier"
                }
            }
        }
    ]
}
```

Note

该政策引用了两种不同的策略 AWS 账户 IDs 来支持跨账户密钥管理方案：

- 111122223333- 这是拥有和存储秘密的账户。第一条语句授予从该账户检索机密值的权限。
- 444455556666-这是拥有用于加密密钥的 AWS KMS 密钥的账户。第二条语句授予使用该账户的密钥解密密密 AWS KMS 钥的权限。

此配置在企业环境中很常见，其中：

- 在一个账户（账户 1）中集中管理密钥
- 加密密钥由单独的安全账户或共享服务账户（账户 2）管理
- 账户 2 中的 AWS KMS 密钥策略还必须允许账户 1 中的服务使用该密钥进行 encryption/decryption 操作

该`kms:EncryptionContext:SecretARN`条件可确保 AWS KMS 密钥只能用于解密此特定机密，从而为跨账户访问提供额外的安全保护。

使用加密的机密的权限策略 AWS 托管式密钥

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
```

```
        {
            "Action": "secretsmanager:GetSecretValue",
            "Resource": "arn:aws:secretsmanager:us-
east-1:111122223333:secret:secretIdentifier",
            "Effect": "Allow"
        }
    ]
}
```

4. 将每个 *placeholder* 替换为您自己的信息。
5. 继续按照使用自定义信任策略创建角色（控制台）步骤创建角色。

创建用于从 S3 存储桶读取代码的服务角色（PySpark 分析模板角色）

AWS Clean Rooms 使用 PySpark 分析模板时，使用服务角色从协作成员的指定 S3 存储桶中读取代码。

创建服务角色以从 S3 存储桶读取代码

1. 使用自定义信任策略创建角色。有关更多信息，请参阅《AWS Identity and Access Management 用户指南》中的使用自定义信任策略创建角色（控制台）过程。
2. 根据使用自定义信任策略创建角色（控制台）步骤使用以下自定义信任策略。

   JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "ForAnyValue:ArnEquals": {
                    "aws:SourceArn": [
                        "arn:aws:cleanrooms:us-
east-1:111122223333:membership/jobRunnerMembershipId",
                        "arn:aws:cleanrooms:us-
east-1:444455556666:membership/analysisTemplateOwnerMembershipId"
```

```
                    ]
                }
            }
        }
    ]
}
```

> **Note**
>
> 此信任策略引用了两种不同的策略 AWS 账户 IDs 来支持多方 AWS Clean Rooms 协作场景：
>
> - 111122223333-该账户包含负责运行查询的成员资格（“作业运行者”）。该成员资格执行分析作业，需要担任此角色才能访问必要的资源。
> - 444455556666-这是拥有分析模板及其关联成员资格（“分析模板所有者”）的账户。此成员资格定义了可以运行哪些查询，还需要担任此角色来管理和执行分析。
>
> 这种配置在多方参与同一个 AWS Clean Rooms 协作的协作中很常见，每方都有自己的 AWS 账户 成员资格。查询执行者和分析模板所有者都需要访问共享资源。该`aws:SourceArn`条件可确保只有源于这两个特定成员身份的 AWS Clean Rooms 操作才能担任该角色，从而为多方协作提供精确的访问控制。

3. 根据使用自定义信任策略创建角色（控制台）步骤使用以下权限策略。

> **Note**
>
> 以下示例策略支持从 Amazon S3 读取您的代码所需的权限。但是，您可能需要修改此策略，具体取决于您设置 S3 数据的方式。
> 您的 Amazon S3 资源必须与 AWS Clean Rooms 协作资源 AWS 区域 相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
```

```
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetObjectVersion"
            ],
            "Resource": ["arn:aws:s3:::s3Path"],
            "Condition":{
                "StringEquals":{
                    "s3:ResourceAccount":[
                        "s3BucketOwnerAccountId"
                    ]
                }
            }
        }
    ]
}
```

4. 用你自己的信息替换每一个*placeholder*信息：

   - *s3Path*— 您的代码的 S3 存储桶位置。
   - *s3BucketOwnerAccountId*— S3 存储桶所有者的 AWS 账户 ID。
   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *jobRunnerAccountId*— 可以运行查询和作业的成员的 AWS 账户 ID。
   - *jobRunnerMembershipId*— 可以查询和运行作业的成员的会员 ID。可以在协作的详细信息选项卡上找到成员身份 ID。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
   - *analysisTemplateAccountId*-分析模板的 AWS 账户 ID。
   - *analysisTemplateOwnerMembershipId*— 拥有分析模板的成员资格 ID。可以在协作的详细信息选项卡上找到成员身份 ID。

5. 继续按照使用自定义信任策略创建角色（控制台）步骤创建角色。

创建服务角色以写入 PySpark 作业结果

AWS Clean Rooms 使用服务角色将 PySpark 任务的结果写入指定的 S3 存储桶。

创建用于写入 PySpark 作业结果的服务角色

1. 使用自定义信任策略创建角色。有关更多信息，请参阅《AWS Identity and Access Management 用户指南》中的使用自定义信任策略创建角色（控制台）过程。

2. 根据使用自定义信任策略创建角色（控制台）步骤使用以下自定义信任策略。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "ForAnyValue:ArnEquals": {
                    "aws:SourceArn": [
                        "arn:aws:cleanrooms:us-
east-1:111122223333:membership/jobRunnerMembershipId",
                        "arn:aws:cleanrooms:us-
east-1:444455556666:membership/rrMembershipId"
                    ]
                }
            }
        }
    ]
}
```

> **Note**
>
> 该信任策略引用了两种不同的策略 AWS 账户 IDs 来支持具有不同运营角色的 AWS Clean Rooms 协作：
>
> - 111122223333-该账户包含负责运行分析作业的成员资格（“作业运行者”）。此成员资格负责执行计算工作负载，并且需要担任此角色才能访问处理资源。
> - 444455556666-该账户包含具有结果接收者 (RR) 责任的成员资格。该成员资格有权接收和访问分析作业的输出，并且需要角色访问权限才能将结果写入指定位置。
>
> 这种配置支持一方运行计算分析，而另一方接收并管理结果的 AWS Clean Rooms 场景。通过相同的执行角色，这两个角色都需要不同但互补的权限。该`aws:SourceArn`条件可

确保只有源自这两个特定成员资格的 AWS Clean Rooms 操作才能担任该角色，从而在支持分布式作业执行和结果管理工作流程的同时维护安全性。

3. 根据使用自定义信任策略创建角色（控制台）步骤使用以下权限策略。

> **Note**
>
> 以下示例策略支持写入 Amazon S3 所需的权限。但是，您可能需要修改此策略，具体取决于您设置 S3 的方式。
> 您的 Amazon S3 资源必须与 AWS Clean Rooms 协作资源 AWS 区域 相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:PutObject"
            ],
            "Resource": "arn:aws:s3:::bucket/optionalPrefix/*",
            "Condition":{
                "StringEquals":{
                    "s3:ResourceAccount":[
                        "s3BucketOwnerAccountId"
                    ]
                }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetBucketLocation",
                "s3:ListBucket"
            ],
            "Resource": "arn:aws:s3:::bucket",
            "Condition":{
                "StringEquals":{
                    "s3:ResourceAccount":[
```

```
                    "s3BucketOwnerAccountId"
                ]
            }
        }
    }
]
}
```

4. 用你自己的信息替换每一个*placeholder*信息：

   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *jobRunnerAccountId*— S3 存储桶所在的 AWS 账户 ID。
   - *jobRunnerMembershipId*— 可以查询和运行作业的成员的会员 ID。可以在协作的详细信息选项卡上找到成员身份 ID。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
   - *rrAccountId*— S3 存储桶所在的 AWS 账户 ID。
   - *rrMembershipId*— 可以接收结果的成员的会员 ID。可以在协作的详细信息选项卡上找到成员身份 ID。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
   - *bucket*— S3 存储桶的名称和位置。
   - *optionalPrefix*— 如果要将结果保存在特定的 S3 前缀下，则为可选前缀。
   - *s3BucketOwnerAccountId*— S3 存储桶所有者的 AWS 账户 ID。

5. 继续按照使用自定义信任策略创建角色（控制台）步骤创建角色。

创建服务角色来接收结果

> **Note**
>
> 如果您是只能接收结果的成员（在控制台中，您的成员能力为仅接收结果），请按照以下步骤操作。
> 如果您是既能查询也能接收结果的成员（在控制台中，您的成员能力既是查询又是接收结果），则可以跳过此步骤。

对于只能接收结果的协作成员， AWS Clean Rooms 使用服务角色将协作中查询数据的结果写入指定的 S3 存储桶。

可以使用两种方法创建此服务角色：

- 如果您拥有创建服务角色所必需的 IAM 权限，请使用 AWS Clean Rooms 控制台创建服务角色。
- 如果您没有`iam:CreateRole`、`iam:CreatePolicy`、`iam:AttachRolePolicy`权限或想要手动创建 IAM 角色，请执行以下任一操作：
  - 使用以下步骤使用自定义信任策略创建服务角色。
  - 要求管理员使用以下步骤创建服务角色。

> **Note**
>
> 只有在您没有使用 AWS Clean Rooms 控制台创建服务角色的必要权限时，您或您的 IAM 管理员才应遵循此过程。

使用自定义信任策略创建用于接收结果的服务角色

1. 使用自定义信任策略创建角色。有关更多信息，请参阅《AWS Identity and Access Management 用户指南》中的使用自定义信任策略创建角色（控制台）过程。
2. 根据使用自定义信任策略创建角色（控制台）步骤使用以下自定义信任策略。

   JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowIfExternalIdMatches",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "ArnLike": {
                    "sts:ExternalId":
 "arn:aws:*:region:*:dbuser:*/a1b2c3d4-5678-90ab-cdef-EXAMPLEaaaaa*"
                }
            }
        },
```

```
        {
            "Sid": "AllowIfSourceArnMatches",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "ForAnyValue:ArnEquals": {
                    "aws:SourceArn": [
                        "arn:aws:cleanrooms:us-
east-1:555555555555:membership/a1b2c3d4-5678-90ab-cdef-EXAMPLEaaaaa"
                    ]

                }
            }
        }
    ]
}
```

3. 根据使用自定义信任策略创建角色（控制台）步骤使用以下权限策略。

> **Note**
>
> 以下示例策略支持读取 AWS Glue 元数据及其相应的 Amazon S3 数据所需的权限。但是，您可能需要修改此策略，具体取决于您设置 S3 数据的方式。
> 您的 AWS Glue 资源和底层 Amazon S3 资源必须与 AWS Clean Rooms 协作 AWS 区域相同。

JSON

```
{

    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetBucketLocation",
                "s3:ListBucket"
```

```
            ],
            "Resource": [
                "arn:aws:s3:::bucket_name"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceAccount":"accountId"
                }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::bucket_name/optional_key_prefix/*"
            ],
            "Condition": {
                "StringEquals": {
                    "aws:ResourceAccount":"accountId"
                }
            }
        }
    ]
}
```

4. 用你自己的信息替换每一个*placeholder*信息：

   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *a1b2c3d4-5678-90ab-cdef-EXAMPLEaaaaa*— 可以查询的成员的会员 ID。可以在协作的详细信息选项卡上找到成员身份 ID。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
   - *arn:aws:cleanrooms:us-east-1:555555555555:membership/a1b2c3d4-5678-90ab-cdef-EXAMPLEaaaaa*— 可以查询的成员的单一会员 ARN。可以在协作的详细信息选项卡上找到成员身份 ARN。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
   - *bucket_name*— S3 存储桶的亚马逊资源名称 (ARN)。Amazon 资源名称 (ARN) 可在 Amazon S3 存储桶的属性选项卡上找到。
   - *accountId*— S3 存储桶所在的 AWS 账户 ID。

bucket_name/optional_key_prefix— 亚马逊 S3 中结果目标的亚马逊资源名称 (ARN)。Amazon 资源名称 (ARN) 可在 Amazon S3 存储桶的属性选项卡上找到。

5. 继续按照使用自定义信任策略创建角色（控制台）步骤创建角色。

为 AWS Clean Rooms ML 设置服务角色

执行相似建模所需的角色与使用自定义模型所需的角色不同。以下各节描述了执行每项任务所需的角色。

主题

- 为相似建模设置服务角色
- 为自定义建模设置服务角色

为相似建模设置服务角色

主题

- 创建服务角色以读取训练数据
- 创建服务角色以写入相似细分
- 创建服务角色以读取种子数据

创建服务角色以读取训练数据

AWS Clean Rooms 使用服务角色读取训练数据。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 `CreateRole` 权限，请要求您的管理员创建服务角色。

创建服务角色以训练数据集

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择创建策略。
4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略。

> **Note**
>
> 以下示例策略支持读取 AWS Glue 元数据及其相应的 Amazon S3 数据所需的权限。但是，您可能需要修改此策略，具体取决于您设置 S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
>
> 您的 AWS Glue 资源和底层 Amazon S3 资源必须与 AWS Clean Rooms 协作 AWS 区域相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "glue:GetDatabase",
                "glue:GetDatabases",
                "glue:GetTable",
                "glue:GetTables",
                "glue:GetPartitions",
                "glue:GetPartition",
                "glue:BatchGetPartition",
                "glue:GetUserDefinedFunctions"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:database/databases",
                "arn:aws:glue:us-east-1:111122223333:table/databases/tables",
                "arn:aws:glue:us-east-1:111122223333:catalog",
                "arn:aws:glue:us-east-1:111122223333:database/default"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "glue:CreateDatabase"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:database/default"
```

```
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:ListBucket",
                "s3:GetBucketLocation"
            ],
            "Resource": [
                "arn:aws:s3:::bucket"
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "111122223333"
                    ]
                }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject"
            ],
            "Resource": [
                "arn:aws:s3:::bucketFolders/*"
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "111122223333"
                    ]
                }
            }
        }
    ]
}
```

如果您需要使用 KMS 密钥解密数据，请将以下 AWS KMS 语句添加到之前的模板中：

```
{
            "Effect": "Allow",
```

```
            "Action": [
                "kms:Decrypt",
            ],
            "Resource": [
                "arn:aws:kms:region:accountId:key/keyId"
            ],
            "Condition": {
                "ArnLike": {
                        "kms:EncryptionContext:aws:s3:arn":
 "arn:aws:s3:::bucketFolders*"
                }
            }
        }
    ]
}
```

5. 用你自己的信息替换每一个*placeholder*信息：

   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *accountId*— S3 存储桶所在的 AWS 账户 ID。
   - *database/databasestable/databases/tables*、*catalog*、和 *database/default* — AWS Clean Rooms 需要访问的训练数据的位置。
   - *bucket*— S3 存储桶的亚马逊资源名称 (ARN)。Amazon 资源名称 (ARN) 可在 Amazon S3 存储桶的属性选项卡上找到。
   - *bucketFolders*— S3 存储桶中 AWS Clean Rooms 需要访问的特定文件夹的名称。

6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。

9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。

10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowAssumeRole",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEqualsIfExists": {
                    "aws:SourceAccount": ["111122223333"]
                },
                "ArnLikeIfExists": {
                    "aws:SourceArn": "arn:aws:cleanrooms-ml:us-
east-1:111122223333:training-dataset/*"
                }
            }
        }
    ]
}
```

永远SourceAccount是你的 AWS 账户。可以将 SourceArn 限制为特定的训练数据集，但仅在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

*accountId*是 AWS 账户 包含训练数据的 ID。

13. 选择下一步，在添加权限下面，输入您刚刚创建的策略的名称。（您可能需要重新加载页面。）
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

a. 查看选择受信任的实体，并在必要时进行编辑。
b. 在添加权限中查看权限，并在必要时进行编辑。
c. 查看标签，并在必要时添加标签。
d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

创建服务角色以写入相似细分

AWS Clean Rooms 使用服务角色将相似的区段写入存储桶。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 `CreateRole` 权限，请要求您的管理员创建服务角色。

创建服务角色以写入相似细分

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择创建策略。
4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略。

> **Note**
>
> 以下示例策略支持读取 AWS Glue 元数据及其相应的 Amazon S3 数据所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
> 您的 AWS Glue 资源和底层 Amazon S3 资源必须与 AWS Clean Rooms 协作 AWS 区域相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
    {
            "Effect": "Allow",
            "Action": [
```

```
                "s3:ListBucket",
                "s3:GetBucketLocation"
            ],
            "Resource": [
                "arn:aws:s3:::buckets"
            ],
            "Condition":{
                "StringEquals":{
                    "s3:ResourceAccount":[
                        "accountId"
                    ]
                }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::bucketFolders/*"
            ],
            "Condition":{
                "StringEquals":{
                    "s3:ResourceAccount":[
                        "accountId"
                    ]
                }
            }
        }
    ]
}
```

如果您需要使用 KMS 密钥加密数据，请将以下 AWS KMS 语句添加到模板中：

```
{
            "Effect": "Allow",
            "Action": [
                "kms:Encrypt",
                "kms:GenerateDataKey*",
                "kms:ReEncrypt*",
            ],
```

```
            "Resource": [
                "arn:aws:kms:region:accountId:key/keyId"
            ],
            "Condition": {
                "ArnLike": {
                        "kms:EncryptionContext:aws:s3:arn":
 "arn:aws:s3:::bucketFolders*"
                }
            }
        }
  ]
}
```

5. 用你自己的信息替换每一个*placeholder*信息：

   - *buckets*— S3 存储桶的亚马逊资源名称 (ARN)。Amazon 资源名称 (ARN) 可在 Amazon S3 存储桶的属性选项卡上找到。
   - *accountId*— S3 存储桶所在的 AWS 账户 ID。
   - *bucketFolders*— S3 存储桶中 AWS Clean Rooms 需要访问的特定文件夹的名称。
   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *keyId*— 加密数据所需的 KMS 密钥。
6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。
9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。
10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

    JSON

```
{
```

```
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowAssumeRole",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEqualsIfExists": {
                    "aws:SourceAccount": ["111122223333"]

                },
                "ArnLikeIfExists": {
                    "aws:SourceArn": "arn:aws:cleanrooms-ml:us-
east-1:111122223333:configured-audience-model/*"
                }
            }
        }
    ]
}
```

永远SourceAccount是你的 AWS 账户。可以将 SourceArn 限制为特定的训练数据集，但仅在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

13. 选择下一步。
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

a. 查看选择受信任的实体，并在必要时进行编辑。
b. 在添加权限中查看权限，并在必要时进行编辑。
c. 查看标签，并在必要时添加标签。

d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

创建服务角色以读取种子数据

AWS Clean Rooms 使用服务角色读取种子数据。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 `CreateRole` 权限，请要求您的管理员创建服务角色。

创建服务角色以读取存储在 S3 存储桶中的种子数据。

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择 Create policy (创建策略)。
4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略之一。

> **Note**
>
> 以下示例策略支持读取 AWS Glue 元数据及其相应的 Amazon S3 数据所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
> 您的 AWS Glue 资源和底层 Amazon S3 资源必须与 AWS Clean Rooms 协作 AWS 区域相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::buckets"
            ],
```

```
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "accountId"
                    ]
                }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject"
            ],
            "Resource": [
                "arn:aws:s3:::bucketFolders/*"
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "accountId"
                    ]
                }
            }
        }
    ]
}
```

> **Note**
>
> 以下示例策略支持读取 SQL 查询结果并将其用作输入数据所需的权限。但是，您可能需要修改此策略，具体取决于查询的结构。该策略不包含用于解密数据的 KMS 密钥。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowCleanRoomsStartQuery",
```

```
            "Effect": "Allow",
            "Action": [
                "cleanrooms:GetCollaborationAnalysisTemplate",
                "cleanrooms:GetSchema",
                "cleanrooms:StartProtectedQuery"
            ],
            "Resource": "*"
        },
        {
            "Sid": "AllowCleanRoomsGetAndUpdateQuery",
            "Effect": "Allow",
            "Action": [
                "cleanrooms:GetProtectedQuery",
                "cleanrooms:UpdateProtectedQuery"
            ],
            "Resource": [
                "arn:aws:cleanrooms:us-
east-1:111122223333:membership/queryRunnerMembershipId"
            ]
        }
    ]
}
```

如果您需要使用 KMS 密钥解密数据，请将以下 AWS KMS 语句添加到模板中：

```
{
            "Effect": "Allow",
            "Action": [
                "kms:Decrypt",
                "kms:DescribeKey"
            ],
            "Resource": [
                "arn:aws:kms:region:accountId:key/keyId"
            ],
            "Condition": {
                "ArnLike": {
                        "kms:EncryptionContext:aws:s3:arn":
 "arn:aws:s3:::bucketFolders*"
                }
            }
        }
  ]
```

```
}
```

5. 用你自己的信息替换每一个*placeholder*信息：

   - *buckets*— S3 存储桶的亚马逊资源名称 (ARN)。Amazon 资源名称 (ARN) 可在 Amazon S3 存储桶的属性选项卡上找到。
   - *accountId*— S3 存储桶所在的 AWS 账户 ID。
   - *bucketFolders*— S3 存储桶中 AWS Clean Rooms 需要访问的特定文件夹的名称。
   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *queryRunnerAccountId*— 将运行查询的账户的 AWS 账户 ID。
   - *queryRunnerMembershipId*— 可以查询的成员的会员 ID。可以在协作的详细信息选项卡上找到成员身份 ID。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
   - *keyId*— 加密数据所需的 KMS 密钥。
6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。
9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。
10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

    JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowAssumeRole",
            "Effect": "Allow",
            "Principal": {
```

```
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEqualsIfExists": {
                    "aws:SourceAccount": ["111122223333"]

                },
                "ArnLikeIfExists": {
                    "aws:SourceArn": "arn:aws:cleanrooms-ml:us-
east-1:111122223333:audience-generation-job/*"
                }
            }
        }
    ]
}
```

永远SourceAccount是你的 AWS 账户。可以将 SourceArn 限制为特定的训练数据集，但仅在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

13. 选择下一步。
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

a. 查看选择受信任的实体，并在必要时进行编辑。
b. 在添加权限中查看权限，并在必要时进行编辑。
c. 查看标签，并在必要时添加标签。
d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

为自定义建模设置服务角色

主题

为自定义 ML 建模创建服务角色-机器学习配置

AWS Clean Rooms 使用服务角色来控制谁可以创建自定义 ML 配置。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 `CreateRole` 权限，请要求您的管理员创建服务角色。

此角色允许您使用 Pu t MLConfiguration 操作。

创建服务角色以允许创建自定义 ML 配置

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择创建策略。
4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略。

> **Note**
>
> 以下示例策略支持访问和向 S3 存储桶写入数据以及发布 CloudWatch 指标所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
> 您的 Amazon S3 资源必须与 AWS Clean Rooms 协作资源 AWS 区域 相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowS3ObjectWriteForExport",
```

```
            "Effect": "Allow",
            "Action": [
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::bucket/*"
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ResourceAccount": [
                        "111122223333"
                    ]
                }
            }
        },
        {
            "Sid": "AllowS3KMSEncryptForExport",
            "Effect": "Allow",
            "Action": [
                "kms:Encrypt",
                "kms:GenerateDataKey*"
            ],
            "Resource": [
                "arn:aws:kms:us-east-1:111122223333:key/keyId"
            ],
            "Condition": {
                "StringLike": {
                    "kms:EncryptionContext:aws:s3:arn":
"arn:aws:s3:::bucket*"
                }
            }
        },
        {
            "Sid": "AllowCloudWatchMetricsPublishingForTrainingJobs",
            "Action": "cloudwatch:PutMetricData",
            "Resource": "*",
            "Effect": "Allow",
            "Condition": {
                "StringLike": {
                    "cloudwatch:namespace": "/aws/cleanroomsml/*"
                }
            }
        },
        {
```

```
            "Sid": "AllowCloudWatchLogsPublishingForTrainingOrInferenceJobs",
            "Effect": "Allow",
            "Action": [
                "logs:CreateLogGroup",
                "logs:CreateLogStream",
                "logs:DescribeLogStreams",
                "logs:PutLogEvents"
            ],
            "Resource": [
                "arn:aws:logs:us-east-1:111122223333:log-group:/aws/
cleanroomsml/*"
            ]
        }
    ]
}
```

5. 用你自己的信息替换每一个*placeholder*信息：

   - *bucket*— S3 存储桶的亚马逊资源名称 (ARN)。Amazon 资源名称 (ARN) 可在 Amazon S3 存储桶的属性选项卡上找到。
   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *accountId*— S3 存储桶所在的 AWS 账户 ID。
   - *keyId*— 加密数据所需的 KMS 密钥。

6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。

9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。

10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "111122223333"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:cleanrooms:us-
east-1:111122223333:membership/membershipID"
                }
            }
        }
    ]
}
```

永远SourceAccount是你的 AWS 账户。可以将 SourceArn 限制为特定的训练数据集，但仅在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

13. 选择下一步。
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

a. 查看选择受信任的实体，并在必要时进行编辑。

b. 在添加权限中查看权限，并在必要时进行编辑。
c. 查看标签，并在必要时添加标签。
d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

创建服务角色以提供自定义 ML 模型

AWS Clean Rooms 使用服务角色来控制谁可以创建自定义 ML 模型算法。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 `CreateRole` 权限，请要求您的管理员创建服务角色。

此角色允许您使用CreateConfiguredModelAlgorithm操作。

创建服务角色以允许成员提供自定义 ML 模型

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择创建策略。
4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略。

> **Note**
>
> 以下示例策略支持检索包含模型算法的 docker 镜像所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
> 您的 Amazon S3 资源必须与 AWS Clean Rooms 协作资源 AWS 区域 相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowECRImageDownloadForTrainingAndInferenceJobs",
            "Effect": "Allow",
            "Action": [
```

```
                "ecr:BatchGetImage",
                "ecr:BatchCheckLayerAvailability",
                "ecr:GetDownloadUrlForLayer"
            ],
            "Resource": "arn:aws:ecr:us-
east-1:111122223333:repository/repoName"
        }
    ]
}
```

5. 用你自己的信息替换每一个*placeholder*信息：

   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *accountId*— S3 存储桶所在的 AWS 账户 ID。
   - *repoName*— 包含您的数据的存储库的名称。
6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。
9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。
10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

    JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
```

```
        }
    ]
}
```

始终SourceAccount是你 AWS 账户 的。SourceArn可以仅限于特定的训练数据集，但只能在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

13. 选择下一步。
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

    a. 查看选择受信任的实体，并在必要时进行编辑。
    b. 在添加权限中查看权限，并在必要时进行编辑。
    c. 查看标签，并在必要时添加标签。
    d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

创建用于查询数据集的服务角色

AWS Clean Rooms 使用服务角色来控制谁可以查询将用于自定义 ML 建模的数据集。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 CreateRole 权限，请要求您的管理员创建服务角色。

此角色允许您使用 “创建MLInput频道” 操作。

创建服务角色以允许成员查询数据集

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择创建策略。

4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略。

> **Note**
>
> 以下示例策略支持查询将用于自定义 ML 建模的数据集所需的权限。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
> 您的 Amazon S3 资源必须与 AWS Clean Rooms 协作资源 AWS 区域 相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowCleanRoomsStartQueryForMLInputChannel",
            "Effect": "Allow",
            "Action": "cleanrooms:StartProtectedQuery",
            "Resource": "*"
        },
        {
            "Sid":
 "AllowCleanroomsGetSchemaAndGetAnalysisTemplateForMLInputChannel",
            "Effect": "Allow",
            "Action": [
                "cleanrooms:GetSchema",
                "cleanrooms:GetCollaborationAnalysisTemplate"
            ],
            "Resource": "*"
        },
        {
            "Sid": "AllowCleanRoomsGetAndUpdateQueryForMLInputChannel",
            "Effect": "Allow",
            "Action": [
                "cleanrooms:GetProtectedQuery",
                "cleanrooms:UpdateProtectedQuery"
            ],
            "Resource": [
                "arn:aws:cleanrooms:us-
east-1:111122223333:membership/queryRunnerMembershipId"
```

```
            ]
        }
    ]
}
```

5. 用你自己的信息替换每一个*placeholder*信息：

   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *queryRunnerAccountId*— 将运行查询的账户的 AWS 账户 ID。
   - *queryRunnerMembershipId*— 可以查询的成员的会员 ID。可以在协作的详细信息选项卡上找到成员身份 ID。这样可以确保 AWS Clean Rooms 只有当该成员在此协作中运行分析时才担任该角色。
6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。
9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。
10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

    JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
    ]
```

```
}
```

始终SourceAccount是你 AWS 账户 的。SourceArn可以仅限于特定的训练数据集，但只能在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

13. 选择下一步。
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

    a. 查看选择受信任的实体，并在必要时进行编辑。
    b. 在添加权限中查看权限，并在必要时进行编辑。
    c. 查看标签，并在必要时添加标签。
    d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

创建服务角色以创建已配置的表关联

AWS Clean Rooms 使用服务角色来控制谁可以创建已配置的表关联。如果您具有必要的 IAM 权限，则可以使用控制台创建此角色。如果您不具备 CreateRole 权限，请要求您的管理员创建服务角色。

此角色允许您使用 CreateConfiguredTableAssociation 操作。

创建服务角色以允许创建已配置的表关联

1. 使用您的管理员账户登录 IAM 控制台 (https://console.aws.amazon.com/iam/)。
2. 在访问管理下，选择策略。
3. 选择创建策略。
4. 在策略编辑器中，选择 JSON 选项卡，然后复制粘贴以下策略。

> **Note**
>
> 以下示例策略支持创建已配置的表关联。但是，您可能需要修改此政策，具体取决于您设置 Amazon S3 数据的方式。该策略不包含用于解密数据的 KMS 密钥。
> 您的 Amazon S3 资源必须与 AWS Clean Rooms 协作资源 AWS 区域 相同。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": [
                "kms:Decrypt",
                "kms:DescribeKey"
            ],
            "Resource": "arn:aws:kms:us-east-1:111122223333:key/KMS-key-ID",
            "Effect": "Allow"
        },
        {
            "Action": [
                "s3:ListBucket",
                "s3:GetBucketLocation"
            ],
            "Resource": "arn:aws:s3:::bucket-name",
            "Effect": "Allow"
        },
        {
            "Action": "s3:GetObject",
            "Resource": "arn:aws:s3:::bucket-name/*",
            "Effect": "Allow"
        },
        {
            "Action": [
                "glue:GetDatabase",
                "glue:GetDatabases",
                "glue:GetTable",
                "glue:GetTables",
                "glue:GetPartitions",
```

```
                "glue:GetPartition",
                "glue:BatchGetPartition"
            ],
            "Resource": [
                "arn:aws:glue:us-east-1:111122223333:catalog",
                "arn:aws:glue:us-east-1:111122223333:database/Glue database
 name",
                "arn:aws:glue:us-east-1:111122223333:table/Glue database
 name/Glue table name"
            ],
            "Effect": "Allow"
        },
        {
            "Action": [
                "glue:GetSchema",
                "glue:GetSchemaVersion"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

替换占位符资源 ARNs

使用此策略时，必须将占位符资源标识符替换为实际 ARNs 资源的标识符：

- AWS KMS 密钥资源：*KMS-key-ID*替换为加密您的 Amazon S3 数据的实际 AWS KMS 密钥 ID。密钥必须位于拥有目录资源的同一个账户 (111122223333) 中。 AWS Glue
- Amazon S3 存储桶资源：*bucket-name*替换为包含您的 AWS Glue 表数据的 Amazon S3 存储桶的实际名称。请注意，Amazon S3 存储桶 ARNs 不包含账户， IDs 因为存储桶名称是全球唯一的。
- AWS Glue 资源：将以下占位符替换为您的实际资源名称：
  - *Glue database name*-您的 AWS Glue 数据库名称
  - *Glue table name*-你的 AWS Glue 桌子的名字

所有 AWS Glue 资源（目录、数据库和表）必须相同 AWS 账户 (111122223333)，以确保访问权限的一致性。此帐户应与拥有用于数据加密的 AWS KMS 密钥的帐户相同，从而为您的 AWS Clean Rooms 数据资源创建统一的安全边界。

5. 用你自己的信息替换每一个*placeholder*信息：

   - *KMS key used to encrypt the Amazon S3 data*— 用于加密 Amazon S3 数据的 KMS 密钥。要解密数据，您需要提供用于加密数据的 KMS 密钥。
   - *Amazon S3 bucket of AWS Glue table*— 包含包含您的数据的 AWS Glue 表的 Amazon S3 存储桶的名称。
   - *region* - AWS 区域的名称。例如 **us-east-1**。
   - *accountId*— 拥有数据的账户的 AWS 账户 ID。
   - *AWS Glue database name*— 包含您的数据的 AWS Glue 数据库的名称。
   - *AWS Glue table name*-包含您的数据的 AWS Glue 表的名称。
6. 选择下一步。
7. 对于查看并创建，输入策略名称和描述，然后查看摘要。
8. 选择创建策略。

   您已经为创建了策略 AWS Clean Rooms。
9. 在 Access management（访问管理）下，请选择 Roles（角色）。

   通过使用角色，您可以创建短期凭证，建议这样做以提高安全性。您也可以选择用户来创建长期凭证。
10. 选择创建角色。
11. 在创建角色向导中，对于可信实体类型，选择自定义信任策略。
12. 将以下自定义信任策略复制粘贴到 JSON 编辑器中。

    JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
```

```
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
    ]
}
```

始终SourceAccount是你 AWS 账户 的。SourceArn可以仅限于特定的训练数据集，但只能在创建该数据集之后。由于您还不知道训练数据集 ARN，因此在此处指定了通配符。

13. 选择下一步。
14. 选中您创建的策略旁边的复选框，然后选择下一步。
15. 对于命名、查看和创建，输入角色名称和描述。

> **Note**
>
> 角色名称必须与授予可以查询和接收结果的成员和成员角色的 passRole 权限中的模式相匹配。

a. 查看选择受信任的实体，并在必要时进行编辑。
b. 在添加权限中查看权限，并在必要时进行编辑。
c. 查看标签，并在必要时添加标签。
d. 选择创建角色。

您已经为创建了服务角色 AWS Clean Rooms。

中的合作和会员资格 AWS Clean Rooms

协作是一个安全的逻辑边界， AWS Clean Rooms 成员可以在其中对已配置的表进行分析。

中的任何成员 AWS Clean Rooms 都可以创建协作。

协作创建者可以指定一个成员来分析已配置的表并接收结果。但是，协作创建者可能希望阻止可以运行分析的成员访问查询结果。在这种情况下，协作创建者可以指定一个成员可以查询，或者一个成员可以运行查询和作业，另一个成员可以接收结果。

在大多数情况下，可以查询的成员或可以查询和运行作业的成员也是支付计算费用的成员。但是，协作创建者可以将其他成员配置为负责为查询计算费用付费。

有关如何使用创建协作的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

主题

创建协作

有三种方法可以在中创建协作 AWS Clean Rooms。

最基本的形式是查询协作。这种合作侧重于 SQL 查询分析，并保持一个由两个主要角色组成的简单结构：一个成员可以运行查询，另一个成员可以接收结果。这种基本的协作设置非常适合简单的数据分析任务。

第二种形式是查询和作业协作，它通过合并 SQL 查询和 PySpark 作业来扩展功能。这种协作设置保持相同的基本角色结构，但扩展了权限以包括任务执行。一个值得注意的要求是，创建 PySpark 分析模板的成员也必须是接收结果的成员，从而确保分析过程中的明确问责制。

第三种形式，机器学习建模协作，专为机器学习工作流程而构建。此协作设置又增加了两个角色：一个用于需要训练模型结果的用户，另一个用于需要使用这些模型进行预测的结果的用户。这种协作设置可帮助协作成员共同处理复杂的数据项目，同时保持每个人的角色和权限清晰。

以下主题说明了如何为查询、作业和机器学习建模创建协作。

主题

- 为查询创建协作
- 为查询和作业创建协作
- 为 ML 建模创建协作模式

为查询创建协作

在此过程中，您作为协作创建者将执行以下任务：

- 创建协作。
- 邀请一个或多个成员加入协作。
- 为成员分配权限，例如可以查询的成员和可以接收结果的成员。

  如果协作创建者也是可以接收结果的成员，则他们会指定结果的目的地和格式。他们还提供服务角色 Amazon 资源名称 (ARN)，用于将结果写入结果目的地。
- 配置哪位成员负责支付协作中的计算成本。

在开始之前，请确保您已完成以下先决条件：

- 您拥有要邀请参与合作的每位成员的姓名和 AWS 账户 ID。
- 您有权与协作的所有成员共享每个成员的姓名和 AWS 账户 ID。

> **Note**
>
> 创建协作后，您无法添加更多成员。

有关如何使用创建协作的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

为查询创建协作

1. 登录 AWS 管理控制台 并使用将充当协作创建者的AWS Clean Rooms 控制台打开控制台。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 在右上角，选择创建协作。
4. 对于步骤 1: 定义协作，请执行以下操作：

   a. 在详细信息中，输入协作的名称和描述。

      受邀参与协作的协作成员将可以看到这些信息。名称和描述可帮助他们了解协作的意义。

   b. 对于成员：

      i. 对于成员 1: 您，输入您希望在协作中显示的成员显示名称。

      > **Note**
      >
      > 会员 AWS 账户 ID 会自动包含您的 AWS 账户 ID。

      ii. 在“成员 2”中，输入要邀请参与协作的成员的成员显示名称和成员 AWS 账户 ID。

      所有受邀参与协作的人都可以看到成员显示名称和成员 AWS 账户 ID。输入并保存这些字段的值后将不可编辑这些值。

      > **Note**
      >
      > 您必须告知协作成员，协作中所有受邀和活跃的协作者都将看到他们的成员 AWS 账户 ID 和成员显示名称。

      iii. 如果要添加其他成员，请选择添加其他成员。然后，为每位成员输入成员的显示名称和成员 AWS 账户 ID，他们可以提供您想邀请参与协作的数据。

   c. 如果要启用分析日志记录，请选中“启用分析记录”复选框。

      - 选中“支持的日志类型”下的“来自查询的日志”复选框。

        您将在您的 Amazon Logs 账户中收到通过 SQL 查询生成的 CloudWatch 日志。

   d. 如果要启用详细监控，请选中“启用详细监控”复选框。

分析运行者和配置的付款人可以在创建成员资格时选择启用详细指标。启用后，将发布详细的监控指标， CloudWatch 用于对协作进行运营监控，包括查询性能和资源利用率。这些指标将在各自的版本中提供给分析运行者和配置的付款 AWS 账户人。

有关 CloudWatch 定价的更多信息，请参阅CloudWatch 定价。

e. 在 “允许的查询结果区域” 下，选择一个或多个要将查询结果发送到 AWS 区域 的位置。

默认情况下，仅选择当前区域（例如弗吉尼亚北部 us-east-1）。

> **Important**
>
> 启用跨区域查询结果交付后，您的结果可能会在来源区域之外进行处理和存储。

有关区域的更多信息，请参阅中的区域和终端节点AWS 一般参考。

f. （可选）通过配置无需手动批准变更请求即可自动更改哪些设置，通过自动更改请求批准来管理对数据的访问权限。默认情况下，某些设置只能通过提交变更请求来更改，变更请求必须得到所有成员的批准才能生效。

- 授予成员能力-选择无需手动批准即可授予协作成员的能力。成员可以随时贡献数据。
  - 选择技能：
    - 贡献数据（始终启用）
    - 接收结果
  - 自动批准具有这些能力的新成员-如果允许，任何添加了上述所选能力的成员都将立即加入协作。添加了其他技能的成员仍需要手动批准才能加入。
- 可以自动撤消的技能-选择无需手动批准即可撤消的技能。成员可以随时贡献数据。
  - 选择技能：
    - 贡献数据（始终启用）
    - 接收结果

如果选择此选项，则可以通过协作详细信息页面的详细信息选项卡上的更改请求历史记录来跟踪所有协作配置的修改。

g. （可选）如果要启用加密计算功能，请选中 “启用加密计算” 复选框。

i. 选择以下加密覆盖率参数：

- 允许 plaintext 列

  如果您需要完全加密的表，请选择 “否”。

  如果您希望在加密表中允许 cleartext 列，请选择是。

  要在特定列上运行 SUM 或 AVG，这些列必须是 cleartext。

- 保留 NULL 值

  如果您不希望保留 NULL 值，请选择否。NULL 值不会在加密表中显示为 NULL。

  如果您希望保留 NULL 值，请选择是。NULL 值将在加密表中显示为 NULL。

ii. 选择以下指纹识别参数：

- 允许重复

  如果您不希望 fingerprint 列中允许重复条目，请选择否。

  如果您希望 fingerprint 列中允许重复条目，请选择是。

- 允许对具有不同名称的列进行 JOIN

  如果您不希望对具有不同名称的 fingerprint 列进行联接，请选择否。

  如果您希望对具有不同名称的 fingerprint 列进行联接，请选择是。

有关加密计算参数的更多信息，请参阅加密计算参数。

有关如何加密数据以便在中使用的更多信息 AWS Clean Rooms，请参阅使用 Clean Rooms 加密计算准备加密的数据表。

> **Note**
>
> 在完成下一步之前，请仔细验证这些配置。创建协作后，您只能编辑协作名称、描述以及日志是否存储在 Amazon Lo CloudWatch gs 中。

h. 如果要为协作资源启用标签，请选择添加新标签，然后输入键和值对。

i. 选择下一步。

5. 对于步骤 2：指定成员能力，对于使用查询和作业进行分析，在 “支持的分析类型” 下，将 “查询” 复选框保持选中状态，然后根据您的目标采取建议的操作。

您的目标	推荐操作
查询协作中的数据并接收结果	1. 将自己选为可以运行查询的成员。 2. 选择自己作为可以从下拉列表中接收分析结果的成员。
查询协作中的数据并分配其他成员来接收结果	1. 将自己选为可以运行查询的成员。 2. 从下拉列表中选择可以从分析中接收结果的成员。
接收协作中的查询结果并分配其他成员来查询数据	1. 从下拉列表中选择可以运行查询的成员。 2. 选择自己作为可以从下拉列表中接收分析结果的成员。
创建和管理协作，分配其他成员来查询数据，并分配其他成员来接收结果	1. 从下拉列表中选择可以运行查询的成员。 2. 从下拉列表中选择可以从分析中接收结果的成员。

a. 如果您使用的是Clean Rooms ML，则使用专门构建的工作流程进行机器学习建模，

   i. （可选）从下拉列表中选择可以从经过训练的模型接收输出的成员。

   ii. （可选）从下拉列表中选择可以从模型推理中接收输出的成员。

b. 使用查看 ID 解析下的成员能力 AWS Entity Resolution 数据匹配服务。

c. 选择下一步。

6. 对于步骤 3：配置付款，对于使用查询进行分析，请根据您的目标采取以下操作之一。

您的目标	推荐操作
将可以运行查询的成员指定为支付查询计算费用的成员	1. 对于使用查询进行分析，请选择将为查询付费的成员与可以运行查询的成员相同。 2. 选择下一步。
分配其他成员来支付查询计算费用	1. 对于使用查询进行分析，请选择自己作为将为查询付费的成员。

您的目标	推荐操作
	2. 选择下一步。

对于使用专门构建的工作流程进行机器学习建模，配置的相似模型的创建者是将为相似建模付费的成员。

对于 ID 解析 AWS Entity Resolution 数据匹配服务，ID 映射表的创建者是将为 ID 映射表付费的成员。

7. 对于 “步骤 4：配置成员资格”，请选择以下选项之一：

   Yes, join by creating membership now

   1. 对于结果设置的默认设置，对于查询结果设置，如果您是可以接收结果的成员，
      a. 对于 Amazon S3 中的结果目标，输入亚马逊 S3 目标或选择 “浏览 S3” 选择 S3 存储桶。
      b. 对于查询结果格式，请选择 CSV 或 PARQUET。
      c. （仅限 Spark）对于结果文件，请选择 “多个” 或 “单个”。
      d. （可选）对于服务访问权限，如果您想将最长需要 24 小时的查询传送到 S3 目标，请选中 “添加服务角色以支持最长需要 24 小时才能完成的查询” 复选框。

         最长需要 24 小时才能完成的大型查询将传送到您的 S3 目标。

         如果您不选中该复选框，则只有在 12 小时内完成的查询才会发送到您的 S3 位置。
      e. 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。

如果你选择...	操作...
创建并使用新的服务角色	• AWS Clean Rooms 使用此表所需的策略创建服务角色。 • 默认服务角色名称为 `cleanrooms-result-receiver-<timestamp>` 。 • 您必须拥有创建角色并附加策略的权限。

如果你选择...	操作...
使用现有服务角色	i. 从下拉列表中选择一个现有服务角色名称。 如果您有列出角色的权限，则会显示角色列表。 如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。 ii. 通过选择在 IAM 中查看外部链接来查看服务角色。 如果没有现有的服务角色，则使用现有服务角色选项不可用。 默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

> **Note**
>
> - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
> - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
> - 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

2. 在“日志”设置中，为 Amazon Log CloudWatch s 中的日志存储选择以下选项之一：

> **Note**
>
> 如果您选择启用查询日志记录，则会出现“日志设置”部分。

a. 选择 “开启”，与您相关的查询日志将存储在您的 Amazon CloudWatch Logs 账户中。

每个成员只能接收他们发起的查询或包含其数据的查询的日志。

可以接收结果的成员还会收到协作中运行的所有查询的日志，即使查询中未访问他们的数据也是如此。

在 “支持的日志类型” 下，“查询日志” 复选框默认处于启用状态。

> **Note**
>
> 开启查询日志后，可能需要几分钟才能设置日志存储并开始在 Amazon Logs 中接收 CloudWatch 日志。在这段短暂的时间内，可以查询的成员可能会运行实际上并未发送日志的查询。

b. 选择 “关闭”，与您相关的查询日志将不会存储在您的 Amazon CloudWatch Logs 账户中。

3. 如果要为成员身份资源启用标签，请选择添加新标签，然后输入键和值对。
4. 如果您是为 Query 计算付费的成员，请选中 “我同意支付此协作中的计算费用” 复选框，表示您接受。

> **Note**
>
> 必须选中此复选框才能继续。
> 有关如何计算费用的更多信息，请参阅的定价 AWS Clean Rooms。

如果您是支付查询计算费用的会员，但不是可以查询的成员，则建议您使用 AWS Budgets 来配置预算， AWS Clean Rooms 并在达到最高预算后接收通知。有关设置预算的更多信息，请参阅《AWS Cost Management 用户指南》中的使用 AWS Budgets管理成本。有关设置通知的更多信息，请参阅《AWS Cost Management 用户指南》中的针对预算通知创建 Amazon SNS 主题。如果已达到预算上限，您可以联系可以查询的成员或退出协作。如果您退出协作，将不再允许运行查询，因此将不再向您收取查询计算费用。

5. 选择下一步。

同时创建协作和您的成员身份。

您在协作中的状态为活跃。

No, I will create a membership later

1. 选择下一步。

   仅创建协作。

   您在协作中的状态为非活跃。

8. 对于“步骤 5：查看并创建”，请执行以下操作：

   a. 查看您在之前的步骤中所做的选择，并在必要时进行编辑。

   b. 从以下选项中选择一个。

如果您选择了...	则选择...
同步创建成员身份和协作（是，立即通过创建成员身份来加入）	创建协作和成员身份
创建协作，此时不创建成员身份（不，我将稍后创建成员身份）	创建协作

成功创建协作后，您可以在协作下看到协作详细信息页面。

您现在已准备好执行以下操作：

- 准备好要分析的数据表 AWS Clean Rooms。（如果您想分析自己的事件数据或要查询身份数据，则可选。）
- 将配置表与协作关联。（如果您想分析自己的事件数据，则可选。）
- 为配置表添加分析规则。（如果您想分析自己的事件数据，则可选。）
- 创建成员身份并加入协作。（如果您已经创建了成员身份，则是可选的。）
- 邀请成员加入协作。

为查询和作业创建协作

在此过程中，您作为协作创建者将执行以下任务：

- 创建协作。
- 邀请一个或多个成员加入协作。
- 为成员分配权限，例如可以运行查询和作业的成员以及可以接收结果的成员。

  如果协作创建者也是可以接收结果的成员，则他们会指定结果的目的地和格式。他们还提供服务角色 Amazon 资源名称 (ARN)，用于将结果写入结果目的地。
- 配置哪位成员负责支付协作中的查询和作业计算费用。

在开始之前，请确保您已完成以下先决条件：

- 您拥有要邀请参与合作的每位成员的姓名和 AWS 账户 ID。
- 您有权与协作的所有成员共享每个成员的姓名和 AWS 账户 ID。

> **Note**
>
> 创建协作后，您无法添加更多成员。

有关如何使用创建协作的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

为查询和作业创建协作

1. 登录 AWS 管理控制台 并使用将充当协作创建者的AWS Clean Rooms 控制台打开控制台。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 在右上角，选择创建协作。
4. 对于步骤 1: 定义协作，请执行以下操作：

   a. 在详细信息中，输入协作的名称和描述。

      受邀参与协作的协作成员将可以看到这些信息。名称和描述可帮助他们了解协作的意义。

   b. 对于成员：

      i. 对于成员 1: 您，输入您希望在协作中显示的成员显示名称。

> **Note**
>
> 会员 AWS 账户 ID 会自动包含您的 AWS 账户 ID。

ii. 在“成员 2”中，输入要邀请参与协作的成员的成员显示名称和成员 AWS 账户 ID。

所有受邀参与协作的人都可以看到成员显示名称和成员 AWS 账户 ID。输入并保存这些字段的值后将不可编辑这些值。

> **Note**
>
> 您必须告知协作成员，协作中所有受邀和活跃的协作者都将看到他们的成员 AWS 账户 ID 和成员显示名称。

iii. 如果要添加其他成员，请选择添加其他成员。然后，为每位成员输入成员的显示名称和成员 AWS 账户 ID，他们可以提供您想邀请参与协作的数据。

c. 如果要启用分析日志记录，请选中启用分析日志记录复选框，然后选择支持的日志类型。

- 如果要接收从 SQL 查询生成的日志，请选中“来自查询的日志”复选框。
- 如果要使用接收作业生成的日志 PySpark，请选中“来自作业的日志”复选框。

d. 如果要启用详细监控，请选中“启用详细监控”复选框。

分析运行者和配置的付款人可以在创建成员资格时选择启用详细指标。启用后，将发布详细的监控指标， CloudWatch 用于对协作进行运营监控，包括查询性能和资源利用率。这些指标将在各自的版本中提供给分析运行者和配置的付款 AWS 账户人。

有关 CloudWatch 定价的更多信息，请参阅CloudWatch 定价。

e. 在“允许的查询结果区域”下，选择一个或多个要将查询结果发送到 AWS 区域 的位置。

默认情况下，仅选择当前区域（例如弗吉尼亚北部 us-east-1）。

> **Important**
>
> 启用跨区域查询结果交付后，您的结果可能会在来源区域之外进行处理和存储。

有关区域的更多信息，请参阅中的区域和终端节点AWS 一般参考。

f. （可选）通过配置无需手动批准变更请求即可自动更改哪些设置，通过自动更改请求批准来管理对数据的访问权限。默认情况下，某些设置只能通过提交变更请求来更改，变更请求必须得到所有成员的批准才能生效。

- 授予成员能力-选择无需手动批准即可授予协作成员的能力。成员可以随时贡献数据。
  - 选择技能：
    - 贡献数据（始终启用）
    - 接收结果
  - 自动批准具有这些能力的新成员-如果允许，任何添加了上述所选能力的成员都将立即加入协作。添加了其他技能的成员仍需要手动批准才能加入。
- 可以自动撤消的技能-选择无需手动批准即可撤消的技能。成员可以随时贡献数据。
  - 选择技能：
    - 贡献数据（始终启用）
    - 接收结果

如果选择此选项，则可以通过协作详细信息页面的详细信息选项卡上的更改请求历史记录来跟踪所有协作配置的修改。

g. （可选）如果要启用加密计算功能，请选中 “启用加密计算” 复选框。

i. 选择以下加密覆盖率参数：

- 允许 plaintext 列

  如果您需要完全加密的表，请选择 “否”。

  如果您希望在加密表中允许 cleartext 列，请选择是。

  要在特定列上运行 SUM 或 AVG，这些列必须是 cleartext。

- 保留 NULL 值

  如果您不希望保留 NULL 值，请选择否。NULL 值不会在加密表中显示为 NULL。

  如果您希望保留 NULL 值，请选择是。NULL 值将在加密表中显示为 NULL。

ii. 选择以下指纹识别参数：

- 允许重复

如果您不希望 fingerprint 列中允许重复条目，请选择否。

如果您希望 fingerprint 列中允许重复条目，请选择是。

- 允许对具有不同名称的列进行 JOIN

  如果您不希望对具有不同名称的 fingerprint 列进行联接，请选择否。

  如果您希望对具有不同名称的 fingerprint 列进行联接，请选择是。

有关加密计算参数的更多信息，请参阅加密计算参数。

有关如何加密数据以便在中使用的更多信息 AWS Clean Rooms，请参阅使用 Clean Rooms 加密计算准备加密的数据表。

> **Note**
>
> 在完成下一步之前，请仔细验证这些配置。创建协作后，您只能编辑协作名称、描述以及日志是否存储在 Amazon Lo CloudWatch gs 中。

h. 如果要为协作资源启用标签，请选择添加新标签，然后输入键和值对。

i. 选择下一步。

5. 对于“步骤 2：指定成员能力”，请执行以下操作：

   a. 对于使用查询和作业进行分析，在支持的分析类型下，选择作业复选框。

      默认情况下，“查询” 复选框处于选中状态。

      i. 从下拉列表中选择可以运行查询和作业的成员。

      ii. 从下拉列表中选择可以从分析中接收结果的成员。

      > **Note**
      >
      > 创建 PySpark 分析模板的成员也必须是接收结果的成员。

   b. 如果您使用的是Clean Rooms ML，则使用专门构建的工作流程进行机器学习建模，

      i. （可选）从下拉列表中选择可以从经过训练的模型接收输出的成员。

      ii. （可选）从下拉列表中选择可以从模型推理中接收输出的成员。

c. 使用查看 ID 解析下的成员能力 AWS Entity Resolution 数据匹配服务。

d. 选择下一步。

6. 对于第 3 步：配置付款，

a. 对于使用查询和作业进行分析，请选择将为查询和工作付费的成员。

您可以将可以运行查询和作业的成员指定为支付查询和作业计算成本的成员。

您可以分配其他成员来支付查询费用和任务计算费用。

b. 对于使用专门构建的工作流程进行机器学习建模，配置的相似模型的创建者是将为相似建模付费的成员。

c. 对于 ID 解析 AWS Entity Resolution 数据匹配服务，ID 映射表的创建者是将为 ID 映射表付费的成员。

d. 选择下一步。

7. 对于 “步骤 4：配置成员资格”，请选择以下选项之一：

Yes, join by creating membership now

1. 对于结果设置的默认设置，对于查询结果设置，如果您是可以接收结果的成员，

a. 选中 “设置查询的默认设置” 复选框。对于 Amazon S3 中的结果目标，输入亚马逊 S3 目标或选择 “浏览 S3” 选择 S3 存储桶。

b. 对于查询结果格式，请选择 CSV 或 PARQUET。

c. （仅限 Spark）对于结果文件，请选择 “多个” 或 “单个”。

d. （可选）对于服务访问权限，如果您想将最长需要 24 小时的查询传送到 S3 目标，请选中 “添加服务角色以支持最长需要 24 小时才能完成的查询” 复选框。

最长需要 24 小时才能完成的大型查询将传送到您的 S3 目标。

如果您不选中该复选框，则只有在 12 小时内完成的查询才会发送到您的 S3 位置。

e. 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。

如果你选择...	操作...
创建并使用新的服务角色	• AWS Clean Rooms 使用此表所需的策略创建服务角色。

如果你选择...	操作...
	• 默认服务角色名称为 `cleanrooms-result-receiver-<timestamp>` 。 • 您必须拥有创建角色并附加策略的权限。
使用现有服务角色	i. 从下拉列表中选择一个现有服务角色名称。 如果您有列出角色的权限，则会显示角色列表。 如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。 ii. 通过选择在 IAM 中查看外部链接来查看服务角色。 如果没有现有的服务角色，则使用现有服务角色选项不可用。 默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

> **Note**
>
> - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
> - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
> - 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

2. 要查看 Job 结果，

   Example

   例如：**`s3://bucket/prefix`**

   a. 选中 “设置任务的默认设置” 复选框，然后通过输入 S3 目标在 Amazon S3 中指定结果目标，或者选择 “浏览 S3” 从可用 S3 存储桶列表中进行选择。

   b. 通过从下拉列表中选择现有服务角色名称来指定服务访问权限。

3. 在 “日志” 设置中，为 Amazon Log CloudWatch s 中的日志存储选择以下选项之一：

   > **Note**
   >
   > 如果您选择启用查询日志记录，则会出现 “日志设置” 部分。

   a. 选择 “开启”，与您相关的查询日志将存储在您的 Amazon CloudWatch Logs 账户中。

   每个成员只能接收他们发起的查询或包含其数据的查询的日志。

   可以接收结果的成员还会收到协作中运行的所有查询的日志，即使查询中未访问他们的数据也是如此。

   在 “支持的日志类型” 下，从协作创建者选择支持的日志类型中进行选择：

   在 “支持的日志类型” 下，“查询日志” 和 “Job Logs” 复选框默认处于启用状态。

   > **Note**
   >
   > 开启分析日志后，可能需要几分钟才能设置日志存储并开始在 Amazon Logs 中接收 CloudWatch 日志。在这段短暂的时间内，可以查询的成员可能会运行实际上并未发送日志的查询。

   b. 选择 “关闭”，与您相关的查询日志将不会存储在您的 Amazon CloudWatch Logs 账户中。

4. 如果要为成员资源启用成员资格标签，请选择添加新标签，然后输入密钥和值对。

5. 如果您是为 Query 计算或 J ob 计算或两者兼而有之付费的成员，请选中 “我同意支付此协作中的计算费用” 复选框，表示您接受。

> **Note**
>
> 必须选中此复选框才能继续。
> 有关如何计算费用的更多信息，请参阅的定价 AWS Clean Rooms。

如果您是支付查询计算费用的会员，但不是可以查询的成员，则建议您使用 AWS Budgets 来配置预算， AWS Clean Rooms 并在达到最高预算后接收通知。有关设置预算的更多信息，请参阅《AWS Cost Management 用户指南》中的使用 AWS Budgets管理成本。有关设置通知的更多信息，请参阅《AWS Cost Management 用户指南》中的针对预算通知创建 Amazon SNS 主题。如果已达到预算上限，您可以联系可以查询的成员或退出协作。如果您退出协作，将不再允许运行查询，因此将不再向您收取查询计算费用。

6. 选择下一步。

同时创建协作和您的成员身份。

您在协作中的状态为活跃。

No, I will create a membership later

1. 选择下一步。

仅创建协作。

您在协作中的状态为非活跃。

8. 对于“步骤 5：查看并创建”，请执行以下操作：

a. 查看您在之前的步骤中所做的选择，并在必要时进行编辑。

b. 从以下选项中选择一个。

如果您选择了...	则选择...
同步创建成员身份和协作（是，立即通过创建成员身份来加入）	创建协作和成员身份
创建协作，此时不创建成员身份（不，我将稍后创建成员身份）	创建协作

成功创建协作后，您可以在协作下看到协作详细信息页面。

您现在已准备好执行以下操作：

- 准备好要分析的数据表 AWS Clean Rooms。（如果您想分析自己的事件数据或要查询身份数据，则可选。）
- 将配置表与协作关联。（如果您想分析自己的事件数据，则可选。）
- 为配置表添加分析规则。（如果您想分析自己的事件数据，则可选。）
- 创建成员身份并加入协作。（如果您已经创建了成员身份，则是可选的。）
- 邀请成员加入协作。

为 ML 建模创建协作模式

在此过程中，您作为协作创建者将执行以下任务：

- 创建协作。
- 邀请一个或多个成员加入协作。
- 为成员分配能力，例如
  - 可以查询的会员
  - 可以收到结果的会员
  - 可以从经过训练的模型中接收输出结果的成员
  - 可以从模型推理中输出结果的成员

  如果协作创建者也是可以接收结果的成员，则他们会指定结果的目的地和格式。他们还提供服务角色 Amazon 资源名称 (ARN)，用于将结果写入结果目的地。
- 配置哪个成员负责支付协作中的计算成本、模型训练和模型推理成本。

在开始之前，请确保您已完成以下先决条件：

- 您拥有要邀请参与合作的每位成员的姓名和 AWS 账户 ID。
- 您有权与协作的所有成员共享每个成员的姓名和 AWS 账户 ID。

> **Note**
>
> 创建协作后，您无法添加更多成员。

有关如何使用创建协作的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

为机器学习建模创建协作模式

1. 登录 AWS 管理控制台 并使用将充当协作创建者的AWS Clean Rooms 控制台打开控制台。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 在右上角，选择创建协作。
4. 对于步骤 1: 定义协作，请执行以下操作：

   a. 在详细信息中，输入协作的名称和描述。

      受邀参与协作的协作成员将可以看到这些信息。名称和描述可帮助他们了解协作的意义。

   b. 对于成员：

      i. 对于成员 1: 您，输入您希望在协作中显示的成员显示名称。

      > **Note**
      >
      > 会员 AWS 账户 ID 会自动包含您的 AWS 账户 ID。

      ii. 在“成员 2”中，输入要邀请参与协作的成员的成员显示名称和成员 AWS 账户 ID。

      所有受邀参与协作的人都可以看到成员显示名称和成员 AWS 账户 ID。输入并保存这些字段的值后将不可编辑这些值。

      > **Note**
      >
      > 您必须告知协作成员，协作中所有受邀和活跃的协作者都将看到他们的成员 AWS 账户 ID 和成员显示名称。

      iii. 如果要添加其他成员，请选择添加其他成员。然后，为每位成员输入成员的显示名称和成员 AWS 账户 ID，他们可以提供您想邀请参与协作的数据。

   c. 如果要启用 Analysis 日志记录，请选中“启用分析日志记录”复选框，然后在“支持的日志类型”下选择“来自查询的日志”。

   d. 如果要启用详细监控，请选中“启用详细监控”复选框。

分析运行者和配置的付款人可以在创建成员资格时选择启用详细指标。启用后，将发布详细的监控指标， CloudWatch 用于对协作进行运营监控，包括查询性能和资源利用率。这些指标将在各自的版本中提供给分析运行者和配置的付款 AWS 账户人。

有关 CloudWatch 定价的更多信息，请参阅CloudWatch 定价。

e. 在 “允许的查询结果区域” 下，选择一个或多个要将查询结果发送到 AWS 区域 的位置。

默认情况下，仅选择当前区域（例如弗吉尼亚北部 us-east-1）。

> **Important**
>
> 启用跨区域查询结果交付后，您的结果可能会在来源区域之外进行处理和存储。

有关区域的更多信息，请参阅中的区域和终端节点AWS 一般参考。

f. （可选）通过配置无需手动批准变更请求即可自动更改哪些设置，通过自动更改请求批准来管理对数据的访问权限。默认情况下，某些设置只能通过提交变更请求来更改，变更请求必须得到所有成员的批准才能生效。

- 授予成员能力-选择无需手动批准即可授予协作成员的能力。成员可以随时贡献数据。
  - 选择技能：
    - 贡献数据（始终启用）
    - 接收结果
  - 自动批准具有这些能力的新成员-如果允许，任何添加了上述所选能力的成员都将立即加入协作。添加了其他技能的成员仍需要手动批准才能加入。
- 可以自动撤消的技能-选择无需手动批准即可撤消的技能。成员可以随时贡献数据。
  - 选择技能：
    - 贡献数据（始终启用）
    - 接收结果

如果选择此选项，则可以通过协作详细信息页面的详细信息选项卡上的更改请求历史记录来跟踪所有协作配置的修改。

g. （可选）如果要启用加密计算功能，请选中 “启用加密计算” 复选框。

i. 选择以下加密覆盖率参数：

- 允许 plaintext 列

  如果您需要完全加密的表，请选择 “否”。

  如果您希望在加密表中允许 cleartext 列，请选择是。

  要在特定列上运行 SUM 或 AVG，这些列必须是 cleartext。

- 保留 NULL 值

  如果您不希望保留 NULL 值，请选择否。NULL 值不会在加密表中显示为 NULL。

  如果您希望保留 NULL 值，请选择是。NULL 值将在加密表中显示为 NULL。

ii. 选择以下指纹识别参数：

- 允许重复

  如果您不希望 fingerprint 列中允许重复条目，请选择否。

  如果您希望 fingerprint 列中允许重复条目，请选择是。

- 允许对具有不同名称的列进行 JOIN

  如果您不希望对具有不同名称的 fingerprint 列进行联接，请选择否。

  如果您希望对具有不同名称的 fingerprint 列进行联接，请选择是。

有关加密计算参数的更多信息，请参阅加密计算参数。

有关如何加密数据以便在中使用的更多信息 AWS Clean Rooms，请参阅使用 Clean Rooms 加密计算准备加密的数据表。

> **Note**
>
> 在完成下一步之前，请仔细验证这些配置。创建协作后，您只能编辑协作名称、描述以及日志是否存储在 Amazon Lo CloudWatch gs 中。

h. 如果要为协作资源启用标签，请选择添加新标签，然后输入键和值对。

i. 选择下一步。

5. 对于步骤 2：指定成员能力，

a. 对于使用查询和作业进行分析，在支持的分析类型下，将查询复选框保持选中状态。

b. 对于运行查询，选择将启动模型训练的成员

c. 在 “从分析中接收结果” 中，选择一个或多个将接收查询结果的成员。

d. 对于使用专门构建的工作流程进行机器学习建模，

i. 对于接收来自训练模型的输出，请选择将接收经过训练的模型结果（包括模型工件和指标）的成员。

ii. 在 “接收模型推理的输出” 中，选择将接收模型推理结果的成员。

e. 使用查看 ID 解析下的成员能力 AWS Entity Resolution 数据匹配服务。

6. 对于第 3 步：配置付款，

a. 在 “使用查询进行分析” 下，对于 “按查询付费”，执行以下操作之一：

- 要让同一个成员付费并运行查询，请选择您为 “运行查询” 选择的同一个成员。
- 要让其他成员支付查询费用，请选择您的成员账户。

b. 对于使用专门构建的工作流程进行机器学习建模，

- 选择将为模型训练付费的成员。

c. 选择将为推理工作付费的成员。

d. 对于 Pay for look 相似建模，无需采取任何操作。配置的相似模型的创建者是将为相似建模付费的成员。

e. （可选）选择将为合成数据生成付费的成员。

f. 对于使用的 ID 解析 AWS Entity Resolution 数据匹配服务，无需执行任何操作。ID 映射表的创建者是将为 ID 映射表付费的成员。

7. 选择下一步。

8. 对于 “步骤 4：配置成员资格”，在 “协作成员资格” 下，选择以下选项之一：

Yes, join by creating membership now

1. 对于结果设置的默认设置，对于查询结果设置，如果您是可以接收结果的成员，

a. 选中 “设置查询的默认设置” 复选框。

b. 对于 Amazon S3 中的结果目标，输入亚马逊 S3 目标或选择 “浏览 S3” 选择 S3 存储桶。

c. 对于查询结果格式，请选择 CSV 或 PARQUET。

d. （仅限 Spark）对于结果文件，请选择 “多个” 或 “单个”。

e. （可选）如果要将最长 24 小时的查询发送到 S3 目标，请选中 “添加服务角色以支持最长需要 24 小时才能完成的查询” 复选框。

最长需要 24 小时才能完成的大型查询将传送到您的 S3 目标。

如果您不选中该复选框，则只有在 12 小时内完成的查询才会发送到您的 S3 位置。

f. 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。

如果你选择...	操作...
创建并使用新的服务角色	• AWS Clean Rooms 使用此表所需的策略创建服务角色。 • 默认服务角色名称为 `cleanrooms-result-receiver-<timestamp>` 。 • 您必须拥有创建角色并附加策略的权限。
使用现有服务角色	i. 从下拉列表中选择一个现有服务角色名称。 如果您有列出角色的权限，则会显示角色列表。 如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。 ii. 通过选择在 IAM 中查看外部链接来查看服务角色。 如果没有现有的服务角色，则使用现有服务角色选项不可用。 默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

> **Note**
>
> - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
> - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
> - 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

2. 对于 ML 配置，
   a. 选中 “创建 ML 配置” 复选框，然后通过输入 S3 目标来指定 Amazon S3 上的模型输出目标，或者选择 “浏览 S3” 从可用 S3 存储桶列表中进行选择。
   b. 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。
   c. 如果 S3 存储桶已加密，请选中使用 KMS 密钥加密目标存储桶复选框，然后输入AWS KMS key或选择创建 AWS KMS key以创建新的 KMS 密钥。
3. 如果要为成员资源启用成员资格标签，请选择添加新标签，然后输入密钥和值对。
4. 如果您是为 Query 计算付费的成员，请选中 “我同意支付此协作中的计算费用” 复选框，表示您接受。

> **Note**
>
> 必须选中此复选框才能继续。
> 有关如何计算费用的更多信息，请参阅的定价 AWS Clean Rooms。

如果您是支付查询计算费用的会员，但不是可以查询的成员，则建议您使用 AWS Budgets 来配置预算， AWS Clean Rooms 并在达到最高预算后接收通知。有关设置预算的更多信息，请参阅《AWS Cost Management 用户指南》中的使用 AWS Budgets管理成本。有关设置通知的更多信息，请参阅《AWS Cost Management 用户指南》中的针对预算通知创建 Amazon SNS 主题。如果已达到预算上限，您可以联系可以查询的成员或退出协作。如果您退出协作，将不再允许运行查询，因此将不再向您收取查询计算费用。

5. 选择下一步。

同时创建协作和您的成员身份。

您在协作中的状态为活跃。

No, I will create a membership later

1. 选择下一步。

   仅创建协作。

   您在协作中的状态为非活跃。

9. 对于“步骤 5：查看并创建”，请执行以下操作：

   a. 查看您在之前的步骤中所做的选择，并在必要时进行编辑。

   b. 从以下选项中选择一个。

如果您选择了...	则选择...
同步创建成员身份和协作（是，立即通过创建成员身份来加入）	创建协作和成员身份
创建协作，此时不创建成员身份（不，我将稍后创建成员身份）	创建协作

创建成员身份并加入协作

成员身份是成员在 AWS Clean Rooms中加入协作时创建的资源。

您可以以以下身份加入协作

- 可以查询的成员
- 可以运行查询和作业的成员
- 可以接收查询或任务结果的成员
- 为查询计算费用付费的会员
- 为查询和工作付费的会员

所有成员都可以贡献数据。

有关如何使用创建成员资格和加入协作的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

在此步骤中，受邀成员通过创建成员身份资源加入协作。

如果受邀成员是可以接收结果的成员，则他们会指定结果的目的地和格式。它们还提供服务角色 ARN，用于写入结果目标。

如果受邀成员是负责支付计算费用的成员，则他们在加入协作之前接受自己的付款责任。

创建成员身份并加入协作

1. 登录 AWS 管理控制台 并与您的成员一起打开AWS Clean Rooms 控制台 AWS 账户。
2. 在左侧导航窗格中，选择协作。
3. 在可加入选项卡上，对于可供加入的协作，选择协作的名称。
4. 在协作详细信息页面的概述部分中，查看协作详细信息，包括您的成员详细信息和其他成员列表。

   确认每 AWS 账户 IDs 位协作成员都是您打算与之签订协作关系的人。
5. 选择创建成员身份。
6. 在 “创建成员资格” 页面的 “概览” 中，查看协作名称、协作描述、 AWS 账户 协作创建者的 AWS 账户 ID、您的成员详细信息以及将为查询付费的成员的 ID。
7. 如果协作创建者选择启用分析日志，请选择以下选项之一作为 Amazon Logs 中的 CloudWatch 日志存储：

如果选择...	操作...
打开	与您相关的日志存储在 Amazon CloudWatch 日志中。 每个成员只能接收他们发起的查询或包含其数据的查询的日志。 能够接收结果的成员还会收到协作中运行的所有分析的日志，即使分析中未访问他们的数据也是如此。 在 “支持的日志类型” 下，从协作创建者选择支持的日志类型中进行选择：

如果选择...	操作...
	1. 如果要接收从 SQL 查询生成的日志，请选中 “来自查询的日志” 复选框。 2. 如果要使用接收作业生成的日志 PySpark，请选中 “来自作业的日志” 复选框。
关闭	与您相关的查询日志不会存储在您的 Amazon CloudWatch Logs 账户中。

> **Note**
>
> 开启分析日志后，可能需要几分钟才能设置日志存储并开始在 Amazon Logs 中接收 CloudWatch 日志。在这段短暂的时间内，可以查询的成员可能会运行实际上并未发送日志的查询。

8. 如果协作创建者为此协作启用了详细监控，请选择是否要在 CloudWatch 账户中接收详细的可观察性指标。

   要进行详细监控：

   详细的监控选项

选项	描述
打开	AWS Clean Rooms 将在您的账户中发布此 CloudWatch次合作的详细监控指标。您可以使用这些指标进行操作监控，包括查询性能和资源利用率。 将收取额外的 CloudWatch 费用。有关更多信息，请参阅CloudWatch 定价。
关闭	不会将任何详细指标导出到您的 CloudWatch 账户。如果其他成员启用了此选项，他们仍然可以在自己的账户中查看详细的监控指标。

> **Note**
>
> 详细的监控指标仅适用于可以运行查询的成员（分析运行者）和配置为协作付款人的成员。

9. 如果您的成员权限包括接收结果，则结果设置的默认设置为：

   a. 对于查询结果，选中 “设置查询的默认设置” 复选框，然后通过输入 S3 目标在 Amazon S3 中指定结果目标，或者选择 “浏览 S3” 从可用 S3 存储桶列表中进行选择。

      Example

      例如：**`s3://bucket/prefix`**

      i. 对于结果格式，请选择 CSV 或 PARQUE T。

      ii. （仅限 Spark）对于结果文件，请选择 “多个” 或 “单个”。

      iii. （可选）对于服务访问权限，如果您想将最长需要 24 小时的查询传送到 S3 目标，请选中 “添加服务角色以支持最长需要 24 小时才能完成的查询” 复选框。

      最长需要 24 小时才能完成的大型查询将传送到您的 S3 目标。

      如果您不选中该复选框，则只有在 12 小时内完成的查询才会发送到您的 S3 位置。

      > **Note**
      >
      > 您必须选择现有的服务角色或具备创建新服务角色的权限。有关更多信息，请参阅 创建服务角色来接收结果。

      iv. 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。

      Create and use a new service role

      - AWS Clean Rooms 使用此表所需的策略创建服务角色。
      - 默认服务角色名称为 `cleanrooms-result-receiver-<timestamp>`。
      - 您必须拥有创建角色并附加策略的权限。

Use an existing service role

1. 从下拉列表中选择一个现有服务角色名称。

   如果您有列出角色的权限，则会显示角色列表。

   如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。
2. 通过选择在 IAM 中查看外部链接来查看服务角色。

   如果没有现有的服务角色，则使用现有服务角色选项不可用。

   默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

> **Note**
>
> - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
> - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
> - 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

b. 对于任务结果，请选中 “为任务设置默认设置” 复选框，然后通过输入 S3 目标在 Amazon S3 中指定结果目标，或者选择 “浏览 S3” 从可用 S3 存储桶列表中进行选择。

   Example

   例如：**`s3://bucket/prefix`**

   - 通过从下拉列表中选择现有服务角色名称来指定服务访问权限。

10. 如果要为成员身份资源启用标签，请选择添加新标签，然后输入键和值对。
11. 如果协作创建者已将您指定为将支付查询费用或为查询和工作付费的成员，请选中 “我同意支付此协作中的计算费用” 复选框，表示您接受。

> **Note**
>
> 必须选中此复选框才能继续。
> 有关如何计算费用的更多信息，请参阅的定价 AWS Clean Rooms。

如果您是支付查询计算费用的会员，或者是支付查询和作业计算费用的会员，但不是可以查询的成员，则建议您使用 AWS Budgets 来配置预算， AWS Clean Rooms 并在达到最高预算后接收通知。有关设置预算的更多信息，请参阅《AWS Cost Management 用户指南》中的使用 AWS Budgets管理成本。有关设置通知的更多信息，请参阅《AWS Cost Management 用户指南》中的针对预算通知创建 Amazon SNS 主题。如果已达到最高预算，您可以联系可以运行查询和作业或退出协作的成员。如果您退出协作，将不再允许运行查询，因此将不再向您收取查询计算费用。

12. 如果您确定要创建成员身份并加入协作，请选择创建成员身份。

授予您对协作元数据的读取权限。除所有姓名和其他成员的姓名外，还包括合作的显示名称和 AWS 账户 IDs 描述等信息。

您现在已准备好执行以下操作：

- 准备要在 AWS Clean Rooms中查询的数据表 （如果您想查询自己的事件数据或想查询身份数据，则是可选的。）
- 将配置表与协作关联 - 如果您想查询事件数据。
- 为配置表添加分析规则 - 如果您想查询事件数据。
- 如果要@@ 创建 ID 映射表来查询身份数据，请创建并关联新的 ID 命名空间。

有关如何退出协作的信息，请参阅退出协作。

编辑协作

作为协作创建者，您可以编辑协作的不同部分。

有关如何使用 AWS 编辑协作的信息 SDKs，请参阅 AWS C lean Rooms API 参考。

主题

- 编辑协作名称和描述

- 更新协作分析引擎
- 关闭日志存储
- 编辑协作日志设置
- 编辑协作标签
- 编辑成员资身份标签
- 添加新成员
- 编辑现有成员能力
- 编辑协作自动批准设置
- 编辑关联表标签
- 编辑分析模板标签
- 编辑差别隐私策略标签

编辑协作名称和描述

创建协作后，您只能编辑协作名称和描述。

编辑协作名称和描述

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 在协作详细信息页面上，选择操作，然后选择编辑协作。
5. 在编辑协作页面的详细信息中，编辑协作的名称和描述。
6. 选择保存更改。

更新协作分析引擎

创建协作后，您可以将分析引擎从 AWS Clean Rooms SQL 更改为 Spark。

> **Note**
>
> 将分析引擎从 AWS Clean Rooms SQL 更改为 Spark 可能会破坏现有的工作流程。

更新协作分析引擎

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 在协作详细信息页面上，选择操作，然后选择编辑协作。
5. 选择保存更改。

关闭日志存储

如果您启用了分析日志，则可以编辑分析日志是否存储在您的 Amazon CloudWatch Logs 账户中。

关闭日志存储

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择已开启分析记录功能的协作。
4. 在协作详细信息页面上，选择操作，然后选择关闭日志存储。

> **Note**
>
> 将出现一条警告，指示以下内容：
>
> - 新的查询将不再记录到您的 CloudWatch帐户中。
> - 将根据您当前的保留设置保留现有日志。
> - 如果您将来重新激活日志记录，则仅适用于重新激活后进行的查询。
> - 此更改仅影响您的日志，其他团队成员的日志设置保持不变。

5. 选择关闭。

编辑协作日志设置

如果您启用了查询日志，则可以编辑查询日志是否存储在您的 Amazon CloudWatch Logs 账户中。

编辑协作日志设置

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 在协作详情页面上，执行以下操作之一：
   - 选择 “操作”，然后选择 “编辑日志设置”。
   - 在日志选项卡上，选择编辑日志设置。
5. 在 “编辑日志设置” 模式中，对于 Amazon 日志中的 CloudWatch 日志存储：
   - 如果您不希望将与您相关的日志存储在您的 Amazon CloudWatch L ogs 账户中，请选择 “关闭”。
   - 如果您确实希望将与您相关的日志存储在您的 Amazon CloudWatch Logs 账户中，请选择 “开启”。

     您只能接收您发起的查询或包含您的数据的查询的日志。

     可以接收结果的成员还会收到协作中运行的所有查询的日志，即使查询中未访问他们的数据也是如此。

     1. 在 “支持的日志类型” 下，从协作创建者选择支持的日志类型中进行选择：
        - 如果要接收从 SQL 查询生成的日志，请选中 “来自查询的日志” 复选框。
        - 如果要使用接收作业生成的日志 PySpark，请选中 “来自作业的日志” 复选框。
6. 选择保存更改。

> **Note**
>
> 开启日志记录功能后，可能需要几分钟才能设置日志存储并开始在 Amazon Logs 中接收 CloudWatch 日志。在这段短暂的时间内，可以查询的成员可能会运行实际上并未发送日志的查询。

编辑协作标签

作为协作创建者，在创建协作后，您可以管理协作资源上的标签。

编辑协作标签

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 选择下列选项之一：

如果您是...	操作...
协作创建者和协作成员	选择详细信息选项卡。
协作创建者但不是协作的成员	在页面中向下滚动到标签部分。

5. 有关协作详细信息，请选择管理标签。
6. 在管理标签页面上，可以执行以下操作：
   - 要删除标签，请选择移除。
   - 要添加标签，请选择添加新标签。
   - 要保存您的更改，请选择保存更改。

编辑成员资身份标签

作为协作创建者，在创建协作后，您可以管理成员身份资源上的标签。

编辑成员身份标签

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 选择详细信息选项卡。
5. 对于成员身份详细信息，选择管理标签。
6. 在管理成员身份标签页面上，可以执行以下操作：
   - 要删除标签，请选择移除。

- 要添加标签，请选择添加新标签。
- 要保存您的更改，请选择保存更改。

添加新成员

有关更多信息，请参阅 向协作中添加新成员。

编辑现有成员能力

有关更多信息，请参阅 更新现有成员的能力。

编辑协作自动批准设置

有关更多信息，请参阅 编辑协作自动批准设置。

编辑关联表标签

作为协作创建者，在将表与一个协作关联后，您可以管理关联的表资源上的标签。

编辑关联表标签

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 选择 Tables (表) 选项卡。
5. 对于由您关联的表，请选择一个表。
6. 在已配置表的详细信息页面上，对于标签，选择管理标签。

   在管理标签页面上，可以执行以下操作：

   - 要删除标签，请选择移除。
   - 要添加标签，请选择添加新标签。
   - 要保存您的更改，请选择保存更改。

编辑分析模板标签

作为协作创建者，在创建协作后，您可以管理分析模板资源上的标签。

编辑成员身份标签

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 选择 Templates 选项卡。
5. 在您创建的分析模板部分，选择分析模板。
6. 在分析模板表详细信息页面上，向下滚动到标签部分。
7. 选择管理标签。
8. 在管理标签页面上，可以执行以下操作：
   - 要删除标签，请选择移除。
   - 要添加标签，请选择添加新标签。
   - 要保存您的更改，请选择保存更改。

编辑差别隐私策略标签

作为协作创建者，在创建协作后，您可以管理分析模板资源上的标签。

编辑成员身份标签

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择包含您要编辑的差别隐私策略的协作。
4. 选择 Tables (表) 选项卡。
5. 在表选项卡上，选择管理标签。
6. 在管理标签页面上，可以执行以下操作：
   - 要删除标签，请选择移除。

- 要添加标签，请选择添加新标签。
- 要保存您的更改，请选择保存更改。

中的更改请求 AWS Clean Rooms

更改请求允许您对现有协作设置提出更改建议，以供其他协作成员批准。通过更改请求，您可以提交添加新成员、更新现有成员权限和修改协作自动批准设置的请求。所有协作成员都必须批准变更请求才能使提议的变更生效。

变更请求是针对协作的，可以由协作创建者提交。

您可以通过以下方式提交变更请求：

- 向协作中添加新成员
- 更新现有成员的能力
- 编辑协作自动批准设置

> **Note**
>
> 您必须是协作创建者才能提交变更请求。

向协作中添加新成员

要向协作添加新成员，您必须是协作创建者。向协作中添加新成员需要现有协作成员的手动批准，并将提交变更请求。

您可以按照以下步骤向协作中添加新成员：

向协作中添加新成员

1. 登录AWS 管理控制台并与您的成员一起打开AWS Clean Rooms控制台AWS 账户。
2. 在左侧导航窗格中，选择协作。
3. 选择您的协作以导航到您的协作详细信息。
4. 在协作页面上，导航到成员选项卡。
5. 在“成员”表格中，选择“编辑成员”。

6. 选择添加其他成员。
7. 输入新成员信息：
   - 成员显示名称
   - 会员AWS 账户账号
   - 指定成员是否可以接收结果。选中该复选框可授予该成员能力。
8. 选择保存更改。
9. 确认您的变更请求提交。在确认模式中，确认更改并选择提交变更请求。

> **Note**
>
> 如果您的协作支持自动批准的变更类型，则可能不需要手动批准变更请求。您可以在协作的“概述”部分查看哪些变更类型不需要变更请求。有关更多信息，请参阅 the section called “编辑协作自动批准设置”。

更新现有成员的能力

要更新现有的协作成员能力，您必须是协作创建者。更新现有协作成员的能力需要现有成员的手动批准，并将提交变更请求。

可以更新的成员技能有：

- 可以接收结果

您可以按照以下步骤更新现有协作成员的成员权限：

更新现有成员的能力

1. 登录AWS 管理控制台并与您的成员一起打开AWS Clean Rooms控制台AWS 账户。
2. 在左侧导航窗格中，选择协作。
3. 选择您的协作以导航到您的协作详细信息。
4. 在协作页面上，导航到成员选项卡。
5. 在“成员”表格中，选择“编辑成员”。
6. 指定成员的更改能力。

7. 选择保存更改。
8. 确认您的变更请求提交。在确认模式中，确认更改并选择提交变更请求。

> **Note**
>
> 如果您的协作支持自动批准的变更类型，则可能不需要手动批准变更请求。您可以在协作的“概述”部分查看哪些变更类型不需要变更请求。有关更多信息，请参阅 the section called “编辑协作自动批准设置”。

编辑协作自动批准设置

要编辑自动批准的协作设置，您必须是协作创建者并提交更改请求以供其他协作成员批准。

您可以按照以下步骤编辑协作的自动批准设置：

编辑协作自动批准设置

1. 登录AWS 管理控制台并与您的成员一起打开AWS Clean Rooms控制台AWS 账户。
2. 在左侧导航窗格中，选择协作。
3. 选择您的协作以导航到您的协作详细信息。
4. 在协作页面上，选择操作按钮，然后选择编辑自动批准。
5. 要在不手动批准变更请求的情况下向现有成员授予成员权限，请执行以下操作：

    1. 导航到“授予成员能力”部分。
    2. 指定可以自动授予哪些成员能力。

    > **Note**
    >
    > 默认情况下，所有协作成员都可以“贡献数据”。

    3. （可选）要允许新成员在不手动批准变更请求的情况下立即加入具有指定能力的协作，请选择自动批准具有这些能力的新成员。
6. 要允许在没有变更请求的情况下移除现有成员的技能，请执行以下操作：

    1. 导航到“可以自动撤消的技能”部分。
    2. 指定可以自动移除哪些成员技能。

7. 选择保存更改。
8. 确认您的变更请求提交。在确认模式中，确认更改并选择提交变更请求。

删除协作

作为协作创建者，您可以删除您创建的协作。

> **Note**
>
> 在删除协作时，您和所有成员无法运行查询，接收结果或贡献数据。每个协作成员根据其成员身份继续访问自己的数据。

删除协作

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择要删除的协作。
4. 在操作下，选择删除协作。
5. 确认删除，然后选择删除。

查看协作

作为协作创建者，您可以查看自己创建的所有协作。

查看协作

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面的最后使用下，查看最近使用的 5 个协作。
4. 在具有活跃成员身份选项卡上，查看拥有活跃成员身份的协作列表。

   您可以按名称、成员身份创建日期和您的成员详细信息进行排序。

   您可以使用搜索栏搜索协作。
5. 在可供加入选项卡上，查看可供加入的协作列表。

6. 在不再可用选项卡上，查看已删除的协作列表和不再可用的协作成员身份（已删除的成员身份）。

邀请成员参与协作

作为协作创建者，在创建协作后，您可以向“成员”选项卡上列出的成员发送邀请链接。

邀请成员加入协作

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 选择成员选项卡。
5. 在“成员”表中，选择“复制邀请链接”按钮。

   邀请链接已复制。
6. 将邀请链接粘贴到您选择的安全通信方式中，然后将其发送给每位协作成员。

监控成员

作为协作创建者，创建协作后，您可以在成员选项卡上监控所有成员的状态。这有助于确保在整个协作过程中进行适当的访问控制和成本管理。

监控成员的状态

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您创建的协作。
4. 选择成员选项卡。
5. 在成员表中，查看每个成员的状态。
6. 在成员能力表中，查看哪些成员可以查询、接收结果、贡献数据以及执行其他任务。
7. 在付款配置表中，查看哪些成员为查询、ID 映射表和机器学习建模付费。

向协作中添加成员

先决条件

- AWS 账户具有管理协作的权限
- 您要添加的成员 AWS 账户 IDs

向协作中添加成员

1. 登录AWS 管理控制台并打开AWS Clean Rooms控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择要向其添加成员的协作。
4. 选择成员选项卡。
5. 选择 Edit members (编辑成员)。
6. 选择添加其他成员并输入以下信息：
   - 成员显示名称
   - 会员AWS 账户账号
   - 指定他们是否可以接收结果
7. 选择保存更改。
8. 在确认模式中，验证信息是否正确，然后选择提交变更请求。
   - 如果变更请求需要其他成员的批准，则在添加新成员之前，所有现有成员都必须批准变更请求。有关变更请求的更多信息，请参阅the section called “更改请求”。
   - 如果具有指定能力的新成员支持自动批准更改类型，则此更改将立即生效。有关更多信息，请参阅 the section called “编辑协作自动批准设置”。
9. 在协作详情页面的 “成员” 选项卡下，确认已添加成员的成员状态显示为 “已邀请”。

完成这些步骤后，受邀成员可以加入协作。有关加入协作的更多信息，请参阅 创建成员身份并加入协作

从协作中移除成员

> **Note**
>
> 在开始之前，请注意，移除成员：
>
> - 从协作中移除所有关联的数据集
> - 如果成员支付查询计算费用，则此操作将停止协作中的所有查询执行。

如果删除成员，还会从协作中删除成员的所有关联数据集。

先决条件

- 你必须是协作创作者
- 你无法删除自己的账户

从协作中删除成员

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择要修改的协作。
4. 选择成员选项卡。
5. 选择要移除的成员旁边的选项按钮。
6. 选择移除 。
7. 在确认对话框中，键入**confirm**以验证删除。

Note

移除成员后，与其账户关联的所有数据集也将从协作中移除。

Important

如果您移除支付查询计算费用的成员，则在您指定新的付费成员之前，协作中无法运行进一步的查询。

退出协作

作为协作成员，您可以通过删除成员身份来退出协作。如果您是协作创建者，则只能通过删除协作来退出协作。

> **Note**
>
> 当您删除成员身份时，您将退出协作且无法重新加入。如果您是支付查询计算费用的成员并且删除了您的成员身份，则不允许运行其他查询。

退出协作

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台。
2. 在左侧导航窗格中，选择协作。
3. 对于具有活跃成员身份，请选择您所属的协作。
4. 选择操作。
5. 选择删除成员身份。
6. 在对话框中，通过在文本输入字段**confirm**中键入内容来确认退出协作的决定，然后选择清空并删除成员资格。

   您会在控制台上看到一条消息，指出成员身份已删除。

   协作创建者将看到成员状态为退出。

在中准备数据表 AWS Clean Rooms

> **Note**
>
> 准备数据表可以在您加入协作之前或之后进行。准备好表后，只要您针对该表的隐私需求相同，您就可以在多个协作中重复使用该表。

作为协作成员，您必须先准备好数据表，然后才能 AWS Clean Rooms 由可以查询的协作成员进行查询。

您在中用于查询的数据表 AWS Clean Rooms 通常与用于其他应用程序的数据表类型相同。例如，Amazon Athena、Amazon EMR、Amazon Redshift Spectrum 和 Amazon Quick 使用相同类型的数据集。

您可以直接从以下任何数据源中查询原始格式的数据：

- Amazon Simple Storage Service（Amazon S3）
- Amazon Athena
- Snowflake

AWS Clean Rooms 在查询运行时访问数据集，从而确保可以查询的成员始终访问最多的 up-to-date 数据。查询完成后，任何临时读入 AWS Clean Rooms 协作的数据都会被删除。查询结果将写入您的 Amazon S3 存储桶。

如果您的使用案例涉及查询身份数据，请参阅AWS Entity Resolution 数据匹配服务 in AWS Clean Rooms。

主题

的数据格式 AWS Clean Rooms

要分析数据，数据集必须采用 AWS Clean Rooms 支持的格式。

主题

PySpark 作业支持的数据格式

AWS Clean Rooms 支持以下结构化格式来运行 PySpark 作业。

- Parquet
- OpenCSV
- JSON

SQL 查询支持的数据格式

AWS Clean Rooms 支持以下用于运行 SQL 查询的结构化格式。

- Apache Iceberg 表
- Parquet
- OpenCSV
- JSON

> **Note**
>
> 文本文件中的 `timestamp` 值必须采用 `yyyy-MM-dd HH:mm:ss.SSSSSS` 格式。例如：`2017-05-01 11:30:59.000000`。

我们建议使用列式存储文件格式（例如 Apache Parquet）。使用列式存储文件格式，您可以通过仅选择所需的列来最大限度地减少数据移动。为了获得最佳性能，应将大型对象拆分为 100 MB - 1 GB 的对象。

支持的数据类型

AWS Clean Rooms 支持以下数据类型。

- ARRAY
- BIGINT
- BOOLEAN
- BYTE
- CHAR
- DATE
- DECIMAL
- FLOAT
- INTEGER
- INTERVAL
- LONG
- MAP
- REAL
- SHORT
- SMALLINT
- STRUCT
- TIME
- TIMESTAMP_LTZ
- TIMESTAMP_NTZ
- TINYINT
- VARCHAR

有关更多信息，请参阅 AWS Clean Rooms SQL 参考中的数据类型。

的文件压缩类型 AWS Clean Rooms

要减少存储空间、提高性能和最大程度地降低成本，我们强烈建议您压缩数据集。

AWS Clean Rooms 根据文件扩展名识别文件压缩类型，并支持下表所示的压缩类型和扩展名。

压缩算法	文件扩展名
GZIP	.gz
Bzip2	.bz2
Snappy	.snappy

可以在不同的级别应用压缩。最常见的情况是，压缩整个文件或压缩文件中的单个块。在文件级压缩列格式不会产生性能优势。

服务器端加密 AWS Clean Rooms

> **Note**
>
> 对于需要加密计算的使用案例，服务器端加密并不能取代加密计算。

AWS Clean Rooms 透明地解密使用以下加密选项加密的数据集：

- SSE-S3 — 使用由 Amazon S3 管理的 AES-256 加密密钥的服务器端加密
- SSE-KMS — 使用由管理的密钥进行服务器端加密 AWS Key Management Service

要使用 SSE-S3，用于将配置的表与协作关联的 AWS Clean Rooms 服务角色必须具有 KMS-Decrypt 权限。要使用 SSE-KMS，KMS 密钥策略还必须允许 AWS Clean Rooms 服务角色解密。

AWS Clean Rooms 不支持 Amazon S3 客户端加密。有关服务器端加密的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的使用服务器端加密保护数据。

Apache Iceberg 里面的桌子 AWS Clean Rooms

Apache Iceberg 是一种用于数据湖的开源表格式。 AWS Clean Rooms 可以使用存储在中的统计信息 Apache Iceberg 元数据用于优化查询计划并减少无尘室查询处理期间的文件扫描。有关更多信息，请参阅 Apache Iceberg 文档。

AWS Clean Rooms 与 Iceberg 表一起使用时，请考虑以下几点：

- 适用于 S3 的 Apache 冰山表 — Apache Iceberg 必须在 AWS Glue Data Catalog 基于开源胶水目录实现的中定义表。
- 适用于 Athena 的 Apache Iceberg 表 — 欲了解更多信息，请参阅-iceberg.html https://docs.aws.amazon.com/athena/ latest/ug/querying
- 适用于 Snowflake 的 Apache Iceberg 桌子 — 欲了解更多信息，请参阅用户指南/tables-iceberg https://docs.snowflake.com/en/
- Parquet 文件格式 — AWS Clean Rooms 仅支持 Parquet 数据文件格式的 Iceberg 表。
- GZIP 和 Snappy 压缩 — AWS Clean Rooms 支持 Parquet 和 GZIP Snappy 压缩。
- Iceberg 版本 — AWS Clean Rooms 支持对版本 1 和版本 2 的 Iceberg 表运行查询。
- 分区 — 您无需为自己的分区手动添加分区 Apache Iceberg 中的桌子 AWS Glue。 AWS Clean Rooms 检测到中的新分区 Apache Iceberg 自动生成表，无需手动操作即可更新表定义中的分区。Iceberg 分区在 AWS Clean Rooms 表架构中显示为常规列，而不是在配置表架构中单独显示为分区键。
- 限制
  - 仅限全新 Iceberg 表

    Apache Iceberg 表格转换自 Apache Parquet 不支持表。
  - 时间旅行查询

    AWS Clean Rooms 不支持时空旅行查询 Apache Iceberg 桌子。
  - Athena 引擎版本 2

    Iceberg 不支持使用 Athena 引擎版本 2 创建的表。
  - 文件格式

    Avro 且不支持优化的行列 (ORC) 文件格式。
  - 压缩

    Zstandard (Zstd) 压缩适用于 Parquet 不支持。

支持的 Iceberg 表数据类型

AWS Clean Rooms 可以查询 Iceberg 包含以下数据类型的表：

- BOOLEAN
- DATE
- DECIMAL
- DOUBLE
- FLOAT
- INT
- LIST
- LONG
- MAP
- STRING
- STRUCT
- TIMESTAMP WITHOUT TIME ZONE

有关 Iceberg 数据类型的更多信息，请参阅 Apache 文档中的 Schemas for Iceberg。

在中为查询准备数据表 AWS Clean Rooms

如果您的使用案例不需要您自带数据，则可以跳过此步骤。

如果您的使用案例涉及查询身份数据，请参阅AWS Entity Resolution 数据匹配服务 in AWS Clean Rooms。

有关您可以使用的数据格式的更多信息，请参阅的数据格式 AWS Clean Rooms。

主题

- 在 Amazon S3 中准备数据表
- 在 Amazon Athena 中准备数据表
- 在 Snowflake 中准备数据表

在 Amazon S3 中准备数据表

您可以分析已编目 AWS Glue 并存储在 Amazon S3 中的数据表。如果您的数据表已在中编目 AWS Glue，请跳至。在中创建配置表 AWS Clean Rooms

> **Important**
>
> 在 Amazon S3 中准备数据以供使用时 AWS Clean Rooms，必须确保您的 AWS Glue 表位置与存储数据文件的目录路径完全匹配。
> 例如：如果您的数据存储在 `s3://mybucket/folder/subfolder/data.parquet` 中，则您的 AWS Glue 表位置必须指向 `s3://mybucket/folder/subfolder/`。将表位置设置为父目录 (`s3://mybucket/folder/`) 将导致查询时表显示为空。

在 Amazon S3 中准备数据表涉及以下步骤：

主题

步骤 1：完成先决条件

要准备数据表以供使用 AWS Clean Rooms，必须满足以下先决条件：

- 您的数据表将另存为支持的数据格式之一 AWS Clean Rooms。
- 您的数据表已编入目录 AWS Glue ，并使用支持的数据类型。 AWS Clean Rooms
- 您的所有数据表都存储在亚马逊简单存储服务 (Amazon S3) 中， AWS 区域 与创建协作时相同。
- AWS Glue Data Catalog 必须与协作位于同一区域。
- 与 AWS Glue Data Catalog 成员资格 AWS 账户 相同。
- Amazon S3 存储桶未向注册 AWS Lake Formation。

步骤 2：（可选）准备用于加密计算的数据

（可选）如果您使用的是加密计算，并且您的数据表包含要加密的敏感信息，则必须使用 C3R 加密客户端对数据表进行加密。

要为加密计算准备数据，请按照使用 Clean Rooms 加密计算准备加密的数据表中的步骤操作。

步骤 3：将数据表上传到 Amazon S3

> **Note**
>
> 如果您打算在协作中使用加密的数据表，则必须先加密数据以进行加密计算，然后再将数据表上传到 Amazon S3。有关更多信息，请参阅 使用 Clean Rooms 加密计算准备加密的数据表。

将数据表上传到 Amazon S3

1. 登录 AWS 管理控制台 并打开 Amazon S3 控制台，网址为https://console.aws.amazon.com/s3/。
2. 选择桶，然后选择您想要用于存储数据表的桶。
3. 选择上传，然后按照提示进行操作。
4. 选择对象选项卡，查看存储数据的前缀。记下文件夹的名称。

   您可以选择用于查看数据的文件夹。

步骤 4：创建 AWS Glue 表

如果您已经有 AWS Glue 数据表，则可以跳过此步骤。

在此步骤中，您将在中设置一个爬虫 AWS Glue 来抓取 S3 存储桶中的所有文件并创建 AWS Glue 表。有关更多信息，请参阅《AWS Glue 用户指南》中的定义 AWS Glue中的爬网程序。

有关支持 AWS Glue Data Catalog 的数据类型的更多信息，请参阅支持的数据类型。

> **Note**
>
> AWS Clean Rooms 目前不支持向注册的 S3 存储桶。 AWS Lake Formation

以下过程描述了如何创建 AWS Glue 表。如果要使用带有 AWS Key Management Service (AWS KMS) 密钥的加密 AWS Glue Data Catalog 对象，则需要配置 KMS 密钥权限策略以允许访问该加密表。有关更多信息，请参阅《AWS Glue 开发人员指南》中的在 AWS Glue 中设置加密。

创建 AWS Glue 表

1. 按照《AWS Glue 用户指南》中的 “在AWS Glue 控制台上使用抓取工具” 步骤进行操作。
2. 记下 AWS Glue 数据库名称和 AWS Glue 表名。

步骤 5：后续步骤

现在，您已经在 Amazon S3 中准备好了数据表，您已准备好：

- 创建配置表。
- 创建 ML 模型

可以在以下之后查询这些表：

- 协作创建者已在 AWS Clean Rooms中建立了协作。有关更多信息，请参阅 创建协作。
- 协作创建者已将协作 ID 发送给作为协作参与者的您。

在 Amazon Athena 中准备数据表

您可以在 Amazon Athena 中查询已作为 AWS Glue Data Catalog (GDC) 视图创建的数据表。

GDC 视图是一个虚拟表，由一个或多个基础 AWS Glue 表创建。它必须使用 Athena 目录中的 Athena SQL 创建。`AwsGlueCatalog`

在 Amazon Athena 中准备数据表涉及以下步骤：

主题

- 步骤 1：完成先决条件
- 步骤 2：（可选）准备用于加密计算的数据
- 步骤 3：后续步骤

步骤 1：完成先决条件

要准备数据表以供使用 AWS Clean Rooms，必须满足以下先决条件：

- 您的数据表将另存为支持的数据格式之一 AWS Clean Rooms。
- 您的数据表使用支持的数据类型 AWS Clean Rooms。
- 你已经使用 Athena 目录中的 Athena SQL 在你的 AWS Glue 表上创建了 GDC 视图。`AwsDataCatalog`

  该视图将显示在：

  - Athena 控制台（下方）作为`AwsDataCatalog`视图：https://console.aws.amazon.com/athena/
  - 将 AWS Glue 控制台当作 AWS Glue 桌子：https://console.aws.amazon.com/glue/

  有关更多信息，请参阅亚马逊 Athena 用户指南中的在 Athena 中使用数据目录视图。

  > **Note**
  >
  > 您需要适当的权限才能在 Athena 和. 中创建视图。 AWS Glue此外，请确保您可以访问视图定义中引用的基础表。
  > AWS Clean Rooms 仅支持 Athena 的 AWS Glue 目录类型，不支持 Lambda 或 Hive 目录类型。

- 您的数据表或 GDC 视图已编入 AWS Glue 并注册到中。 AWS Lake Formation
- 您已在 Amazon S3 中创建了一个单独的输出存储桶来接收 Athena 结果。
- 您已设置服务角色来读取 Amazon Athena 中的数据。有关更多信息，请参阅 创建服务角色以读取来自亚马逊 Athena 的数据。
  - 该服务角色具有 Lake Formation Select 和 Describe 对 GDC 视图或表格的访问权限。

步骤 2：（可选）准备用于加密计算的数据

（可选）如果您使用的是加密计算，并且您的数据表包含要加密的敏感信息，则必须使用 C3R 加密客户端对数据表进行加密。

要为加密计算准备数据，请按照使用 Clean Rooms 加密计算准备加密的数据表中的步骤操作。

步骤 3：后续步骤

现在，您已经在 Amazon Athena 中准备好了数据表，您已准备好：

- 创建配置表。
- 创建 ML 模型

可以在以下之后查询这些表：

- 协作创建者已在 AWS Clean Rooms中建立了协作。有关更多信息，请参阅 创建协作。
- 协作创建者已将协作 ID 发送给作为协作参与者的您。

在 Snowflake 中准备数据表

您可以查询存储在 Snowflake 数据仓库中的数据表。

在 Snowflake 中准备数据表涉及以下步骤：

主题

- 步骤 1：完成先决条件
- 步骤 2：（可选）准备用于加密计算的数据
- 步骤 3：创建 AWS Secrets Manager 密钥
- 步骤 4：后续步骤

步骤 1：完成先决条件

要准备数据表以供使用 AWS Clean Rooms，必须满足以下先决条件：

- 您被授予 AWS 账户 了读取数据表的适当权限。有关更多信息，请参阅 创建服务角色以从 Snowflake 读取数据。
- 您的数据表将另存为支持的数据格式之一 AWS Clean Rooms。
- 您的数据表使用支持的数据类型 AWS Clean Rooms。
- 您的数据表存储在 Snowflake 仓库中。有关更多信息，请参阅 Snowflake 文档。
- 您已经设置了一个新的 Snowflake 用户，该用户对要与协作关联的 Snowflake 表具有只读权限。

步骤 2：（可选）准备用于加密计算的数据

（可选）如果您使用的是加密计算，并且您的数据表包含要加密的敏感信息，则必须使用 C3R 加密客户端对数据表进行加密。

要为加密计算准备数据，请按照使用 Clean Rooms 加密计算准备加密的数据表中的步骤操作。

步骤 3：创建 AWS Secrets Manager 密钥

要从中连接到 Snowflake AWS Clean Rooms，你需要创建你的 Snowflake 凭据并将其存储在密钥中，然后将该 AWS Secrets Manager 密钥与中的 Snowflake 表关联起来。 AWS Clean Rooms

> **Note**
>
> 我们建议您创建一个专用于的新用户 AWS Clean Rooms。该用户只能拥有对您要访问的数据具有读取权限 AWS Clean Rooms 的角色。

创建密 AWS Secrets Manager 钥

1. 在 Snowflake 中，生成一个用户`snowflakeUser`并设置密钥对身份验证。

   > **Note**
   >
   > 2025 年 11 月，Snowflake 将过渡到仅支持密钥对身份验证。此更改将影响当前与 Snowflake 的 AWS Clean Rooms 集成，Snowflake 使用用户名和密码身份验证。在此日期之后，Snowflake 连接 AWS Clean Rooms 将需要使用 Snowflake 隐私增强邮件 (PEM) 私钥进行密钥对身份验证。

2. 确定该用户将与哪个 Snowflake 仓库进行互动。`snowflakeWarehouse`要么在 Snowflake `snowflakeUser` 中将其设置为 for，要么记住它以备下一步使用。`DEFAULT_WAREHOUSE`

3. 在 AWS Secrets Manager 中，使用您的 Snowflake 凭证创建密钥。要在 Secrets Manager 中创建密钥，请按照AWS Secrets Manager 用户指南中创建 AWS Secrets Manager 密钥中的教程进行操作。创建密钥后，保留密钥名称`secretName`以供下一步使用。

   - 选择键/值对时，请使用密钥为`snowflakeUser`创建一个对。`sfUser`
   - 选择密钥/值对时，请为您的 Snowflake PEM 私钥与密钥创建一对。`pem_private_key`
   - 选择键/值对时，请使用密钥为`snowflakeWarehouse`创建一个对。`sfWarehouse`

     如果在 Snowflake 中设置了默认值，则不需要这样做。
   - 选择键/值对时，请使用密钥为`snowflakeRole`创建一个对。`sfRole`

步骤 4：后续步骤

现在，您已经在 Snowflake 中准备好了数据表，您已准备好：

- 创建配置表。
- 创建 ML 模型

可以在以下之后查询这些表：

- 协作创建者已在 AWS Clean Rooms中建立了协作。有关更多信息，请参阅 创建协作。
- 协作创建者已将协作 ID 发送给作为协作参与者的您。

使用 Clean Rooms 加密计算准备加密的数据表

Clean Rooms(C3R) 的加密计算是一种功能。 AWS Clean Rooms您可以使用 C3R 以加密方式限制任何一方和协作 AWS 中可以学到的内容。 AWS Clean Rooms

在将数据表上传到您的数据源之前，您可以使用 C3R 加密客户端（一种客户端加密工具）对数据表进行加密：亚马逊简单存储服务 (Amazon S3)、Amazon Athena 或 Snowflake。

有关更多信息，请参阅 加密计算 Clean Rooms。

使用 C3R 准备加密的数据表涉及以下步骤：

Steps

- 步骤 1：完成先决条件
- 步骤 2：下载 C3R 加密客户端
- 步骤 3：（可选）查看 C3R 加密客户端中的可用命令。
- 步骤 4：为表格文件生成加密架构
- 步骤 5：创建共享密钥
- 步骤 6：将共享密钥存储在 环境变量中。
- 步骤 7：加密数据
- 步骤 8：验证数据加密
- （可选）创建架构（高级用户）

步骤 1：完成先决条件

要准备数据表以供 C3R 使用，您必须满足以下先决条件：

- 您可以通过以下网址访问Clean Rooms存储库的加密计算： GitHub

  https://github.com/aws/c3r
- 您已经设置了使用 C3R 加密客户端的 AWS 凭据。C3R 加密客户端使用这些凭据进行只读 API 调用， AWS Clean Rooms 以检索协作元数据。有关更多信息，请参阅《AWS Command Line Interface 用户指南版本 2》中的配置 AWS CLI。
- 您的计算机上安装了 Java Runtime Environment (JRE) 11 或更高版本。
  - 推荐Java Runtime Environment的 Amazon Corretto 11 或更高版本可以从 /corretto 下载。https://aws.amazon.com
  - Java Development Kit(JDK) 包括同一个版本的对应 JRE。但是，运行 Clean Rooms 加密计算 (C3R) 加密客户端不需要 JDK 的附加功能。
- 您的表格数据文件 (.csv) 或 Parquet 文件 (.parquet) 保存在本地。
- 您或协作中的其他成员可以创建共享密钥。有关更多信息，请参阅 步骤 5：创建共享密钥。
- 协作创建者创建了一种协作， AWS Clean Rooms 其中启用了加密计算，以实现协作。有关更多信息，请参阅 创建协作。
- 协作创建者已将协作 ID 发送给作为协作参与者的您。发送的邀请中包含协作的 Amazon 资源名称 (ARN)，其中包含协作 ID。

步骤 2：下载 C3R 加密客户端

要从以下网址下载 C3R 加密客户端 GitHub

1. 前往密码计算获取Clean RoomsAWSGitHub存储库：c3r https://github.com/aws/
2. 选择并下载文件。

   源代码、许可证和相关资料可以从 GitHub 存储库的登陆页面克隆或下载为 .zip 文件。（参见存储库内容列表右上角的代码按钮）。

   最新签名的 C3R 加密客户端 Java Executable File（即命令行界面应用程序）位于 GitHub 存储库的发布页面上。

Apache Spark 的 C3R 加密客户端包 (`c3r-cli-spark`) 是 c3r-cli 的一个版本，必须作为作业提交给正在运行的 Apache Spark 服务器。有关更多信息，请参阅在 Apache Spark 上运行 C3R。

步骤 3：（可选）查看 C3R 加密客户端中的可用命令。

使用此过程熟悉 C3R 加密客户端中的可用命令。

查看 C3R 加密客户端中的所有可用命令。

1. 从命令行界面 (CLI)，导航到包含已下载 c3r-cli.jar 文件的文件夹。
2. 运行以下命令：`java -jar c3r-cli.jar`
3. 查看可用命令和选项的列表。

步骤 4：为表格文件生成加密架构

要加密数据，需要一个描述如何使用数据的加密架构。本节介绍 C3R 加密客户端如何帮助为带有标题行的 CSV 文件或 Parquet 文件生成加密架构。

每个文件只需进行一次该操作。架构存在后，可以重复使用它来加密同一个文件（或任何具有相同列名的文件）。如果列名或所需的加密架构发生变化，则必须更新架构文件。有关更多信息，请参阅（可选）创建架构（高级用户）。

> **Important**
>
> 所有协作方都必须使用相同的共享密钥。如果要在查询中对列名进行 JOIN 或以其他方式比较列名是否相等，各协作方还应协调列名，使其相匹配。否则，SQL 查询可能会产生意外或不正确的结果。但是，如果协作创建者在创建协作期间启用了 `allowJoinsOnColumnsWithDifferentNames` 加密设置，则无需这样做。有关加密相关设置的更多信息，请参阅加密计算参数。

在架构模式下运行时，C3R 加密客户端会逐列浏览输入文件，提示您是否以及如何处理该列。如果文件中包含许多加密输出不需要的列，交互式模式生成可能会变得繁琐，因为您必须跳过每个不需要的列。为避免这种情况，您可以手动编写架构，或者创建仅包含所需列的输入文件的简化版本。然后，可以在该简化的文件上运行交互式架构生成器。C3R 加密客户端会输出有关架构文件的信息，并询问您应如何在目标输出中包含或加密源列（如果有）。

对于输入文件中的每个源列，系统会提示：

1. 应该生成多少个目标列
2. 应如何加密每个目标列（如果有）
3. 每个目标列的名称
4. 如果将列作为 sealed 列进行加密，在加密之前应如何填充数据。

> **Note**
>
> 对已加密为 sealed 列的列的数据进行加密时，必须确定哪些数据需要填充。C3R 加密客户端建议在架构生成期间使用默认填充，将列中的所有条目填充到相同的长度。
> 在确定 `fixed` 的长度时，请注意填充以字节为单位，而不是以位为单位。

以下是创建架构的决定表。

架构决定表

决策	源列中的目标列数 <' name-of-column '> ?	目标列类型：[c] cleartext 、[f] fingerprint 或 [s] sealed ?	目标列标题名称 <default 'name-of-column'>	在标题中添加 <suffix> 后缀以指示它是如何加密的，[y] 是或 [n] 否 <default 'yes'>	<' name-of-column _sealed'> 填充类型：[n] 一、[f] 固定或 [m] 最大 <default 'max'>
保持列未加密。	1	c	不适用	不适用	不适用
将列加密为 fingerprint 列。	1	f	选择默认值或输入新的标题名称。	输入 y 选择默认值 (`_fingerprint`) 或输入 n。	不适用

决策	源列中的目标列数 <' name-of-column '> ?	目标列类型：[c] cleartext 、[f] fingerprint 或 [s] sealed ?	目标列标题名称 <default 'name-of-column'>	在标题中添加 <suffix> 后缀以指示它是如何加密的，[y] 是或 [n] 否 <default 'yes'>	<' name-of-column _sealed'> 填充类型：[n] 一、[f] 固定或 [m] 最大 <default 'max'>
将列加密为 sealed 列。	1	s	选择默认值或输入新的标题名称。	输入 y 选择默认值 (_sealed) 或输入 n。	选择填充类型。 有关更多信息，请参阅（可选）创建架构（高级用户）。
将列同时加密为 fingerprint 和 sealed。	2	输入第一个目标列：f。 输入第二个目标列：s。	为每个目标列选择目标标题。	输入 y 选择默认值或输入 n.	选择填充类型（仅适用于 sealed 列）。 有关更多信息，请参阅（可选）创建架构（高级用户）。

以下是如何创建加密架构的两个示例。交互的具体内容取决于输入文件和您提供的响应。

示例

- 示例：为 fingerprint 列和 cleartext 列生成加密架构
- 示例：生成带有 sealed、fingerprint 和 cleartext 列的加密架构

示例：为 fingerprint 列和 cleartext 列生成加密架构

在此示例中，对于 ads.csv，只有两列：username 和 ad_variant。对于这些列，我们需要以下内容：

- 将 username 列加密为 fingerprint 列
- 将 ad_variant 列加密为 cleartext 列

为 fingerprint 列和 cleartext 列生成加密架构

1. （可选）要确保 c3r-cli.jar 文件和要加密的文件存在，请执行以下操作：

   a. 导航到所需的目录并运行 ls（如果使用 Mac 或 Unix/Linux）或 dir（如果使用 Windows）。

   b. 查看表格数据文件（例如 .csv）列表，然后选择要加密的文件。

   在此示例中，ads.csv 是我们要加密的文件。

2. 在 CLI 中，运行以下命令以交互方式创建架构。

   ```
   java -jar c3r-cli.jar schema ads.csv --interactive --output=ads.json
   ```

   > **Note**
   >
   > - 您可以运行 java --jar PATH/TO/c3r-cli.jar。或者，如果您已将 PATH/TO/c3r-cli.jar 添加到 CLASSPATH 环境变量中，也可以运行该类名。C3R 加密客户端将在 CLASSPATH 中查找以找到它（例如 java com.amazon.psion.cli.Main）。
   > - --interactive 标志选择开发架构的交互模式。这将引导用户完成创建架构的向导。具有高级技能的用户无需使用向导即可创建自己的架构 JSON。有关更多信息，请参阅（可选）创建架构（高级用户）。
   > - --output 标志设置输出名称。如果不包含 --output 标志，则 C3R 加密客户端会尝试选择默认输出名称（例如 <input>.out.csv 或架构的 <input>.json）。

3. 对于 Number of target columns from source column ‘username’?，输入 **1**，然后按 Enter。

4. 对于 Target column type: [c]leartext, [f]ingerprint, or [s]ealed?，输入 **f**，然后按 Enter。

5. 对于 `Target column headername <default 'username'>`，按 Enter。

   默认名称为“`username`”。

6. 对于 `Add suffix '_fingerprint' to header to indicate how it was encrypted, [y]es or [n]o <default 'yes'>`，输入 **y**，然后按 Enter。

   > **Note**
   >
   > 交互模式建议在加密的列标题中添加的后缀（fingerprint 列添加 `_fingerprint`，sealed 列添加 `_sealed`）。当您执行诸如将数据上传到 AWS 服务 或创建 AWS Clean Rooms 协作之类的任务时，后缀可能会有所帮助。这些后缀可以帮助指示对每列中的加密数据可以做些什么。例如，如果您将列加密为 sealed 列 (`_sealed`) 并尝试对其进行 JOIN 或尝试反向操作，则会出现问题。

7. 对于 `Number of target columns from source column ‘ad_variant’?`，输入 **1**，然后按 Enter。

8. 对于 `Target column type: [c]leartext, [f]ingerprint, or [s]ealed?`，输入 **c**，然后按 Enter。

9. 对于 `Target column headername <default 'username'>`，按 Enter。

   默认名称为“`ad_variant`”。

   架构被写入名为 `ads.json` 的新文件。

   > **Note**
   >
   > 您可以通过在任何文本编辑器（例如 Windows 上的 Notepad 或 macOS 上的 TextEdit）中打开架构来查看架构。

10. 现在，您可以加密数据了。

示例：生成带有 sealed、fingerprint 和 cleartext 列的加密架构

在此示例中，对于 `sales.csv`，有三列：`username`、`purchased` 和 `product`。对于这些列，我们需要以下内容：

- 将 `product` 列加密为 `sealed` 列
- 将 `username` 列加密为 `fingerprint` 列

- 将 purchased 列加密为 cleartext 列

生成带有 sealed、fingerprint 和 cleartext 列的加密架构

1. （可选）要确保 c3r-cli.jar 文件和要加密的文件存在，请执行以下操作：

   a. 导航到所需的目录并运行 ls（如果使用 Mac 或 Unix/Linux）或 dir（如果使用 Windows）。

   b. 查看表格数据文件 (.csv) 列表并选择要加密的文件。

      在此示例中，sales.csv 是我们要加密的文件。

2. 在 CLI 中，运行以下命令以交互方式创建架构。

   ```
   java -jar c3r-cli.jar schema sales.csv --interactive --
   output=sales.json
   ```

   > **Note**
   >
   > - --interactive 标志选择开发架构的交互模式。这将引导用户完成创建架构的指导性工作流程。
   > - 如果您是高级用户，则无需使用指导性工作流程即可创建自己的架构 JSON。有关更多信息，请参阅 （可选）创建架构（高级用户）。
   > - 对于没有列标题的 .csv 文件，请参阅 CLI 中可用的架构命令的 --noHeaders 标志。
   > - --output 标志设置输出名称。如果不包含 --output 标志，则 C3R 加密客户端会尝试选择默认输出名称（例如 <input>.out 或架构的 <input>.json）。

3. 对于 Number of target columns from source column 'username'?，输入 **1**，然后按 Enter。

4. 对于 Target column type: [c]leartext, [f]ingerprint, or [s]ealed?，输入 **f**，然后按 Enter。

5. 对于 Target column headername <default 'username'>，按 Enter。

   默认名称为“username”。

6. 对于 Add suffix '_fingerprint' to header to indicate how it was encrypted, [y]es or [n]o <default 'yes'>，输入 **y**，然后按 Enter。

7. 对于 Number of target columns from source column 'purchased'?，输入 **1**，然后按 Enter。
8. 对于 Target column type: [c]leartext, [f]ingerprint, or [s]ealed?，输入 **c**，然后按 Enter。
9. 对于 Target column headername <default 'purchased'>，按 Enter。

   默认名称为“purchased”。
10. 对于 Number of target columns from source column 'product'?，输入 **1**，然后按 Enter。
11. 对于 Target column type: [c]leartext, [f]ingerprint, or [s]ealed?，输入 **s**，然后按 Enter。
12. 对于 Target column headername <default 'product'>，按 Enter。

    默认名称为“product”。
13. 对于 'product_sealed' padding type: [n]one, [f]ixed, or [m]ax <default 'max'?>，按 Enter 选择默认值。
14. 对于 Byte-length beyond max length to pad cleartext to in 'product_sealed' <default '0'>?，按 Enter 选择默认值。

    架构被写入名为 sales.json 的新文件。
15. 现在，您可以加密数据了。

步骤 5：创建共享密钥

要加密数据表，协作参与者必须同意并安全地共享共享密钥。

共享密钥必须至少为 256 位（32 字节）。您可以指定更大的密钥，但它不会为您提供任何额外的安全性。

> **Important**
>
> 请记住，所有协作参与者用于加密和解密的密钥和协作 ID 必须相同。

以下各节提供了控制台命令的示例，这些命令用于生成作为 secret.key 保存在相应终端当前工作目录中的共享密钥。

主题

示例：使用 OpenSSL 生成密钥

对于常见的通用密码库，请运行以下命令创建共享密钥。

```
openssl rand 32 > secret.key
```

如果您使用 Windows 但尚未安装 OpenSSL，则可以使用示例：使用 PowerShell 时在 Windows 上生成密钥中描述的示例生成密钥。

示例：使用 PowerShell 在 Windows 上生成密钥

对于 Windows 上可用的终端应用程序 PowerShell，请运行以下命令来创建共享密钥。

```
$bs = New-Object Byte[](32);
[Security.Cryptography.RandomNumberGenerator]::Create().GetBytes($bs); Set-
Content 'secret.key' -Encoding Byte -Value $bs
```

步骤 6：将共享密钥存储在 环境变量中。

环境变量是一种方便且可扩展的方式，用户可以从各种密钥存储库中提供密钥，例如 AWS Secrets Manager 并将其传递给 C3R 加密客户端。

AWS 服务 如果您使用将这些密钥存储在相关的环境变量中，则 C3R 加密客户端可以使用中存储的密钥。 AWS CLI 例如，C3R 加密客户端可以使用中的密钥。 AWS Secrets Manager有关更多信息，请参阅《AWS Secrets Manager 用户指南》中的使用 AWS Secrets Manager创建和管理密钥。

> **Note**
>
> 但是，在使用 AWS 服务 诸如 AWS Secrets Manager 来保存 C3R 密钥之前，请验证您的用例是否允许。某些用例可能需要隐瞒 AWS密钥。这是为了确保加密的数据和密钥永远不会由同一个第三方持有。

共享密钥的唯一要求是，共享密钥必须经过 base64 编码并存储在环境变量 C3R_SHARED_SECRET 中。

以下各节介绍用于将 secret.key 文件转换为 base64 并将其存储为环境变量的控制台命令。secret.key 文件可能由 步骤 5：创建共享密钥 中列出的任何命令生成，并且只是一个示例源。

在 Windows 上使用 PowerShell 将密钥存储到环境变量中

要在 Windows 上使用 PowerShell 转换为 base64 并设置环境变量，请运行以下命令。

```
$Bytes=[IO.File]::ReadAllBytes((Get-Location).ToString()+'\secret.key');
 $env:C3R_SHARED_SECRET=[Convert]::ToBase64String($Bytes)
```

在 Linux 或 macOS 上将密钥存储到环境变量中

要在 Linux 或 macOS 上转换为 base64 并设置环境变量，请运行以下命令。

```
export C3R_SHARED_SECRET="$(cat secret.key | base64)"
```

步骤 7：加密数据

要执行此步骤，您必须获取协 AWS Clean Rooms 作 ID 和共享密钥。有关更多信息，请参阅先决条件。

在以下示例中，我们使用我们创建的名为 ads.json 的架构在 ads.csv 上运行加密。

加密数据

1. 将协作的共享密钥存储在 步骤 6：将共享密钥存储在 环境变量中。 中。
2. 在命令行中，输入以下命令。

   ```
   java -jar c3r-cli.jar encrypt <name of input .csv file> --schema=<name of schema .json file> --id=<collaboration id> --output=<name of output.csv file> <optional flags>
   ```
3. 对于*<name of input .csv file>*，输入输入.csv 文件的名称。
4. 对于 schema=，输入 .json 加密架构文件的名称。
5. 对于 id=，输入协作 ID。
6. 对于 output=，输入输出文件的名称（例如，ads-output.csv）。
7. 包括 加密计算参数 和 Clean Rooms 加密计算中的可选标志 中描述的任何命令行标志。

8. 运行命令。

在 ads.csv 的示例中，我们运行以下命令。

java -jar c3r-cli.jar encrypt ***ads.csv*** --schema=***ads.json*** --id=***123e4567-e89b-42d3-a456-556642440000*** --output=***ads-output.csv***

在 sales.csv 的示例中，我们运行以下命令。

java -jar c3r-cli.jar encrypt ***sales.csv*** --schema=***sales.json*** --id=***123e4567-e89b-42d3-a456-556642440000***

> **Note**
>
> 在此示例中，我们没有指定输出文件名 (--output=*sales-output.csv*)。结果，生成了默认的输出文件名 name-of-file.out.csv。

现在，您可以验证加密的数据了。

步骤 8：验证数据加密

验证数据是否已加密

1. 查看加密的数据文件（例如 sales-output.csv）。
2. 验证以下列：
   a. 列 1 — 已加密（例如 username_fingerprint）。

      对于 fingerprint 列 (HMAC)，在版本和类型前缀（例如 01:hmac:）之后，有 44 个字符的 base64 编码数据。
   b. 列 2 — 未加密（例如 purchased）。
   c. 列 3 — 已加密（例如 product_sealed）。

      对于已加密 (SELECT) 列，cleartext 的长度加上版本和类型前缀（例如 01:enc:）后的任何填充，与加密后的 cleartext 的长度成正比。也就是说，长度等于输入的大小加上大约 33% 的编码开销。

您现在已准备好执行以下操作：

1. 将加密的数据上传到 S3。
2. 创建 AWS Glue 表。
3. 在 AWS Clean Rooms中创建配置表。

C3R 加密客户端将创建不包含未加密数据的临时文件（除非这些数据在最终输出中也未加密）。但是，某些加密值可能无法正确填充。即使协作设置 `allowRepeatedFingerprintValue` 为 `false`，指纹列也可能包含重复的值。之所以出现此问题，是因为临时文件是在检查正确的填充长度和重复项删除属性之前写入的。

如果 C3R 加密客户端失败或在加密过程中中断，则它可能会在写入临时文件之后但在检查这些属性和删除临时文件之前停止。因此，这些临时文件可能仍在磁盘上。在这种情况下，这些文件中的内容对明文数据的保护程度将不如输出文件。特别是，这些临时文件可能会向统计分析揭示明文数据，而这些数据不会对最终输出产生影响。用户应删除这些文件（尤其是 SQLite 数据库），以防止这些文件落入未经授权的人手中。

（可选）创建架构（高级用户）

手动创建架构适用于高级用户。

以下是对带或不带列标题的输入文件的 JSON 架构文件格式的描述。如果需要，高级用户可以直接编写或修改架构。

> **Note**
>
> C3R 加密客户端可通过 示例：生成带有 sealed、fingerprint 和 cleartext 列的加密架构 中描述的交互式流程或通过创建存根模板协助您创建架构。

映射和定位表架构

以下部分描述了两种表架构：

- 映射表架构 — 此架构用于加密带有标题行的 .csv 文件和 Apache Parquet 文件。
- 位置表架构 — 此架构用于加密没有标题行的 .csv 文件。

C3R 加密客户端可以加密表格文件以进行协作。为此，它必须具有相应的架构文件，该文件指定应如何从输入中导出加密输出。

C3R 加密客户端可以通过在命令行运行 C3R 加密客户端架构命令来帮助为 INPUT 文件生成架构。命令的一个示例是 `java -jar c3r-cli.jar schema --interactive INPUT`。

架构指定以下信息：

1. 哪些源列通过标题名称（映射架构）或位置（位置架构）映射到输出文件中哪些已转换的列
2. 哪些目标列要保留 cleartext
3. 要对哪些目标列进行加密以进行 SELECT 查询
4. 要对哪些目标列进行加密以进行 JOIN 查询

这些信息在特定于表的 JSON 架构文件中编码，该文件由一个对象组成，其 `headerRow` 字段是一个布尔值。对于有标题行的 Parquet 文件和 .csv 文件，该值必须为 `true`，否则为 `false`。

映射表架构

映射架构具有以下形状。

```
{
  "headerRow": true,
  "columns": [
    {
      "sourceHeader": STRING,
      "targetHeader": STRING,
      "type": TYPE,
      "pad": PAD
    },
    ...
  ]
}
```

如果 `headerRow` 为 `true`，对象中的下一个字段就是 `columns`，其中包含一个将源标题映射到目标标题的列架构（即描述输出列应包含内容的 JSON 对象）数组。

- `sourceHeader` — 数据来源于的源列的 `STRING` 标题名称。

> **Note**
>
> 同一个源列可以用于多个目标列。
>
> 输入文件中未列为架构中任意位置的 `sourceHeader` 列不会出现在输出文件中。

- `targetHeader` — 输出文件中相应列的 STRING 标题名称。

> **Note**
>
> 对于映射架构，此字段为可选项。如果省略此字段，输出中的标题名称将重复使用 `sourceHeader`。如果输出列分别为 fingerprint 列或 sealed 列，则附加 `_fingerprint` 或 `_sealed`。

- `type` — 输出文件中目标列的 TYPE。即 `cleartext`、`sealed` 或 `fingerprint` 其中之一，具体取决于该列在协作中的使用方式。
- `pad` — 列架构对象的字段，仅当 TYPE 为 `sealed` 才存在。其对应的 PAD 值是一个对象，用于描述数据在加密前应如何填充。

```
{
  "type": PAD_TYPE,
  "length": INT
}
```

要指定加密前的填充，请按如下方式使用 `type` 和 `length`：

- PAD_TYPE 为 `none` — 不对列的数据进行填充，`length` 字段不适用（即省略）。
- PAD_TYPE 为 `fixed` — 将列的数据填充到指定的字节 `length`。
- PAD_TYPE 为 `max` — 列的数据填充到最长值的字节长度加上额外的 `length` 字节。

以下是映射架构的示例，每个类型都有一列。

```
{
  "headerRow": true,
  "columns": [
    {
      "sourceHeader": "FullName",
      "targetHeader": "name",
      "type": "cleartext"
    },
    {
      "sourceHeader": "City",
      "targetHeader": "city_sealed",
      "type": "sealed",
      "pad": {
        "type": "max",
```

```
        "length": 16
      }
    },
    {
      "sourceHeader": "PhoneNumber",
      "targetHeader": "phone_number_fingerprint",
      "type": "fingerprint"
    },
    {
      "sourceHeader": "PhoneNumber",
      "targetHeader": "phone_number_sealed",
      "type": "sealed",
      "pad": {
        "type": "fixed",
        "length": 20
      }
    }
  ]
}
```

作为一个更复杂的示例，以下是带有标题的 .csv 文件示例。

```
FirstName,LastName,Address,City,State,PhoneNumber,Title,Level,Notes
Jorge,Souza,12345 Mills Rd,Anytown,SC,703-555-1234,CEO,10,
Paulo,Santos,0 Street,Anytown,MD,404-555-111,CIO,9,This is a really long note that
 could really be a paragraph
Mateo,Jackson,1 Two St,Anytown,NY,304-555-1324,COO,9,""
Terry,Whitlock4 N St,Anytown,VA,407-555-8888,EA,7,Secret notes
Diego,Ramirez,9 Hollows Rd,Anytown,VA,407-555-1222,SDE I,4,null
John,Doe,8 Hollows Rd,Anytown,VA,407-555-4321,SDE I,4,Jane's younger brother
Jane,Doe,8 Hollows Rd,Anytown,VA,407-555-4322,SDE II,5,John's older sister
```

在以下映射架构示例中，列 `FirstName` 和 `LastName` 是 `cleartext` 列。`State` 列作为 `fingerprint` 列和 `sealed` 列进行加密，填充为 none。其余列均省略。

```
{
  "headerRow": true,
  "columns": [
    {
      "sourceHeader": "FirstName",
      "targetHeader": "GivenName",
      "type": "cleartext"
    },
```

```
    {
      "sourceHeader": "LastName",
      "targetHeader": "Surname",
      "type": "cleartext"
    },
    {
      "sourceHeader": "State",
      "targetHeader": "State_Join",
      "type": "fingerprint"
    },
    {
      "sourceHeader": "State",
      "targetHeader": "State",
      "type": "sealed",
      "pad": {
        "type": "none"
      }
    }
  ]
}
```

以下是映射架构生成的 .csv 文件。

```
givenname,surname,state_fingerprint,state
John,Doe,01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv
+1Mk=,01:enc:FQ3n3Ahv9BQQNWQGcugeHzHYzEZE1vapHa2Uu4SRgSAtZ3qObjPA4TcsHt
+BOkMKBcnHWI13BeGG/SBqmj7vKpI=
Paulo,Santos,01:hmac:CHF4eIrtTNgAooU9v4h9Qjc
+txBnMidQTjdjWuaDTTA=,01:enc:KZ5n5GtaXACco65AXk48BQO2durDNR2ULc4YxmMC8NaZZKKJiksU1IwFadAvV4iBQ1
Mateo,Jackson,01:hmac:iIRnjfNBzryusIJ1w35lgNzeY1RQ1bSfq6PDHW8Xrbk=,01:enc:mLKpS5HIOSgphdEsrzhEd
eN9nBO2gAbIygt4OFn4LalYn9Xyj/XUWXlmn8zFe2T4kyDTD8kGOvpQEUGxAUFk=
Diego,Ramirez,01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv+1Mk=,01:enc:rmZhT98Zm
+IIGw1UTjMIJP4IrW/AAltBLMXcHvnYfRgmWP623VFQ6aUnhsb2MDqEw4G5Uwg5rKKZepUxx5uKbfk=
Jorge,Souza,01:hmac:3BxJdXiFFyZ8HBbYNqqEhBVqhNOd7s2ZiKUe7QiTyo8=,01:enc:vVaqWC1VRbhvkf8gnuR7q0z
Terry,Whitlock01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv+1Mk=,01:enc:3c9VEWbODO/
xbQjdGuccLvI7oZTBdPU+SyrJIyr2kudfAxbuMQ2uRdU/q7rbgyJjxZS8M2U35ILJf/lDgTyg7cM=
Jane,Doe,01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv+1Mk=,01:enc:9RWv46YLveykeNZ/
G0NdlYFg+AVdOnu05hHyAYTQkPLHnyX+0/jbzD/g9ZT8GCgVE9aB5bV4ooJIXHGBVMXcjrQ=
```

位置表架构

映射架构具有以下形状。

```
{
  "headerRow": false,
  "columns": [
    [
      {
        "targetHeader": STRING,
        "type": TYPE,
        "pad": PAD
      },
      {
        "targetHeader": STRING,
        "type": TYPE,
        "pad": PAD
      }
    ],
    [],
    ...
  ]
}
```

如果 headerRow 是 false，则对象中的下一个字段是 columns，其中包含一个条目数组。每个条目本身就是一个由零个或多个位置列架构（无 sourceHeader 字段）组成的数组，这些架构是描述输出应包含内容的 JSON 对象。

- sourceHeader — 数据来源于的源列的 STRING 标题名称。

> **Note**
>
> 在位置架构中必须省略此字段。在位置架构中，源列由架构文件中该列的相应索引推断出来。

- targetHeader — 输出文件中相应列的 STRING 标题名称。

> **Note**
>
> 对于位置架构，此字段为必填字段。

- type — 输出文件中目标列的 TYPE。即 cleartext、sealed 或 fingerprint 其中之一，具体取决于该列在协作中的使用方式。

- pad — 列架构对象的字段，仅当 TYPE 为 sealed 才存在。其对应的 PAD 值是一个对象，用于描述数据在加密前应如何填充。

```
{
  "type": PAD_TYPE,
  "length": INT
}
```

要指定加密前的填充，请按如下方式使用 type 和 length：

- PAD_TYPE 为 none — 不对列的数据进行填充，length 字段不适用（即省略）。
- PAD_TYPE 为 fixed — 将列的数据填充到指定的字节 length。
- PAD_TYPE 为 max — 列的数据填充到最长值的字节长度加上额外的 length 字节。

> **Note**
>
> 如果您提前知道列数据的字节大小的上限，则 fixed 很有用。如果该列中的任何数据长于指定的 length，则会引发错误。
> 当输入数据的确切大小未知时，max 很方便，因为无论数据的大小如何，它都能正常工作。但是，由于它会对数据进行两次加密，因此 max 需要额外的处理时间。max 在读入临时文件时对数据进行一次加密，在已知列中最长的数据条目之后加密一次。
> 此外，最长值的长度不会在两次调用客户端之间保存。如果您计划分批加密数据或定期加密新数据，请注意生成的加密文字长度可能因批次而异。

以下是位置架构的示例。

```
{
  "headerRow": false,
  "columns": [
    [
      {
        "targetHeader": "name",
        "type": "cleartext"
      }
    ],
    [
      {
        "targetHeader": "city_sealed",
        "type": "sealed",
```

```
        "pad": {
          "type": "max",
          "length": 16
        }
      }
    ],
    [
      {
        "targetHeader": "phone_number_fingerprint",
        "type": "fingerprint"
      },
      {
        "targetHeader": "phone_number_sealed",
        "type": "sealed",
        "pad": {
          "type": "fixed",
          "length": 20
        }
      }
    ]
  ]
}
```

举一个复杂的示例，以下是一个 .csv 文件示例，前提是它的第一行没有标题。

```
Jorge,Souza,12345 Mills Rd,Anytown,SC, 703 -555 -1234,CEO, 10,
Paulo,Santos, 0 Street,Anytown,MD, 404-555-111,CIO, 9,This is a really long note that
 could really be a paragraph
Mateo,Jackson, 1 Two St,Anytown,NY, 304-555-1324,COO, 9, ""
Terry,Whitlock, 4 N St,Anytown,VA, 407-555-8888,EA, 7,Secret notes
Diego,Ramirez, 9 Hollows Rd,Anytown,VA, 407-555-1222,SDE I, 4,null
John,Doe, 8 Hollows Rd,Anytown,VA, 407-555-4321,SDE I, 4,Jane's younger brother
Jane,Doe, 8 Hollows Rd,Anytown,VA, 407-555-4322,SDE II, 5,John's older sister
```

映射架构具有以下形式。

```
{
  "headerRow": false,
  "columns": [
    [
      {
        "targetHeader": "GivenName",
        "type": "cleartext"
```

```
      }
    ],
    [
      {
        "targetHeader": "Surname",
        "type": "cleartext"
      }
    ],
    [],
    [],
    [
      {
        "targetHeader": "State_Join",
        "type": "fingerprint"
      },
      {
        "targetHeader": "State",
        "type": "sealed",
        "pad": {
          "type": "none"
        }
      }
    ],
    [],
    [],
    [],
    []
  ]
}
```

上述架构生成以下输出文件，该文件具有包含指定目标标题的标题行。

```
givenname,surname,state_fingerprint,state
Mateo,Jackson,01:hmac:iIRnjfNBzryusIJ1w35lgNzeY1RQ1bSfq6PDHW8Xrbk=,01:enc:ENS6QD3cMVl9vQEGfe9MN
Q8m/Y5SA89dJwKpT5rGPp8e36h6klwDoslpFzGvU0=
Jorge,Souza,01:hmac:3BxJdXiFFyZ8HBbYNqqEhBVqhNOd7s2ZiKUe7QiTyo8=,01:enc:LKo0zirq2+
+XEIIIMNRjAsGMdyWUDwYaum0B+IFP+rUf1BNeZDJjtFe1Z+zbZfXQWwJy52Rt7HqvAb2WIK1oMmk=
Paulo,Santos,01:hmac:CHF4eIrtTNgAooU9v4h9Qjc
+txBnMidQTjdjWuaDTTA=,01:enc:MyQKyWxJ9kvK1xDQQtXlUNwv3F+yrBRr0xrUY/1BGg5KFgOn9pK+MZ7g
+ZNqZEPcPz4lht1u0t/wbTaqzOCLXFQ=
Jane,Doe,01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv+1Mk=,01:enc:Pd8sbITBfb0/
ttUB4svVsgoYkDfnDvgkvxzeci0Yxq54rLSwccy1o3/B50C3cpkkn56dovCwzgmmPNwrmCmYtb4=
```

```
Terry,Whitlock01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv
+1Mk=,01:enc:Qmtzu3B3GAXKh2KkRYTiEAaMopYedsSdF2e/
ADUiBQ9kv2CxKPzWyYTD3ztmKPMka19dHre5VhUHNpO3O+j1AQ8=
Diego,Ramirez,01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv+1Mk=,01:enc:ysdg
+GHKdeZrS/geBIooOEPLHG68MsWpx1dh3xjb+fG5rmFmqUcJLNuuYBHhHAlxchM2WVeV1fmHkBX3mvZNvkc=
John,Doe,01:hmac:UK8s8Cn/WR2JO/To2dTxWD73aDEe2ZUXeSHy3Tv+1Mk=,01:enc:9uX0wZuO7kAPAx
+Hf6uvQownkWqFSKtWS7gQIJSe5aXFquKWCK6yZN0X5Ea2N3bn03Uj1kh0agDWoiP9FRZGJA4=
```

使用 C3R 加密客户端解密数据表

对于使用加密计算的协作，请按照以下步骤操作 Clean Rooms 以及用于加密数据表的 C3R 加密客户端。在协作中查询数据后，请使用此过程。

此过程需要共享密钥和协作 ID。

能够接收结果的成员使用用于加密协作数据的共享密钥和协作 ID 来解密数据。

> **Note**
>
> AWS Clean Rooms 协作已经限制了谁可以执行和查看查询结果。要执行解密，任何有权访问这些结果的人都需要使用与加密数据相同的共享密钥和协作 ID。

解密已加密的数据表

1. （可选）在 C3R 加密客户端中查看可用命令。
2. （可选）导航到所需的目录并运行 `ls` (macOS) 或 `dir` (Windows).
   - 确认 c3r-cli.jar 文件和加密的查询结果数据文件位于所需的目录中。

     > **Note**
     >
     > 如果查询结果是从 AWS Clean Rooms 控制台界面下载的，则可能位于您的用户帐户的 “下载” 文件夹中。（例如，您的用户目录中的 “下载” 文件夹 Windows 以及 macOS。）我们建议您将查询结果文件移到与查询结果文件相同的文件夹 c3r-cli.jar.
3. 将共享密钥存储在 `C3R_SHARED_SECRET` 环境变量中。有关更多信息，请参阅 步骤 6：将共享密钥存储在 环境变量中。。
4. 从 AWS Command Line Interface (AWS CLI) 中运行以下命令。

java -jar c3r-cli.jar decrypt ***<name of input .csv file>*** --id=***<collaboration id>*** --output=***<output file name>***

5. 用你自己的信息替换每一个*user input placeholder*信息：

   a. 对于 `id=`，输入协作 ID。

   b. 对于 `output=`，输入输出文件的名称（例如，`results-decrypted.csv`）。

      如果不指定输出名称，则终端中会显示默认名称。

   c. 使用您的首选 CSV 查看指定输出文件中的解密数据，或者 Parquet 查看应用程序（例如 Microsoft Excel、文本编辑器或其他应用程序）。

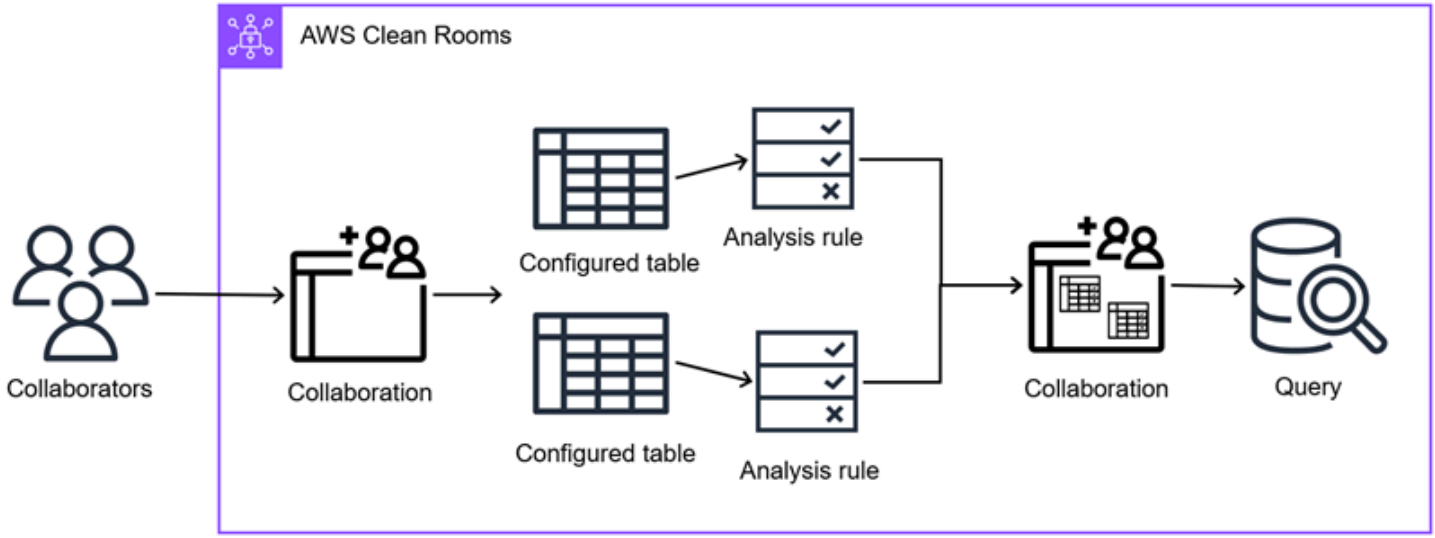
中配置的表 AWS Clean Rooms

配置的表是对数据源中现有表的引用。它包含一个分析规则，用于确定如何查询数据， AWS Clean Rooms 并且可以包括用于控制表使用情况的数据访问预算。配置表可以与一个或多个协作关联。

使用 AWS Clean Rooms，您可以对事件数据进行聚合分析，例如购买数量与购买数量的比较。您还可以对事件数据执行列表分析，例如从细分数据到 CRM 数据丰富重叠的客户数据。您还可以对事件数据（例如观众数据和细分属性）执行自定义查询并设置差别隐私。对于其中任何一种分析类型，您都可以设置数据访问预算，以监控和控制通过查询访问的数据量。

首先，您可以在中创建协作 AWS Clean Rooms 并添加要邀请的人，或者通过创建成员资格来加入受邀加入的协作。 AWS 账户 接下来，您和协作中的其他成员创建配置表。您既可以向配置的表（聚合、列表或自定义）添加分析规则，也可以选择设置数据访问预算。然后，将配置的表与协作关联起来。最后，可以查询的成员跨两个数据表运行查询，在执行查询时消耗数据访问预算。

下图总结了如何在中使用事件数据 AWS Clean Rooms。



主题

- 在中创建配置表 AWS Clean Rooms
- 为配置表添加分析规则。
- 将配置表与协作关联
- 配置数据访问预算
- 为配置表添加协作分析规则。
- 配置差别隐私策略（可选）

- 查看表格和分析规则
- 编辑已配置的表
- 编辑配置表标签
- 编辑配置的表分析规则
- 删除已配置的表分析规则
- 配置表不允许的列
- 编辑配置表关联
- 取消关联已配置的表

在中创建配置表 AWS Clean Rooms

配置的表是对数据源中现有表的引用。它包含一个分析规则，用于确定如何在 AWS Clean Rooms中查询数据。配置表可以与一个或多个协作关联。

有关如何使用创建已配置表的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

主题

- 创建已配置的表-Amazon S3 数据源
- 创建已配置表-Amazon Athena 数据源
- 创建已配置的表 — Snowflake 数据源

创建已配置的表-Amazon S3 数据源

在此步骤中，成员将执行以下任务：

- 配置现有 AWS Glue 表以在中使用。 AWS Clean Rooms（除非使用 Clean Rooms 加密计算，否则此步骤可以在加入协作之前或之后完成。）

> **Note**
>
> AWS Clean Rooms 支持 AWS Glue 表格。有关获取数据的更多信息 AWS Glue，请参阅步骤 3：将数据表上传到 Amazon S3。

- 为配置表命名，并选择要在协作中使用的列。

以下步骤假设：

- 协作成员已将其数据表上传到 Amazon S3 并创建了一个 AWS Glue 表。

> **Note**
>
> Amazon S3 中的结果目标不能与任何数据源位于同一 S3 存储桶中。

- （可选）仅对于加密数据表，协作成员已经使用 C3R 加密客户端准备了加密数据表。

您可以使用提供的统计数据生成 AWS Glue 来计算表的列级统计数据。 AWS Glue Data Catalog 为数据目录中的表 AWS Glue 生成统计数据后，Amazon Redshift Spectrum 会自动使用这些统计数据来优化查询计划。有关使用计算列级统计信息的更多信息 AWS Glue，请参阅AWS Glue 用户指南中的使用列统计信息优化查询性能。有关更多信息 AWS Glue，请参阅 AWS Glue 开发人员指南。

创建已配置的表-Amazon S3 数据源

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 在右上角，选择配置新表。
4. 对于数据源，在AWS 数据源下，选择 Amazon S3。
5. 在 Amazon S3 表格下：

   a. 选择托管 S3 表的区域。

   默认情况下，选择当前区域（例如弗吉尼亚北部 us-east-1）。

   > **Warning**
   >
   > 当您的 Amazon S3 数据源位于与您的处理位置不同的区域时，数据处理可能会暂时在源区域之外进行。在继续操作之前，请验证跨区域数据移动是否符合您的数据主权要求、监管合规政策和数据治理标准。

   有关区域的更多信息，请参阅中的区域和终端节点AWS 一般参考。

   b. 从下拉列表中选择数据库。

c. 从下拉列表中选择要配置的表。

> **Note**
>
> 要验证是否是正确的表，请执行以下任一操作：
>
> - 选择 “查看方式” AWS Glue。
> - 打开 “查看来自的架构” AWS Glue以查看架构。

> **Important**
>
> 对于数据采用 CSV 格式的 AWS Glue 表，Glue 架构中的列名和顺序必须与 CSV 数据完全匹配。如果它们不对齐，则可能无法正确执行已配置表的允许列列表。

6. 对于协作中允许的列和分析方法，

   a. 您想在协作中允许哪些专栏？

      - 选择所有列以允许在协作中查询所有列。
      - 选择自定义列表以允许在协作中查询 “指定允许的列” 下拉列表中的一个或多个列。

   b. 对于允许的分析方法，

      i. 选择 “直接查询” 以允许直接在此表上运行 SQL 查询

      ii. 选择 Direct job 以允许直接在此表上运行 PySpark 作业。

   Example示例

   例如，如果要允许协作成员在所有列上同时运行直接 SQL 查询和 PySpark 作业，请选择 “所有列”、“直接查询” 和 “直接作业”。

7. 对于已配置表的详细信息，

   a. 为已配置的表输入名称。

      您可以使用默认名称或重命名此表。

   b. 输入表的描述。

该描述有助于区分其他具有相似名称的已配置表。

8. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。
9. 选择配置新表。

现在您已经创建了一个配置表，您已准备好：

- 为配置表添加分析规则
- 将配置表与协作关联

创建已配置表-Amazon Athena 数据源

Amazon Athena 数据源选项允许您查询存储在 Amazon S3 中、在数据目录或联合目录 AWS Glue 中编目以及通过控制访问的数据。 AWS Lake Formation同时支持表格和 AWS Glue Data Catalog 视图。Lake Formation 资源链接可用于在加入 AWS Clean Rooms 协作的 AWS Clean Rooms 成员账户之间 AWS 区域 共享表格 AWS 账户 和视图。

> **Note**
>
> 只有基于 Amazon S3 的数据集才能通过 Athena 数据源集成进行查询。

在此步骤中，成员将执行以下任务：

- 在中配置现有表或视图以 AWS Glue Data Catalog 供使用 AWS Clean Rooms
- 为配置表命名，并选择要在协作中使用的列。

以下步骤假设：

- 协作成员已经创建了 AWS Glue Data Catalog 数据库和表或 GDC 视图。

创建已配置表-Athena 数据源

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。

3. 在右上角，选择配置新表。
4. 对于数据源，在AWS 数据源下，选择 Amazon Athena。
5. 在亚马逊 Athena 表格下：

   a. 选择托管 Amazon Athena 表的区域。

   默认情况下，选择当前区域（例如弗吉尼亚北部 us-east-1）。

   > **Warning**
   >
   > 当您的 Amazon Athena 数据源位于与您的处理地点不同的区域时，数据处理可能会暂时在源区域之外进行。在继续操作之前，请验证跨区域数据移动是否符合您的数据主权要求、监管合规政策和数据治理标准。

   有关区域的更多信息，请参阅中的区域和终端节点AWS 一般参考。

   b. 从下拉列表中选择目录。

   默认情况下，“AWS Glue 数据目录” 处于选中状态。

   - AWS Glue 数据目录-中表格的默认目录 AWS Glue。
   - 联合目录-如果您已将目录联合配置为连接到远程 Apache Iceberg REST AWS Glue 目录，则可用。有关更多信息，请参阅《AWS Lake Formation 开发人员指南》中的目录联合。

   c. 从下拉列表中选择数据库。

   d. 从下拉列表中选择要配置的表。

   > **Note**
   >
   > 要验证是否是正确的表，请执行以下任一操作：
   >
   > - 选择查看方式 AWS Glue或查看方式 AWS Lake Formation（取决于您的目录类型）。
   > - 打开 “查看来自的架构” AWS Glue以查看架构。

6. 对于亚马逊 Athena 配置，

   a. 从下拉列表中选择一个工作组。

   b. 对于 S3 输出位置，请根据以下情况之一选择建议的操作。

场景	推荐操作
您的工作组没有默认输出位置。	输入 S 3 输出位置或选择 “浏览 S3”。
您的工作组强制使用您的默认输出位置。	S 3 输出位置是自动选择的，您无法对其进行更改。
您的工作组不强制使用您的默认输出位置。	输入 S 3 输出位置或选择 “浏览 S3”。

7. 对于协作中允许的列，请根据您的目标选择一个选项。

您的目标	建议的选项
允许在中使用所有列 AWS Clean Rooms （视分析规则而定）	所有列
允许 “指定允许的列” 下拉列表中的一个或多个列	自定义列表

8. 对于已配置表的详细信息，

   a. 为已配置的表输入名称。

      您可以使用默认名称或重命名此表。

   b. 输入表的描述。

      该描述有助于区分其他具有相似名称的已配置表。

   c. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。

9. 选择配置新表。

现在您已经创建了一个配置表，您已准备好：

- 为配置表添加分析规则
- 将配置表与协作关联

创建已配置的表 — Snowflake 数据源

在此步骤中，成员将执行以下任务：

- 配置现有的 Snowflake 表以在中使用。 AWS Clean Rooms（除非使用 Clean Rooms 加密计算，否则此步骤可以在加入协作之前或之后完成。）
- 为配置表命名，并选择要在协作中使用的列。

以下步骤假设：

- 协作成员已经将其数据表上传到 Snowflake。
- （可选）仅对于加密数据表，协作成员已经使用 C3R 加密客户端准备了加密数据表。

创建已配置的表-Snowflake 数据源

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 在右上角，选择配置新表。
4. 对于数据源，在第三方云和数据源下，选择 Snowflake。
5. 使用现有密钥 ARN 或存储此表的新密钥指定 Sn owflake 凭证。

   Use existing secret ARN

   1. 如果您有秘密 ARN，请将其输入到秘密 ARN 字段。

      您可以通过选择 “前往” 来查找您的秘密 ARN。 AWS Secrets Manager
   2. 如果您已有来自其他表的密钥，请选择从现有表中导入密钥 ARN。

   > **Note**
   >
   > 秘密 ARN 可以是跨账户的。

   Store a new secret for this table

   1. 输入以下 Snowflake 凭据：

- 雪花用户名
- 雪花仓库
- 雪花角色
- Snowflake 隐私增强邮件 (PEM) 私钥

2. 要进行加密，请执行以下任一操作：
   - 要使用 AWS 托管式密钥 （默认），请清除 “自定义加密设置” 复选框。
   - 要使用自定义，请执行 AWS KMS key以下操作：
     - 选中 “自定义加密设置” 复选框。
     - 对于 KMS 密钥，请输入密钥 ARN 或从列表中选择一个密钥。
3. 输入 S ecret 名称以帮助你以后找到你的凭证。

6. 有关 Snowflake 表和架构的详细信息，请手动输入详细信息或自动导入详细信息。

   Enter the details manually

   1. 输入 Sn owflake 账户标识符。

      有关更多信息，请参阅 Snowflake 文档中的账户标识符。

      您的账户标识符必须采用 Snowflake 驱动程序使用的格式。你需要用连字符 (-) 替换句点 (.)，这样标识符的格式就是。**<orgname>-<account_name>**
   2. 进入雪花数据库。

      有关更多信息，请参阅 Snowflake 文档中的 Snowflake 数据库。
   3. 输入 Snowflake 架构名称。
   4. 输入 Sn owflake 表的名称。

      有关更多信息，请参阅 Snowflake 文档中的了解 Snowflake 表结构。
   5. 对于架构，输入列名并从下拉列表中选择数据类型。
   6. 选择 “添加列” 以添加更多列。
      - 如果选择对象数据类型，请指定对象架构。

        Example对象架构示例

```
name STRING,
location OBJECT(
    x INT,
```

```
    y INT,
    metadata OBJECT(uuid STRING)
),
history ARRAY(TEXT)
```

- 如果选择数组数据类型，请指定数组架构。

  Example数组架构示例

  ```
  OBJECT(x INT, y INT)
  ```

- 如果选择地图数据类型，请指定地图架构。

  Example地图架构示例

  ```
  STRING, OBJECT(x INT, y INT)
  ```

Automatically import the details

1. 将你的 “列” 视图从 Snowflake 导出为 CSV 文件。

   有关 Snowflake COLUMNS 视图的更多信息，请参阅 Snowflake 文档中的列视图。

2. 选择 “从文件导入” 以导入 CSV 文件并指定任何其他信息。

   将自动导入数据库名称、架构名称、表名、列名和数据类型。

   - 如果选择对象数据类型，请指定对象架构。
   - 如果选择数组数据类型，请指定数组架构。
   - 如果选择地图数据类型，请指定地图架构。

3. 输入 Sn owflake 账户标识符。

   有关更多信息，请参阅 Snowflake 文档中的账户标识符。

> **Note**
>
> 只有编入目录的 S3 表 AWS Glue 才能用于自动检索表架构。

7. 对于协作中允许的列，请根据您的目标选择一个选项。

您的目标	建议的选项
允许在中使用所有列 AWS Clean Rooms （视分析规则而定）	所有列
允许 “指定允许的列” 下拉列表中的一个或多个列	自定义列表

8. 对于已配置表的详细信息，

   a. 为已配置的表输入名称。

      您可以使用默认名称或重命名此表。

   b. 输入表的描述。

      该描述有助于区分其他具有相似名称的已配置表。

   c. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。

9. 选择配置新表。

现在您已经创建了一个配置表，您已准备好：

- 为配置表添加分析规则
- 将配置表与协作关联

为配置表添加分析规则。

以下各节介绍了如何为您的配置表添加分析规则。通过定义分析规则，您可以授权可以查询的成员运行与 AWS Clean Rooms支持的特定分析规则匹配的查询。

AWS Clean Rooms 支持以下类型的分析规则：

- 聚合分析规则
- 列表分析规则
- 中的自定义分析规则 AWS Clean Rooms

每个配置表只能有一个分析规则。您可以在将配置表与协作关联之前随时配置分析规则。

> **Important**
>
> 如果您在协作中使用 Clean Rooms 加密计算且有加密数据表，则添加到加密配置表的分析规则应与数据的加密方式一致。例如，如果您为 SELECT（聚合分析规则）加密了数据，则不应添加 JOIN（列表分析规则）的分析规则。

主题

为表添加聚合分析规则（引导流程）

聚合分析规则支持使用 COUNT、SUM 和 AVG 函数按可选维度聚合统计数据而不会泄露行级信息的查询。

此过程描述了使用 AWS Clean Rooms 控制台中的引导流程选项为配置表添加聚合分析规则的过程。

> **Note**
>
> 使用非 S3 数据源的配置表仅支持自定义分析规则。

为表添加聚合分析规则（引导流程）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择配置表。
4. 在配置表详细信息页面上，选择配置分析规则。
5. 在 “步骤 1：选择分析规则类型” 下，在 “分析规则类型” 下，选择 “聚合” 选项。

6. 在创建方法下，选择引导流程，然后选择下一步。
7. 在步骤 2: 指定查询控制下，对于聚合函数：
   a. 从下拉列表中选择一个聚合函数：
      - COUNT
      - COUNT DISTINCT
      - SUM
      - SUM DISTINCT
      - AVG
   b. 从列下拉列表中选择哪些列可以用于聚合函数。
   c. （可选）选择添加其他函数以添加另一个聚合函数，并将一个或多个列与该函数相关联。

      > **Note**
      > 至少需要一个聚合函数。

   d. （可选）选择移除以删除聚合函数。
8. 对于联接控制，
   a. 为允许单独查询表选择一个选项：

如果选择...	操作...
否，只能查询重叠	只有在联接到可以查询的成员拥有的表时，才能对表进行查询。
是	表可以单独查询，也可以在与其他表联接后进行查询。

   b. 在指定联接列下，选择要允许在 INNER JOIN 语句中使用的列。

      如果您在上一步中选择了是，则这是可选的。
   c. 在指定允许的匹配运算符下，选择哪些运算符（如果有）可用于在多个联接列上进行匹配。如果您选择两列或更多 JOIN 列，则需要其中一个运算符。

如果选择...	操作...
AND	您可以在 INNER JOIN 匹配条件中包含 AND，在表之间将一列联接到另一列。
或者	您可以在 INNER JOIN 匹配条件中包含 OR，在表之间将一列与另一列进行匹配。此逻辑运算符对于获得更高的匹配率很有用。

9. （可选）对于维度控制，在指定维度列下拉列表中，选择要允许在 SELECT 语句中使用的列，以及查询的 WHERE、GROUP BY 和 ORDER BY 部分。

> **Note**
>
> 聚合函数或联接列不能用作维度列。

10. 对于标量函数，请为要允许哪些标量函数？选择一个选项。

如果选择...	操作...
所有目前都支持 AWS Clean Rooms	您允许 AWS Clean Rooms当前支持的所有标量函数。 • 您可以选择查看列表以查看 AWS Clean Rooms中支持的标量函数的完整列表。
自定义列表	您可以自定义允许哪些标量函数。 • 从指定允许的标量函数下拉列表中选择一个或多个选项。
无	您不想允许任何标量函数。

有关更多信息，请参阅 标量函数。

11. 选择下一步。
12. 在步骤 3: 指定查询结果控制下，为聚合约束：

a. 选择每个列名称的下拉列表。

b. 选择应用 COUNT DISTINCT 函数后返回的每个输出行必须满足的每个不同值的最小数量的下拉列表。

c. 选择添加约束，添加更多聚合约束。

d. （可选）选择移除以删除聚合约束。

13. 对于应用于输出的其他分析，请根据您的目标选择一个选项。

您的目标	建议的选项
仅允许对该表进行直接查询。拒绝对查询结果运行其他分析。该表只能用于直接查询。	不允许
允许但不要求对该表进行直接查询和其他分析。	允许
要求该表只能用于通过所需的其他分析之一进行处理的直接查询。对该表进行的直接查询必须经过进一步处理才能返回。	必填

14. 选择下一步。

15. 在步骤 4: 查看并配置下，查看您在之前的步骤中所做的选择，必要时进行编辑，然后选择配置分析规则。

您将看到一条确认消息，指出您成功为表配置了聚合分析规则。

为表添加列表分析规则（引导流程）

列表分析规则支持输出关联表与可查询成员的表之间重叠情况行级列表的查询。

此过程描述了使用 AWS Clean Rooms 控制台中的 “引导流程” 选项将列表分析规则添加到配置的表中的过程。

> **Note**
>
> 使用非 S3 数据源的配置表仅支持自定义分析规则。

为表添加列表分析规则（引导流程）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择配置表。
4. 在配置表详细信息页面上，选择配置分析规则。
5. 在 “步骤 1：选择分析规则类型” 下，在 “分析规则类型” 下，选择 “列表” 选项。
6. 在创建方法下，选择引导流程，然后选择下一步。
7. 在步骤 2: 指定查询控制下，对于联接控制：

   a. 在指定联接列下，选择要允许在 INNER JOIN 语句中使用的列。

   b. 在指定允许的匹配运算符下，选择哪些运算符（如果有）可用于在多个联接列上进行匹配。如果您选择两列或更多 JOIN 列，则需要其中一个运算符。

如果选择...	操作...
AND	您可以在 `INNER JOIN` 匹配条件中包含 `AND`，在表之间将一列联接到另一列。
或者	您可以在 `INNER JOIN` 匹配条件中包含 `OR`，在表之间将一列与另一列进行匹配。此逻辑运算符对于获得更高的匹配率很有用。

8. （可选）对于列表控制，在指定列表列下拉列表中，选择要允许在查询输出中使用（即在 SELECT 语句中使用）或用于筛选结果（即 WHERE 语句）的列。
9. 选择下一步。
10. 在步骤 3: 指定查询结果控制下，对于应用于输出的其他分析，请根据您的目标选择一个选项。

您的目标	建议的选项
仅允许对该表进行直接查询。拒绝对查询结果运行其他分析。该表只能用于直接查询。	不允许
允许但不要求对该表进行直接查询和其他分析。	允许

您的目标	建议的选项
要求该表只能用于通过所需的其他分析之一进行处理的直接查询。对该表进行的直接查询必须经过进一步处理才能返回。	必填

11. 在步骤 4: 查看并配置下，查看您在之前的步骤中所做的选择，必要时进行编辑，然后选择配置分析规则。

您将看到一条确认消息，指出您成功为表配置了列表分析规则。

为表添加自定义分析规则（引导流程）

自定义分析规则允许对已配置的表进行自定义 SQL 查询或 PySpark 作业。如果使用下列项，则需要自定义分析规则：

- 分析模板允许一组特定的预先批准的 SQL 查询或 PySpark 作业，或者一组可以提供使用您的数据的查询的特定帐户。
- AWS Clean Rooms 差异隐私，可防止用户识别尝试。
- 非 S3 数据源，例如 Amazon Athena 或 Snowflake。

此过程描述了使用 AWS Clean Rooms 控制台中的 “引导流程” 选项将自定义分析规则添加到配置的表中的过程。

为表添加自定义分析规则（引导流程）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择配置表。
4. 在配置表详细信息页面上，选择配置分析规则。
5. 在 “步骤 1：选择分析规则类型” 下，在 “分析规则类型” 下，选择 “自定义” 选项。
6. 在创建方法下，选择引导流程，然后选择下一步。
7. 在 “步骤 2：指定分析控件” 下，对于直接分析控件，根据您的目标选择一个选项。

您的目标	推荐操作
在允许在此配置的表上运行每个新分析之前，请先对其进行检查	1. 在允许运行的分析模板下，选择添加分析模板。 2. 从下拉列表中选择相应的协作和分析模板。 3. 选择下一步。
允许特定的合作者无需查看此表即可对所选类型进行任何分析	1. 在 “分析类型” 下， a. 选择 “任意查询” 以允许由 AWS 账户 您指定的创建的任何查询。 b. 选择 “任意查询” 以允许 AWS 账户 您指定创建的任何作业。 2. 在 “AWS 账户 允许创建任何分析” 下，选择 “添加” AWS 账户。 3. 输入 AWS 账户 或从下拉列表中选择 AWS 账户 ID.。 4. （可选）选择 “添加另一个” AWS 账户以添加另一个 AWS 账户。 5. 选择下一步。

8. 在 “步骤 3：指定分析结果控件” 下，

   a. 对于 Job 结果控件，请注意，不支持其他结果控件。

   b. 在 “查询结果控件” 下，对于 “输出中不允许的列”，根据您的目标选择要允许在查询输出中使用的列。

您的目标	推荐操作
允许在查询输出中返回所有列	1. 选择 “无” 2. 继续执行应用于输出的其他分析。
不允许在查询输出中返回某些列	1. 选择 “自定义” 列表 2. 在指定不允许的列下，选择要从查询输出中删除的列。

c. 对于应用于输出的其他分析，请根据您的目标选择是否可以将其他分析应用于查询输出。

您的目标	建议的选项
• 仅允许对该表进行直接查询。 • 拒绝对查询结果运行其他分析。 • 该表只能用于直接查询。	不允许
允许但不要求对该表进行直接查询和其他分析。	允许
• 要求该表只能用于通过所需的其他分析之一进行处理的直接查询。 • 对该表进行的直接查询必须经过进一步处理才能返回。	必填

d. 选择下一步。

9. （可选）在“步骤 4：设置差异隐私”下，确定是要开启还是关闭差异隐私。

差别隐私是一种经过数学验证的技术，可以保护您的数据以免受到重新识别攻击。

> **Note**
>
> AWS Clean Rooms 差异隐私仅适用于数据存储在 Amazon S3 中的协作。

对于差异隐私，请根据您的目标选择是开启还是关闭差分隐私。

您的目标	推荐操作
• 您不需要针对重新识别尝试的保护 • 您的表中没有用户级数据	1. 选择关闭。 2. 选择下一步。
• 您需要防范重新识别尝试 • 您的表包含用户级数据	1. 选择打开。 2. 选择包含用户唯一标识符的用户标识符`user_id`列，例如要保护其隐私的列。

您的目标	推荐操作
	要为协作中的两个或更多表开启差别隐私，您必须在两个分析规则中配置与用户标识符列相同的列，以在表之间保持一致的用户定义。如果未正确进行配置，可以查询的成员将收到一条错误消息，指出具有两列可供选择，以便在运行查询时计算用户贡献数量（例如，用户生成的广告展示次数）。 3. 选择下一步。

10. 在步骤 5: 查看并配置下，查看您在之前的步骤中所做的选择，必要时进行编辑，然后选择配置分析规则。

您将看到一条确认消息，指出您成功为表配置了自定义分析规则。

为表添加分析规则（JSON 编辑器）

以下过程说明如何使用 AWS Clean Rooms 控制台中的 JSON 编辑器选项向表中添加分析规则。

> **Note**
>
> 使用非 S3 数据源的配置表仅支持自定义分析规则。

为表添加聚合、列表或自定义分析规则（JSON 编辑器）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择配置表。
4. 在配置表详细信息页面上，选择配置分析规则。
5. 在 “步骤 1：选择分析规则类型” 下，在 “分析规则类型” 下，选择 “聚合”、“列表” 或 “自定义” 选项。
6. 在创建方法下，选择 JSON 编辑器，然后选择下一步。
7. 在步骤 2: 指定控制下，您可以选择插入查询结构（插入模板）或插入文件（从文件导入）。

如果选择...	操作...
插入模板	1. 在分析规则定义中为所选分析规则指定参数。 2. 您可以按 Ctrl + 空格键启用自动完成。 有关聚合分析规则参数的更多信息，请参阅聚合分析规则 — 查询控制。 有关列表分析规则参数的更多信息，请参阅列表分析规则 — 查询控制。
从文件导入	1. 从本地驱动器中选择您的 JSON 文件。 2. 选择打开。 分析规则定义显示上传文件中的分析规则。

8. 选择下一步。
9. 在步骤 3: 查看并配置下，查看您在之前的步骤中所做的选择，必要时进行编辑，然后选择配置分析规则。

您将收到一条确认消息，指出您成功为表配置了分析规则。

后续步骤

现在，您已经为配置表配置了分析规则，您已准备好：

- 将配置表与协作关联
- 查询数据表（以可以查询的成员身份）

将配置表与协作关联

创建已配置的表并向其添加分析规则后，可以将其与协作关联并授予 AWS Clean Rooms 服务角色来访问您的 AWS Glue 表。

> **Note**
>
> 此服务角色拥有对表的权限。服务角色只能由 AWS Clean Rooms 承担，代表可以查询的成员运行允许的查询。任何协作成员（数据所有者除外）都无法访问协作中的底层表。数据所有者可以开启差别隐私，以使其表可供其他成员查询。

数据访问预算

关联已配置的表时，可以应用数据访问预算。数据访问预算控制协作中表可用于查询、作业和机器学习输入渠道的次数。这些预算通过限制表格使用来帮助组织管理资源利用率和控制成本。

每次在查询、作业或 ML 输入渠道中使用表时，该表的预算都会减少一个。当预算达到零时，该表不能用于 SQL 查询、Pyspark 作业，也不能用作从该表派生的 ML 输入通道的一部分。

您可以制定定期刷新的每期预算、总使用量的生命周期预算，或两者兼而有之。默认情况下，表格使用量不受限制。

- 每期预算 — 一种可续期的分配，它限制了该表在指定时间段内的使用次数。您可以将时段设置为每天、每周或每月。可以将此预算设置为每天、每周或每月自动刷新。
- 生命周期预算 — 一种连续分配，用于限制使用此表的总次数。

关联已配置的表

以下主题介绍如何使用 AWS Clean Rooms 控制台关联已配置的表以及如何将数据访问预算应用于协作。

有关如何使用将配置的表格与协作关联的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

步骤 1：完成先决条件

要关联已配置的表，必须满足以下先决条件：

- 指向 Amazon S3 文件夹位置的 AWS Glue 表（不是单个文件）
- 对于加密 AWS Glue 表：
  - 具有使用 AWS KMS 密钥解 AWS Glue 密表的权限的服务角色
  - 对于 AWS KMS加密的 Amazon S3 数据集：服务角色还必须有权使用密 AWS KMS 钥解密 Amazon S3 数据

有关配置加密的信息，请参阅《AWS Glue 开发人员指南》 AWS Glue中的设置加密。

要验证您的 AWS Glue 餐桌位置，请执行以下操作：

1. 在以下位置打开 AWS Glue 控制台 https://console.aws.amazon.com/glue/
2. 查看您的表格详细信息并确认位置指向 S3 文件夹

步骤 2：关联已配置的表

关联已配置的表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 选择关联表的方法：
   a. 在已配置的表格详细信息页面上：
      i. 在左侧导航窗格中，选择 表。
      ii. 选择配置表。
      iii. 在配置表详细信息页面上，选择与协作关联。
      iv. 在将表与协作关联对话框中，从下拉列表中选择协作。
   b. 在协作详情页面上：
      i. 在左侧导航窗格中，选择协作。
      ii. 选择协作。
      iii. 在表选项卡上，选择关联表。
3. 在 “关联” 表格页面上，执行以下任一操作：
   - 选择现有的已配置表-从下拉列表中选择要与协作关联的已配置表名。
   - 配置新表-选择 “配置新表”，然后按照 “配置新表” 页面上的提示进行操作。
   - 查看已配置表的架构和分析规则-打开 “查看架构和分析规则”。
4. 对于表关联详细信息，
   a. 输入关联表的名称。

      您可以使用默认名称或重命名此表。
   b. （可选）输入表的描述。

该描述有助于编写查询。

5. 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。

> **Note**
>
> 如果您要关联由 Amazon Athena 支持的已配置表，请从下拉列表中选择现有服务角色名称。确保服务角色具有数据集的 IAM 权限，如果需要，还有 Lake Formation 权限。

如果选择...	操作...
创建并使用新的服务角色	• AWS Clean Rooms 使用此表所需的策略创建服务角色。 • 默认服务角色名称为 `cleanrooms-<timestamp>` 。 • 您必须拥有创建角色并附加策略的权限。 • 如果您的输入数据已加密，则可以选择此数据使用 KMS 密钥加密，然后输入将用于解密数据输入的 AWS KMS key 。
使用现有服务角色	1. 从下拉列表中选择一个现有服务角色名称。 如果您有列出角色的权限，则会显示角色列表。 如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。 2. 通过选择在 IAM 中查看外部链接来查看服务角色。 如果没有现有的服务角色，则使用现有服务角色选项不可用。 默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

如果选择...	操作...
	3. （可选）选中为该角色添加具有必要权限的预配置策略复选框以向该角色添加必要的附加权限。您必须拥有修改角色并创建策略的权限。

> **Note**
>
> - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
> - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
> - 如果您无法修改角色策略，则会收到一条错误消息，指出 AWS Clean Rooms 找不到该服务角色的策略。

6. 如果要为已配置的表关联资源启用已配置的表关联标签，请选择添加新标签，然后输入密钥和值对。
7. 选择下一步。
8. 在配置协作分析规则页面上，选择以下选项之一：

   - 是的，立即创建协作分析规则 — 将您的表格与此协作关联并创建协作分析规则
   - 不，我稍后将创建协作分析规则 — 仅将您的表格与该协作关联起来。您可以稍后创建协作分析规则。

9. 如果您选择 “是”，请立即创建协作分析规则，对于结果交付，请从下拉列表中选择 “允许成员接收查询输出结果”。
10. 选择下一步。
11. 在添加数据访问预算页面上，为数据访问预算配置选择以下选项之一：

    - 是的，立即添加数据访问预算 — 将您的表格与此协作关联并添加数据访问预算。您可以选择期间预算、生命周期预算，或两者兼而有之。
    - 不，我稍后会添加数据访问预算 — 仅将您的表格与本次协作关联起来。您可以稍后添加数据访问预算。

      如果您选择 “否”，我稍后将添加数据访问预算，请跳至步骤 15。

12. 如果您选择 “是，立即添加数据访问预算”，请选择以下预算配置之一：

仅限每期预算	仅限终身预算	每个周期和生命周期的预算
1. 将按期间添加预算保留为选中状态。 2. 输入介于 1 到 1,000,000 之间的每期预算金额。 3. 在 “周期” 中，选择 “每日”、“每周” 或 “每月”。 4. （可选）选中 “每周自动刷新预算” 以续订分配。 5. 清除 “添加生命周期预算”。	1. 清除 “添加按期间预算”。 2. 选择 “添加生命周期预算”。 3. 输入介于 1 到 1,000,000 之间的终身预算金额。	1. 将按期间添加预算保留为选中状态。 2. 输入介于 1 到 1,000,000 之间的每期预算金额。 3. 在 “周期” 中，选择 “每日”、“每周” 或 “每月”。 4. 将 “每周自动刷新预算” 保留为选中状态。 5. 选择 “添加生命周期预算”。 6. 输入介于 1 到 1,000,000 之间的终身预算金额。

13. 在 “数据访问预算摘要” 下查看您的选择。

    Example示例

    例如，如果您选择了每周期预算金额为 1,000，将周期设置为每周，将 “每周自动刷新预算” 复选框保持选中状态，并将生命周期预算设置为 1,000,000，则访问预算摘要将显示以下消息：每周，此表最多可使用 1,000 次来运行查询或作业。该预算设置为每周日 00:00 UTC 自动刷新，并将继续刷新，直到该表达到其使用寿命预算，即 1,000,000 次使用。

14. （可选）如果要为访问预算资源启用数据访问预算标签，请选择添加新标签并输入密钥和值对。
15. 选择下一步。
16. 在 “查看并创建” 页面上查看信息。

    a. 如果您需要编辑任何部分，请选择编辑。

    b. 编辑您的配置，然后选择 “下一步”。

17. 选择关联表。

步骤 3：后续步骤

现在，您已将配置数据表与协作关联，您已准备好：

- 添加协作分析规则至配置表
- 编辑协作（如果您是协作创建者）
- 查询数据表（以可以查询的成员身份）

配置数据访问预算

协作者可以查看、添加、编辑和删除数据访问预算，从而限制表在工作流程中的使用次数。使用这些预算来管理数据和成本。

每次使用从表中派生的 ML 输入通道查询表或运行作业时，该表的预算都会减少一个。当预算达到零时，无法查询表，也无法使用从表中派生的机器学习输入通道运行机器学习作业。

您可以制定定期刷新的每期预算、总使用量的生命周期预算，或两者兼而有之。默认情况下，表格使用量不受限制。

- 每期预算 — 一种可续期的分配，它限制了该表在指定时间段内的使用次数。您可以将时段设置为每天、每周或每月。可以将此预算设置为每天、每周或每月自动刷新。
- 生命周期预算 — 一种连续分配，用于限制使用此表的总次数。

主题

- 查看数据访问预算
- 向现有关联表添加数据访问预算
- 编辑数据访问预算
- 删除数据访问预算

查看数据访问预算

您可以从 “表” 选项卡或表格详细信息页面查看数据访问预算。

查看数据访问预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。

5. 请执行以下操作之一：

   - 在 “剩余数据访问预算” 列下，选择预算以查看详细信息。
   - 选择一个表格，然后在表格详细信息页面上，向下滚动以查看数据访问预算详细信息部分。

向现有关联表添加数据访问预算

作为协作成员，您可以向现有的关联表添加数据访问预算。

向现有关联表添加数据访问预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择要向其添加数据访问预算的表格旁边的选项按钮。
6. 从 “操作” 下拉列表的 “数据访问预算” 下，选择 “添加”（如果还没有预算）。
7. 选择以下预算配置之一：

仅限每期预算	仅限终身预算	每个周期和生命周期的预算
1. 将按期间添加预算保留为选中状态。 2. 输入介于 1 到 1,000,000 之间的每期预算金额。 3. 在 “周期” 中，选择 “每日”、“每周” 或 “每月”。 4. （可选）选中 “每周自动刷新预算” 以续订分配。 5. 清除添加生命周期预算。	1. 清除 “添加按期间预算”。 2. 选择添加生命周期预算。 3. 输入介于 1 到 1,000,000 之间的终身预算金额。	1. 将按期间添加预算保留为选中状态。 2. 输入介于 1 到 1,000,000 之间的每期预算金额。 3. 在 “周期” 中，选择 “每日”、“每周” 或 “每月”。 4. 将 “每周自动刷新预算” 保留为选中状态。 5. 选择添加生命周期预算。 6. 输入介于 1 到 1,000,000 之间的终身预算金额。

8. 在 “数据访问预算摘要” 下查看您的选择。

9. Example示例

   例如，如果您选择了每周期预算金额为 1,000，将周期设置为每周，将 “每周自动刷新预算” 复选框保持选中状态，并将生命周期预算设置为 1,000,000，则访问预算摘要将显示以下消息：每周，此表最多可使用 1,000 次来运行查询或作业。该预算设置为每周日 00:00 UTC 自动刷新，并将继续刷新，直到该表达到其使用寿命预算，即 1,000,000 次使用。

10. （可选）如果要为访问预算资源启用数据访问预算标签，请选择添加新标签并输入密钥和值对。
11. 选择添加数据访问预算。

编辑数据访问预算

作为协作成员，您可以编辑数据访问预算。当您编辑数据访问预算时，它会重置当前的预算余额。

您可以从 “表” 选项卡或表格详细信息页面编辑数据访问预算。

Tables tab

通过 “表” 选项卡编辑数据访问预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择要编辑的表格旁边的选项按钮。
6. 从 “操作” 下拉列表的 “数据访问预算” 下，选择 “编辑”。
7. 在编辑数据访问预算页面上，更新每周期预算或生命周期预算信息。
8. 查看数据访问预算摘要，以验证您所做的编辑是否正确。
9. 选择保存更改。

Table details page

从表格详细信息页面编辑数据访问预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择一个表。
6. 在表格详细信息页面上，向下滚动至数据访问预算详细信息部分。
7. 从操作下拉列表中，选择编辑。
8. 在编辑数据访问预算页面上，更新每周期预算或生命周期预算信息。
9. 选择保存更改。

删除数据访问预算

您可以从 “表” 选项卡或表格详细信息页面中删除数据访问预算。

Tables tab

从 “表” 选项卡中删除数据访问预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择要删除的表格旁边的选项按钮。
6. 从 “操作” 下拉列表的 “数据访问预算” 下，选择 “删除”。

> **Important**
>
> 您无法撤消此操作，您的数据访问预算将重置为无限制。

7. 如果您确定要删除数据访问预算，请选择删除。

Table details page

从表格详细信息页面删除数据访问预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择一个表。
6. 在表格详细信息页面上，向下滚动至数据访问预算详细信息部分。
7. 从操作下拉列表中，选择删除。

> **Important**
> 您无法撤消此操作，您的数据访问预算将重置为无限制。

8. 如果您确定要删除数据访问预算，请选择删除。

为配置表添加协作分析规则。

协作分析规则支持您指定特定于此协作的控制。这些控制与配置表分析规则配合使用，用于确定如何在此协作中分析该表。

在创建配置表、添加分析规则并将其与协作关联后，可以将协作分析规则添加到配置表中。如果将表配置为支持直接分析或允许进行其他分析，则需要添加协作分析规则。

- 直接分析 - 该表可用于直接对其进行分析的查询。例如，在输出聚合测量分析或激活标识符列表的查询中。
- 其他分析 - 除了直接分析表的查询外，还可以将该表用作其他分析的输入。例如，该表可用于查询中，该查询是相似机器学习模型的种子，或者是自定义机器学习模型的机器学习输入通道。

为表添加协作分析规则

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。

2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在表选项卡中的您关联的表下，查看您已与协作关联的配置表。
   - 如果直接分析状态或其他分析状态的状态为准备就绪，则该表已准备就绪，可供查询。
   - 如果直接分析状态或其他分析状态的状态为尚未准备就绪，请选中该状态，然后在对话框中选择配置。
5. 在配置协作分析规则页面上，展开查看配置表分析规则来查看详细信息。
6. 对于允许的其他分析，请根据您的目标选择相应选项。

您的目标	建议的选项
允许对表进行任何其他分析。	任何
仅允许特定成员对表进行其他分析。	由特定成员任意
仅允许对表进行特定分析。	自定义列表

7. 对于结果交付，请在允许接收查询输出结果的成员下拉列表中指定可以接收结果的成员。
8. 选择配置分析规则。

配置差别隐私策略（可选）

> **Note**
>
> AWS Clean Rooms 差异隐私仅适用于数据存储在 Amazon S3 中的协作。

此过程描述了使用 AWS Clean Rooms 控制台中的 “引导流程” 选项在协作中配置差异隐私策略的过程。对于所有具有差别隐私保护的表来说，这是一次性步骤。

配置差别隐私设置（引导流程）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。

3. 选择协作。
4. 在协作页面的表选项卡上，选择配置差别隐私策略。
5. 在配置差别隐私策略页面上，选择以下属性的值：

   - 隐私预算
   - 每月刷新隐私预算
   - 每个查询添加的噪声

   您可以使用默认值，或输入支持您的特定使用案例的自定义值。在选择隐私预算和每个查询添加的噪声值后，您可以根据数据的所有查询中可能进行的聚合数量预览产生的效用。
6. 选择配置。

您将看到一条确认消息，指出您成功为协作配置了差别隐私策略。

您现在配置了差别隐私，您已准备好：

- 查询数据表（以可以查询的成员身份）
- 协作（如果您是协作创建者）

查看差别隐私使用情况日志

作为使用差别隐私保护数据的协作成员，在创建具有差别隐私的协作后，您可以监控隐私预算的使用情况。

查看运行了多少聚合以及使用了多少隐私预算

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择查看使用情况日志（蓝色文本）。
6. 查看使用情况详细信息，包括隐私预算和提供了多少效用。

编辑差别隐私策略

在配置差别隐私策略后，您可以随时更新该策略以更好地反映您的隐私需求。

编辑差别隐私策略

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在协作页面的表选项卡上，在由您关联的表下面选择编辑。
5. 在编辑差别隐私页面上，为以下属性选择新的值：

   - 隐私预算 - 移动滑块以在协作期间随时增加或减少预算。在可以查询的成员开始查询您的数据后，您无法减少预算。如果增加隐私预算， AWS Clean Rooms 将继续使用现有预算，直到其用完为止，然后再使用新增加的隐私预算。
   - 每个查询添加的噪声 - 移动滑块以在协作期间随时增加或减少每个查询添加的噪声。

   > **Note**
   >
   > 您可以选择交互式示例以了解隐私预算和每个查询添加的噪声的不同值如何影响您可以运行的聚合函数数量。

   您无法更改隐私预算刷新的值。要更改您选择的值，您必须删除差别隐私策略并创建一个新策略。
6. 选择保存更改。

您会看到一条确认消息，指出您已成功编辑差别隐私策略。

删除差别隐私策略

您可以从协作的表选项卡中删除差别隐私策略。

删除差别隐私策略

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在协作页面的表选项卡上的差别隐私策略旁边，选择删除。
5. 如果您确定要删除差别隐私策略，请选择删除。

在删除差别隐私策略后，您无法访问该策略的隐私预算使用情况日志。如果删除了差别隐私策略，将无法查询开启了差别隐私的表。

查看计算的差别隐私参数

对于具有差异隐私专业知识的用户，您可以从协作的 “分析” 选项卡中查看计算出的差异隐私参数。

查看计算的差别隐私参数

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在 “分析” 选项卡的 “结果” 部分中，选择 “查看计算得出的差异隐私参数”。

在计算得出的差别隐私参数表中，您可以看到聚合函数的灵敏度值，该值定义为在添加、删除或修改单个用户的记录时函数结果可能发生的最大变化量。该列表包括以下差别隐私参数：

- 用户贡献限制 (UCL) 是用户在 SQL 查询中贡献的最大行数。例如，如果您想计算指定广告系列中每位用户可以获得多次曝光的匹配曝光总数，则 AWS Clean Rooms 差异隐私需要绑定单个用户的曝光量，以确保差异隐私计算的准确性。换句话说，如果任何用户的曝光量超过了界限，则 AWS Clean Rooms 会根据计算出的 UCL 值自动随机抽取该用户的展示次数的统一随机样本，并在执行查询时排除该用户的剩余展示次数。如果您计算唯一用户数，则 UCL 值等于 1。这是因为添加、删除或修改单个用户最多可以将不同用户的计数更改 1。
- 最小值是聚合函数（例如 `sum()`）中使用的表达式的下限。例如，如果表达式是名为 `purchase_value` 的列，则最小值是该列的下限。
- 最大值是聚合函数（例如 `sum()`）中使用的表达式的上限。例如，如果表达式是名为 `purchase_value` 的列，则最大值是该列的上限。

在计算得出的差别隐私参数表中，您可以使用这些参数更好地了解查询结果中的总噪声量。例如，当配置的每个查询添加的噪音为 30 个用户并且正在运行`COUNT DISTINCT (user_id)`查询时， AWS Clean Rooms 差异隐私会添加介于 -30 和 30 之间的随机噪声，且可能性很高，因为的灵敏度`COUNT DISTINCT`为 1。对于具有相同配置的 `COUNT` 查询， AWS Clean Rooms Differential Privacy 添加按用户贡献限制扩展的统计噪声，因为单个用户可能为查询结果贡献多个行。对于像所有列值均为正值这样的`SUMSUM (purchase_value)`查询，总噪音按用户贡献限制乘以最大值进行缩放。 AWS Clean Rooms Difersial Privacy 会自动计算灵敏度参数以在查询运行时执行噪声加法，从而耗尽隐私预算。由于灵敏度参数依赖于数据，因此，需要耗尽隐私预算。

查看表格和分析规则

查看与协作和分析规则关联的表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择 Tables (表) 选项卡。
5. 选择下列选项之一：
   a. 要查看协作中关联的表，请针对由您关联的表，选择一个表（蓝色文本）。
   b. 要查看协作中关联的其他表，请针对由协作者关联的表，选择一个表（蓝色文本）。
6. 在表详细信息页面查看表的详细信息和分析规则。

编辑已配置的表

先决条件：

- 可以 AWS 账户 访问的 AWS Clean Rooms

以下各节说明如何编辑 Amazon S3、Amazon Athena 和 Snowflake 数据源的表的名称、描述和配置细节。

有关如何使用编辑已配置表的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

编辑已配置的表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择您创建的配置表。
4. 在已配置的表格详细信息页面上，选择编辑。
5. 编辑您的配置。
6. 选择保存更改。

编辑配置表标签

作为协作成员，在创建已配置表后，您可以在配置表选项卡上管理配置表资源上的标签。

编辑配置表标签

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择您创建的配置表。
4. 在配置表详细信息页面上，向下滚动到标签部分。
5. 选择管理标签。
6. 在管理标签页面上，可以执行以下操作：
   - 要删除标签，请选择移除。
   - 要添加标签，请选择添加新标签。
   - 要保存您的更改，请选择保存更改。

编辑配置的表分析规则

编辑配置表的分析规则

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

2. 在左侧导航窗格中，选择 表。
3. 选择您创建的配置表。
4. 在配置表详细信息页面上，向下滚动到聚合分析规则、列表分析规则或自定义分析规则部分。（您的选择取决于您为配置表选择的分析规则类型。）
5. 选择编辑。
6. 在编辑分析规则页面上，您可以：
   - 通过以下方式修改分析规则定义：
     - 修改 JSON 编辑器。
     - 选择从文件导入以上传新的分析规则定义。
   - 从以下选项中进行选择，预览成员将在协作中看到的内容：
     - 表视图
     - JSON
     - 查询示例
7. 选择保存更改以保存您的更改。

删除已配置的表分析规则

> **Warning**
>
> 此操作无法撤消，并且会影响所有相关资源。

删除配置表分析规则

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 选择您创建的配置表。
4. 在配置表详细信息页面上，向下滚动到聚合分析规则、列表分析规则或自定义分析规则部分。（您的选择取决于您为配置表选择的分析规则类型。）
5. 选择删除。
6. 如果您确定要删除分析规则，请选择删除。

配置表不允许的列

不允许的输出列配置是 AWS Clean Rooms 自定义分析规则中的一个控件，它允许您定义不允许在查询结果中投影的列列表（如果有）。此列表中引用的列被视为“不允许的输出列”。这意味着通过转换、别名或其他方式对此类列的任何引用都可能不会出现在查询的最终 SELECT（投影）中。

虽然该功能禁止在输出中直接投影列，但它并不会完全阻止通过其他机制间接推断出基础值。这些列仍然可以在投影子句（例如子查询或公用表表达式 (CTE)）中使用，前提是它们在最终投影中没有被引用。

不允许的输出列配置使您可以灵活地在表上应用和编纂控制，并根据使用案例和相应的隐私要求执行分析模板级审查。

有关如何设置此配置的更多信息，请参阅为表添加自定义分析规则（引导流程）。

示例

以下示例显示了如何应用不允许的输出列控制。

- 成员 A 与成员 B 协作。
- 成员 B 是可以运行查询的成员。
- 成员 A 使用 age、gender、email 和 name 列定义表 users。age 和 name 列是不允许的输出列。
- 成员 B 用一组相似的列来定义表 pets：age、gender 和 owner_name。但是，其没有对输出列设置任何限制，这就表示可以在查询中自由投影表中的所有列。

如果成员 B 运行以下查询，则会被阻止，因为无法直接投影不允许的输出列：

```
SELECT
  age
FROM
  users
```

如果成员 B 运行以下查询，则会被阻止，因为无法通过投影星号隐式投影不允许的输出列：

```
SELECT
  *
FROM
  users
```

如果成员 B 运行以下查询，则会被阻止，因为无法投影不允许的输出列的转换：

```
SELECT
  COUNT(age)
FROM
  users
```

如果成员 B 运行以下查询，则会被阻止，因为无法使用别名在最终投影中引用不允许的输出列：

```
SELECT
  count_age
FROM
  (SELECT COUNT(age) AS count_age FROM users)
```

如果成员 B 运行以下查询，则会被阻止，因为在输出中投影了转换的受限制列：

```
SELECT
  CONCAT(name, email)
FROM
  users
```

如果成员 B 运行以下查询，则会被阻止，因为无法在最终投影中引用 CTE 中定义的不允许的输出列：

```
WITH cte AS (
  SELECT
    age AS age_alias
  FROM
    users
)
SELECT age_alias FROM cte
```

如果成员 B 运行以下查询，则会被阻止，因为在最终投影中无法将不允许的输出列用作排序键或分区键：

```
SELECT
```

```
  LISTAGG(gender) WITHIN GROUP (ORDER BY age) OVER (PARTITION BY age)
FROM
  users
```

如果成员 B 运行以下查询，则会成功，因为属于不允许的输出列的列仍然可以在查询中的其他构造中使用，例如在联接或筛选子句中。

```
SELECT
  u.name,
  p.gender,
  p.age
FROM
  users AS u
JOIN
  pets AS p
ON
  u.name = p.owner_name
```

在同一场景中，成员 B 还可以使用 users 中的 name 列作为筛选器或排序键：

```
SELECT
  u.email,
  u.gender
FROM
  users AS u
WHERE
  u.name = 'Mike'
ORDER BY
  u.name
```

此外，用户不允许的输出列可用于中间投影，例如子查询和 CTEs，例如：

```
WTIH cte AS (
 SELECT
   u.gender,
   u.id,
   u.first_name
 FROM
```

```
   users AS u
)
SELECT
  first_name
FROM
  (SELECT cte.gender, cte.id, cte.first_name FROM cte)
```

编辑配置表关联

作为协作成员，您可以编辑已创建的已配置表关联。

编辑配置表关联

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择表选项卡。
5. 对于由您关联的表，请选择一个表。
6. 在表详细信息页面上，向下滚动以查看表关联详细信息。
7. 选择编辑。
8. 在编辑已配置的表关联页面上，更新描述或服务访问信息。
9. 选择保存更改。

取消关联已配置的表

作为协作成员，您可以取消已配置的表与协作的关联。此操作可阻止可以查询的成员查询表。

取消关联配置表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择表选项卡。

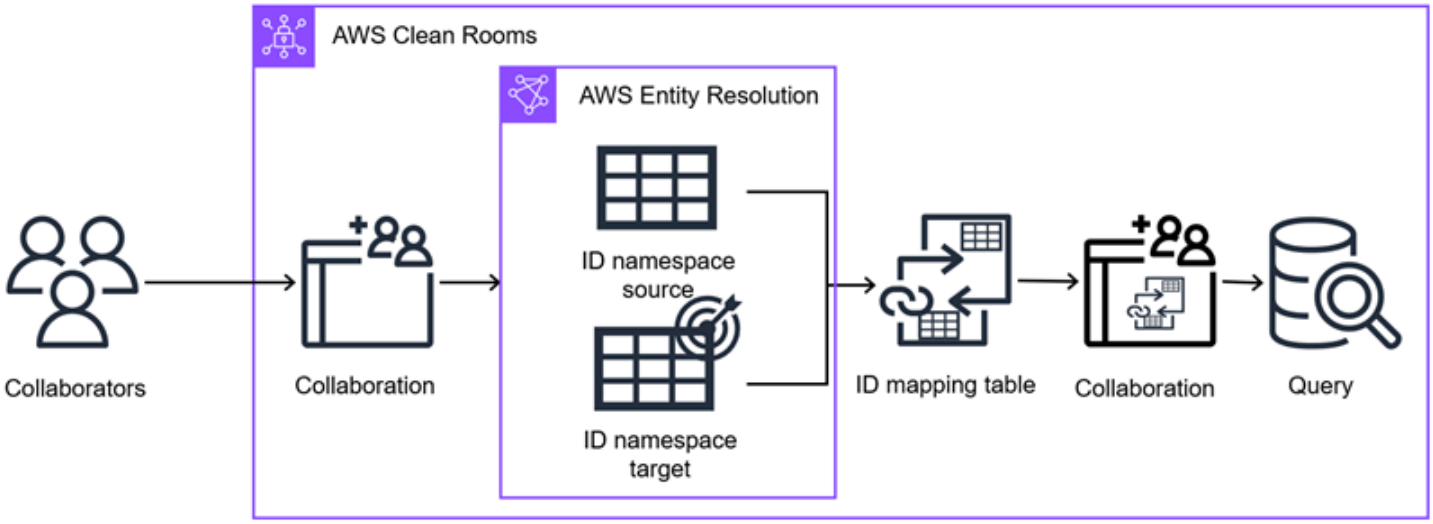
5. 对于由您关联的表，选择要取消关联的表旁边的选项按钮。
6. 选择取消关联。
7. 在对话框中，确认取消关联配置表的决定，并通过选择取消关联来阻止可以查询的成员查询该表。

AWS Entity Resolution 数据匹配服务 in AWS Clean Rooms

使用 AWS Entity Resolution 数据匹配服务 中 AWS Clean Rooms，您可以将数据从源转换为目标，使用转换后的数据填充 ID 映射表，并查询数据。

首先，您可以在中创建协作 AWS Clean Rooms 并添加要邀请的人，或者通过创建成员资格来加入受邀加入的协作。 AWS 账户 接下来，对两个数据表执行 ID 映射。为此，您可以关联现有 ID 命名空间源或在 AWS Entity Resolution 数据匹配服务创建新的 ID 命名空间源。协作的另一个成员关联现有 ID 命名空间目标或创建新的 ID 命名空间目标。然后，您根据两个关联的 ID 命名空间创建并填充一个 ID 映射表。最后，可以查询的成员通过联接 ID 映射表，在两个数据表中运行查询。

下图总结了如何在 AWS Entity Resolution 数据匹配服务 中使用 AWS Clean Rooms。



> **Note**
>
> 目前支持的转码服务提供商是 LiveRamp，可在以下地区使用 AWS 区域：美国东部（弗吉尼亚北部）、美国东部（俄亥俄州）和美国西部（俄勒冈）。

主题

- ID 中的命名空间 AWS Clean Rooms
- 中的 ID 映射表 AWS Clean Rooms

ID 中的命名空间 AWS Clean Rooms

ID 命名空间是一个围绕身份表的包装程序，它使您能够提供元数据，解释您的数据集以及如何在 ID 映射工作流程中使用该数据集。ID 映射工作流程是一种数据处理作业，它根据指定的 ID 映射方法将数据从输入数据源映射到输入数据目标。它会生成一个 ID 映射表。

ID 命名空间有两种类型：源和目标。源包含将在 ID 映射工作流程中处理的源数据的配置。目标包含所有源都将解析到其中的目标数据的配置。要定义要跨两个集合解析的输入数据 AWS 账户，请创建一个 ID 命名空间源和一个 ID 命名空间目标，以将数据从一个集（源）转换为另一个集（目标）。

您可以创建新的 ID 命名空间，也可以关联现有命名空间。有关如何在中创建 ID 命名空间的更多信息 AWS Entity Resolution 数据匹配服务，请参阅AWS Entity Resolution 数据匹配服务 用户指南中的使用 ID 命名空间定义输入数据。

主题

创建并关联新的 ID 命名空间

在创建 ID 映射表以查询身份数据之前，每个协作成员都必须创建并关联 ID 命名空间源或 ID 命名空间目标。

如果您已经在中创建了 ID 命名空间 AWS Entity Resolution 数据匹配服务，请跳至关联现有 ID 命名空间。

创建和关联新的 ID 命名空间

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在实体解析选项卡上，选择关联 ID 命名空间。
5. 在关联 ID 命名空间页面上，对于实体解析数据，请选择创建 ID 命名空间。

AWS Entity Resolution 数据匹配服务 控制台出现在新选项卡中。

6. 按照 AWS Entity Resolution 数据匹配服务 控制台中创建 ID 命名空间页面上的提示进行操作。

   a. 在详细信息中，输入 ID 命名空间名称、描述，然后选择 ID 命名空间类型（源或目标）。

   b. 对于 ID 命名空间方法，如果要进行基于规则的匹配，请选择基于规则的方法，或者如果要进行第三方转码，请选择提供商服务。

   c. 根据您选择的 ID 命名空间方法，指定数据输入类型。

   d. 选择创建 ID 命名空间。

7. 返回 AWS Clean Rooms 控制台。

8. 在关联 ID 命名空间页面上，对于实体解析数据，从下拉列表中选择要与协作关联的 AWS Entity Resolution 数据匹配服务 ID 命名空间源或目标。

9. 对于关联详细信息，请执行以下步骤。

   a. 输入关联的 ID 命名空间的名称。

      您可以使用默认名称或重命名此 ID 命名空间。

   b. （可选）输入 ID 命名空间的描述。

      该描述有助于编写查询。

10. 通过选择一个选项并采取建议的操作来指定 AWS Clean Rooms 访问 的权限。

Option	推荐操作
AWS Clean Rooms 允许添加和管理权限策略	AWS Clean Rooms 使用该关联所需的策略创建服务角色。
手动添加和管理权限	请执行以下操作之一： • 查看资源策略并向该策略添加必要的权限。 • 通过选择添加策略声明来使用现有策略。 您必须拥有修改角色并创建策略的权限。

Option	推荐操作
	Note 如果您无法修改角色策略，则会收到一条错误消息，指出 AWS Clean Rooms 找不到该服务角色的策略。

11. （可选）对于高级 ID 映射表配置，请修改来自 ID 命名空间的列的默认保护。

    默认情况下，ID 映射表配置为仅允许对 `sourceID` 列和 `targetID` 列执行 `INNER JOIN`。您可以修改此配置，以便可以在查询中的任何位置允许来自此 ID 命名空间（`sourceID` 或 `targetID`）的列。

您的目标	建议的选项
将该列归类为“联接列”，并且仅允许在 `INNER JOIN` 子句中使用该列	是
将该列归类为“维度列”，并允许它出现在查询中的任何位置，包括查询的 `JOIN` 子句、`SELECT`、`WHERE` 和 `GROUP BY` 语句。	不，允许出现在查询中的任何位置

12. （可选）如果要为 ID 命名空间资源启用标签，请选择添加新标签，然后输入键和值对。
13. 选择关联 。
14. 在实体解析选项卡上的关联 ID 命名空间表下，查看关联的 ID 命名空间并验证 ID 命名空间类型是否正确（源或目标）。

协作中的所有成员都关联其 ID 命名空间后，您可以创建 ID 映射表并查询数据。

关联现有 ID 命名空间

在此步骤中，每个成员在协作中关联其现有 ID 命名空间源或 ID 命名空间目标。

关联现有 ID 命名空间

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在实体解析选项卡上，选择关联 ID 命名空间。
5. 在关联 ID 命名空间页面上，对于实体解析数据，从下拉列表中选择要与协作关联的 AWS Entity Resolution 数据匹配服务 ID 命名空间源或目标。
6. 对于关联详细信息，请执行以下步骤。

   a. 输入关联的 ID 命名空间的名称。

      您可以使用默认名称或重命名此 ID 命名空间。

   b. （可选）输入 ID 命名空间的描述。

      该描述有助于编写查询。
7. 通过选择一个选项并采取建议的操作来指定 AWS Clean Rooms 访问 的权限。

Option	推荐操作
AWS Clean Rooms 允许添加和管理权限策略	AWS Clean Rooms 使用该关联所需的策略创建服务角色。
手动添加和管理权限	请执行以下操作之一： • 查看资源策略并向该策略添加必要的权限。 • 通过选择添加策略声明来使用现有策略。 您必须拥有修改角色并创建策略的权限。 **Note** 如果您无法修改角色策略，则会收到一条错误消息，指出 AWS Clean Rooms 找不到该服务角色的策略。

8. （可选）对于高级 ID 映射表配置，请修改来自 ID 命名空间的列的默认保护。

   默认情况下，ID 映射表配置为仅允许对 sourceID 列和 targetID 列执行 INNER JOIN。您可以修改此配置，以便可以在查询中的任何位置允许来自此 ID 命名空间（sourceID 或 targetID）的列。

您的目标	建议的选项
将该列归类为“联接列”，并且仅允许在 INNER JOIN 子句中使用该列。	是
将该列归类为“维度列”，并允许它出现在查询中的任何位置，包括查询的 JOIN 子句、SELECT、WHERE 和 GROUP BY 语句。	不，允许出现在查询中的任何位置

9. （可选）如果要为 ID 命名空间资源启用标签，请选择添加新标签，然后输入键和值对。
10. 选择关联。
11. 在实体解析选项卡上的关联 ID 命名空间表下，查看关联的 ID 命名空间并验证 ID 命名空间类型是否正确（源或目标）。

协作中的所有成员都关联其 ID 命名空间后，您可以创建 ID 映射表并查询数据。

编辑 ID 命名空间关联

作为协作成员，您可以编辑已创建的 ID 命名空间关联。

编辑 ID 命名空间关联

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择实体解析选项卡。
5. 对于关联的 ID 命名空间，选择 ID 命名空间。
6. 在 ID 命名空间详细信息页面上，向下滚动以查看ID 命名空间关联详细信息。
7. 选择编辑。

8. 在编辑 ID 命名空间关联页面上，编辑以下任何内容：

   a. 对于关联详细信息，请更新名称或描述。

   b. （可选）对于高级 ID 映射表配置，请修改来自 ID 命名空间的列的默认保护。

      默认情况下，ID 映射表配置为仅允许对 sourceID 列和 targetID 列执行 INNER JOIN。您可以修改此配置，以便可以在查询中的任何位置允许来自此 ID 命名空间（sourceID 或 targetID）的列。

您的目标	建议的选项
将该列归类为“联接列”，并且仅允许在 INNER JOIN 子句中使用该列	是
将该列归类为“维度列”，并允许它出现在查询中的任何位置，包括查询的 JOIN 子句、SELECT、WHERE 和 GROUP BY 语句。	不，允许出现在查询中的任何位置

9. 选择保存更改。

取消 ID 命名空间关联

作为协作成员，您可以取消 ID 命名空间与协作的关联。此操作可阻止可以查询的成员查询表。

> **Warning**
>
> 取消 ID 命名空间与协作的关联会删除派生 ID 映射表中的任何数据，使其不可查询。例如，如果您的 ID 命名空间关联被用作三个不同的 ID 映射表中的 SOURCE，那么当您取消关联您的 ID 命名空间关联时，这些 ID 映射表中的所有数据都将被删除。

取消关联 ID 命名空间关联

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在左侧导航窗格中，选择协作。

3. 选择协作。
4. 选择实体解析选项卡。
5. 对于关联的 ID 命名空间，选择要取消关联的 ID 命名空间旁边的选项按钮。
6. 选择取消关联。
7. 在对话框中，选择取消关联，确认您决定断开与 ID 命名空间的连接。此操作可阻止任何可以查询的成员访问 ID 映射表。

   如果某个协作成员移除了其中一个 ID 命名空间，则如果源已离开协作，您将无法重新填充 ID 映射表。

   尽管 ID 映射表之前已填充，但解除 ID 命名空间的关联意味着您无法再对该表运行查询。

中的 ID 映射表 AWS Clean Rooms

ID 映射表是一种在协作中 AWS Clean Rooms 启用多方身份映射的资源。

在创建 ID 映射表之前，必须先将源数据和目标数据配置为 ID 命名空间。

创建 ID 映射表后，您可以使用 ID 映射工作流程将源 ID 命名空间转换为目标 ID 命名空间。您可以使用基于规则的方法或提供商服务转码方法来执行此操作。

ID 映射工作流程是一种数据处理作业，它根据指定的 ID 映射工作流程方法将数据从输入数据源映射到输入数据目标。此工作流程将填充 ID 映射表。

> **Note**
>
> 只能根据存储在 Amazon S3 中并爬到 AWS Glue 表中的数据集创建 ID 映射表。

有两种 ID 映射工作流程方法：基于规则的 ID 映射或提供商服务 ID 映射：

- 基于规则的 ID 映射 - 您可以使用匹配规则将第一方数据从源转换为目标。
- 提供商服务 ID 映射-您可以使用 LiveRamp 提供者服务将第三方数据从源转换为目标。

> **Note**
>
> 当前支持的转码服务提供商是 LiveRamp。协作中任何 LiveRamp 通过订阅的成员 AWS Data Exchange 都可以创建 ID 映射表。如果您已经订阅 LiveRamp了但尚未订阅 AWS

Data Exchange，请联系 LiveRamp 以获取私人优惠。有关更多信息，请参阅《AWS Entity Resolution 数据匹配服务 用户指南》中的在 AWS Data Exchange上订阅提供商服务。

主题

- 创建并填充新的 ID 映射表
- 填充现有 ID 映射表
- 编辑 ID 映射表
- 删除 ID 映射表

创建并填充新的 ID 映射表

先决条件

在创建 ID 映射表之前，请确保：

- 关联的 ID 命名空间源和目标
- 为基于规则的 ID 映射或提供商服务 ID 映射配置的 ID 命名空间

您可以创建两种类型的 ID 映射表：

- 基于规则-使用匹配规则翻译第一方数据
- 提供者服务 — LiveRamp 用于翻译 Ramp IDs

创建 ID 映射表后，您可以通过运行 ID 映射工作流程立即填充该表，也可以等待稍后再填充该表。

成功填充 ID 映射表后，您可以在 ID 映射表上运行多表联接查询，将 `sourceId` 与 `targetId` 联接并分析数据。

主题

- 创建 ID 映射表（基于规则）
- 创建 ID 映射表（提供商服务）

创建 ID 映射表（基于规则）

本主题介绍创建使用匹配规则将第一方数据从源转换为目标 ID 映射表的过。

创建基于规则的 ID 映射表时，您可以通过启用增量处理来选择仅处理工作流中的新记录、更新记录或已删除记录。

使用基于规则的方法创建和填充新的 ID 映射表

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在导航窗格中，选择协作。
3. 选择协作，然后选择实体分辨率选项卡。
4. 选择创建 ID 映射表。
5. 在 ID 映射设置下，执行以下操作之一：
   - 要创建新工作流程，请将创建新 ID 映射工作流程保持选中状态。
   - 要使用现有工作流程，请清除该复选框并从下拉列表中选择一个工作流程，然后跳至步骤 9。
6. 在 “身份数据” 下，查看或配置源和目标。
   - 对于单个 ID 命名空间对：查看预先选择的源和目标。
   - 对于多个 ID 命名空间：从下拉列表中选择源和目标。
7. （可选）选择 “启用增量处理”，仅处理工作流中新的、已更新或已删除的记录。

   AWS Entity Resolution 数据匹配服务 仅处理源 ID 或目标 ID 命名空间中的新记录、已更新记录或已删除记录，而不是重新创建整个 ID 映射表。

   如果未选中此选项，则会在 ID 映射表上 AWS Entity Resolution 数据匹配服务 运行默认的批处理 ID 映射工作流。
8. 在规则参数下，配置以下内容：
   - 规则控件-选择目标还是源提供匹配的规则。

     您可以通过打开显示规则来查看规则。

     规则控制在源 ID 命名空间和目标 ID 命名空间之间必须兼容，才能在 ID 映射工作流程中使用。例如，如果源 ID 命名空间将规则限制于目标，但目标 ID 命名空间将规则限制于源，则会导致错误。

- 比较类型会自动设置为多个输入字段。

  这是因为两个参与者之前都选择了此选项。
- 记录匹配
  - 一个源到一个目标 — 为每个目标存储一条匹配的记录
  - 多个源指向一个目标 — 存储每个目标的所有匹配记录

> **Note**
>
> 为源 ID 命名空间和目标 ID 命名空间指定的限制必须兼容。

9. 有关 ID 映射的详细信息，请配置以下内容：

   a. 输入 ID 映射表名称或保留默认名称。

   b. （可选）输入 ID 映射表的描述。

      该描述有助于编写查询。

10. 要AWS Clean Rooms 访问权限，请选择一个：

- AWS Clean Rooms 允许添加和管理权限策略-自动创建服务角色。
- 手动添加和管理权限-查看和修改资源策略或选择添加策略声明。

> **Note**
>
> 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

11. 要AWS Entity Resolution 数据匹配服务 访问权限，请选择一个：

    此部分仅在创建新 ID 映射表时可见。

- 创建并使用新的服务角色
  - 默认服务角色名称为 `entityresolution-id-mapping-workflow-<timestamp>`
  - （可选）对于加密数据，请选择 “此数据由 KMS 密钥加密”，然后输入AWS KMS 密钥。
- 使用现有服务角色
  - 从下拉列表中选择现有服务角色名称或输入角色 ARN。

    如果您有列出角色的权限，则会显示角色列表。

通过选择在 IAM 中查看外部链接来查看服务角色。

如果没有现有的服务角色，则使用现有服务角色选项不可用。

默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

- （可选）选择向该角色添加具有必要权限的预配置策略，以便为该角色附加必要的权限。

您必须拥有修改角色并创建策略的权限。

12. （可选）在“其他设置”下，配置：

a. ID 映射表设置

- 要启用自定义加密，请选择自定义加密设置并输入密 AWS KMS 钥。

> **Note**
>
> 此 KMS 密钥需要授予在中使用所需的权限 AWS Entity Resolution 数据匹配服务才能`cleanrooms.amazonaws.com`使用 KMS 密钥策略。有关使用 ID 映射工作流程进行加密所需的权限的更多详细信息，请参阅《AWS Entity Resolution 数据匹配服务 用户指南》中的为 AWS Entity Resolution 数据匹配服务创建工作流程作业角色。

- 要添加标签，请选择添加新标签并输入键值对

b. ID 映射工作流程设置（仅限新工作流程）：

- 要使用不同的名称，请清除“保持相同的 ID 映射表名称和描述”，然后输入新值。
- 要添加标签，请选择添加新标签并输入键值对

13. 选择下列选项之一：

- 创建 ID 映射表-创建可稍后填充的空表 () 填充现有 ID 映射表
- 创建并填充 ID 映射表-创建并立即填充表（可能需要几个小时）

ID 映射工作流程过程开始。在此过程中，ID 映射表中填充了已翻译的内容 IDs。ID 映射表工作流程可能需要几个小时才能完成整个过程。

成功填充 ID 映射表后，您可以查询 ID 映射表以将 `sourceId` 与 `targetId` 联接并分析数据。

创建 ID 映射表（提供商服务）

本主题介绍使用提供者服务 (LiveRamp) 创建 ID 映射表的过程。 LiveRamp 提供者服务使用维护或派生的 Ramp 将一组源 Ramp IDs 转换为另一组源 R IDs amp。

使用提供商服务方法创建新的 ID 映射表

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在导航窗格中，选择协作。
3. 选择协作，然后选择实体分辨率选项卡。
4. 选择创建 ID 映射表。
5. 在 ID 映射设置下，执行以下操作之一：
   - 要创建新工作流程，请将创建新 ID 映射工作流程保持选中状态。
   - 要使用现有工作流程，请清除该复选框并从下拉列表中选择一个工作流程，然后跳至步骤 9。
6. 在“身份数据”下，查看或配置源和目标。
   - 对于单个 ID 命名空间对：查看预先选择的源和目标。
   - 对于多个 ID 命名空间：从下拉列表中选择源和目标。
7. 在“方法”下，验证所选的 ID 映射工作流方法是否正在LiveRamp 转码。
8. 要进行LiveRamp 配置，请执行以下任一操作：
   - 输入 LiveRamp ID 管理器 ARN 和LiveRamp 密钥管理器 AR N。
   - 选择从现有工作流程导入。
9. 有关 ID 映射的详细信息，请配置以下内容：
   a. 输入 ID 映射表名称或保留默认名称。
   b. （可选）输入 ID 映射表的描述。

      该描述有助于编写查询。
10. 要AWS Clean Rooms 访问权限，请选择一个：
    - AWS Clean Rooms 允许添加和管理权限策略-自动创建服务角色。
    - 手动添加和管理权限-查看和修改资源策略或选择添加策略声明。

> **Note**
>
> 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

11. 要AWS Entity Resolution 数据匹配服务 访问权限，请选择一个：

    此部分仅在创建新 ID 映射表时可见。

    - 创建并使用新的服务角色
      - 默认服务角色名称为 `entityresolution-id-mapping-workflow-<timestamp>`
      - （可选）对于加密数据，请选择 “此数据由 KMS 密钥加密”，然后输入AWS KMS 密钥。
    - 使用现有服务角色
      - 从下拉列表中选择现有服务角色名称或输入角色 ARN。

        如果您有列出角色的权限，则会显示角色列表。

        通过选择在 IAM 中查看外部链接来查看服务角色。

        如果没有现有的服务角色，则使用现有服务角色选项不可用。

        默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。
      - （可选）选择向该角色添加具有必要权限的预配置策略，以便为该角色附加必要的权限。

        您必须拥有修改角色并创建策略的权限。

12. （可选）在 “其他设置” 下配置：

    a. ID 映射表设置

       - 要启用自定义加密，请选择自定义加密设置并输入密 AWS KMS 钥。

       > **Note**
       >
       > 此 KMS 密钥需要授予在中使用所需的权限 AWS Entity Resolution 数据匹配服务才能`cleanrooms.amazonaws.com`使用 KMS 密钥策略。有关使用 ID 映射工作流程进行加密所需的权限的更多详细信息，请参阅《AWS Entity Resolution 数据匹

配服务 用户指南》中的为 AWS Entity Resolution 数据匹配服务创建工作流程作业角色。

- 要添加标签，请选择添加新标签并输入键值对。

b. ID 映射工作流程设置（仅限新工作流程）：

- 要使用不同的名称，请清除 “保持相同的 ID 映射表名称和描述”，然后输入新值。
- 要添加标签，请选择添加新标签并输入键值对。

13. 选择下列选项之一：

- 创建 ID 映射表-创建可稍后填充的空表 () 填充现有 ID 映射表
- 创建并填充 ID 映射表-创建并立即填充表（可能需要几个小时）

ID 映射工作流程过程开始。在此过程中，ID 映射表中填充了已翻译的内容 IDs。ID 映射表工作流程可能需要几个小时才能完成整个过程。

成功填充 ID 映射表后，您可以查询 ID 映射表以将 `sourceId` 与 `targetId` 联接并分析数据。

填充现有 ID 映射表

向 ID 命名空间添加新数据时，请使用此工作流程。如果您在创建 ID 映射表时选择启用增量处理，则只能处理源 ID 或目标 ID 命名空间中的新记录、更新记录或已删除记录，而不必重新创建整个 ID 映射表。

填充现有 ID 映射表

1. 登录 AWS 管理控制台 并打开 AWS Clean Rooms 控制台，网址为https://console.aws.amazon.com/cleanrooms/。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 转到 “实体解析” 选项卡。
5. 在 ID 映射表部分下，选择一个 ID 映射表。
6. 如果您尚未为此 ID 映射表启用增量处理，请选择一个 ID 映射表，然后选择填充。
7. 如果您已为此 ID 映射表启用增量处理，请选择 Pop ulate w ith，然后选择以下选项之一：

- 增量处理-仅处理源 ID 或目标 ID 命名空间中的新记录、更新记录或已删除记录。

建议用于频繁更新、每日运行或实时数据同步。

- Batch 处理-处理整个 ID 映射表。

  建议在初始设置、定期进行完全刷新或源和目标 ID 命名空间发生重大更改时使用。

- 仅删除处理-仅处理从源 ID 命名空间中删除的记录，并相应地更新目标 ID 命名空间。

  建议用于快速同步移除操作。

8. ID 映射工作流程过程开始。

   在此过程中，会在 ID 映射表中填充转码后的内容 IDs。ID 映射表工作流程可能需要几个小时才能完成整个过程。

成功填充 ID 映射表后，您可以查询 ID 映射表以将 `sourceId` 与 `targetId` 联接。

编辑 ID 映射表

作为协作成员，您可以编辑已创建的 ID 映射表。

编辑 ID 映射表

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择实体解析选项卡。
5. 对于 ID 映射表，请选择一个表。
6. 在 ID 映射表详细信息页面上，向下滚动以查看 ID 映射表详细信息。
7. 选择编辑。
8. 在编辑 ID 映射表页面上，更新描述或服务访问信息。
9. 选择保存更改。

删除 ID 映射表

作为协作成员，您可以删除已创建的 ID 映射表。此操作可阻止可以查询的成员查询表。

> **Warning**
>
> 删除映射表会永久删除所有已填充的数据。

删除 ID 映射表

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 选择实体解析选项卡。
5. 对于 ID 映射表，请选择一个表。
6. 在 ID 映射表详细信息页面上，向下滚动以查看 ID 映射表。
7. 选择一个 ID 映射表，然后选择删除。
8. 如果您确定要删除 ID 映射表，请选择删除。

中的分析模板 AWS Clean Rooms

分析模板可与 中的自定义分析规则 AWS Clean Rooms 配合使用。使用分析模板，您可以定义参数来帮助您重复使用相同的查询。 AWS Clean Rooms 支持带有字面值的参数化子集。

分析模板针对协作。对于每个协作，成员只能看到该协作中的查询。如果您计划在协作中使用差别隐私，应确保您的分析模板与 AWS Clean Rooms Differential Privacy 的通用查询结构兼容。

您可以通过两种方式创建分析模板：使用 SQL 代码或使用适用于 Spark 的 Python 代码。

主题

SQL 分析模板

SQL 分析模板使您能够在协作中跨不同数据集查询和分析数据。您可以使用这些模板执行各种类型的分析，例如识别受众重叠和计算汇总指标。

使用 SQL 分析模板，您可以：

- 编写标准 SQL 查询
- 添加参数以使您的查询动态化
- 控制对特定列和表的访问权限
- 为敏感数据设置聚合要求
- 为自定义机器学习 (ML) 模型定义用于生成隐私增强型合成数据集的输入数据

主题

创建 SQL 分析模板

先决条件

在创建 SQL 分析模板之前，您必须具备以下条件：

- 积极 AWS Clean Rooms 合作
- 访问协作中至少一个已配置的表

  有关在中配置表的信息 AWS Clean Rooms，请参见在中创建配置表 AWS Clean Rooms。
- 创建分析模板的权限
- SQL 查询语法的基础知识

以下过程描述了使用AWS Clean Rooms 控制台创建 SQL 分析模板的过程。

有关如何使用创建 SQL 分析模板的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

创建 SQL 分析模板

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台，该控制台将充当协作创建者。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在模板选项卡上，转到您创建的分析模板部分。
5. 选择创建分析模板。
6. 在 “创建分析模板” 页面上，对于详细信息，
   a. 输入分析模板的名称。
   b. （可选）输入描述。
   c. 对于 “格式”，将 SQL 选项保留为选中状态。
7. 对于表，查看与协作关联的配置表。
8. 对于定义，
   a. 输入分析模板的定义。
   b. 选择导入自以导入定义。
   c. （可选）在 SQL 编辑器中通过在参数名称前输入冒号 (:) 来指定参数。

      例如：

      ```
      WHERE table1.date + :date_period > table1.date
      ```

9. 如果您之前添加了参数，请在参数 - 可选下，为每个参数名称选择类型和默认值（可选）。
10. 对于合成数据，如果要生成用于模型训练的合成数据，请选中 “要求分析模板输出为合成数据” 复选框。

    有关更多信息，请参阅隐私增强型合成数据集生成。

    a. 对于列分类，请从下拉列表中选择一个列。至少需要五列。

        i. 从下拉列表中选择一个分类。这标识了每列的数据类型。

            分类类型包括：

            - 数值-连续数值，例如测量值或计数
            - 分类- 离散值或类别，例如标签或类型

        ii. 要删除列，请选择删除。

        iii. 要添加另一列，请选择添加另一列。从下拉列表中选择 “列” 和 “分类”。

        iv. 对于 “预测值”，请从下拉列表中选择一列。这是自定义模型在合成数据集上训练后用于预测的列。

    b. 高级设置允许您设置隐私级别和隐私阈值。调整设置以满足您的需求。

        - 在 “隐私级别” 中，输入 epsilon 值以确定合成模型为保护生成的数据集中的隐私而添加了多少噪点。该值必须介于 0.0001 和 10 之间。

            - 较低的值会增加更多的噪音，从而提供更强的隐私保护，但可能会降低根据这些数据训练的下游自定义模型的效用。
            - 值越高，噪音越少，精度越高，但可能会降低隐私保护。

            在隐私阈值中，输入成员资格推断攻击可以识别原始数据集成员的最大允许概率。该值必须介于 50.0 和 100 之间。

            - 50% 的分数表明成员资格推断攻击无法比随机猜测更好地区分成员和非成员。
            - 如果没有隐私限制，请输入 100%。

            最佳值取决于您的具体使用案例和隐私需求。如果超过隐私阈值，则机器学习输入通道的创建将失败，并且您无法使用合成数据集来训练模型。

> **Warning**
>
> 合成数据生成可以防止推断出个人属性，无论特定个体存在于原始数据集中，还是存在这些个体的学习属性。但是，它并不能阻止原始数据集中的文字值，包括个人身份信息 (PII) 出现在合成数据集中。
>
> 我们建议避免输入数据集中仅与一个数据主体关联的值，因为这些值可能会重新识别数据主体。例如，如果只有一个用户居住在邮政编码中，则合成数据集中存在该邮政编码将确认该用户位于原始数据集中。诸如截断高精度值或用其他目录替换不常见的目录之类的技术可以用来降低这种风险。这些转换可以是用于创建 ML 输入通道的查询的一部分。

11. 如果要为资源启用标签，请选择添加新标签，然后输入密钥和值对。
12. 选择创建。
13. 现在，您可以通知您的协作成员他们可以查看分析模板。（如果您想查询自己的数据，则是可选的。）

查看 SQL 分析模板

协作成员创建 SQLanalysis 模板后，您可以对其进行审核和批准。分析模板获得批准后，便可以在中的查询中使用该模板 AWS Clean Rooms。

> **Note**
>
> 在协作中使用分析代码时，请注意以下几点：
>
> - AWS Clean Rooms 不验证或保证分析代码的行为。
>   - 如果您需要确保某些行为，请直接查看合作伙伴的守则，或者与值得信赖的第三方审计机构合作进行审查。
> - 在共享安全模型中：
>   - 您（客户）应对环境中运行的代码的安全性负责。
>   - AWS Clean Rooms 负责环境安全，确保
>   - 只有经过批准的代码才能运行
>   - 只有指定的配置表才可访问
>   - 唯一的输出目的地是结果接收器的 S3 存储桶。

使用 AWS Clean Rooms 控制台查看 SQL 分析模板

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台，该控制台将充当协作创建者。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在模板选项卡上，转到其他成员创建的分析模板部分。
5. 选择可以运行状态为否 - 需要您的审核的分析模板。
6. 选择审核。
7. 审核分析规则概述、定义和参数（如果有）。
8. 审核定义中引用的表格下列出的已配置表。

   每个表旁边的状态将显示为不允许使用模板。
9. 选择一个表。

如果您	则选择…
批准分析模板	允许在表格上使用模板。选择允许，确认您的批准。
不批准分析模板	不允许

现在，您可以使用 SQL 分析模板查询已配置的表了。有关更多信息，请参阅 运行 SQL 查询。

PySpark 分析模板

PySpark 分析模板需要 Python 用户脚本和可选的虚拟环境才能使用自定义和开源库。这些文件称为工件。

在创建分析模板之前，您需要先创建项目，然后将项目存储在 Amazon S3 存储桶中。 AWS Clean Rooms 在运行分析作业时使用这些工件。 AWS Clean Rooms 仅在运行作业时访问工件。

在 PySpark 分析模板上运行任何代码之前，请通过以下方式 AWS Clean Rooms 验证构件：

- 检查创建模板时使用的特定 S3 对象版本
- 验证工件的 SHA-256 哈希值

- 任何修改或移除工件的任务都失败了

> **Note**
>
> 中给定 PySpark 分析模板的所有组合工件的最大大小 AWS Clean Rooms 为 1 GB。

PySpark 分析模板的安全性

为了保护安全的计算环境， AWS Clean Rooms 使用两层计算架构将用户代码与系统操作隔离开来。该架构基于 Amazon EMR Serverless 细粒度访问控制技术，也称为薄膜。有关更多信息，请参阅膜——在存在命令式代码的情况下，Apache Spark 中的安全和高性能数据访问控制。

计算环境组件分为单独的用户空间和系统空间。用户空间执行 PySpark 分析模板中的 PySpark 代码。AWS Clean Rooms 使用系统空间使作业能够运行，包括使用客户提供的服务角色读取数据以运行作业，以及实现列允许名单。由于这种架构，影响系统空间的客户 PySpark 代码（可能包括少量 Spark SQL 和 PySpark DataFrames APIs）会被阻止。

PySpark 的局限性 AWS Clean Rooms

当客户提交经批准的 PySpark 分析模板时， AWS Clean Rooms 将在自己的安全计算环境中运行该模板，任何客户都无法访问。计算环境实现具有用户空间和系统空间的计算架构，以保护安全的计算环境。有关更多信息，请参阅 PySpark 分析模板的安全性。

在中使用 PySpark 之前，请考虑以下限制 AWS Clean Rooms。

限制

- 仅支持 DataFrame 输出
- 每次任务执行一次 Spark 会话

不支持的功能

- 数据管理
  - 冰山表格式
  - LakeFormation 托管表
  - 弹性分布式数据集（RDD）

  - Spark 流
  - 嵌套列的访问控制
- 自定义函数和扩展
  - 用户定义的表函数 (UDTFs)
  - Hive UDFs
  - 用户定义函数中的自定义类
  - 自定义数据来源
  - 其他 JAR 文件用于：
    - Spark 扩展
    - 连接器
    - 元数据仓配置
- 监测和分析
  - 火花日志
  - 火花用户界面
  - `ANALYZE TABLE` 命令

> **Important**
>
> 设置这些限制是为了保持用户和系统空间之间的安全隔离。
> 无论协作配置如何，所有限制都适用。
> 未来的更新可能会根据安全评估增加对其他功能的支持。

最佳实践

在创建 PySpark 分析模板时，我们建议采用以下最佳实践。

- 在设计分析模板时要PySpark 的局限性 AWS Clean Rooms牢记这一点。
- 请先在开发环境中测试您的代码。
- 仅使用支持的 DataFrame 操作。
- 规划您的输出结构，使其 DataFrame 不受限制地使用。

我们推荐以下管理构件的最佳实践

- 将所有 PySpark 分析模板项目保存在专用的 S3 存储桶或前缀中。
- 对不同的构件版本使用清晰的版本命名。
- 需要更新对象时创建新的分析模板。
- 维护一份清单，列出哪些模板使用哪些构件版本。

有关如何编写 Spark 代码的更多信息，请参阅以下内容：

- Apache 火花示例
- 在《亚马逊 EMR 发布指南》中编写 Spark 应用程序
- 教程：在AWS Glue 用户指南中 AWS Glue 编写 for Spark 脚本

以下主题说明了如何在创建和查看分析模板之前创建 Python 用户脚本和库。

主题

- 创建用户脚本
- 使用 PySpark 分析模板中的参数
- 创建虚拟环境（可选）
- 在 S3 中存储用户脚本和虚拟环境
- 创建 PySpark 分析模板
- 查看 PySpark 分析模板

创建用户脚本

用户脚本必须包含入口点函数（换句话说，处理程序）。您可以使用任何有效的 Python 文件名来命名用户脚本文件。

以下过程介绍如何创建用户脚本来定义 PySpark 分析的核心功能。

先决条件

- PySpark 1.0（对应于 Python 3.11 和 Spark 3.5.3）
- Amazon S3 中的数据集只能在您定义的 Spark 会话中作为已配置的表关联进行读取。
- 你的代码无法直接调用 Amazon S3 而且 AWS Glue
- 你的代码无法进行网络调用

创建用户脚本

1. 打开您选择的文本编辑器或集成开发环境 (IDE)。

   你可以使用任何支持 Python 文件的文本编辑器或 IDE（例如 Visual Studio Code 或 Notepad++）。 PyCharm

2. 使用您选择的名称创建一个新的 Python 文件（例如，**my_analysis.py**）。

3. 定义一个接受上下文对象参数的入口点函数。

   ```
   def entrypoint(context)
   ```

   context对象参数是一个字典，用于访问基本的 Spark 组件、引用的表和分析参数。其中包含：

   通过 Spark 访问会话 context['sparkSession']

   通过以下方式引用了表 context['referencedTables']

   通过分析参数context['analysisParameters']（如果参数是在模板中定义的）

4. 定义入口点函数的结果：

   ```
   return results
   ```

   results必须向输出 DataFrame返回一个包含文件名结果字典的对象。

   > **Note**
   >
   > AWS Clean Rooms 自动将 DataFrame 对象写入结果接收器的 S3 存储桶。

5. 您现在已准备好执行以下操作：

   a. 将此用户脚本存储在 S3 中。有关更多信息，请参阅 在 S3 中存储用户脚本和虚拟环境。

   b. 创建可选的虚拟环境以支持用户脚本所需的任何其他库。有关更多信息，请参阅 创建虚拟环境（可选）。

Example示例 1

<caption>The following example demonstrates a generic user script for a PySpark analysis template.</caption>

```
# File name: my_analysis.py

def entrypoint(context):
    try:
        # Access Spark session
        spark = context['sparkSession']

        # Access input tables
        input_table1 = context['referencedTables']['table1_name']
        input_table2 = context['referencedTables']['table2_name']

        # Example data processing operations
        output_df1 = input_table1.select("column1", "column2")
        output_df2 = input_table2.join(input_table1, "join_key")
        output_df3 = input_table1.groupBy("category").count()

        # Return results - each key creates a separate output folder
        return {
            "results": {
                "output1": output_df1,        # Creates output1/ folder
                "output2": output_df2,        # Creates output2/ folder
                "analysis_summary": output_df3 # Creates analysis_summary/ folder
            }
        }

    except Exception as e:
        print(f"Error in main function: {str(e)}")
        raise e
```

此示例的文件夹结构如下所示：

```
analysis_results/
#
### output1/ # Basic selected columns
# ### part-00000.parquet
# ### _SUCCESS
#
### output2/ # Joined data
# ### part-00000.parquet
# ### _SUCCESS
#
### analysis_summary/ # Aggregated results
### part-00000.parquet
```

```
### _SUCCESS
```

Example示例 2

<caption>The following example demonstrates a more complex user script for a PySpark analysis template.</caption>

```
def entrypoint(context):
    try:
        # Get DataFrames from context
        emp_df = context['referencedTables']['employees']
        dept_df = context['referencedTables']['departments']

        # Apply Transformations
        emp_dept_df = emp_df.join(
            dept_df,
            on="dept_id",
            how="left"
        ).select(
            "emp_id",
            "name",
            "salary",
            "dept_name"
        )

        # Return Dataframes
        return {
            "results": {
                "outputTable": emp_dept_df
            }
        }

    except Exception as e:
        print(f"Error in entrypoint function: {str(e)}")
        raise e
```

使用 PySpark 分析模板中的参数

参数允许在作业提交时提供不同的值，从而提高 PySpark 分析模板的灵活性。参数可通过传递给入口点函数的上下文对象进行访问。

> **Note**
>
> 参数是用户提供的字符串，可以包含任意内容。
>
> - 查看代码以确保参数得到安全处理，以防止分析中出现意外行为。
> - 无论提交时提供什么参数值，设计参数处理都要安全运行。

访问参数

参数可在`context['analysisParameters']`字典中找到。所有参数值均为字符串。

Example安全访问参数

```
def entrypoint(context):
    # Access parameters from context
    parameters = context['analysisParameters']
    threshold = parameters['threshold']
    table_name = parameters['table_name']

    # Continue with analysis using parameters
    spark = context['sparkSession']
    input_df = context['referencedTables'][table_name]

    # Convert threshold value
    threshold_val = int(threshold)

    # Use parameter in DataFrame operation
    filtered_df = input_df.filter(input_df.amount > threshold_val)

    return {
        "results": {
            "output": filtered_df
        }
    }
```

参数安全的最佳实践

> **Warning**
>
> 参数是用户提供的字符串，可以包含任意内容。您必须安全地处理参数，以防分析代码中存在安全漏洞。

要避免的不安全参数处理模式：

- 将参数作为代码执行 — 切勿exec()在参数值上使用 o eval() r

```
# UNSAFE - Don't do this
eval(parameters['expression'])  # Can execute arbitrary code
```

- SQL 字符串插值 — 切勿将参数直接连接到 SQL 字符串中

```
# UNSAFE - Don't do this
sql = f"SELECT * FROM table WHERE column = '{parameters['value']}'"  # SQL injection
 risk
```

- 不安全的文件路径操作 — 未经验证，切勿在文件系统操作中直接使用参数

```
# UNSAFE - Don't do this
file_path = f"/data/{parameters['filename']}"  # Path traversal risk
```

安全的参数处理模式：

- 在 DataFrame 操作中使用参数 — Spark DataFrames 可以安全地处理参数值

```
# SAFE - Use parameters in DataFrame operations
threshold = int(parameters['threshold'])
filtered_df = input_df.filter(input_df.value > threshold)
```

- 验证参数值-使用前检查参数是否符合预期格式

```
# SAFE - Validate parameters before use
def validate_date(date_str):
    try:
        from datetime import datetime
```

```
        datetime.strptime(date_str, '%Y-%m-%d')
        return True
    except ValueError:
        return False

date_param = parameters['date_filter'] or '2024-01-01'
if not validate_date(date_param):
    raise ValueError(f"Invalid date format: {date_param}")
```

- 对参数值使用许可名单-如果可能，请根据已知的正确值验证参数

```
# SAFE - Use allowlists
allowed_columns = ['column1', 'column2', 'column3']
column_param = parameters['column_name']
if column_param not in allowed_columns:
    raise ValueError(f"Invalid column: {column_param}")
```

- 带错误处理的类型转换-将字符串参数安全地转换为预期类型

```
# SAFE - Convert with error handling
try:
    batch_size = int(parameters['batch_size'] or '1000')
    if batch_size <= 0 or batch_size > 10000:
        raise ValueError(f"Batch size must be between 1 and 10000")
except ValueError as e:
    print(f"Invalid parameter: {e}")
    raise
```

> **⚠ Important**
>
> 请记住，当作业运行器提供不同的值时，参数会绕过代码审查。无论提供什么参数值，都要设计出可以安全运行的参数处理。

完整参数示例

Example在 PySpark 脚本中安全地使用参数

```
def entrypoint(context):
    try:
        # Access Spark session and tables
```

```
        spark = context['sparkSession']
        input_table = context['referencedTables']['sales_data']

        # Access parameters - fail fast if analysisParameters missing
        parameters = context['analysisParameters']

        # Validate and convert numeric parameter (handles empty strings with default)
        try:
            threshold = int(parameters['threshold'] or '100')
            if threshold <= 0:
                raise ValueError("Threshold must be positive")
        except (ValueError, TypeError) as e:
            print(f"Invalid threshold parameter: {e}")
            raise

        # Validate date parameter (handles empty strings with default)
        date_filter = parameters['start_date'] or '2024-01-01'
        from datetime import datetime
        try:
            datetime.strptime(date_filter, '%Y-%m-%d')
        except ValueError:
            raise ValueError(f"Invalid date format: {date_filter}")

        # Use parameters safely in DataFrame operations
        filtered_df = input_table.filter(
            (input_table.amount > threshold) &
            (input_table.date >= date_filter)
        )

        result_df = filtered_df.groupBy("category").agg(
            {"amount": "sum"}
        )

        return {
            "results": {
                "filtered_results": result_df
            }
        }

    except Exception as e:
        print(f"Error in analysis: {str(e)}")
        raise
```

创建虚拟环境（可选）

如果您的用户脚本需要任何其他库，则可以选择创建虚拟环境来存储这些库。如果您不需要其他库，可以跳过此步骤。

使用具有本机扩展的库时，请记住 PySpark ，in 在具有 ARM64 架构的 Linux 上 AWS Clean Rooms 运行。

以下过程演示如何使用基本 CLI 命令创建虚拟环境。

创建虚拟环境

1. 打开终端或命令提示符。
2. 添加以下内容：

```
# create and activate a python virtual environment
python3 -m venv pyspark_venvsource
source pyspark_venvsource/bin/activate

# install the python packages
pip3 install pandas # add packages here

# package the virtual environment into an archive
pip3 install venv-pack
venv-pack -f -o pyspark_venv.tar.gz


# optionally, remove the virtual environment directory
deactivate
rm -fr pyspark_venvsource
```

3. 现在，您可以将此虚拟环境存储在 S3 中。有关更多信息，请参阅 在 S3 中存储用户脚本和虚拟环境。

有关使用 Docker 和 Amazon ECR 的更多信息，请参阅亚马逊 ECRUser 指南。

在 S3 中存储用户脚本和虚拟环境

以下过程说明如何在 Amazon S3 中存储用户脚本和可选的虚拟环境。在创建 PySpark 分析模板之前完成此步骤。

Important

创建分析模板后，请勿修改或移除对象（用户脚本或虚拟环境）。
这样做将：

- 导致所有使用此模板的 future 分析作业失败。
- 需要使用新构件创建新的分析模板。
- 不影响之前完成的分析作业

先决条件

- AWS 账户 具有适当权限的
- 用户脚本文件（例如`my_analysis.py`）
- （可选，如果存在）虚拟环境包（`.tar.gz`文件）
- 创建或修改 IAM 角色的权限

Console

要使用控制台在 S3 中存储用户脚本和虚拟环境，请执行以下操作：

1. 登录 AWS 管理控制台 并打开 Amazon S3 控制台，网址为https://console.aws.amazon.com/s3/。
2. 创建新的 S3 存储桶或使用现有的 S3 存储桶。
3. 为存储桶启用版本控制。
   a. 选择您的存储桶。
   b. 选择属性。
   c. 在 “存储桶版本控制” 部分，选择 “编辑”。
   d. 选择 “启用” 并保存更改。
4. 上传您的工件并启用 SHA-256 哈希。
   a. 导航到您的存储桶。
   b. 选择上传。
   c. 选择添加文件并添加您的用户脚本文件。

d. （可选，如果存在）添加您的 .tar.gz 文件。

e. 展开 “属性”。

f. 在 “校验和” 下，对于 “校验和函数”，选择。SHA256

g. 选择上传。

5. 现在，您可以创建 PySpark 分析模板了。

CLI

要在 S3 中存储用户脚本和虚拟环境，请使用以下命令 AWS CLI：

1. 运行如下命令：

```
aws s3 cp --checksum-algorithm sha256 pyspark_venv.tar.gz s3://ARTIFACT-BUCKET/
EXAMPLE-PREFIX/
```

2. 现在，您可以创建 PySpark 分析模板了。

> **Note**
>
> 如果您需要更新脚本或虚拟环境：
>
> 1. 将新版本作为单独的对象上传。
> 2. 使用新构件创建新的分析模板。
> 3. 弃用旧模板。
> 4. 如果仍需要旧模板，请将原始工件保留在 S3 中。

创建 PySpark 分析模板

> **Note**
>
> 参数是用户提供的字符串，可以包含任意内容。
>
> - 查看代码以确保参数得到安全处理，以防止分析中出现意外行为。
> - 无论提交时提供什么参数值，设计参数处理都要安全运行。

先决条件

在创建 PySpark 分析模板之前，您必须具备以下条件：

- 积极 AWS Clean Rooms 协作的成员资格
- 访问活动协作中至少一个已配置的表
- 创建分析模板的权限
- 在 S3 中创建和存储的 Python 用户脚本和虚拟环境
  - S3 存储桶已启用版本控制。有关更多信息，请参阅在 S3 存储桶中使用版本控制
  - S3 存储桶可以计算已上传项目的 SHA-256 校验和。有关更多信息，请参阅使用校验和
- 从 S3 存储桶读取代码的权限

  有关创建所需服务角色的信息，请参阅创建用于从 S3 存储桶读取代码的服务角色（PySpark 分析模板角色）。

以下过程描述了使用AWS Clean Rooms 控制台创建 PySpark 分析模板的过程。它假设您已经创建了用户脚本和虚拟环境文件，并将您的用户脚本和虚拟环境文件存储在 Amazon S3 存储桶中。

> **Note**
>
> 创建 PySpark 分析模板的成员也必须是接收结果的成员。

有关如何使用创建 PySpark 分析模板的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

创建 PySpark 分析模板

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台，该控制台将充当协作创建者。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在模板选项卡上，转到您创建的分析模板部分。
5. 选择创建分析模板。
6. 在“创建分析模板”页面上，对于详细信息，

   a. 输入分析模板的名称。

b. （可选）输入描述。

c. 对于“格式”，选择PySpark选项。

7. 对于定义，

a. 在继续操作之前，请查看先决条件并确保满足每个先决条件。

b. 对于入口点文件，输入 S3 存储桶或选择“浏览 S3”。

c. （可选）对于库文件，输入 S3 存储桶或选择“浏览 S3”。

8. 对于参数-可选，如果您想添加参数以使分析模板可重复使用：

a. 选择 Add parameter (添加参数)。

b. 输入参数名称。

参数名称必须以字母或下划线开头，后跟字母数字字符或下划线。

c. 对于“类型”，系统会自动选择 STRING 作为 PySpark 分析模板唯一支持的类型。

d. （可选）输入参数的默认值。

如果您提供默认值，则作业运行器可以在运行作业时使用此值，而无需明确提供参数值。

e. 要添加更多参数，请选择添加其他参数并重复前面的步骤。

> **Note**
>
> 每个 PySpark 分析模板最多可以定义 50 个参数。每个参数值最多可包含 1,000 个字符。

9. 对于定义中引用的表，

- 如果定义中引用的所有表都已与协作关联：
  - 将“定义中引用的所有表都已关联到协作”复选框保持选中状态。
  - 在与协作关联的表格下，选择定义中引用的所有关联表。
- 如果定义中引用的所有表都未与协作关联：
  - 清除“定义中引用的所有表都已关联到协作”复选框。
  - 在与协作关联的表格下，选择定义中引用的所有关联表。
  - 在稍后将关联的表下，输入表名。
  - 选择“列出其他表”以列出另一个表。

10. 对于错误消息配置，请选择以下选项之一：

- 基本错误消息-返回基本错误消息而不暴露基础数据。建议用于生产工作负载。
- 详细的错误消息-返回详细的错误消息，以便更快地进行故障排除。建议在开发和测试环境中使用。可能会暴露敏感数据，包括个人身份信息 (PII)。

> **Note**
>
> 使用详细错误消息时，所有数据提供者成员都必须批准模板的此设置。

11. 通过从下拉列表中选择现有服务角色名称来指定服务访问权限。

    1. 如果您有列出角色的权限，则会显示角色列表。

       如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。

    2. 通过选择在 IAM 中查看外部链接来查看服务角色。

       如果没有现有的服务角色，则使用现有服务角色选项不可用。

       默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

    > **Note**
    >
    > - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
    > - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
    > - 如果您无法修改角色策略，则会收到一条错误消息，指出 AWS Clean Rooms 找不到该服务角色的策略。

12. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。
13. 选择创建。
14. 现在，您可以通知您的协作成员他们可以查看分析模板。（如果您想查询自己的数据，则是可选的。）

Important

创建分析模板后，请勿修改或移除构件（用户脚本或虚拟环境）。
这样做将：

- 导致所有使用此模板的 future 分析作业失败。
- 需要使用新构件创建新的分析模板。
- 不影响先前完成的分析作业。

查看 PySpark 分析模板

当其他成员在您的协作中创建分析模板时，您必须对其进行审核和批准才能使用。

以下过程向您展示如何查看 PySpark 分析模板，包括其规则、参数和参考表。作为协作成员，您将评估模板是否符合您的数据共享协议和安全要求。

分析模板获得批准后，即可将其用于中的作业 AWS Clean Rooms。

Note

在协作中使用分析代码时，请注意以下几点：

- AWS Clean Rooms 不验证或保证分析代码的行为。
  - 如果您需要确保某些行为，请直接查看合作伙伴的守则，或者与值得信赖的第三方审计机构合作进行审查。
- AWS Clean Rooms 保证 PySpark分析模板中列出的代码的 SHA-256 哈希值与 PySpark 分析环境中运行的代码相匹配。
- AWS Clean Rooms 不会对您引入环境的其他库进行任何审计或安全分析。
- 在共享安全模型中：
  - 您（客户）应对环境中运行的代码的安全性负责。
  - 您（客户）负责为环境设置相应的错误消息配置。
  - AWS Clean Rooms 负责环境安全，确保
    - 只有经过批准的代码才能运行
    - 只有指定的配置表才可访问
    - 唯一的输出目的地是结果接收器的 S3 存储桶。

AWS Clean Rooms 生成用户脚本和虚拟环境的 SHA-256 哈希值供您查看。但是，实际的用户脚本和库无法在其中直接访问 AWS Clean Rooms。

要验证共享的用户脚本和库是否与分析模板中引用的相同，您可以创建共享文件的 SHA-256 哈希值，并将其与创建的分析模板哈希值进行比较 AWS Clean Rooms。代码运行的哈希值也将出现在作业日志中。

先决条件

- Linux/Unix 操作系统或 Linux 版 Windows 子系统 (WSL)
- 要哈希处理的用户脚本文件
  - 请求分析模板创建者通过安全渠道共享文件。
- 由创建的分析模板哈希值 AWS Clean Rooms

使用 AWS Clean Rooms 控制台查看 PySpark 分析模板

1. 登录 AWS 管理控制台 并打开AWS Clean Rooms 控制台，该控制台将充当协作创建者。 AWS 账户
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在模板选项卡上，转到其他成员创建的分析模板部分。
5. 选择可以运行状态为否 - 需要您的审核的分析模板。
6. 选择审核。
7. 审核分析规则概述、定义和参数（如果有）。

> **Note**
>
> 参数允许分析运行者在提交时提交不同的值。如果分析模板支持参数，请查看合作伙伴代码中如何使用参数值，以确保其满足您的要求。

8. 验证共享用户脚本和库是否与分析模板中引用的脚本和库相同。

   a. 创建共享文件的 SHA-256 哈希值，并将其与创建的分析模板哈希值进行比较 AWS Clean Rooms。

      您可以通过导航到包含用户脚本文件的目录，然后运行以下命令来生成哈希值：

```
sha256sum your_script_filename.py
```

输出示例：

```
e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855 my_analysis.py
```

b. 或者，您可以使用 Amazon S3 校验和功能。有关更多信息，请参阅 Amazon S3 用户指南中的在 Amazon S 3 中检查对象完整性。

c. 另一种选择是在作业日志中查看已执行代码的哈希值。

9. 审核定义中引用的表格下列出的已配置表。

每个表旁边的状态将显示为不允许使用模板。

10. 选择一个表。

a. 要批准分析模板，请选择 “允许在表格上使用模板”。选择允许，确认您的批准。

b. 要拒绝批准，请选择 “不允许”。

如果您选择批准分析模板，则可以运行 PySpark 作业的成员现在可以使用 PySpark 分析模板在已配置的表上运行作业。有关更多信息，请参阅 正在运行的 PySpark 作业。

疑难解答 PySpark 分析模板

使用 PySpark 分析模板运行作业时，您可能会在任务初始化或执行过程中遇到故障。这些故障通常与脚本配置、数据访问权限或环境设置有关。

有关 PySpark 限制的更多信息，请参阅PySpark 的局限性 AWS Clean Rooms。

主题

对代码进行故障排除

为了帮助您开发代码并对其进行故障排除，我们建议您启用详细错误消息并使用自己的 AWS 账户 测试数据运行作业，从而自己进行模拟 AWS Clean Rooms 。

您也可以按照以下步骤 PySpark AWS Clean Rooms 在 Amazon EMR Serverless 中进行模拟。它与 AWS Clean Rooms PySpark 中的差异很小，但主要涵盖代码的运行方式。

在 EMR Ser PySpark verl AWS Clean Rooms ess 中进行模拟

1. 在 Amazon S3 中创建数据集，将其编入目录 AWS Glue Data Catalog，然后设置 Lake Formation 权限。
2. 使用自定义角色向 Lake Formation 注册 S3 地点。
3. 如果你还没有 Amazon EMR Studio 实例，请创建一个（使用亚马逊 EMR Studio 需要使用 Amazon EMR Studio 才能使用 Amazon EMR Serverless）。
4. 创建 EMR 无服务器应用程序
   - 选择发行版 emr-7.7.0。
   - 选择 ARM64 架构。
   - 选择 “使用自定义设置”。
   - 禁用预初始化的容量。
   - 如果您打算进行交互式工作，请选择交互式终端节点 > 为 EMR studio 启用终端节点。
   - 选择其他配置 > 使用 Lake Formation 进行精细的访问控制。
   - 创建应用程序。
5. 通过 EMR-Studio 笔记本或 API 使用 EMR-S。`StartJobRun`

分析模板作业无法启动

常见原因

由于三个主要的配置问题，分析模板作业可能会在启动时立即失败：

- 脚本命名不正确，与所需格式不符
- 用户脚本中的入口点函数缺失或格式不正确
- 虚拟环境中的 Python 版本不兼容

解决方案

要解决这个问题，请执行以下操作：

1. 验证您的用户脚本：
   - 检查您的用户脚本是否具有有效的 Python 文件名。

     有效的 Python 文件名使用小写字母、下划线分隔单词，并使用.py 扩展名。
2. 验证入口点函数。如果您的用户脚本没有入口点函数，请添加一个。
   a. 打开您的用户脚本。
   b. 添加以下入口点函数：

```
def entrypoint(context):
    # Your analysis code here
```

   c. 确保函数名称的拼写完全相同。`entrypoint`
   d. 验证函数是否接受该`context`参数。
3. 检查 Python 版本兼容性：
   a. 验证您的虚拟环境是否使用 Python 3.9 或 3.11。
   b. 要检查您的版本，请运行：`python --version`
   c. 如果需要，请更新您的虚拟环境：

```
conda create -n analysis-env python=3.9
conda activate analysis-env
```

预防措施

- 使用提供的包含正确文件结构的分析模板起始代码。
- 使用 Python 3.9 或 3.11 为所有分析模板设置专用的虚拟环境。
- 提交作业之前，请使用模板验证工具在本地测试您的分析模板。
- 实施 CI/CD 检查以验证脚本命名和入口点函数要求。

分析模板作业开始但在处理过程中失败

常见原因

出于以下安全和格式原因，分析作业可能会在执行过程中失败：

- 未经授权的直接访问诸如 Amazon S3 之类的 AWS 服务或 AWS Glue
- 以不符合要求 DataFrame 规格的错误格式返回输出
- 由于执行环境中的安全限制，网络呼叫被阻止

解决方案

要解决

1. 删除直接 AWS 服务访问权限：
    a. 在您的代码中搜索直接 AWS 服务导入和调用。
    b. 使用提供的 Spark 会话方法替换直接访问 S3。
    c. 通过协作界面仅使用预先配置的表格。
2. 正确格式化输出：
    a. 确认所有输出均为 Spark DataFrames。
    b. 更新您的退货声明以匹配以下格式：

```
return {
    "results": {
        "output1": dataframe1
    }
}
```

    c. 移除所有不DataFrame 返回的对象。
3. 移除网络通话：
    a. 识别并移除任何外部 API 调用。
    b. 移除任何 urllib、请求或类似的网络库。
    c. 移除所有套接字连接或 HTTP 客户端代码。

预防措施

- 使用提供的代码 linter 检查是否存在未经授权的 AWS 导入和网络调用。
- 在安全限制与生产相匹配的开发环境中测试作业。
- 在部署作业之前，请按照输出架构验证过程进行操作。
- 查看安全指南，了解经批准的服务访问模式。

虚拟环境设置失败

常见原因

虚拟环境配置失败通常是由于：

- 开发环境和执行环境之间的 CPU 架构不匹配
- 阻碍正确环境初始化的 Python 代码格式化问题
- 容器设置中的基础映像配置不正确

解决方案

要解决

1. 配置正确的架构：
    a. 使用以下命令检查您当前的架构 `uname -m`.
    b. 更新你的 Dockerfile 以指定： ARM64

```
FROM --platform=linux/arm64 public.ecr.aws/amazonlinux/amazonlinux:2023-minimal
```

    c. 使用以下方法重建您的容器 `docker build --platform=linux/arm64.`
2. 修复 Python 缩进：
    a. 像在代码文件上一样运行 Pyth `black` on 代码格式化程序。
    b. 验证空格或制表符的使用是否一致（不能两者兼而有之）。
    c. 检查所有代码块的缩进：

```
def my_function():
    if condition:
```

```
        do_something()
    return result
```

d. 使用带有 Python 缩进突出显示功能的 IDE。

3. 验证环境配置：

   a. 运行`python -m py_compile your_script.py`以检查语法错误。
   b. 部署前在本地测试环境。
   c. 验证中列出了所有依赖关系`requirements.txt`。

预防措施

- 使用 Visual Studio 代码或 Py PyCharm thon 格式化插件
- 配置预提交挂钩以自动运行代码格式化程序
- 使用提供的 ARM64 基础镜像在本地构建和测试环境
- 在您的 CI/CD 管道中实现自动代码风格检查

在协作中分析数据

在中 AWS Clean Rooms，您可以通过运行查询或作业来分析数据。

查询是一种使用一组支持的函数、类和变量在协作中访问和分析已配置表的方法。中当前支持的查询语言 AWS Clean Rooms 是 SQL。有 3 种方法可以运行查询 AWS Clean Rooms：编写 SQL 代码、使用经批准的 SQL 分析模板或使用分析生成器用户界面。

作业是一种使用一组支持的函数、类和变量在协作中访问和分析已配置表的方法。中当前支持的作业类型 AWS Clean Rooms 是 PySpark。有一种方法可以运行作业 AWS Clean Rooms：使用经批准的 PySpark 分析模板。

以下主题介绍如何 AWS Clean Rooms 通过运行 SQL 查询或 PySpark作业来分析中的数据。

主题

运行 SQL 查询

> **Note**
>
> 只有当负责支付查询计算费用的成员以活跃成员的身份加入协作时，您才能运行查询。

作为可以查询的成员，您可以通过以下方式运行 SQL 查询：

- 使用 SQL 代码编辑器手动构建 SQL 查询。
- 使用经批准的 SQL 分析模板。
- 使用分析生成器用户界面无需编写 SQL 代码即可生成查询。

当可以查询的成员对协作中的表运行 SQL 查询时，AWS Clean Rooms 将扮演相关角色来代表他们访问这些表。 AWS Clean Rooms 根据需要将分析规则应用于输入查询及其输出。

分析规则和输出约束是自动强制执行的。 AWS Clean Rooms 仅返回符合定义的分析规则的结果。

AWS Clean Rooms 支持可能与其他查询引擎不同的 SQL 查询。有关规范，请参阅 AWS Clean Rooms SQL 参考。如果要对受差别隐私保护的数据表运行查询，您应该确保查询与 AWS Clean Rooms Differential Privacy 的通用查询结构兼容。

> **Note**
>
> 使用 Clean Rooms 加密计算时，并非所有 SQL 操作都会生成有效的结果。例如，您可以对加密列执行 COUNT，但是对加密的数字执行 SUM 会导致错误。此外，查询还可能产生错误的结果。例如，SUM 密封列的查询会产生错误。但是，对密封列的 GROUP BY 查询似乎成功了，但生成的组与通过对 cleartext 的 GROUP BY 查询生成的组不同。

为查询计算费用付费的成员要对协作中运行的查询付费。

可以查询的成员可以选择多个可以接收结果的成员，以接收来自单个查询的结果。有关更多信息，请参阅 使用 SQL 代码编辑器查询配置表。有关接收查询结果的一般信息，请参见接收和使用分析结果。

先决条件

在运行 SQL 查询之前，请确保具备以下条件：

- AWS Clean Rooms 合作中的活跃会员
- 访问协作中至少一个已配置的表
- 确认负责查询计算成本的成员是活跃的协作成员

有关如何通过直接调用 AWS Clean Rooms StartProtectedQuery API 操作或使用来查询数据或查看查询的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

有关查询日志记录的信息，请参阅分析登录 AWS Clean Rooms。

> **Note**
>
> 如果您对加密数据表运行查询，则加密列的结果将被加密。

SQL 查询的 Spark 属性配置

AWS Clean Rooms 使您可以选择通过为 SQL 查询配置支持的 Spark 属性来自定义 Spark 运行时行为。这些属性允许您微调性能、内存使用情况和查询执行参数。借助此功能，您可以更好地控制基于 Spark 的查询的处理方式，从而可以根据您的特定工作负载要求进行优化。

现在，您可以直接从 AWS Clean Rooms 控制台调整诸如随机分区、广播加入阈值和自适应查询执行参数之类的设置。此功能对于复杂的查询或大型数据集特别有用，其中默认配置可能不是最佳配置。通过微调这些 Spark 属性，您可以提高查询性能、减少资源消耗，并更好地管理基于 Spark 的协作分析的内存使用情况。

要利用此功能，你可以在查询界面中找到一个新的 Spark 属性部分。您可以从支持的属性列表中进行选择并指定自定义值。您也可以使用 StartProtectedQuery API 以编程方式配置 Spark 属性。这种高级配置选项使数据分析师和工程师能够优化查询，从而提高效率和可扩展性。

有关 Spark 属性的更多信息，包括默认值，请参阅 Apache Spark 文档中的 Spark 属性。

以下主题介绍如何使用 AWS Clean Rooms 控制台在协作中查询数据。

主题

- 使用 SQL 代码编辑器查询配置表
- 使用 SQL 代码编辑器查询 ID 映射表
- 使用 SQL 分析模板查询已配置的表
- 使用分析构建器查询
- 查看差别隐私的影响
- 查看最近的查询
- 查看查询详细信息

使用 SQL 代码编辑器查询配置表

作为可以查询的成员，您可以通过在 SQL 代码编辑器中编写 SQL 代码来手动生成查询。SQL 代码编辑器位于 AWS Clean Rooms 控制台中 “分析” 选项卡的 “分析” 部分。

默认情况下显示 SQL 代码编辑器。如果要使用分析构建器来生成查询，请参阅使用分析构建器查询。

> **Important**
>
> 如果您开始在代码编辑器中编写 SQL 查询，然后打开分析构建器用户界面，则不会保存您的查询。

AWS Clean Rooms 支持许多 SQL 命令、函数和条件。有关更多信息，请参阅 AWS Clean Rooms SQL 参考。

> **Tip**
>
> 如果查询运行时发生计划的维护，查询会终止并回滚。必须重新开始查询。

使用 SQL 代码编辑器查询配置表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为运行查询的协作。
4. 在分析选项卡的表格下，查看表列表及其关联的分析规则类型（聚合分析规则、列表分析规则或自定义分析规则）。

   > **Note**
   >
   > 如果您没有在列表中看到所期望的表，可能是由于以下原因：
   >
   > - 这些表尚未关联。
   > - 这些表没有配置分析规则。

5. （可选）要查看表的架构和分析规则控制，请选择加号图标 (+) 展开表。
6. 在“分析”部分下，在“分析”模式下，选择“编写 SQL 代码”。

   > **Note**
   >
   > 只有在可以接收结果的成员和负责支付查询计算费用的成员作为活跃成员加入协作时，才会显示分析部分。

7. 通过在 SQL 代码编辑器中键入查询来构建查询。

   有关支持的 SQL 命令和函数的更多信息，请参阅 AWS Clean Rooms SQL 参考。

   您也可以使用以下选项来构建查询。

   Use an example query

   使用示例查询

   1. 选择表名称旁边的三个垂直点。
   2. 在在编辑器中插入下，选择查询示例。

      > **Note**
      > 插入示例查询会将其附加到编辑器中已有的查询中。

      此时将显示查询示例。表下列出的所有表都包含在查询中。
   3. 编辑查询中的占位符值。

   Insert column names or functions

   插入列名或函数

   1. 选择列旁边的三个垂直点。
   2. 在在编辑器中插入下，选择列名。
   3. 要手动插入列上允许的函数，
      a. 选择列旁边的三个垂直点。
      b. 选择 “在编辑器中插入”。
      c. 选择允许的函数的名称（例如INNER JOIN、SUMSUM DISTINCT、或COUNT）。
   4. 按 Ctrl + 空格键可在代码编辑器中查看表架构。

      > **Note**
      > 可以查询的成员可以查看和使用每个配置表关联中的分区列。确保将分区列标记为已配置 AWS Glue 表下方的表中的分区列。

5. 编辑查询中的占位符值。

8. 指定支持的工作器类型和工作人员人数。

   您可以选择运行您的 SQL 查询的实例类型和实例（工作程序）数量。

   对于 CR.1X，您最多可以选择 128 名工作人员或至少 4 名工作人员。

   对于 CR.4X，您最多可以选择 32 名工作人员或至少 4 名工作人员。

   使用下表来确定您的用例所需的工作人员类型和人数。

Worker 类型	vCPU	内存（GB）	存储（GB）	工作线程数	洁净室处理单元总数 (CRPU)
CR.1X（默认）	4	30	100	4	8
				128	256
CR.4X	16	120	400	4	32
				32	256

> **Note**
>
> 不同的工作人员类型和人数会产生相关成本。要了解有关定价的更多信息，请参阅AWS Clean Rooms 定价。

9. 在“将结果发送给”中，指定谁可以接收结果。

> **Note**
>
> 要接收结果，必须将协作成员配置为结果接收者，并且必须是协作的活跃参与者（状态：活跃）

10. （仅限可以查询的成员）默认情况下，“使用您的默认结果设置”复选框处于选中状态。如果要保留默认结果设置，请将其选中。

如果要为此查询指定不同的结果设置，请清除 “使用默认结果设置” 复选框，然后选择以下选项。

a. 结果格式（CSV 或 PAR QU ET）

b. 结果文件（单个或多个）

c. 亚马逊 S3 中的结果目的地

每个可以接收结果的成员都可以在 A mazon S3 中指定不同的结果格式、结果文件和结果目标。

11. 要指定 Spark 属性：

a. 展开 Spark 属性。

b. 选择 “添加 Spark 属性”。

c. 在 Spark 属性对话框中，从下拉列表中选择一个属性名称并输入值。

下表提供了每个属性的定义。

有关 Spark 属性的更多信息，请参阅 Apache Spark 文档中的 Spark 属性。

属性名称	说明	默认值
spark.task.maxFa	控制任务在失败之前可以连续失败多少次。需要一个大于或等于 1 的值。允许的重试次数等于该值减去 1。如果任何尝试成功，则失败计数将重置。不同任务的失败不会累积到这个极限。	4
spark.sql.files。maxPartitionBytes	设置从基于文件的源（例如 Parquet、JSON 和 ORC）读取数据时要打包到单个分区的最大字节数。	128MB
spark.hadoop.fs.s3 .maxRetries	设置 Amazon S3 文件操作的最大重试次数。	
spark.network.	设置所有网络交互的默认超时时间。如果未配置，则覆盖以下超时设置：	120

属性名称	说明	默认值
	• Spark.storage。 blockManagerHeartbeatTimeoutMs • spark.shuffle.io.connectionT • Spark.rpc.askTimeout • spark.rpc.lookupTim	
spark.rdd.com	指定是否使用 spark.io.compression.codec 压缩序列化的 RDD 分区。适用于 Java 和 Scala 中的 StorageLevel .MEMORY_ONLY_SER，或 Python 中的.MEMORY_ONLY。 StorageLevel减少存储空间，但需要额外的 CPU 处理时间。	FALSE
Spark.shuffle.spill.compress	指定是否使用 spark.io.compression.codec 压缩随机播放数据。	TRUE
spark.sql.自适应。advisoryPartitionSizeInBytes	当 spark.sql.adaptive.enabled 为真时，设置自适应优化期间洗牌分区的目标大小（以字节为单位）。控制合并小分区或拆分倾斜分区时的分区大小。	（spark.sql.adaptive.shuffle 的值。targetPostShuffleInputSize)
spark.sql.自适应。autoBroadcastJoin阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。仅适用于自适应框架。使用与 spark.sql 相同的默认值。 autoBroadcastJoin阈值。设置为 -1 可禁用广播。	（无）
Spark.sql.adaptive.coalescePartitions.enabled	指定是否根据 spark.sql.adaptive 合并连续的洗牌分区。 advisoryPartitionSizeInBytes 以优化任务规模。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive.coalescePartive.coalinitialPartitionNum	定义合并前随机分区的初始数量。需要同时启用 spark.sql.adaptive.enabled 和 spark.sql.adaptive.coalescePartitions.enabled 才能成真。默认为 spark.sql.shuffle.partitions 的值。	（无）

属性名称	说明	默认值
Spark.sql.adaptive .coalescePartive.coal minPartitionSize	设置合并后的随机分区的最小大小，以防止自适应优化期间分区变得太小。	1 MB
Spark.sql.adaptive .coalescePartition s.parallelism Firs	指定是否根据集群并行度而不是 spark.sql .adaptive 来计算分区大小。 advisoryP artitionSizeInBytes 在分区合并期间。生成的分区大小小于配置的目标大小，以最大限度地提高并行度。我们建议在繁忙的群集上将其设置为 false，以通过防止过多的小任务来提高资源利用率。	TRUE
sql.adaptive.enabled	指定是否启用自适应查询执行，以便在查询执行期间根据准确的运行时统计数据重新优化查询计划。	TRUE
spark.sql.自适应。 forceOptimizeSkewe d加入	指定是否强制启用， OptimizeSkewedJoin 即使它引入了额外的随机播放。	FALSE
spark.sql.自适应。 localShuffleReader. 已启用	指定在不需要随机分区时（例如从排序合并联接转换为广播哈希联接之后）是否使用本地随机播放读取器。需要 spark.sql.adaptive .enabled 才为真。	TRUE
spark.sql.自适应。 maxShuffledHashJoi nLocalMapThreshold	设置用于构建本地哈希映射的最大分区大小（以字节为单位）。在以下情况下，优先考虑洗牌后的哈希联接而不是排序合并联接： • 此值等于或超过 spark.sql.adaptive。 advisoryPartitionSizeInBytes • 所有分区大小均在此限制范围内 覆盖 spark.sql.join。 preferSortMerge加入设置。	0 字节

属性名称	说明	默认值
spark.sql.自适应。optimizeSkewsInRebalancePartitions.enabled	指定是否通过基于 spark.sql.adaptive 将倾斜的随机分区拆分为较小的分区来优化这些分区。 advisoryPartitionSizeInBytes。需要 spark.sql.adaptive.enabled 才为真。	TRUE
spark.sql.自适应。rebalancePartitionsSmallPartitionFactor	定义拆分期间合并分区的大小阈值系数。小于此因子的分区乘以 spark.sql.adaptive。advisoryPartitionSizeInBytes 已合并。	0.2
Spark.sql.adaptive.skewjoin.enable	指定是否通过拆分和可选复制倾斜的分区来处理洗牌联接中的数据倾斜。适用于排序合并和洗牌哈希联接。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive.skewJoin。skewedPartitionFactor	确定决定分区偏斜的大小系数。当分区的大小超过两个分区时，分区就会出现偏差： • 该因子乘以分区大小中位数 • spark.sql.adaptive.skewJoin 的值。skewedPartitionThresholdInBytes	5
Spark.sql.adaptive.skewJoin。skewedPartitionThresholdInBytes	设置用于识别偏斜分区的大小阈值（以字节为单位）。当分区的大小超过两个分区时，分区就会出现偏差： • 这个门槛 • 分区大小中位数乘以 spark.sql.adaptive.skewJoin。 skewedPartitionFactor 我们建议将此值设置为大于 spark.sql.adaptive。 advisoryPartitionSizeInBytes。	256MB
spark.sql。autoBroadcastJoin阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。设置为 -1 可禁用广播。	10MB

属性名称	说明	默认值
sql.broadcastTimeout	控制广播加入期间广播操作的超时时间（以秒为单位）。	300 秒
spark.sql.cbo.enabled	指定是否为计划统计数据估算启用基于成本的优化 (CBO)。	FALSE
spark.sql.cbo.joinreorder.dp.star.Filter	指定是否在基于开销的联接枚举期间应用星型联接过滤器启发式算法。	FALSE
spark.sql.cbo.joinreorder.dp.Thresh	设置动态规划算法中允许的最大连接节点数。	12
Spark.sql.cbo.joinreorder.enabled	指定是否在基于成本的优化 (CBO) 中启用联接重新排序。	FALSE
Spark.sql.cbo.planstats.enabled	指定在逻辑计划生成期间是否从目录中提取行数和列统计信息。	FALSE
spark.sql.cbo。starSchemaDetection	指定是否启用基于星型架构检测的联接重新排序。	FALSE
spark.sql.files。maxPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最大数量。当初始计数超过此值时，重新缩放分区。这是建议的目标，而不是保证的上限。	（无）
spark.sql.files。maxRecordsPer文件	设置写入单个文件的最大记录数。如果设置为零或负值，则不适用任何限制。	0
spark.sql.files。minPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最小数量。默认为 spark.sql。 leafNodeDefault并行性。这是建议的目标，而不是保证的上限。	（无）

属性名称	说明	默认值
spark.sql。inMemoryColumnarStorage.batchSize	控制列式缓存的批次大小。增加大小可以提高内存利用率和压缩率，但会增加 out-of-memory出错的风险。	10000
spark.sql。inMemoryColumnar存储. 已压缩	指定是否根据数据统计信息自动为列选择压缩编解码器。	TRUE
spark.sql。inMemoryColumnar存储。 enableVectorizedReader	指定是否为列式缓存启用矢量化读取。	TRUE
Spark.sql.legacy。allowHashOnMapType	指定是否允许对地图类型数据结构进行哈希操作。此传统设置保持了与旧版 Spark 地图类型处理的兼容性。	
Spark.sql.legacy。allowNegativeScaleOfDecimal	指定是否允许在十进制类型定义中使用负比例值。此传统设置保持了与支持负十进制小数位数的旧 Spark 版本的兼容性。	
Spark.sql.legacy。castComplexTypesToString.enabled	指定是否启用将复杂类型转换为字符串的传统行为。保持与旧版 Spark 的类型转换规则的兼容性。	
Spark.sql.legacy。charVarcharAs字符串	指定是否将 CHAR 和 VARCHAR 类型视为字符串类型。此传统设置提供了与旧版 Spark 的字符串类型处理的兼容性。	
Spark.sql.legacy。createEmptyCollectionUsingStringType	指定是否使用字符串类型元素创建空集合。此传统设置保持了与旧版 Spark 的集合初始化行为的兼容性。	

属性名称	说明	默认值
Spark.sql.legacy。exponentLiteralAsdecimal. 已启用	指定是否将指数文字解释为十进制类型。此传统设置保持了与旧版 Spark 的数字文字处理的兼容性。	
spark.sql.legacy.json。 allowEmptyString.enabled	指定是否允许在 JSON 处理中使用空字符串。此传统设置保持了与旧版 Spark 的 JSON 解析行为的兼容性。	
spark.sql.legacy.parquet.int96RebaseModelRead	指定在读取 Parquet 文件时是否使用传统 INT96 的时间戳变基模式。此传统设置保持了与旧版 Spark 的时间戳处理的兼容性。	
Spark.sql.legacy。timeParserPolicy	控制时间解析行为以实现向后兼容。此传统设置决定了如何从字符串中解析时间戳和日期。	
Spark.sql.legacy.typeCorcion。datetimeToString.enabled	指定在将日期时间值转换为字符串时是否启用传统类型强制行为。保持与旧版 Spark 版本的日期时间转换规则的兼容性。	
spark.sql。maxSinglePartition字节	设置最大分区大小（以字节为单位）。规划器为较大的分区引入了洗牌操作以提高并行度。	128m
Spark.sql.metadataCache TTLSeconds	控制元数据缓存的 time-to-live (TTL)。适用于分区文件元数据和会话目录缓存。需要： • 大于零的正值 • Spark.sql.catalog实现设置为蜂巢 • spark.sql.hive。 filesourcePartitionFileCacheSize 大于零 • spark.sql.hive。 manageFilesourcePartitions 设置为 true	-1000 毫秒

属性名称	说明	默认值
火花.sql.optimizer。collapseProjectAlways内联	指定是否折叠相邻的投影和行内表达式，即使这会导致重复。	FALSE
火花.sql.optimizer。dynamicPartitionPruning.enabled	指定是否为用作联接键的分区列生成谓词。	TRUE
火花.sql.optimizer。enableCsvExpression优化	指定是否通过从 from_csv 操作中删除不必要的列来优化 SQL 优化器中的 CSV 表达式。	TRUE
火花.sql.optimizer。enableJsonExpression优化	通过以下方式指定是否优化 SQL 优化器中的 JSON 表达式： • 从 from_json 操作中删除不必要的列 • 简化 from_json 和 to_json 的组合 • 优化 named_struct 操作	TRUE
spark.sql.Optimizer.ExcludedR	定义要禁用的优化器规则，由逗号分隔的规则名称标识。某些规则无法禁用，因为它们是正确性所必需的。优化器会记录哪些规则已成功禁用。	（无）
spark.sql.Optimizer.runtime.bloomFilapplicationSideScanSizeThreshold	设置在应用程序端注入 Bloom 过滤器所需的最小聚合扫描大小（以字节为单位）。	10GB
spark.sql.Optimizer.runtime.bloomFilcreationSideThreshold	定义在创建端注入 Bloom 滤镜的最大大小阈值。	10MB

属性名称	说明	默认值
Spark.sql.Optimizer.runtime.bloomFilter.enable	指定当随机连接的一侧具有选择性谓词时，是否插入布隆过滤器以减少随机播放数据。	TRUE
spark.sql.Optimizer.runtime.bloomFil expectedNumItems	定义运行时 Bloom 过滤器中预期项目的默认数量。	1000000
spark.sql.Optimizer.runtime.bloomFil maxNumBits	设置运行时 Bloom 过滤器中允许的最大位数。	67108864
spark.sql.Optimizer.runtime.bloomFil maxNumItems	设置运行时 Bloom 过滤器中允许的最大预期项目数。	4000000
spark.sql.Optimizer.runtime.bloomFilter.number.	限制每次查询允许的非 DPP 运行时过滤器的最大数量，以防止驱动程序 out-of-memory出错。	10
Spark.sql.Optimizer.runtime.bloomfilter.numbit	定义运行时 Bloom 过滤器中使用的默认位数。	8388608
Spark.sql.optimizer.runtime rowlevelOperationGroup过滤器. 已启用	指定是否为行级操作启用运行时组筛选。允许数据源： • 使用数据源筛选器修剪整组数据（例如文件或分区） • 执行运行时查询以识别匹配的记录 • 丢弃不必要的组以避免昂贵的重写 限制： • 并非所有表达式都可以转换为数据源筛选器 • 有些表达式需要 Spark 求值（例如子查询）	TRUE

属性名称	说明	默认值
Spark.sql.Optimizer.runtimeFiltersemiJoinReduction.enabled	指定当随机连接的一侧具有选择性谓词时，是否插入半联接以减少随机播放数据。	FALSE
spark.sql.parquet.AgregatePus	指定是否将聚合向下推送到 Parquet 进行优化。支持： • 布尔型、整数、浮点型和日期类型的最小值和最大值 • 所有数据类型的计数 如果任何 Parquet 文件页脚中缺少统计信息，则会抛出异常。	FALSE
sql.parquet。columnarReaderBatch大小	控制每个 Parquet 矢量化阅读器批次中的行数。选择一个平衡性能开销和内存使用量的值，以防止 out-of-memory出错。	4096
Spark.sql.session.time	定义会话时区，用于处理字符串文字中的时间戳和 Java 对象转换。接受： • 以地区为基础 IDs 的 area/city 格式（例如美国/洛杉矶） • 区域偏移量采用 (+/-) HH、(+/-) HH: mm 或 (+/-) HH: mm: ss 格式（例如 -08 或 + 01:00） • UTC 或 Z 作为 + 00:00 的别名	（当地时区的值）
spark.sql.shuffle.part	设置联接或聚合期间用于数据洗牌的默认分区数。无法在结构化流式查询从同一检查点位置重新启动之间进行修改。	200

属性名称	说明	默认值
spark.sql。shuffledHashJoin因子	定义用于确定 shuffle 哈希加入资格的乘法系数。当小边数据大小乘以此系数小于大边数据大小时，将选择随机哈希联接。	3
火花.sql.sources。parallelPartitionDiscovery. 阈值	使用基于文件的源（Parquet、JSON 和 ORC）设置驱动端文件列表的最大路径数。如果在分区发现期间超出限制，则使用单独的 Spark 分布式作业列出文件。	32
spark.sql.statistics.histics.h	指定是否在列统计数据计算期间生成等高直方图以提高估计精度。除了基本列统计数据所需的扫描之外，还需要进行额外的表扫描。	FALSE

12. 选择运行。

> **Note**
>
> 如果可以接收结果的成员尚未配置查询结果设置，您将无法运行查询。

13. 查看结果。

    有关更多信息，请参阅 接收和使用分析结果。

14. 继续调整参数并再次运行查询，或者选择 + 按钮在新选项卡中开始新查询。

> **Note**
>
> AWS Clean Rooms 旨在提供清晰的错误消息。如果错误消息中没有足够的详细信息来帮助您进行故障排除，请联系客户团队。向他们说明错误情况和错误信息（包括任何标识符）。有关更多信息，请参阅 故障排除 AWS Clean Rooms。

使用 SQL 代码编辑器查询 ID 映射表

以下过程介绍了如何在 ID 映射表上运行多表联接查询，以将 `sourceId` 与 `targetId` 联接。

在查询 ID 映射表之前，必须成功填充 ID 映射表。

使用 SQL 代码编辑器查询 ID 映射表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为运行查询的协作。
4. 在 “分析” 选项卡上，转到 “分析” 部分。

> **Note**
>
> 只有在可以接收结果的成员和负责支付查询计算费用的成员作为活跃成员加入协作时，才会显示分析部分。

5. 在 “分析” 选项卡的 “表” 下，查看 ID 映射表列表（在 “管理者” 下 AWS Clean Rooms）及其关联的分析规则类型（ID 映射表分析规则）。

> **Note**
>
> 如果您没有在列表中看到所需 ID 映射表，可能是由于 ID 映射表没有成功填充。有关更多信息，请参阅 填充现有 ID 映射表。

6. 通过在 SQL 代码编辑器中键入查询来构建查询。

（可选）如果要使用示例查询	（可选）如果要插入表名
1. 选择表名称旁边的三个垂直点。 2. 在在编辑器中插入下，选择示例 JOIN 语句。 **Note** 插入示例 JOIN 语句会附加编辑器中已有的查询。 将出现示例 JOIN 语句。 3. 编辑查询中的占位符值。	1. 选择列旁边的三个垂直点。 2. 在在编辑器中插入下，选择表名称。 3. 编辑查询中的占位符值。

7. 指定支持的工作器类型和工作人员人数。

   使用下表来确定您的用例所需的工作人员类型和人数。

Worker 类型	vCPU	内存（GB）	存储（GB）	工作线程数	洁净室处理单元总数 (CRPU)
CR.1X（默认）	4	30	100	4	8
				128	256
CR.4X	16	120	400	4	32
				32	256

> **Note**
>
> 不同的工作人员类型和人数会产生相关成本。要了解有关定价的更多信息，请参阅AWS Clean Rooms 定价。

8. 选择运行。

> **Note**
>
> 如果可以接收结果的成员尚未配置查询结果设置，您将无法运行查询。

9. 查看结果。

   有关更多信息，请参阅 接收和使用分析结果。

10. 继续调整参数并再次运行查询，或者选择 + 按钮在新选项卡中开始新查询。

> **Note**
>
> AWS Clean Rooms 旨在提供清晰的错误消息。如果错误消息中没有足够的详细信息来帮助您进行故障排除，请联系客户团队。向他们说明错误情况和错误信息（包括任何标识符）。有关更多信息，请参阅 故障排除 AWS Clean Rooms。

使用 SQL 分析模板查询已配置的表

此过程演示如何使用 AWS Clean Rooms 控制台中的分析模板通过自定义分析规则查询已配置的表。

使用 SQL 分析模板通过自定义分析规则查询已配置的表

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为运行查询的协作。
4. 在分析选项卡的表格部分下，查看表格及其关联的分析规则类型（自定义分析规则）。

> **Note**
>
> 如果您没有在列表中看到所期望的表，可能是由于以下原因：
>
> - 这些表尚未关联。
> - 这些表没有配置分析规则。

5. 在 “分析” 部分下，在 “分析” 模式下，选择 “运行分析模板”，然后从下拉列表中选择分析模板。
6. SQL 分析模板中的参数将自动填充到定义中。
7. 指定支持的工作器类型和工作人员人数。

   使用下表来确定您的用例所需的工作人员类型和人数。

Worker 类型	vCPU	内存（GB）	存储（GB）	工作线程数	洁净室处理单元总数 (CRPU)
CR.1X（默认）	4	30	100	4	8
				128	256
CR.4X	16	120	400	4	32
				32	256

> **Note**
>
> 不同的工作人员类型和人数会产生相关成本。要了解有关定价的更多信息，请参阅AWS Clean Rooms 定价。

8. 指定支持的 Spark 属性。

   a. 选择 “添加 Spark 属性”。

   b. 在 Spark 属性对话框中，从下拉列表中选择一个属性名称并输入值。

   下表提供了每个属性的定义。

   有关 Spark 属性的更多信息，请参阅 Apache Spark 文档中的 Spark 属性。

属性名称	说明	默认值
spark.task.maxFa	控制任务在失败之前可以连续失败多少次。需要一个大于或等于 1 的值。允许的重试次数等于该值减去 1。如果任何尝试成功，则失败计数将重置。不同任务的失败不会累积到这个极限。	4
spark.sql.files。maxPartitionBytes	设置从基于文件的源（例如 Parquet、JSON 和 ORC）读取数据时要打包到单个分区的最大字节数。	128MB
spark.hadoop.fs.s3 .maxRetries	设置 Amazon S3 文件操作的最大重试次数。	
spark.network.	设置所有网络交互的默认超时时间。如果未配置，则覆盖以下超时设置： • Spark.storage。 blockManagerHeartb eatTimeoutMs • spark.shuffle.io.connectionT • Spark.rpc.askTimeout	120

属性名称	说明	默认值
	• spark.rpc.lookupTim	
spark.rdd.com	指定是否使用 spark.io.compression.codec 压缩序列化的 RDD 分区。适用于 Java 和 Scala 中的 StorageLevel .MEMORY_ONLY_SER，或 Python 中的.MEMORY_ONLY。 StorageLevel减少存储空间，但需要额外的 CPU 处理时间。	FALSE
Spark.shuffle.spill.compress	指定是否使用 spark.io.compression.codec 压缩随机播放数据。	TRUE
spark.sql.自适应。advisoryPartitionSizeInBytes	当 spark.sql.adaptive.enabled 为真时，设置自适应优化期间洗牌分区的目标大小（以字节为单位）。控制合并小分区或拆分倾斜分区时的分区大小。	（spark.sql.adaptive.shuffle 的值。targetPostShuffleInputSize)
spark.sql.自适应。autoBroadcastJoin阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。仅适用于自适应框架。使用与 spark.sql 相同的默认值。 autoBroadcastJoin阈值。设置为 -1 可禁用广播。	（无）
Spark.sql.adaptive.coalescePartitions.enabled	指定是否根据 spark.sql.adaptive 合并连续的洗牌分区。 advisoryPartitionSizeInBytes 以优化任务规模。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive.coalescePartive.coalinitialPartitionNum	定义合并前随机分区的初始数量。需要同时启用 spark.sql.adaptive.enabled 和 spark.sql.adaptive.coalescePartitions.enabled 才能成真。默认为 spark.sql.shuffle.partitions 的值。	（无）
Spark.sql.adaptive.coalescePartive.coalminPartitionSize	设置合并后的随机分区的最小大小，以防止自适应优化期间分区变得太小。	1 MB

属性名称	说明	默认值
Spark.sql.adaptive .coalescePartition s.parallelism Firs	指定是否根据集群并行度而不是 spark.sql .adaptive 来计算分区大小。 advisoryP artitionSizeInBytes 在分区合并期间。生成的分区大小小于配置的目标大小，以最大限度地提高并行度。我们建议在繁忙的群集上将其设置为 false，以通过防止过多的小任务来提高资源利用率。	TRUE
sql.adaptive.enabled	指定是否启用自适应查询执行，以便在查询执行期间根据准确的运行时统计数据重新优化查询计划。	TRUE
spark.sql.自适应。 forceOptimizeSkewe d加入	指定是否强制启用， OptimizeSkewedJoin 即使它引入了额外的随机播放。	FALSE
spark.sql.自适应。 localShuffleReader .enabled	指定在不需要随机分区时（例如从排序合并联接转换为广播哈希联接之后）是否使用本地随机播放读取器。需要 spark.sql.adaptive .enabled 才为真。	TRUE
spark.sql.自适应。 maxShuffledHashJoi nLocalMapThreshold	设置用于构建本地哈希映射的最大分区大小（以字节为单位）。在以下情况下，优先考虑洗牌后的哈希联接而不是排序合并联接： • 此值等于或超过 spark.sql.adaptive。 advisoryPartitionSizeInBytes • 所有分区大小均在此限制范围内 覆盖 spark.sql.join。 preferSortMerge加入设置。	0 字节

属性名称	说明	默认值
spark.sql.自适应。optimizeSkewsInRebalancePartitions.enabled	指定是否通过基于 spark.sql.adaptive 将倾斜的随机分区拆分为较小的分区来优化这些分区。 advisoryPartitionSizeInBytes。需要 spark.sql.adaptive.enabled 才为真。	TRUE
spark.sql.自适应。rebalancePartitionsSmallPartitionFactor	定义拆分期间合并分区的大小阈值系数。小于此因子的分区乘以 spark.sql.adaptive。advisoryPartitionSizeInBytes 已合并。	0.2
Spark.sql.adaptive.skewjoin.enable	指定是否通过拆分和可选复制倾斜的分区来处理洗牌联接中的数据倾斜。适用于排序合并和洗牌哈希联接。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive.skewJoin。skewedPartitionFactor	确定决定分区偏斜的大小系数。当分区的大小超过两个分区时，分区就会出现偏差： • 该因子乘以分区大小中位数 • spark.sql.adaptive.skewJoin 的值。skewedPartitionThresholdInBytes	5
Spark.sql.adaptive.skewJoin。skewedPartitionThresholdInBytes	设置用于识别偏斜分区的大小阈值（以字节为单位）。当分区的大小超过两个分区时，分区就会出现偏差： • 这个门槛 • 分区大小中位数乘以 spark.sql.adaptive.skewJoin。 skewedPartitionFactor 我们建议将此值设置为大于 spark.sql.adaptive。 advisoryPartitionSizeInBytes。	256MB
spark.sql。autoBroadcastJoin阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。设置为 -1 可禁用广播。	10MB

属性名称	说明	默认值
sql.broadcastTimeout	控制广播加入期间广播操作的超时时间（以秒为单位）。	300 秒
spark.sql.cbo.enabled	指定是否为计划统计数据估算启用基于成本的优化 (CBO)。	FALSE
spark.sql.cbo.joinreorder.dp.star.Filter	指定是否在基于开销的联接枚举期间应用星型联接过滤器启发式算法。	FALSE
spark.sql.cbo.joinreorder.dp.Thresh	设置动态规划算法中允许的最大连接节点数。	12
Spark.sql.cbo.joinreorder.enabled	指定是否在基于成本的优化 (CBO) 中启用联接重新排序。	FALSE
Spark.sql.cbo.planstats.enabled	指定在逻辑计划生成期间是否从目录中提取行数和列统计信息。	FALSE
spark.sql.cbo。starSchemaDetection	指定是否启用基于星型架构检测的联接重新排序。	FALSE
spark.sql.files。maxPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最大数量。当初始计数超过此值时，重新缩放分区。这是建议的目标，而不是保证的上限。	（无）
spark.sql.files。maxRecordsPer文件	设置写入单个文件的最大记录数。如果设置为零或负值，则不适用任何限制。	0
spark.sql.files。minPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最小数量。默认为 spark.sql。 leafNodeDefault并行性。这是建议的目标，而不是保证的上限。	（无）

属性名称	说明	默认值
spark.sql。inMemoryColumnarStorage.batchSize	控制列式缓存的批次大小。增加大小可以提高内存利用率和压缩率，但会增加 out-of-memory出错的风险。	10000
spark.sql。inMemoryColumnar存储. 已压缩	指定是否根据数据统计信息自动为列选择压缩编解码器。	TRUE
spark.sql。inMemoryColumnar存储。 enableVectorizedReader	指定是否为列式缓存启用矢量化读取。	TRUE
Spark.sql.legacy。allowHashOnMapType	指定是否允许对地图类型数据结构进行哈希操作。此传统设置保持了与旧版 Spark 地图类型处理的兼容性。	
Spark.sql.legacy。allowNegativeScaleOfDecimal	指定是否允许在十进制类型定义中使用负比例值。此传统设置保持了与支持负十进制小数位数的旧 Spark 版本的兼容性。	
Spark.sql.legacy。castComplexTypesToString.enabled	指定是否启用将复杂类型转换为字符串的传统行为。保持与旧版 Spark 的类型转换规则的兼容性。	
Spark.sql.legacy。charVarcharAs字符串	指定是否将 CHAR 和 VARCHAR 类型视为字符串类型。此传统设置提供了与旧版 Spark 的字符串类型处理的兼容性。	
Spark.sql.legacy。createEmptyCollectionUsingStringType	指定是否使用字符串类型元素创建空集合。此传统设置保持了与旧版 Spark 的集合初始化行为的兼容性。	

属性名称	说明	默认值
Spark.sql.legacy。exponentLiteralAsdecimal. 已启用	指定是否将指数文字解释为十进制类型。此传统设置保持了与旧版 Spark 的数字文字处理的兼容性。	
spark.sql.legacy.json。 allowEmptyString.enabled	指定是否允许在 JSON 处理中使用空字符串。此传统设置保持了与旧版 Spark 的 JSON 解析行为的兼容性。	
spark.sql.legacy.parquet.int96RebaseModelRead	指定在读取 Parquet 文件时是否使用传统 INT96 的时间戳变基模式。此传统设置保持了与旧版 Spark 的时间戳处理的兼容性。	
Spark.sql.legacy。timeParserPolicy	控制时间解析行为以实现向后兼容。此传统设置决定了如何从字符串中解析时间戳和日期。	
Spark.sql.legacy.typeCorcion。datetimeToString.enabled	指定在将日期时间值转换为字符串时是否启用传统类型强制行为。保持与旧版 Spark 版本的日期时间转换规则的兼容性。	
spark.sql。maxSinglePartition字节	设置最大分区大小（以字节为单位）。规划器为较大的分区引入了洗牌操作以提高并行度。	128m
Spark.sql.metadataCache TTLSeconds	控制元数据缓存的 time-to-live (TTL)。适用于分区文件元数据和会话目录缓存。需要： • 大于零的正值 • Spark.sql.catalog实现设置为蜂巢 • spark.sql.hive。 filesourcePartitionFileCacheSize 大于零 • spark.sql.hive。 manageFilesourcePartitions 设置为 true	-1000 毫秒

属性名称	说明	默认值
火花.sql.optimizer。collapseProjectAlways内联	指定是否折叠相邻的投影和行内表达式，即使这会导致重复。	FALSE
火花.sql.optimizer。dynamicPartitionPruning.enabled	指定是否为用作联接键的分区列生成谓词。	TRUE
火花.sql.optimizer。enableCsvExpression优化	指定是否通过从 from_csv 操作中删除不必要的列来优化 SQL 优化器中的 CSV 表达式。	TRUE
火花.sql.optimizer。enableJsonExpression优化	通过以下方式指定是否优化 SQL 优化器中的 JSON 表达式： • 从 from_json 操作中删除不必要的列 • 简化 from_json 和 to_json 的组合 • 优化 named_struct 操作	TRUE
spark.sql.Optimizer.ExcludedR	定义要禁用的优化器规则，由逗号分隔的规则名称标识。某些规则无法禁用，因为它们是正确性所必需的。优化器会记录哪些规则已成功禁用。	（无）
spark.sql.Optimizer.runtime.bloomFilapplicationSideScanSizeThreshold	设置在应用程序端注入 Bloom 过滤器所需的最小聚合扫描大小（以字节为单位）。	10GB
spark.sql.Optimizer.runtime.bloomFilcreationSideThreshold	定义在创建端注入 Bloom 滤镜的最大大小阈值。	10MB

属性名称	说明	默认值
Spark.sql.Optimizer.runtime.bloomFilter.enable	指定当随机连接的一侧具有选择性谓词时，是否插入布隆过滤器以减少随机播放数据。	TRUE
spark.sql.Optimizer.runtime.bloomFil expectedNumItems	定义运行时 Bloom 过滤器中预期项目的默认数量。	1000000
spark.sql.Optimizer.runtime.bloomFil maxNumBits	设置运行时 Bloom 过滤器中允许的最大位数。	67108864
spark.sql.Optimizer.runtime.bloomFil maxNumItems	设置运行时 Bloom 过滤器中允许的最大预期项目数。	4000000
spark.sql.Optimizer.runtime.bloomFilter.number.	限制每次查询允许的非 DPP 运行时过滤器的最大数量，以防止驱动程序 out-of-memory出错。	10
Spark.sql.Optimizer.runtime.bloomfilter.numbit	定义运行时 Bloom 过滤器中使用的默认位数。	8388608
Spark.sql.optimizer.runtime rowlevelOperationGroup过滤器. 已启用	指定是否为行级操作启用运行时组筛选。允许数据源： • 使用数据源筛选器修剪整组数据（例如文件或分区） • 执行运行时查询以识别匹配的记录 • 丢弃不必要的组以避免昂贵的重写 限制： • 并非所有表达式都可以转换为数据源筛选器 • 有些表达式需要 Spark 求值（例如子查询）	TRUE

属性名称	说明	默认值
Spark.sql.Optimizer.runtimeFilter semiJoinReduction.enabled	指定当随机连接的一侧具有选择性谓词时，是否插入半联接以减少随机播放数据。	FALSE
spark.sql.parquet.AgregatePus	指定是否将聚合向下推送到 Parquet 进行优化。支持： • 布尔型、整数、浮点型和日期类型的最小值和最大值 • 所有数据类型的计数 如果任何 Parquet 文件页脚中缺少统计信息，则会抛出异常。	FALSE
sql.parquet。columnarReaderBatch大小	控制每个 Parquet 矢量化阅读器批次中的行数。选择一个平衡性能开销和内存使用量的值，以防止 out-of-memory出错。	4096
Spark.sql.session.time	定义会话时区，用于处理字符串文字中的时间戳和 Java 对象转换。接受： • 以地区为基础 IDs 的 area/city 格式（例如美国/洛杉矶） • 区域偏移量采用 (+/-) HH、(+/-) HH: mm 或 (+/-) HH: mm: ss 格式（例如 -08 或 + 01:00） • UTC 或 Z 作为 + 00:00 的别名	（当地时区的值）
spark.sql.shuffle.part	设置联接或聚合期间用于数据洗牌的默认分区数。无法在结构化流式查询从同一检查点位置重新启动之间进行修改。	200

属性名称	说明	默认值
spark.sql。shuffledHashJoin因子	定义用于确定 shuffle 哈希加入资格的乘法系数。当小边数据大小乘以此系数小于大边数据大小时，将选择随机哈希联接。	3
火花.sql.sources。parallelPartitionDiscovery. 阈值	使用基于文件的源（Parquet、JSON 和 ORC）设置驱动端文件列表的最大路径数。如果在分区发现期间超出限制，则使用单独的 Spark 分布式作业列出文件。	32
spark.sql.statistics.histics.h	指定是否在列统计数据计算期间生成等高直方图以提高估计精度。除了基本列统计数据所需的扫描之外，还需要进行额外的表扫描。	FALSE

9. 选择运行。

> **Note**
>
> 如果可以接收结果的成员尚未配置查询结果设置，您将无法运行查询。

10. 继续调整参数并再次运行查询，或者选择 + 按钮在新选项卡中开始新查询。

使用分析构建器查询

您无需编写 SQL 代码即可使用分析构建器来构建查询。使用分析构建器，您可以为具有以下特征的协作构建查询：

- 单个使用聚合分析规则且不需要 JOIN 的表
- 两个使用聚合分析规则的表（每个成员一个）
- 两个使用列表分析规则的表（每个成员一个）
- 两个使用聚合分析规则的表（每个成员一个）和两个使用列表分析规则的表（每个成员一个）

如果要手动编写 SQL 查询，请参阅使用 SQL 代码编辑器查询配置表。

分析生成器在 AWS Clean Rooms 控制台的 “分析” 选项卡的 “分析” 部分中显示为 “分析生成器” 用户界面选项。

> **Important**
>
> 如果您打开分析构建器用户界面，开始在分析构建器中构建查询，然后关闭分析构建器用户界面，则不会保存您的查询。

> **Tip**
>
> 如果查询运行时发生计划的维护，查询会终止并回滚。必须重新开始查询。

以下主题介绍分析构建器的使用。

主题

使用分析构建器查询单个表（聚合）

此过程演示如何使用 AWS Clean Rooms 控制台中的 Analys is Builder 用户界面来生成查询。该查询适用于具有单个表的协作，该表使用聚合分析规则且无需 JOIN。

使用分析构建器查询单个表

登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

1. 在左侧导航窗格中，选择协作。
2. 选择您的成员能力状态为运行查询的协作。
3. 在 “分析” 选项卡的 “表” 下，查看表格及其关联的分析规则类型。（分析规则类型应为聚合分析规则。）

> **Note**
>
> 如果您没有看到所期望的表，可能是由于以下原因：
>
> - 该表尚未关联。

- 该表没有配置分析规则。

4. 在分析部分下，打开分析构建器用户界面。
5. 构建查询。

   如果要查看所有聚合指标，请跳至步骤 9。

   a. 对于选择指标，请查看默认情况下预先选择的聚合指标，并在需要时删除任何指标。

   b. （可选）对于添加分段 - 可选，请选择一个或多个参数。

   > **Note**
   >
   > 只有在为表指定维度时才会显示添加分段 - 可选。

   c. （可选）对于添加筛选条件 - 可选，请选择添加筛选条件，然后选择参数、运算符和值。

   要添加更多筛选条件，请选择再添加一个筛选条件。

   要删除筛选条件，请选择移除。

   > **Note**
   >
   > ORDER BY 不支持聚合查询。
   > 筛选条件仅支持 AND 运算符。

   d. （可选）对于添加描述 - 可选，请输入描述以帮助识别查询列表中的查询。
6. 展开预览 SQL 代码。

   a. 查看分析构建器生成的 SQL 代码。

   b. 要复制 SQL 代码，请选择复制。

   c. 要编辑 SQL 代码，请选择在 SQL 代码编辑器中编辑。
7. 选择运行。

   > **Note**
   >
   > 如果可以接收结果的成员尚未配置查询结果设置，您将无法运行查询。

8. 继续调整参数并再次运行查询，或者选择 + 按钮在新选项卡中开始新查询。

> **Note**
>
> AWS Clean Rooms 旨在提供清晰的错误消息。如果错误消息中没有足够的详细信息来帮助您进行故障排除，请联系客户团队。向他们说明错误情况和错误信息（包括任何标识符）。有关更多信息，请参阅 故障排除 AWS Clean Rooms。

使用分析构建器查询两个表（聚合或列表）

此过程介绍如何使用 AWS Clean Rooms 控制台中的分析生成器为具有以下特征的协作生成查询：

- 两个使用聚合分析规则的表（每个成员一个）
- 两个使用列表分析规则的表（每个成员一个）
- 两个使用聚合分析规则的表（每个成员一个）和两个使用列表分析规则的表（每个成员一个）

使用分析构建器查询两个表

登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

1. 在左侧导航窗格中，选择协作。
2. 选择处于 “您的成员权限” 状态为 “运行查询” 的协作...
3. 在 “分析” 选项卡的 “表” 下，查看两个表及其关联的分析规则类型（聚合分析规则或列表分析规则）。

   > **Note**
   >
   > 如果您没有在列表中看到所期望的表，可能是由于以下原因：
   >
   > - 这些表尚未关联。
   > - 这些表没有配置分析规则。

4. 在分析部分下，打开分析构建器用户界面。
5. 构建查询。

   如果协作包含两个使用聚合分析规则的表和两个使用列表分析规则的表，请先选择聚合或列表，然后根据所选分析规则按照提示进行操作。

如果两个表使用聚合分析规则	如果两个表使用列表分析规则
1. 对于选择指标，请查看默认情况下预先选择的聚合指标，并在需要时删除任何指标。 2. 对于匹配记录，请选择一个或多个记录。 **Note** 使用分析构建器时，只能对一对列进行匹配。 3. （可选）对于添加分段 - 可选，请选择一个或多个参数。 **Note** 只有在为表指定维度时才会显示添加分段 - 可选。 4. （可选）对于添加筛选条件 - 可选，请选择添加筛选条件，然后选择参数、运算符和值。 要添加更多筛选条件，请选择再添加一个筛选条件。 要删除筛选条件，请选择移除。 **Note** ORDER BY 不支持聚合查询。 筛选条件仅支持 AND 运算符。 5. （可选）对于添加描述 - 可选，请输入描述以帮助识别最近查询列表中的查询。	1. 对于选择属性，请查看默认情况下预先选择的列表属性，并在需要时删除任何指标。 2. 对于匹配记录，请选择一个或多个记录。 **Note** 使用分析构建器时，只能对一对列进行匹配。 3. （可选）对于添加筛选条件 - 可选，请选择添加筛选条件，然后选择参数、运算符和值。 要添加更多筛选条件，请选择再添加一个筛选条件。 要删除筛选条件，请选择删除。 **Note** LIMIT 不支持列表查询。 筛选条件仅支持 AND 运算符。 4. （可选）对于添加描述 - 可选，请输入描述以帮助识别最近查询列表中的查询。

6. 展开预览 SQL 代码。

a. 查看分析构建器生成的 SQL 代码。

b. 要复制 SQL 代码，请选择复制。

c. 要编辑 SQL 代码，请选择在 SQL 代码编辑器中编辑。

7. 选择运行。

> **Note**
>
> 如果可以接收结果的成员尚未配置查询结果设置，您将无法运行查询。

8. 继续调整参数并再次运行查询，或者选择 + 按钮在新选项卡中开始新查询。

> **Note**
>
> AWS Clean Rooms 旨在提供清晰的错误消息。如果错误消息中没有足够的详细信息来帮助您进行故障排除，请联系客户团队。向他们说明错误情况和错误信息（包括任何标识符）。有关更多信息，请参阅 故障排除 AWS Clean Rooms。

查看差别隐私的影响

一般来说，在开启差别隐私后，编写和运行查询不会发生变化。不过，如果剩余的隐私预算不足，则无法运行查询。在您运行查询并使用隐私预算时，您可以大致了解可以运行的聚合数量以及这可能会如何影响将来的查询。

查看差别隐私在协作中的影响

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员详细信息状态为运行查询的协作。
4. 在 “分析” 选项卡的 “表” 下，查看剩余的隐私预算。这显示为估计的剩余聚合函数数量和使用的效用（显示为百分比）。

> **Note**
>
> 仅为可以查询的成员显示估计的剩余聚合函数数量和使用的效用百分比。

5. 选择查看影响以查看在结果中注入了多少噪声以及您大约可以运行多少个聚合函数。

查看最近的查询

您可以在 “分析” 选项卡上查看过去 90 天内运行的查询。

> **Note**
>
> 如果您唯一的成员权限是 Contrib ute 数据，而您不是支付查询计算费用的会员，则控制台上不会显示 “分析” 选项卡。

查看最近的查询

登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

1. 在左侧导航窗格中，选择协作。
2. 选择协作。
3. 在 “分析” 选项卡的 “分析” 下，从下拉列表中选择 “所有查询”，然后查看过去 90 天内运行的查询。
4. 要按状态对最近的查询进行排序，请从所有状态下拉列表中选择一个状态。

   状态为：已提交、已开始、已取消、成功、失败和超时。

查看查询详细信息

您可以以能够运行查询的成员或能够接收结果的成员的身份查看查询的详细信息。

查看查询的详细信息

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在 “分析” 选项卡上，执行以下任一操作：
   - 选择要查看的特定查询对应的选项按钮，然后选择查看详细信息。
   - 选择受保护的查询 ID。
5. 在查询详细信息页面上，
   - 如果您是可以运行查询的成员，请查看查询详细信息、SQL 文本和结果。

     您会看到一条消息，确认查询结果已发送给可以接收结果的成员。
   - 如果您是可以接收结果的成员，请查看查询详细信息和结果。

正在运行的 PySpark 作业

作为可以查询的成员，您可以使用已批准的 PySpark 分析模板在已配置的表上运行 PySpark 作业。

先决条件

在运行 PySpark 作业之前，你必须：

- AWS Clean Rooms 合作中的活跃会员
- 在协作中至少访问一个分析模板
- 访问协作中至少一个已配置的表
- 将 PySpark 任务结果写入指定 S3 存储桶的权限

  有关创建所需服务角色的信息，请参阅创建服务角色以写入 PySpark 作业结果。
- 负责支付计算费用的成员已作为活跃成员加入协作

有关如何通过直接调用 AWS Clean Rooms `StartProtectedJob` API 操作或使用来查询数据或查看查询的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

有关作业记录的信息，请参阅分析登录 AWS Clean Rooms。

有关接收作业结果的信息，请参阅接收和使用分析结果。

以下主题说明了如何使用 AWS Clean Rooms 控制台在协作中对已配置的表运行 PySpark 作业。

主题

使用 PySpark 分析模板在已配置的表上运行 PySpark 作业

此过程演示如何使用 AWS Clean Rooms 控制台中的 PySpark 分析模板通过自定义分析规则分析已配置的表。

使用 PySpark 分析模板在已配置的表上运行 PySpark 作业

登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

1. 在左侧导航窗格中，选择协作。
2. 选择处于 “您的成员权限” 状态为 “运行作业” 的协作。
3. 在分析选项卡的表格部分下，查看表格及其关联的分析规则类型（自定义分析规则）。

> **Note**
>
> 如果您没有在列表中看到所期望的表，可能是由于以下原因：
>
> - 这些表尚未关联。
> - 这些表没有配置分析规则。

4. 在 “分析” 部分下，在 “分析” 模式下，选择 “运行分析模板”。
5. 从 “ PySpark 分析模板” 下拉列表中选择分析模板。

   PySpark 分析模板中的参数将自动填充到定义中。
6. 如果分析模板定义了参数，请在 “参数” 下提供参数值：

   a. 查看每个参数的参数名称和默认值（如果已配置）。

   b. 为要覆盖的每个参数输入一个值。

> **Note**
>
> 如果您未提供值但存在默认值，则将使用默认值。

> **Important**
>
> 参数值最多可包含 1,000 个字符，并且支持 UTF-8 编码。所有参数值都被视为字符串，并通过上下文对象传递给您的用户脚本。
> 确保您的用户脚本能够安全地验证和处理参数值。有关安全参数处理的更多信息，请参阅使用 PySpark 分析模板中的参数。

7. 指定支持的工作器类型和工作人员人数。

   使用下表来确定您的用例所需的工作人员类型和人数。

Worker 类型	vCPU	内存（GB）	存储（GB）	工作线程数	洁净室处理单元总数 (CRPU)
CR.1X（默认）	4	30	100	4	8
				128	256
CR.4X	16	120	400	4	32
				32	256

> **Note**
>
> 不同的工作人员类型和人数会产生相关成本。要了解有关定价的更多信息，请参阅AWS Clean Rooms 定价。

8. 选择运行。

> **Note**
>
> 如果可以接收结果的成员尚未配置作业结果设置，则无法运行作业。

9. 继续调整参数并重新运行作业，或者选择 + 按钮在新选项卡中开始新作业。

查看最近的工作

您可以在分析选项卡上查看过去 90 天内运行的作业。

> **Note**
>
> 如果您唯一的成员权限是 Contrib ute 数据，而您不是支付工作计算费用的会员，则控制台上不会显示 “分析” 选项卡。

查看最近的工作

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在 “分析” 选项卡的 “分析” 下，从下拉列表中选择 “所有作业”，然后查看过去 90 天内运行的作业。
5. 要按状态对最近的作业进行排序，请从 “所有状态” 下拉列表中选择一个状态。

   状态为：已提交、已开始、已取消、成功、失败和超时。

查看任务详细信息

您可以以可以运行作业的成员或可以接收结果的成员的身份查看作业详细信息。

查看任务的详细信息

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。

2. 在左侧导航窗格中，选择协作。
3. 选择协作。
4. 在 “分析” 选项卡的 “分析” 下，从下拉列表中选择 “所有作业”，然后执行以下操作之一：
   - 选择要查看的特定作业的选项按钮，然后选择 “查看详细信息”。
   - 选择受保护的作业 ID。
5. 在 Job 详情页面上，
   - 如果您是可以运行作业的成员，请查看 Job 详细信息、Job 和结果。

     您会看到一条消息，确认工作结果已发送给可以接收结果的成员。
   - 如果您是可以接收结果的成员，请查看 Job 详情和结果。

接收和使用分析结果

能够接收结果的成员可以在 AWS Clean Rooms 控制台或他们在加入协作时指定的 Amazon S3 存储桶中查看查询结果。

> **Note**
>
> 仅对于加密数据表，可以接收结果的成员通过在解密模式下运行 C3R 加密客户端来解密查询结果。
> Amazon S3 中的结果目标不能与任何数据源位于同一 S3 存储桶中。

以下主题说明如何使用 AWS Clean Rooms 控制台接收分析结果。

主题

有关如何通过直接调用 AWS Clean Rooms API 或使用来查询数据或查看查询的信息 AWS SDKs，请参阅 AWS Clean Rooms API 参考。

有关查询日志记录的信息，请参阅分析登录 AWS Clean Rooms。

> **Note**
>
> 如果您对加密数据表运行查询，则加密列的结果将被加密。

接收查询结果

> **Note**
>
> Amazon S3 中的结果目标不能与任何数据源位于同一 S3 存储桶中。

查询结果位于 AWS Clean Rooms 控制台中 “分析” 选项卡的 “结果设置默认值” 部分。

接收查询结果

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （ 如果您尚未这样做 ）。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为接收结果的协作。
4. 要直接从中接收查询结果 AWS Clean Rooms，请在 “分析” 选项卡的 “分析” 下拉列表中选择 “所有查询”，然后在 “受保护的查询 ID” 列下选择查询。
5. 在查询详细信息页面的结果下，执行以下任一操作：

如果要...	则选择...
复制结果。	复制
下载结果。	下载 **Note** 默认情况下，下载的文件的名称是在AWS Clean Rooms中运行查询时显示的相应 `Query id`。
在 Amazon S3 中查看结果。	在 Amazon S3 中查看 这将在单独的选项卡中打开 Amazon S3 控制台。

6. 如果您使用的是加密数据，则现在可以解密数据表。

   有关更多信息，请参阅 使用 C3R 加密客户端解密数据表。

接收工作结果

> **Note**
>
> Amazon S3 中的结果目标不能与任何数据源位于同一 S3 存储桶中。

作业的结果位于 AWS Clean Rooms 控制台中 “分析” 选项卡的 “结果设置默认值” 部分。

接收工作成绩

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （ 如果您尚未这样做 ）。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为接收结果的协作。
4. 要直接从中接收作业结果 AWS Clean Rooms，请在 “分析” 选项卡的 “分析” 下拉列表中选择 “所有作业”，然后在 “受保护的作业 ID” 列下选择作业。
5. 在任务详情页面的结果下，复制作业 ID。

   返回分析选项卡并展开结果设置默认值。

   在 “结果目标” 下，选择链接以在 Amazon S3 中查看结果。

   这将在单独的选项卡中打开 Amazon S3 控制台。

   在 Amazon S3 中，将作业 ID 粘贴到搜索栏中，然后按 Enter。

   将出现包含结果的文件夹。选择要查看作业结果的文件夹。

编辑查询结果设置的默认值

> **Note**
>
> Amazon S3 中的结果目标不能与任何数据源位于同一 S3 存储桶中。

作为可以接收结果的成员，您可以在 AWS Clean Rooms 控制台中编辑查询结果设置的默认值。

编辑查询结果设置的默认值

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为接收结果的协作。
4. 在分析选项卡的结果设置默认值下，选择编辑。
5. 在编辑结果设置默认值页面上，根据需要修改以下任意内容：

   a. 在 “查询结果” 下，修改 Amazon S3 中的结果目标、结果格式或结果文件。

   b. （可选）对于服务访问权限，如果您想将最长需要 24 小时的查询传送到 S3 目标，请选中 “添加服务角色以支持最长需要 24 小时才能完成的查询” 复选框。

      最长需要 24 小时才能完成的大型查询将传送到您的 S3 目标。

      如果您不选中该复选框，则只有在 12 小时内完成的查询才会发送到您的 S3 位置。

      - 通过选择创建并使用新的服务角色或使用现有服务角色来指定服务访问权限。

        Create and use a new service role

        - AWS Clean Rooms 使用此表所需的策略创建服务角色。
        - 默认服务角色名称为 `cleanrooms-query-receiver-<timestamp>`。
        - 您必须拥有创建角色并附加策略的权限。

        Use an existing service role

        1. 从下拉列表中选择一个现有服务角色名称。

           如果您有列出角色的权限，则会显示角色列表。

           如果您没有列出角色的权限，可以输入要使用的角色的 Amazon 资源名称 (ARN)。
        2. 通过选择在 IAM 中查看外部链接来查看服务角色。

           如果没有现有的服务角色，则使用现有服务角色选项不可用。

           默认情况下， AWS Clean Rooms 不会尝试更新现有角色策略以添加必要的权限。

> **Note**
>
> - AWS Clean Rooms 需要权限才能根据分析规则进行查询。有关权限的更多信息 AWS Clean Rooms，请参阅AWS 的托管策略 AWS Clean Rooms。
> - 如果该角色没有足够的权限 AWS Clean Rooms，则会收到一条错误消息，指出该角色没有足够的权限 AWS Clean Rooms。必须先添加角色策略，然后才能继续。
> - 如果您无法修改角色策略，则会收到一条错误消息，指出找 AWS Clean Rooms 不到该服务角色的策略。

6. 选择保存更改。
7. 更新后的查询结果设置显示在协作详细信息页面上。

编辑作业结果设置的默认值

> **Note**
>
> Amazon S3 中的结果目标不能与任何数据源位于同一 S3 存储桶中。

作为可以接收结果的成员，您可以在 AWS Clean Rooms 控制台中编辑作业结果设置的默认值。

编辑作业结果设置的默认值

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 选择您的成员能力状态为接收结果的协作。
4. 在分析选项卡的结果设置默认值下，选择编辑。
5. 在编辑结果设置默认值页面上，根据需要修改以下任意内容：
   a. 在 “任务结果” 下，修改 Amazon S3 中的结果目标。
   b. 在 “服务访问权限” 下，修改现有服务角色名称。
6. 选择保存更改。

7. 更新后的 Job 结果设置显示在协作详情页面上。

在其他中使用查询输出 AWS 服务

SQL 查询输出可用于 Clean Rooms ML 模型的种子数据。有关更多信息，请参阅 AWS Clean Rooms ML。

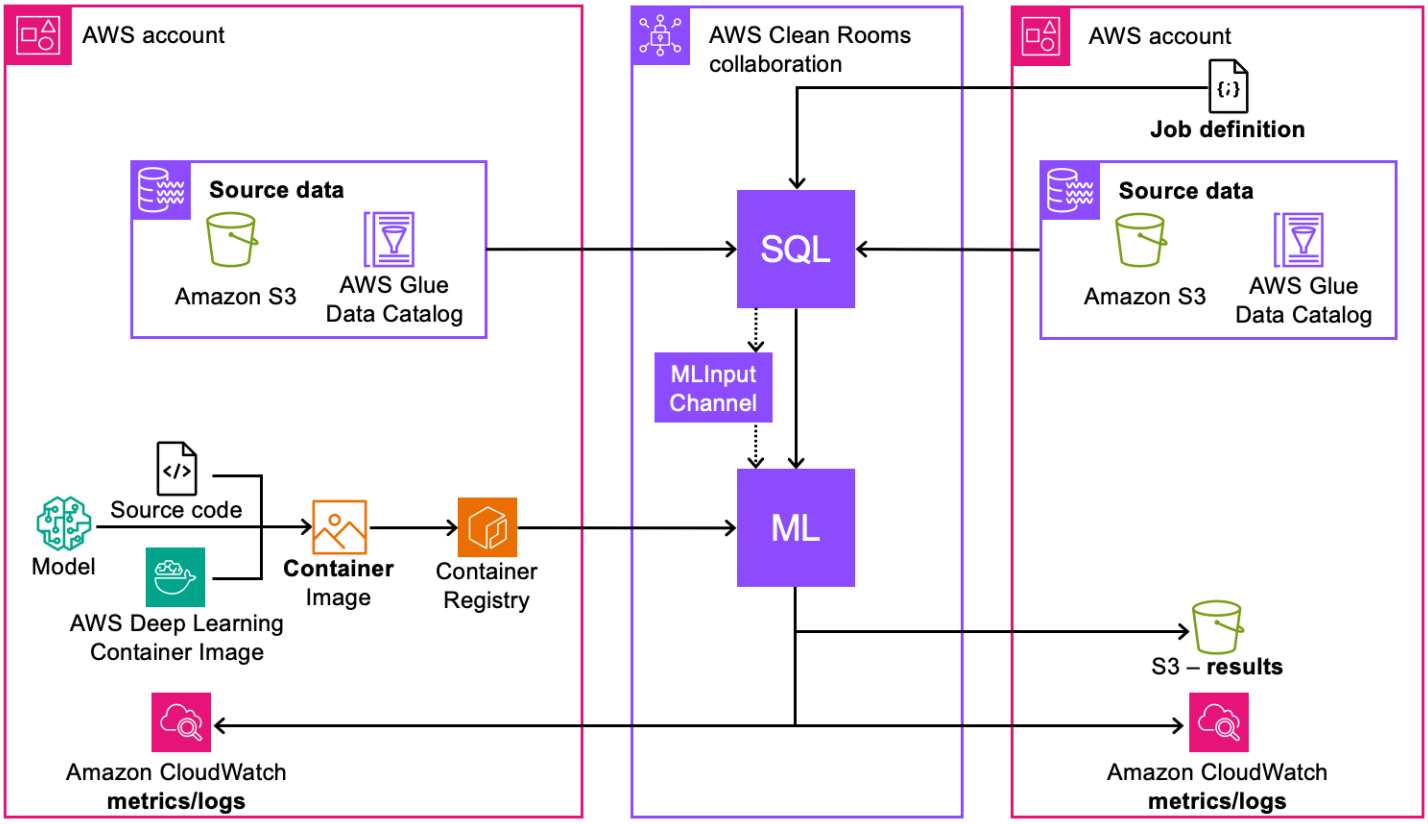
的查询输出可在控制台上找到 AWS Clean Rooms（如果使用控制台运行查询），并下载到指定的 Amazon S3 存储桶中。然后，您可以在其他服务中使用查询输出 AWS 服务，例如 Amazon Quick 和 Amazon SageMaker AI，具体取决于这些服务如何使用来自 Amazon S3 的数据。

有关 Amazon Quick 的更多信息，请参阅亚马逊快速文档。

有关亚马逊 A SageMaker I 的更多信息，请参阅亚马逊 A SageMaker I 文档。

AWS Clean Rooms ML 中的自定义建模

从技术角度来看，下图描述了自定义 ML 建模在 AWS Clean Rooms ML 中的工作原理。



以下是自定义 ML 建模在 Clean Rooms ML 中的工作原理：

1. 数据源配置
   - 源数据可以存储在 Amazon S3 目录中 AWS Glue Data Catalog、或 Snowflake 中
   - AWS Glue Data Catalog 用于整理和编目
   - 多个数据 AWS 账户 可以在同一个协作中使用
2. SQL 查询和数据处理
   - SQL 查询用于访问和处理源数据
   - 查询在 AWS Clean Rooms 协作边界内运行
   - 处理后的数据馈入 ML 输入通道以进行模型训练
3. 机器学习模型开发
   - 可以使用 AWS 深度学习容器映像开发模型的源代码
   - 必须创建自定义容器镜像并将其存储在 Amazon 弹性容器注册表中

4. 基础架构组件
   - Amazon 弹性容器注册表存储和管理 ML 模型容器
   - 机器学习处理发生在安全的 AWS Clean Rooms 协作环境中
5. 监控和日志记录
   - Amazon CloudWatch 为双方合作方提供指标和日志
   - 协作 AWS 账户 参与者均可进行监控
   - 相关方可以访问性能指标和操作日志
6. 结果管理
   - 对结果的访问权限根据协作权限进行控制

在开始之前，请参阅自定义 ML 建模先决条件和，了解训练容器的模型创作指南更多信息。

隐私增强型合成数据集生成

合成数据集与其所基于的原始数据集具有相似的统计特性，但不包含原始数据集中存在的真实观测结果。通过使用隐私增强型合成数据集，您可以解锁新的机器学习 (ML) 模型训练用例，而这些用例是数据隐私问题以前阻止的。创建 ML 输入通道时，可以在训练 ML 模型时生成合成数据以保护敏感信息。

使用合成数据创建模板时，您必须：

- 要求模板输出是合成的
- 将输出架构列分类为数字列或分类列
- 根据组织需求自定义合成数据
- 调整隐私设置：
  - 设置隐私级别 (epsilon)
  - 配置隐私阈值

> **Warning**
>
> 合成数据生成可以防止推断出个人属性，无论特定个体存在于原始数据集中，还是存在这些个体的学习属性。但是，它并不能阻止原始数据集中的文字值，包括个人身份信息 (PII) 出现在合成数据集中。
> 我们建议避免输入数据集中仅与一个数据主体关联的值，因为这些值可能会重新识别数据主体。例如，如果只有一个用户居住在邮政编码中，则合成数据集中存在该邮政编码将确认该用

户位于原始数据集中。诸如截断高精度值或用其他目录替换不常见的目录之类的技术可以用来降低这种风险。这些转换可以是用于创建 ML 输入通道的查询的一部分。

有关如何为自定义模型训练生成合成数据的更多信息，请参阅创建 SQL 分析模板。

带有合成输出的分析模板只能用于创建 ML 输入通道。有关更多信息，请参阅 在 AWS Clean Rooms ML 中创建机器学习输入通道。

合成数据生成的注意事项

借助 AWS Clean Rooms ML，协作成员可以创建一个合成数据集，该数据集可以不可逆转地将原始数据集的主体从其集体数据集中去识别出来，从而训练自定义机器学习模型。创建协作时，您必须配置付款信息，以指定谁为合成数据生成付费。以下是生成合成数据集和训练自定义机器学习模型的高级步骤：

1. 协作成员创建的分析模板包括：
   - 需要使用 SQL 来定义要合成的数据集。
   - 与隐私相关的配置，用于确保合成数据符合数据提供者的合规性要求。
2. 一旦所有数据提供者都批准了分析模板，协作查询运行器就会使用该模板创建一个机器学习 (ML) 输入通道。
3. Clean Rooms ML 生成合成数据集并验证其是否符合分析模板中指定的隐私阈值。
4. 如果满足所有阈值，则使用合成数据集填充 ML 输入通道。
5. 然后，客户可以使用此机器学习输入渠道来训练与协作相关的自定义 ML 模型。

重要注意事项：

- 在 Clean Rooms ML 中生成的合成数据不会删除、编辑、混淆或消毒任何个人值，包括在原始数据集中找到的个人身份信息 (PII)。合成数据集是由原始数据集中的采样值生成的，但不是整个记录生成的。
- 如果原始数据集包含相似的行，则合成数据可能包含看起来与原始数据集中的行完全相同的行。

数据集准备：

- 避免使用类别分布严重不平衡的列。这对于预测值或 “Y” 列尤其重要。极端的不平衡会降低合成数据集的整体隐私。

- Clean Rooms ML 不支持根据时间序列数据生成合成数据，在这些数据中，保持顺序记录之间的相关性非常重要。
- Clean Rooms ML 不支持从文本或非结构化数据生成合成数据。
- 支持以下数据类型：

数据类型名称
BIGINT
BOOLEAN
CHAR
DATE
DECIMAL
FLOAT
INTEGER
LONG
REAL
SHORT
SMALLINT
TIME
TIMESTAMP_LTZ

数据类型名称
TIMESTAMP_NTZ
TINYINT
VARCHAR

限制：

- 对于合成数据生成，预测列的最大数目为 1。
- 如果目标列是分类列，则原始数据集中的最大类别数为 100。
- 在原始数据集中，行数必须介于 1,500 到 250 万之间，最大列数为 1.000。对于目标列中的非空值，最小行数为 1,000。

隐私指标：

- Clean Rooms ML 提供了一个隐私分数，用于衡量生成的合成数据对成员资格推断攻击的保护程度（MIAs）。该服务保留了合成过程中原始数据的5%来计算该分数。
- 接近 50% 的分数被认为是不错的；分数越高表示防御能力越差 MIAs。分数明显低于 50% 的情况很少见，这可能是由于合成数据中未显示原始数据的模式。

下游自定义模型：

- 在 Clean Rooms ML 中生成的合成数据最适合训练二元分类模型和最多包含五个类别的多类分类模型。
- 根据均方根误差 (RMSE) 的测量，使用在 Clean Rooms ML 中生成的合成数据训练回归模型可能会导致模型精度降低。

在 AWS Clean Rooms ML 中创建和加入合作

协作创建者负责创建协作、邀请成员和分配其角色。根据协作的设置方式，受邀成员加入协作并指定结果设置、训练模型工件目的地设置并承担付款责任。

为机器学习创建协作模式

以下过程演示如何创建机器学习协作、邀请一个或多个成员以及分配可以开始模型训练、接收结果、接收经过训练的模型结果（包括模型工件和指标）以及接收模型推理结果的成员。协作创建者还会分配一名成员，该成员将支付查询计算、模型训练和模型推理费用。

Console

为机器学习创建协作模式（控制台）

1. 创建协作并邀请一个或多个成员加入协作
2. 使用查询和作业为分析分配以下成员权限：
   - 将 Run 查询分配给将开始模型训练的成员。
   - 将接收分析结果分配给将接收查询结果的成员。
3. 使用专门构建的工作流程为机器学习建模分配以下成员能力：
   - 将接收训练模型的输出分配给将接收训练模型结果的成员，包括模型工件和指标。
   - 将接收模型推理的输出分配给将接收模型推理结果的成员。
4. 对于配置付款，请指定将支付查询计算、模型训练、模型推理和合成数据集生成成本的成员。

   合成数据集与其所基于的原始数据集具有相似的统计特性，但不包含原始数据集中存在的真实观测结果。通过使用隐私增强型合成数据集，您可以解锁以前数据隐私问题阻止的新机器学习模型训练用例。有关更多信息，请参阅隐私增强型合成数据生成。

   这些费用中的每一项都可以分配给相同或不同的成员。如果受邀成员是负责支付付款费用的会员，则他们必须在加入合作之前承担付款责任。
5. 对于配置成员资格，协作创建者可以决定立即加入成员资格或稍后创建成员资格。然后，协作创建者必须设置 ML 配置。

   a. 如果协作创建者也是结果接收者，则他们还必须在 “结果” 设置默认值中指定查询结果的目标和格式。

b. 机器学习配置为 Clean Rooms ML 提供了向发布指标的角色 AWS 账户。如果协作创建者也在接收经过训练的模型项目，他们可以指定用于接收结果的 Amazon S3 存储桶。

c. 在“机器学习配置”部分，选择“创建 ML 配置”，然后在 Amazon S3 上指定模型输出目标以及访问此位置所需的服务访问角色。

d. 如果协作创建者是负责支付费用的成员，则他们必须在创建协作之前接受自己的付款责任。

API

为机器学习 (API) 创建协作模式

1. 创建协作并邀请一个或多个成员加入协作
2. 为协作成员分配以下角色：
   - CAN_QUERY-分配给将开始模型训练和推理的成员。
   - CAN_RECEIVE_MODEL_OUTPUT-分配给将获得经过训练的模型结果的成员。
   - CAN_RECEIVE_INFERENCE_OUTPUT-分配给将接收模型推理结果的成员。

   如果协作创建者也是结果接收者，则他们还必须在创建协作期间指定查询结果的目标和格式。他们还为服务角色提供亚马逊资源名称 (ARN)，用于将结果写入查询结果目的地。
3. 指定将支付查询计算、模型训练和模型推理成本的成员。这些费用中的每一项都可以分配给相同或不同的成员。如果受邀成员是负责支付付款费用的会员，则他们必须在加入合作之前承担付款责任。
4. 以下代码创建协作，邀请可以运行查询和接收结果的成员，并将协作创建者指定为模型构件接收者。

```
import boto3
acr_client= boto3.client('cleanrooms')

collaboration = a_acr_client.create_collaboration(
    members=[
        {
         'accountId': 'invited_member_accountId',
         'memberAbilities':["CAN_QUERY","CAN_RECEIVE_RESULTS"],
         'displayName': 'member_display_name'
        }
    ],
    name='collaboration_name',
    description=collaboration_description,
```

```
    creatorMLMemberAbilities= {
        'customMLMemberAbilities':["CAN_RECEIVE_MODEL_OUTPUT",
 "CAN_RECEIVE_INFERENCE_OUTPUT"],
    },
    creatorDisplayName='creator_display_name',
    queryLogStatus="ENABLED",
    analyticsEngine="SPARK",
    creatorPaymentConfiguration={
        "queryCompute": {
            "isResponsible": True
        },
        "machineLearning": {
            "modelTraining": {
                "isResponsible": True
            },
            "modelInference": {
                "isResponsible": True
            }
        }
    }
)

collaboration_id = collaboration['collaboration']['id']
print("collaborationId: {collaboration_id}")

member_membership = a_acr_client.create_membership(
    collaborationIdentifier = collaboration_id,
    queryLogStatus = 'ENABLED',
    paymentConfiguration={
        "queryCompute": {
            "isResponsible": True
        },
        "machineLearning": {
            "modelTraining": {
                "isResponsible": True
            },
            "modelInference": {
                "isResponsible": True
            }
        }
    }
)
```

5. 然后，协作创建者必须设置 ML 配置。机器学习配置为 Clean Rooms ML 提供了一个向发布指标和日志的角色 AWS 账户。如果协作创建者也在接收结果（模型工件或推理结果），他们可以指定用于接收结果的 Amazon S3 存储桶。

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.put_ml_configuration(
        membershipIdentifier=membership_id,
        defaultOutputLocation={
           'roleArn':'arn:aws:iam::account:role/roleName',
           'destination':{
                's3Destination':{
                    's3Uri':"s3://bucketName/prefix"
                }
            }
        }
    )
```

加入协作

协作创建者完成任务后，受邀成员必须完成任务。

Console

创建成员资格并加入协作（控制台）

1. 受邀成员创建成员资格并加入协作。
2. 如果受邀成员是负责支付费用的成员，包括查询计算、模型训练和模型推理费用，则他们必须在加入协作之前接受自己的付款责任。
3. 受邀成员设置机器学习配置，该配置为 Clean Rooms ML 提供了向发布模型指标的角色 AWS 账户。如果他们也是接收经过训练的模型构件的成员，则他们必须提供一个用于存储经过训练的模型工件的 Amazon S3 存储桶。

API

创建成员资格并加入协作 (API)

1. 如果受邀成员也是可以接收结果的成员，则他们会指定查询结果的目标和格式。它们还提供服务角色 ARN，允许服务写入查询结果目标

   如果受邀成员是负责支付费用的成员，包括查询计算、模型训练和模型推理费用，则他们必须在加入协作之前接受自己的付款责任。

   如果受邀成员是负责为自定义建模支付模型训练和模型推理费用的成员，则他们必须在加入协作之前接受自己的付款责任。

   以下代码创建启用了查询日志记录的成员资格。

```
import boto3
acr_client= boto3.client('cleanrooms')

acr_client.create_membership(
    membershipIdentifier='membership_id',
    queryLogStatus='ENABLED'
)
```

2. 受邀成员设置机器学习配置，该配置为 Clean Rooms ML 提供了向发布模型指标的角色 AWS 账户。如果他们也是接收经过训练的模型构件的成员，则他们必须提供用于存储经过训练的模型工件的 Amazon S3 存储桶。

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.put_ml_configuration(
        membershipIdentifier='membership_id',
        defaultOutputLocation={
           'roleArn':"arn:aws:iam::account:role/role_name",
           'destination':{
                's3Destination':{
                    's3Uri':"s3://bucket_name/prefix"
                }
            }
        }
    )
```

在 AWS Clean Rooms ML 中贡献训练数据

在协作创建者创建协作并且受邀成员加入后，您就可以为协作贡献训练数据了。任何成员都可以贡献训练数据。

Console

贡献训练数据（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择 表。
3. 在 “表” 页面上，选择 “配置新表”。
4. 对于配置新表，对于数据源，选择 A mazon S3、Amazon A thena 或 Snowflake，然后根据您的数据源完成以下步骤：

如果你正在使用	那么
Amazon S3	1. 从下拉列表中选择一个数据库，然后从数据库中选择表。 2. 对于协作中允许的列，选择所有列或自定义列表。 3. 有关已配置表的详细信息，请提供该表的名称和可选描述。 4. 如果要报告模型指标，请输入指标的名称和将搜索输出日志以查找指标的 Regex 语句。 5. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。
Amazon Athena	1. 从下拉列表中选择一个数据库，然后从数据库中选择表。 2. 对于协作中允许的列，选择所有列或自定义列表。 3. 有关已配置表的详细信息，请提供该表的名称和可选描述。

如果你正在使用	那么
	4. 如果要报告模型指标，请输入指标的名称和将搜索输出日志以查找指标的 Regex 语句。 5. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。
Snowflake	1. 使用现有密钥 ARN 或存储此表的新密钥指定 Sn owflake 凭证。 2. 要获取 Snowflake 表和架构的详细信息，请手动输入详细信息或自动导入详细信息。 3. 对于架构，输入列名并从下拉列表中选择数据类型。 4. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。

5. 选择配置新表。
6. 在表格详细信息页面上，选择配置分析规则，为该表配置自定义分析规则。自定义分析规则限制对数据的访问权限。您可以允许对您的数据进行一组特定的预授权查询，也可以允许一组特定的账户查询您的数据。

   a. 对于分析规则类型，选择自定义，对于创建方法，选择引导流程。

   b. 选择下一步。

   c. 在 “指定分析控件” 中，在 “查看每项新分析” 和 “允许特定合作者进行任何分析” 之间进行选择。

   d. 选择下一步。

   e. （可选）对于指定分析结果控件，对于输出中不允许的列，请指定是否要从输出中排除任何列。如果选择 “无”，则不会从输出中排除任何列。如果选择 “自定义列表”，则可以指定将从输出中删除的某些列。

   f. 对于应用于输出的其他分析，请指定在生成结果之前是允许、拒绝还是要求进行其他分析。

   g. 选择下一步。

   h. （可选）在 “设置差异隐私” 中，选择 “关闭”。

i. 选择下一步。

j. 查看查看和配置页面上的信息，然后选择配置分析规则。

7. 在表格详细信息页面中，选择关联到协作。
8. 在 “关联表” 对话框中，选择要将此表格关联到的协作，然后选择 “选择协作”。
9. 在 “关联表” 页面上，查看并验证表关联详细信息、服务访问权限和标签中的信息。选择关联表。
10. 在 “与您关联的表” 表中，选择刚才关联的表旁边的单选按钮。从 “操作” 菜单中，选择 “协作分析规则” 组中的 “配置”。
11. 在 “配置协作分析规则” 页面上，在 “允许的其他分析” 中，选择是否有任何协作成员或特定协作成员可以执行其他分析。

    对于结果交付，请选择允许哪些成员接收来自查询输出的结果。
12. 选择配置分析规则。

API

贡献训练数据 (API)

1. AWS Clean Rooms 通过提供 AWS Glue 表和可以使用的列，配置现有表以供在中使用。

   使用您的特定参数运行以下代码。

```
import boto3
acr_client= boto3.client('cleanrooms')

acr_client.create_configured_table(
    name='configured_table_name',
    tableReference= {
        'glue': {
            'tableName': 'glue_table_name',
            'databaseName': 'glue_database_name'
        }
    },
    analysisMethod="DIRECT_QUERY",
    allowedColumns=["column1", "column2", "column3",...]
)
```

2. 配置限制对数据的访问的自定义分析规则。您可以允许对您的数据进行一组特定的预授权查询，也可以允许一组特定的账户查询您的数据。

使用您的特定参数运行以下代码。

```
import boto3
acr_client= boto3.client('cleanrooms')

acr_client.create_configured_table_analysis_rule(
    configuredTableIdentifier='configured_table_id',
    analysisRuleType='CUSTOM',
    analysisRulePolicy= {
        'v1': {
            'custom': {
                'allowedAnalyses': ['ANY_QUERY'],
                'allowedAnalysisProviders': ['query_runner_account'],
                'additionalAnalyses': "REQUIRED"
            }
        }
    }
)
```

在此示例中，允许特定账户对数据运行任何查询，并且需要进行额外的分析。

3. 将已配置的表与协作关联，并为这些 AWS Glue 表提供服务访问角色。

使用您的特定参数运行以下代码。

```
import boto3
acr_client= boto3.client('cleanrooms')

acr_client.create_configured_table_association(
    name='configured_table_association_name',
    membershipIdentifier='membership_id',
    configuredTableIdentifier='configured_table_id',
    roleArn='arn:aws:iam::account:role/role_name'
)
```

> **Note**
>
> 此服务角色拥有对表的权限。只有代表可以查询的成员运行 AWS Clean Rooms 允许的查询时，才可以假设服务角色。任何协作成员（数据所有者除外）都无法访问协作中的底层表。数据所有者可以关闭差异隐私，使其表可供其他成员查询。

4. 最后，向配置的表关联添加分析规则。

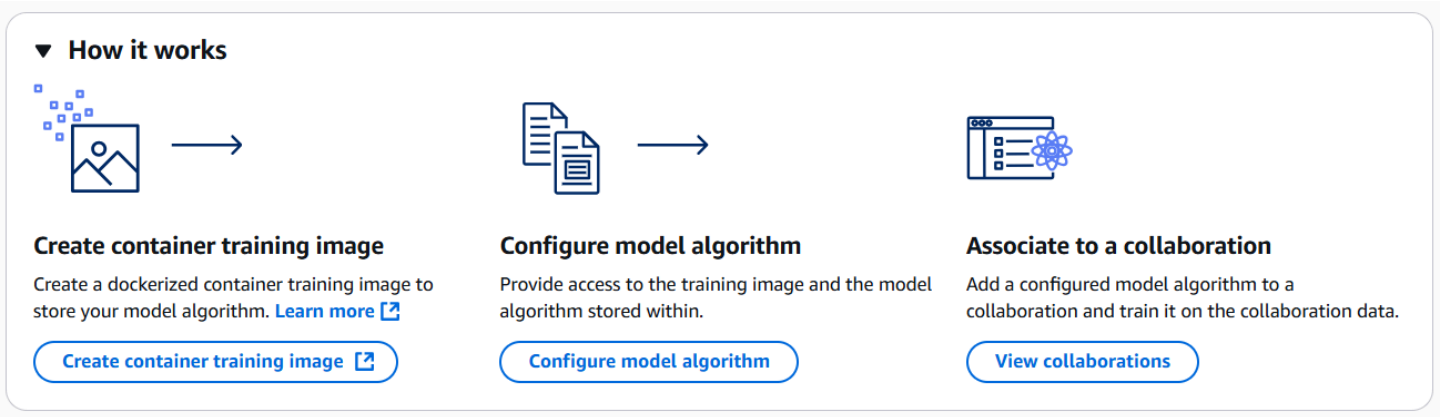
   使用您的特定参数运行以下代码。

```
import boto3
acr_client= boto3.client('cleanrooms')

acr_client.create_configured_table_association_analysis_rule(

 configuredTableAssociationIdentifier='configured_table_association_identifier',
    membershipIdentifier='membership_id',
    configuredTableIdentifier='configured_table_id',
    analysisRuleType = 'CUSTOM',
    analysisRulePolicy= {
        'v1': {
            'custom': {
                'allowedAdditionalAnalyses':
 ['configured_model_algorithm_association_arns'],
                'allowedResultReceivers': ['query_runner_account']
            }
        }
    }
)
```

在 AWS Clean Rooms ML 中配置模型算法

创建容器训练镜像后，必须配置模型算法。配置模型算法使其可用于关联到协作。

下图显示了将模型算法配置为在创建容器训练映像之后以及将其与协作关联之前发生的步骤。



Console

配置自定义 ML 模型算法（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择自定义 ML 模型。
3. 在自定义 ML 模型页面上，选择配置模型算法。
4. 在配置模型算法页面上，要了解模型算法的详细信息，请输入名称和可选的描述。
5. 如果要执行模型训练，有关训练图像 ECR 容器的详细信息，
   a. 选中 “指定训练图像 URI” 复选框。
   b. 从下拉列表中选择包含训练模型和/或推理容器的存储库。
   c. 选择图片。
   d. （可选）输入用于访问训练图像的入口点的值。
   e. （可选）输入参数的值。
6. （可选）如果要报告模型指标，请为训练指标输入指标的名称和正则表达式语句，该语句将搜索输出日志以查找指标。
7. 如果要执行模型推理，有关推理图像 ECR 容器的详细信息，
   a. 选中 “指定推理图像 URI” 复选框。
   b. 从下拉列表中选择存储库。
   c. 选择图片。
8. 对于服务访问，选择将用于访问该表的现有服务角色名称。
9. 对于加密，选择自定义加密设置以指定您自己的 KMS 密钥和相关信息。否则，Clean Rooms ML 将管理加密。
10. 如果要启用标签，请选择添加新标签，然后输入密钥和值对。
11. 选择 “配置模型算法”。

API

配置自定义 ML 模型算法 (API)

1. 创建与 A SageMaker I 兼容的 docker 镜像。Clean Rooms ML 仅支持与 SageMaker AI 兼容的 docker 镜像。

2. 创建与 A SageMaker I 兼容的 docker 镜像后，使用 Amazon ECR 创建训练镜像。按照 Amazon Elastic Container Registry 用户指南中的说明创建容器训练镜像。
3. 配置模型算法以在 Clean Rooms ML 中使用。您必须提供以下信息：
   - Amazon ECR 存储库链接以及用于训练模型和运行推理的其他参数。Clean Rooms ML 支持在推理容器上运行批量转换作业。
   - 允许 Clean Rooms ML 访问存储库的服务访问角色。
   - （可选）推理容器。尽管您可以在单独配置的模型算法中提供该算法，但我们建议您在此步骤中提供该算法，以便将训练和推理容器作为同一资源的一部分进行管理。

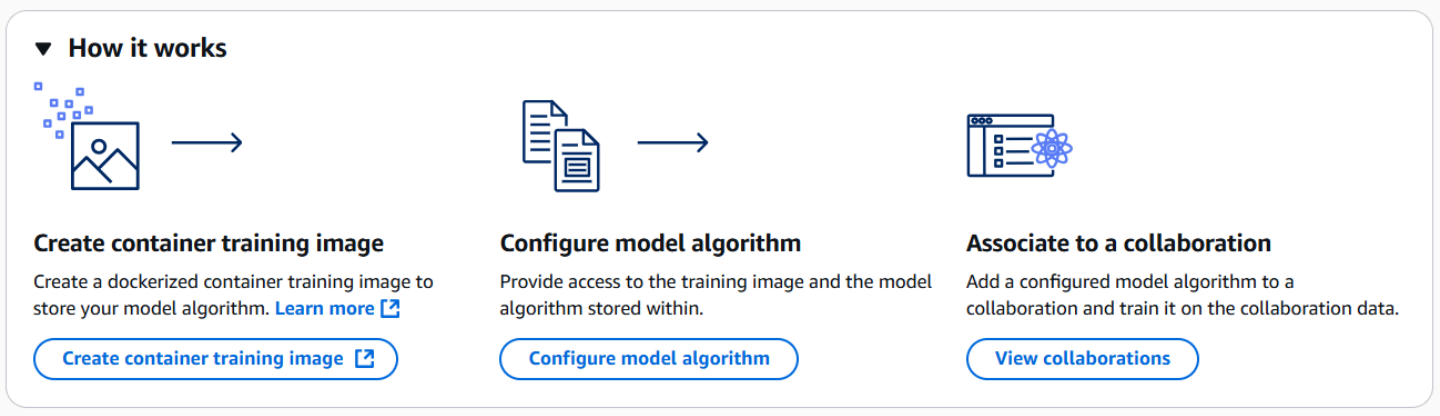
   使用您的特定参数运行以下代码。

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.create_configured_model_algorithm(
    name='configured_model_algorithm_name',
    trainingContainerConfig={
        'imageUri': 'account.dkr.ecr.region.amazonaws.com/image_name:tag',
        'metricDefinitions': [
            {
                'name': 'custom_metric_name_1',
                'regex': 'custom_metric_regex_1'
            }
        ]
    },
    inferenceContainerConfig={
        'imageUri':'account.dkr.ecr.region.amazonaws.com/image_name:tag',
    }
    roleArn='arn:aws:iam::account:role/role_name'
)
```

在 AWS Clean Rooms ML 中关联配置的模型算法

配置模型算法后，就可以将模型算法与协作关联了。关联模型算法后，该模型算法可供协作的所有成员使用。

下图显示了在创建容器训练镜像并配置模型算法之后，将配置的模型算法关联为最后一步。



Console

> **Note**
>
> 关联模型算法后，便无法对其进行编辑。要进行更改，可以删除关联的模型算法并关联新的算法。

关联自定义 ML 模型算法（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择自定义 ML 模型。
3. 在自定义机器学习模型页面上，选择要与协作关联的已配置模型算法，然后选择关联到协作。
4. 在 “关联已配置的模型算法” 窗口中，选择要关联的协作。
5. 选择选择协作。
6. 在 “关联模型算法” 页面上，输入模型算法关联的详细信息，输入名称和可选的描述。
7. 对于模型算法，选择已配置的模型算法。
8. 对于经过训练的模型导出隐私配置，
   a. 要导出模型文件，请选中模型文件复选框。
   b. 要导出输出文件，请选中 “输出文件” 复选框。
   c. 输入导出数据的最大大小值。该值必须介于 0.01 和 10 之间。
9. （可选）如果您想向成员发送完整的错误日志或更短的错误摘要，请在训练有素的模型推理作业隐私配置下，

a. 在 “完整日志” 下， IDs从下拉列表中选择一个或多个帐户。
b. （可选）如果要发送与筛选模式匹配的日志，请输入筛选模式。
c. （可选）如果要添加其他帐户和可选的筛选模式，请选择添加日志策略。
d. 在 “错误摘要” 下， IDs从下拉列表中选择一个或多个账户。
e. （可选）选择一个或多个要密文的实体，以指定将从错误日志或错误摘要中删除哪些实体。

- PII — 编辑个人身份信息
- 数字 — 编辑数字
- 自定义 — 根据自定义密文模式进行密文

i. 如果您在上一步中选择了自定义，请输入自定义密文模式。这将记录与此模式匹配的信息。
ii. （可选）如果要添加其他自定义密文图案，请选择添加其他自定义图案。

10. （可选）如果要配置经过训练的模型指标，请在 “训练模型指标配置” 下，从下拉列表中选择噪音水平。

    您可以选择 “无”、“低”、“中” 和 “高”。
11. （可选）如果要设置最大构件大小，请在构件配置下输入最大工件大小值。该值必须介于 0.01 和 10 之间。
12. （可选）如果要启用标签，请选择添加新标签，然后输入密钥和值对。
13. 选择关联 。

API

关联自定义 ML 模型算法 (API)

使用您的特定参数运行以下代码。

您还提供了一份隐私政策，该政策定义了谁有权访问不同的日志，允许客户定义正则表达式，以及可以从训练模型输出或推理结果中导出多少数据。

> **Note**
>
> 配置的模型算法关联是不可变的。

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.create_configured_model_algorithm_association(
    name='configured_model_algorithm_association_name',
    description='purpose of the association',
    configuredModelAlgorithmArn='arn:aws:cleanrooms-ml:region:account:membership/
membershipIdentifier/configured-model-algorithm/identifier',
    privacyConfiguration={
        "policies": {
            "trainedModelExports": {
                "filesToExport": ['files to export'],
                "containerLogs": [
                    {
                        "allowedAccountIds": ['member_account_id'],
                        "filterPattern": ['filter pattern'],
                        "logRedactionConfiguration": {
                            "entitiesToRedact": [
                                'ALL_PERSONALLY_IDENTIFIABLE_INFORMATION',
                                'NUMBERS',
                                'CUSTOM'
                            ],
                            "customEntityConfig": {
                                "customDataIdentifiers": [
                                    'custom_regex_1',
                                    'custom_regex_2'
                                ]
                            }
                        }
                    }
                ],
                "containerMetrics": {
                    "noiseLevel": 'noise value'
                },
                "maxArtifactSize": {
                    "unit": 'unit',
                    "value": 'number'
                }
            },
            "trainedModelInferenceJobs": {
                "containerLogs": [
                    {
                        "allowedAccountIds": ['member_account_id'],
```

```
                        "filterPattern": ['filter pattern'],
                        "logRedactionConfiguration": {
                            "entitiesToRedact": [
                                'ALL_PERSONALLY_IDENTIFIABLE_INFORMATION',
                                'NUMBERS',
                                'CUSTOM'
                            ],
                            "customEntityConfig": {
                                "customDataIdentifiers": [
                                    'custom_regex_1',
                                    'custom_regex_2'
                                ]
                            }
                        }
                    }
                ],
                "maxOutputSize": {
                    "unit": 'unit',
                    "value": 'number'
                }
            }
        }
    },
    tags={
        'tag': 'tag'
    }
)
```

将配置的模型算法与协作关联后，训练数据提供者必须向其表中添加协作分析规则。此规则允许配置的模型算法关联访问其配置的表。所有提供培训数据的提供者都必须运行以下代码：

```
import boto3
acr_client= boto3.client('cleanrooms')

acr_client.create_configured_table_association_analysis_rule(
    membershipIdentifier= 'membership_id',
    configuredTableAssociationIdentifier= 'configured_table_association_id',
    analysisRuleType= 'CUSTOM',
    analysisRulePolicy = {
        'v1': {
            'custom': {
                'allowedAdditionalAnalyses': ['arn:aws:cleanrooms-
ml:region:*:membership/*/configured-model-algorithm-association/*''],
```

```
                'allowedResultReceivers': []
            }
        }
    }
)
```

> **Note**
>
> 由于配置的模型算法关联是不可变的，因此我们建议想要将模型列入许可名单的训练数据提供者在自定义模型配置的前几次迭代中`allowedAdditionalAnalyses`使用通配符。这样，模型提供者就可以对其代码进行迭代，而无需其他训练提供者在使用数据训练更新后的模型代码之前重新关联。

在 AWS Clean Rooms ML 中创建机器学习输入通道

先决条件：

- 可以 AWS 账户 访问的 AWS Clean Rooms
- 您要在 AWS Clean Rooms 其中创建 ML 输入通道的协作设置
- 在协作中查询数据和创建机器学习输入通道的权限。
- （可选）用于与 ML 输入通道关联的现有模型算法，或创建新模型的权限
- （可选）包含可以针对您的指定模型运行的分析规则的表。
- （可选）用于生成数据集的现有 SQL 查询或分析模板
- （可选）具有相应权限的现有服务角色，或创建新服务角色的权限
- （可选）如果您想使用自己的加密 AWS KMS 密钥，请使用自定义密钥
- 在协作中创建和管理机器学习模型的适当权限

机器学习输入通道是根据特定数据查询创建的数据集。能够查询数据的成员可以通过创建 ML 输入通道为训练和推理做好数据准备。创建 ML 输入通道允许在同一个协作中将这些数据用于不同的训练模型。您应该为训练和推理创建单独的 ML 输入通道。

要创建 ML 输入通道，必须指定用于查询输入数据和创建 ML 输入通道的 SQL 查询。此查询的结果永远不会与任何成员共享，并且保持在 Clean Rooms ML 的范围内。在接下来的步骤中，将使用引用 Amazon 资源名称 (ARN) 来训练模型或运行推理。

Console

创建 ML 输入频道（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面上，选择要在其中创建机器学习输入渠道的协作。
4. 协作打开后，选择机器学习模型选项卡。
5. 在 “自定义 ML 模型” 下的 “机器学习输入通道” 部分中，选择 “创建 ML 输入通道”。
6. 在 “创建 ML 输入频道” 页面上，要获取 ML 输入频道的详细信息，请执行以下操作：

   a. 在 “名称” 中，输入频道的唯一名称。

   b. （可选）在描述中，输入您的频道的描述。

   c. 在 “关联模型算法” 中，选择要使用的算法。

      选择 “关联模型算法” 以添加新算法。
7. 对于数据集，选择一种生成训练数据集的方法：

   - 选择 SQL 查询可将 SQL 查询的结果用作训练数据集。

     如果您选择了 SQL 查询，请在 SQL 查询字段中输入您的查询。

     （可选）要导入您最近使用的查询，请选择从最近的查询中导入。
   - 选择分析模板以使用分析模板的结果作为训练数据集。

     > **Warning**
     >
     > 合成数据生成可以防止推断出个人属性，无论特定个体存在于原始数据集中，还是存在这些个体的学习属性。但是，它并不能阻止原始数据集中的文字值，包括个人身份信息 (PII) 出现在合成数据集中。
     > 我们建议避免输入数据集中仅与一个数据主体关联的值，因为这些值可能会重新识别数据主体。例如，如果只有一个用户居住在邮政编码中，则合成数据集中存在该邮政编码将确认该用户位于原始数据集中。诸如截断高精度值或用其他目录替换不常见的目录之类的技术可以用来降低这种风险。这些转换可以是用于创建 ML 输入通道的查询的一部分。

a. 如果没有关联表，请选择 “关联表” 以添加具有可针对指定模型运行的分析规则的表。

b. 选择创建此数据通道时要使用的工作器类型。默认的工作器类型为 CR.1X。指定要使用的员工人数。默认工作人员编号为 16。要指定火花属性，请执行以下操作：

   i. 展开 Spark 属性。

   ii. 选择 “添加 Spark 属性”。

   iii. 在 Spark 属性对话框中，从下拉列表中选择属性名称并输入值。

下表提供了每个属性的定义。

有关 Spark 属性的更多信息，请参阅 Apache Spark 文档中的 Spark 属性。

属性名称	说明	默认值
spark.task.maxFa	控制任务在失败之前可以连续失败多少次。需要一个大于或等于 1 的值。允许的重试次数等于该值减去 1。如果任何尝试成功，则失败计数将重置。不同任务的失败不会累积到这个极限。	4
spark.sql.files。maxPartitionBytes	设置从基于文件的源（例如 Parquet、JSON 和 ORC）读取数据时要打包到单个分区的最大字节数。	128MB
spark.hadoop.fs.s3.maxRetries	设置 Amazon S3 文件操作的最大重试次数。	
spark.network.	设置所有网络交互的默认超时时间。如果未配置，则覆盖以下超时设置： • Spark.storage。 blockManagerHeartbeatTimeoutMs • shuffle.io.connectionTimeout • Spark.rpc.askTimeout • spark.rpc.lookupTim	120

属性名称	说明	默认值
spark.rdd.com	指定是否使用 spark.io.compressi on.codec 压缩序列化的 RDD 分区。适用于 Java 和 Scala 中的 StorageLe vel .MEMORY_ONLY_SER，或 Python 中的.MEMORY_ONLY。 StorageLevel减少存储空间，但需要额外的 CPU 处理时间。	FALSE
Spark.shuffle.spil l.compress	指定是否使用 spark.io.compressi on.codec 压缩随机播放数据。	TRUE
spark.sql.自适应。 advisoryPartitionS izeInBytes	当 spark.sql.adaptive.enabled 为真时，设置自适应优化期间洗牌分区的目标大小（以字节为单位）。控制合并小分区或拆分倾斜分区时的分区大小。	（spark.sql.adaptiv e.shuffle 的值。 targetPostShuffleI nputSize)
spark.sql.自适应。 autoBroadcastJoin 阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。仅适用于自适应框架。使用与 spark.sql 相同的默认值。 autoBroadcastJoin阈值。设置为 -1 可禁用广播。	（无）
Spark.sql.adaptive .coalescePartition s.enabled	指定是否根据 spark.sql.adaptive 合并连续的洗牌分区。 advisoryPartitionS izeInBytes 以优化任务规模。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive .coalescePartive.c oal initialPa rtitionNum	定义合并前随机分区的初始数量。需要同时启用 spark.sql.adaptive.enabled 和 spark.sql.adaptive.coalescePartition s.enabled 才能成真。默认为 spark.sql .shuffle.partitions 的值。	（无）

属性名称	说明	默认值
Spark.sql.adaptive .coalescePartive.c oal minPartit ionSize	设置合并后的随机分区的最小大小，以防止自适应优化期间分区变得太小。	1 MB
Spark.sql.adaptive .coalescePartition s.parallelism Firs	指定是否根据集群并行度而不是 spark.sql .adaptive 来计算分区大小。 advisoryP artitionSizeInBytes 在分区合并期间。生成的分区大小小于配置的目标大小，以最大限度地提高并行度。我们建议在繁忙的群集上将其设置为 false，以通过防止过多的小任务来提高资源利用率。	TRUE
sql.adaptive.enabl ed	指定是否启用自适应查询执行，以便在查询执行期间根据准确的运行时统计数据重新优化查询计划。	TRUE
spark.sql.自适应。 forceOpti mizeSkewed加入	指定是否强制启用， OptimizeS kewedJoin 即使它引入了额外的随机播放。	FALSE
spark.sql.自适应。 localShuffleReader . 已启用	指定在不需要随机分区时（例如从排序合并联接转换为广播哈希联接之后）是否使用本地随机播放读取器。需要 spark.sql .adaptive.enabled 才为真。	TRUE

属性名称	说明	默认值
spark.sql.自适应。 maxShuffledHashJoinLocalMapThreshold	设置用于构建本地哈希映射的最大分区大小（以字节为单位）。在以下情况下，优先考虑洗牌后的哈希联接而不是排序合并联接： • 此值等于或超过 spark.sql.adaptive。advisoryPartitionSizeInBytes • 所有分区大小均在此限制范围内 覆盖 spark.sql.join。 preferSortMerge加入设置。	0 字节
spark.sql.自适应。 optimizeSkewsInRebalancePartitions.enabled	指定是否通过基于 spark.sql.adaptive 将倾斜的随机分区拆分为较小的分区来优化这些分区。 advisoryPartitionSizeInBytes。需要 spark.sql.adaptive.enabled 才为真。	TRUE
spark.sql.自适应。 rebalancePartitionsSmallPartitionFactor	定义拆分期间合并分区的大小阈值系数。小于此因子的分区乘以 spark.sql.adaptive。 advisoryPartitionSizeInBytes 已合并。	0.2
Spark.sql.adaptive.skewjoin.enable	指定是否通过拆分和可选复制倾斜的分区来处理洗牌联接中的数据倾斜。适用于排序合并和洗牌哈希联接。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive.skewJoin。 skewedPartitionFactor	确定决定分区偏斜的大小系数。当分区的大小超过两个分区时，分区就会出现偏差： • 该因子乘以分区大小中位数 • spark.sql.adaptive.skewJoin 的值。 skewedPartitionThresholdInBytes	5

属性名称	说明	默认值
Spark.sql.adaptive .skewJoin。 skewedPartitionThr esholdInBytes	设置用于识别偏斜分区的大小阈值（以字节为单位）。当分区的大小超过两个分区时，分区就会出现偏差： • 这个门槛 • 分区大小中位数乘以 spark.sql.adaptive .skewJoin。 skewedPartitionFactor 我们建议将此值设置为大于 spark.sql .adaptive。 advisoryPartitionSizeInByte s。	256MB
spark.sql。 autoBroadcastJoin 阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。设置为 -1 可禁用广播。	10MB
sql.broad castTimeout	控制广播加入期间广播操作的超时时间（以秒为单位）。	300 秒
spark.sql.cbo.enab led	指定是否为计划统计数据估算启用基于成本的优化 (CBO)。	FALSE
spark.sql.cbo.join reorder.dp.star.Fi lter	指定是否在基于开销的联接枚举期间应用星型联接过滤器启发式算法。	FALSE
spark.sql.cbo.join reorder.dp.Thresh	设置动态规划算法中允许的最大连接节点数。	12
Spark.sql.cbo.join reorder.enabled	指定是否在基于成本的优化 (CBO) 中启用联接重新排序。	FALSE
Spark.sql.cbo.plan stats.enabled	指定在逻辑计划生成期间是否从目录中提取行数和列统计信息。	FALSE

属性名称	说明	默认值
spark.sql.cbo。starSchemaDetection	指定是否启用基于星型架构检测的联接重新排序。	FALSE
spark.sql.files。maxPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最大数量。当初始计数超过此值时，重新缩放分区。这是建议的目标，而不是保证的上限。	（无）
spark.sql.files。maxRecordsPer文件	设置写入单个文件的最大记录数。如果设置为零或负值，则不适用任何限制。	0
spark.sql.files。minPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最小数量。默认为 spark.sql。 leafNodeDefault 并行性。这是建议的目标，而不是保证的上限。	（无）
spark.sql。inMemoryColumnarStorage.batchSize	控制列式缓存的批次大小。增加大小可以提高内存利用率和压缩率，但会增加 out-of-memory出错的风险。	10000
spark.sql。inMemoryColumnar存储. 已压缩	指定是否根据数据统计信息自动为列选择压缩编解码器。	TRUE
spark.sql。inMemoryColumnar存储。 enableVectorizedReader	指定是否为列式缓存启用矢量化读取。	TRUE

属性名称	说明	默认值
Spark.sql.legacy。allowHash OnMapType	指定是否允许对地图类型数据结构进行哈希操作。此传统设置保持了与旧版 Spark 地图类型处理的兼容性。	
Spark.sql.legacy。allowNega tiveScale OfDecimal	指定是否允许在十进制类型定义中使用负比例值。此传统设置保持了与支持负十进制小数位数的旧 Spark 版本的兼容性。	
Spark.sql.legacy。castCompl exTypesTo String.enabled	指定是否启用将复杂类型转换为字符串的传统行为。保持与旧版 Spark 的类型转换规则的兼容性。	
Spark.sql.legacy。charVarcharAs字符串	指定是否将 CHAR 和 VARCHAR 类型视为字符串类型。此传统设置提供了与旧版 Spark 的字符串类型处理的兼容性。	
Spark.sql.legacy。createEmp tyCollectionUsingS tringType	指定是否使用字符串类型元素创建空集合。此传统设置保持了与旧版 Spark 的集合初始化行为的兼容性。	
Spark.sql.legacy。exponentLiteralAsd ecimal. 已启用	指定是否将指数文字解释为十进制类型。此传统设置保持了与旧版 Spark 的数字文字处理的兼容性。	
spark.sql.legacy.j son。 allowEmpt yString. 已启用	指定是否允许在 JSON 处理中使用空字符串。此传统设置保持了与旧版 Spark 的 JSON 解析行为的兼容性。	
spark.sql.legacy.p arquet.int96 RebaseModelRead	指定在读取 Parquet 文件时是否使用传统 INT96 的时间戳变基模式。此传统设置保持了与旧版 Spark 的时间戳处理的兼容性。	

属性名称	说明	默认值
Spark.sql.legacy。timeParserPolicy	控制时间解析行为以实现向后兼容。此传统设置决定了如何从字符串中解析时间戳和日期。	
Spark.sql.legacy.typeCorcion。datetimeToString.已启用	指定在将日期时间值转换为字符串时是否启用传统类型强制行为。保持与旧版 Spark 版本的日期时间转换规则的兼容性。	
spark.sql。maxSinglePartition字节	设置最大分区大小（以字节为单位）。规划器为较大的分区引入了洗牌操作以提高并行度。	128m
Spark.sql.metadataCacheTTLSeconds	控制元数据缓存的 time-to-live (TTL)。适用于分区文件元数据和会话目录缓存。需要： • 大于零的正值 • Spark.sql.catalogCatalog实现设置为蜂巢 • spark.sql.hive。 filesourcePartitionFileCacheSize 大于零 • spark.sql.hive。 manageFilesourcePartitions 设置为 true	-1000 毫秒
sql.optimizer。collapseProjectAlways内联	指定是否折叠相邻的投影和行内表达式，即使这会导致重复。	FALSE
sql.optimizer。dynamicPartitionPruning. 已启用	指定是否为用作联接键的分区列生成谓词。	TRUE
sql.optimizer。enableCsvExpression优化	指定是否通过从 from_csv 操作中删除不必要的列来优化 SQL 优化器中的 CSV 表达式。	TRUE

属性名称	说明	默认值
sql.optimizer。enableJsonExpression优化	通过以下方式指定是否优化 SQL 优化器中的 JSON 表达式： • 从 from_json 操作中删除不必要的列 • 简化 from_json 和 to_json 的组合 • 优化 named_struct 操作	TRUE
spark.sql.Optimizer.ExcludedR	定义要禁用的优化器规则，由逗号分隔的规则名称标识。某些规则无法禁用，因为它们是正确性所必需的。优化器会记录哪些规则已成功禁用。	（无）
spark.sql.Optimizer.runtime.bloomFilapplicationSideScanSizeThreshold	设置在应用程序端注入 Bloom 过滤器所需的最小聚合扫描大小（以字节为单位）。	10GB
spark.sql.Optimizer.runtime.bloomFilcreationSideThreshold	定义在创建端注入 Bloom 滤镜的最大大小阈值。	10MB
Spark.sql.Optimizer.runtime.bloomFilter.enable	指定当随机连接的一侧具有选择性谓词时，是否插入布隆过滤器以减少随机播放数据。	TRUE
spark.sql.Optimizer.runtime.bloomFilexpectedNumItems	定义运行时 Bloom 过滤器中预期项目的默认数量。	1000000
spark.sql.Optimizer.runtime.bloomFilmaxNumBits	设置运行时 Bloom 过滤器中允许的最大位数。	67108864

属性名称	说明	默认值
spark.sql.Optimizer.runtime.bloomFilmaxNumItems	设置运行时 Bloom 过滤器中允许的最大预期项目数。	4000000
spark.sql.Optimizer.runtime.bloomFilter.number.	限制每次查询允许的非 DPP 运行时过滤器的最大数量，以防止驱动程序 out-of-memory出错。	10
Spark.sql.Optimizer.runtime.bloomfilter.numbit	定义运行时布隆过滤器中使用的默认位数。	8388608
spark.sql.optimizer.runtim rowlevelOperationGroup过滤器. 已启用	指定是否为行级操作启用运行时组筛选。允许数据源： • 使用数据源筛选器修剪整组数据（例如文件或分区） • 执行运行时查询以识别匹配的记录 • 丢弃不必要的组以避免昂贵的重写 限制： • 并非所有表达式都可以转换为数据源筛选器 • 有些表达式需要 Spark 求值（例如子查询）	TRUE
Spark.sql.Optimizer.runtimeFiltersemiJoinReduction. 已启用	指定当随机连接的一侧具有选择性谓词时，是否插入半联接以减少随机播放数据。	FALSE

属性名称	说明	默认值
spark.sql.parquet.AgregatePus	指定是否将聚合向下推送到 Parquet 进行优化。支持： • 布尔型、整数、浮点型和日期类型的最小值和最大值 • 所有数据类型的计数 如果任何 Parquet 文件页脚中缺少统计信息，则会抛出异常。	FALSE
sql.parquet。columnarReaderBatch大小	控制每个 Parquet 矢量化阅读器批次中的行数。选择一个平衡性能开销和内存使用量的值，以防止 out-of-memory出错。	4096
Spark.sql.session.time	定义会话时区，用于处理字符串文字中的时间戳和 Java 对象转换。接受： • 以地区为基础 IDs 的 area/city 格式（例如美国/洛杉矶） • 区域偏移量采用 (+/-) HH、(+/-) HH: mm 或 (+/-) HH: mm: ss 格式（例如 -08 或 + 01:00） • UTC 或 Z 作为 + 00:00 的别名	（当地时区的值）
spark.sql.shuffle.part	设置联接或聚合期间用于数据洗牌的默认分区数。无法在结构化流式查询从同一检查点位置重新启动之间进行修改。	200
spark.sql。shuffledHashJoin因子	定义用于确定 shuffle 哈希加入资格的乘法系数。当小边数据大小乘以此系数小于大边数据大小时，将选择随机哈希联接。	3

属性名称	说明	默认值
火花.sql.sources。parallelPartitionDiscovery. 阈值	使用基于文件的源（Parquet、JSON 和 ORC）设置驱动端文件列表的最大路径数。如果在分区发现期间超出限制，则使用单独的 Spark 分布式作业列出文件。	32
spark.sql.statistics.histics.h	指定是否在列统计数据计算期间生成等高直方图以提高估计精度。除了基本列统计数据所需的扫描之外，还需要进行额外的表扫描。	FALSE

c. 对于以天为单位的数据保留，请输入数据的保留天数。

d. 对于结果格式，选择 CSV 或 Par q uet 作为 ML 输入通道应使用的数据格式。

8. 对于服务访问权限，请选择将用于访问此表的现有服务角色名称，或者选择创建并使用新的服务角色。
9. 对于加密，选择使用自定义 KMS 密钥加密密钥以指定您自己的 KMS 密钥和相关信息。否则，Clean Rooms ML 将管理加密。
10. 选择 “创建 ML 输入通道”。

    创建 ML 输入通道需要几分钟。您可以在 “机器学习模型” 选项卡上查看机器学习输入通道列表。

> **Note**
>
> 创建 ML 输入通道后，您无法对其进行编辑。

API

创建 ML 输入频道 (API)

使用您的特定参数运行以下代码：

```
import boto3
acr_client = boto3.client('cleanroomsml')

acr_client.create_ml_input_channel(
```

```
    name="ml_input_channel_name",
    membershipIdentifier='membership_id',

 configuredModelAlgorithmAssociations=[configured_model_algorithm_association_arn],
    retentionInDays=1,
    inputChannel={
        "dataSource": {
            "protectedQueryInputParameters": {
                "sqlParameters": {
                    "queryString": "select * from table",
                    "computeConfiguration": {
                        "worker": {
                            "type": "CR.1X",
                            "number": 16,
                            "properties": {
                                "spark": {
                                    "spark configuration key": "spark configuration
 value",
                                }
                            }
                        },
                    "resultFormat": "PARQUET"
                }
            }
        },
        "roleArn": "arn:aws:iam::111122223333:role/role_name"
    }
)
channel_arn = resp['ML Input Channel ARN']
```

在 AWS Clean Rooms ML 中创建经过训练的模型

先决条件：

- 可以 AWS 账户 访问的 AWS Clean Rooms
- 合作建立在 AWS Clean Rooms
- 与协作相关的已配置模型算法
- 至少有一个已配置的 ML 输入通道
- 在协作中创建和管理机器学习模型的适当权限

将配置的模型算法与协作关联，然后创建并配置机器学习输入通道后，就可以创建经过训练的模型了。协作成员使用经过训练的模型来共同分析其数据。

您可以使用以下步骤创建经过训练的模型。

或者，您可以使用增量训练使用新数据改进现有模型，或者使用分布式训练来跨多个计算实例训练模型。

主题

- 在 AWS Clean Rooms ML 中使用增量训练
- 在 AWS Clean Rooms ML 中使用分布式培训

Console

创建经过训练的模型（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面上，选择要在其中创建训练模型的协作。
4. 协作打开后，选择机器学习模型选项卡。
5. 在 “自定义 ML 模型” 下的 “经过训练的模型” 部分中，选择创建经过训练的模型。
6. 在创建训练模型页面上，为关联模型算法指定算法。
7. 有关训练模型的详细信息，请输入以下内容：

   a. 在名称中，输入协作中模型的唯一名称。

   b. （可选）在描述中，输入训练模型的描述。

   c. 对于训练数据输入模式，请选择以下选项之一：

      - 如果您的数据集较小，可以容纳机器学习存储卷，并且您更喜欢使用传统的文件系统访问权限来访问训练脚本，请选择 “文件”。
      - 对于大型数据集，选择 Pip e 可以直接从 S3 流式传输数据，无需将所有内容下载到磁盘，这样可以提高训练速度并降低存储需求。
      - FastFile如果要将从 S3 进行流式传输的优势与文件系统访问相结合，尤其是在顺序读取数据或处理较少文件以缩短启动时间时，请选择此选项。
8. 要了解 ML 输入通道的详细信息，请执行以下操作：

a. 对于 ML 输入通道，请指定为模型算法提供数据的 ML 输入通道。

要添加其他频道，请选择添加另一个 ML 输入频道。您最多可以添加 19 个额外的 ML 输入通道。

b. 在频道名称中，输入 ML 输入频道的名称。

c. 对于 Amazon S3 数据分配类型，请选择以下选项之一：

- 选择 “完全复制”，为每个训练实例提供数据集的完整副本。当您的数据集足够小以容纳内存时，或者当每个实例都需要访问所有数据时，这种方法效果最好。
- 选择 “按 S3 密钥分片”，根据 S3 密钥将您的数据集划分到训练实例。每个实例接收大约 S3 对象总数的 1/n，其中 'n' 是实例数。这最适合您想要并行处理的大型数据集。

> **Note**
>
> 选择分布类型时，请考虑您的数据集大小和训练要求。完全复制可提供完整的数据访问权限，但需要更多存储空间，而 Sharded by S3 密钥支持对大型数据集进行分布式处理。

9. 在最长训练持续时间中，选择要训练模型的最大时间。
10. 对于超参数，请指定任何特定于算法的参数及其预期值。超参数特定于正在训练的模型，用于微调模型训练。
11. 对于环境变量，请指定任何特定于算法的变量及其预期值。环境变量是在 Docker 容器中设置的。
12. 对于加密，要使用自定义密钥 AWS KMS key，请选中使用自定义 KMS 密钥加密密钥复选框。
13. 对于 EC2 资源配置，请指定有关用于模型训练的计算资源的信息。

    a. 在实例类型中，选择要运行的实例类型。

    b. 在实例数中，输入实例数。

    c. 对于以 GB 为单位的卷大小，请输入 ML 存储卷大小。
14. 选择创建经过训练的模型。

API

创建经过训练的模型 (API)

能够训练模型的成员通过选择 ML 输入通道和模型算法开始训练。

使用您的特定参数运行以下代码：

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.create_trained_model(
    membershipIdentifier= 'membership_id',
    configuredModelAlgorithmAssociationArn = 'arn:aws:cleanrooms-
ml:region:account:membership/membershipIdentifier/configured-model-algorithm-
association/identifier',
    name='trained_model_name',
    resourceConfig={
        'instanceType': "ml.m5.xlarge",
        'volumeSizeInGB': 1
    },
    dataChannels=[
        {
            "mlInputChannelArn": channel_arn_1,
            "channelName": "channel_name"
        },
        {
            "mlInputChannelArn": channel_arn_2,
            "channelName": "channel_name"
        }
    ]
)
```

> **Note**
>
> 创建训练后的模型后，您无法对其进行编辑。要进行更改，请删除经过训练的模型并创建一个新模型。

在 AWS Clean Rooms ML 中使用增量训练

先决条件：

- 可以 AWS 账户 访问的 AWS Clean Rooms
- 协作中现有的经过训练的模型

- 用于增量训练的新数据集或更新的数据集
- 在协作中创建和管理机器学习模型的适当权限
- 熟悉现有模型的超参数和配置

通过增量训练，您可以使用现有模型的构件和更新的数据集来训练新模型。增量训练可节省时间和资源。

使用增量训练可以：

- 使用扩展的数据集训练新模型，该数据集具有早期训练中未考虑的底层模式。
- 训练模型的多个变体，要么使用不同的超参数，要么使用不同的数据集。

Console

运行增量训练作业（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面上，选择要用于增量训练的模型构件所在的协作。
4. 协作打开后，选择机器学习模型选项卡。
5. 在 “自定义 ML 模型” 下的 “训练模型” 部分中，选择要增量训练的已训练模型旁边的单选按钮。
6. 在 “概述” 页面的 “版本” 下，
   a. 选择要进行增量训练的训练模型旁边的单选按钮。
   b. 从版本中选择火车。
7. 在从版本创建训练模型页面上，对于经过训练的模型版本，选择版本。

   将自动选择基本模型版本。如果存在其他版本，则可以更改此版本。
8. 有关训练模型的详细信息，请输入以下内容：
   a. 在名称中，输入协作中模型的唯一名称。
   b. （可选）在描述中，输入训练模型的描述。
   c. 对于训练数据输入模式，请选择以下选项之一：

- 如果您的数据集较小，可以容纳机器学习存储卷，并且您更喜欢使用传统的文件系统访问权限来访问训练脚本，请选择 “文件”。
- 对于大型数据集，选择 Pip e 可以直接从 S3 流式传输数据，无需将所有内容下载到磁盘，这样可以提高训练速度并降低存储需求。
- FastFile如果要将从 S3 进行流式传输的优势与文件系统访问相结合，尤其是在顺序读取数据或处理较少文件以缩短启动时间时，请选择此选项。

d. 在增量训练频道名称中，输入增量训练频道的名称

> **Note**
>
> 如果您指定增量训练通道名称但没有版本 ID，则系统将使用基础模型进行增量训练。

9. 要了解 ML 输入通道的详细信息，请执行以下操作：

a. 对于 ML 输入通道，请指定为模型算法提供数据的 ML 输入通道。

要添加其他频道，请选择添加另一个 ML 输入频道。您最多可以添加 19 个额外的 ML 输入通道。

b. 在频道名称中，输入 ML 输入频道的名称。

c. 对于 Amazon S3 数据分配类型，请选择以下选项之一：

- 选择 “完全复制”，为每个训练实例提供数据集的完整副本。当您的数据集足够小以容纳内存时，或者当每个实例都需要访问所有数据时，这种方法效果最好。
- 选择 “按 S3 密钥分片”，根据 S3 密钥将您的数据集划分到训练实例。每个实例接收大约 S3 对象总数的 1/n，其中 'n' 是实例数。这最适合您想要并行处理的大型数据集。

> **Note**
>
> 选择分布类型时，请考虑您的数据集大小和训练要求。完全复制可提供完整的数据访问权限，但需要更多存储空间，而 Sharded by S3 密钥支持对大型数据集进行分布式处理。

10. 在最长训练持续时间中，选择要训练模型的最大时间。
11. 对于超参数，请指定任何特定于算法的参数及其预期值。超参数特定于正在训练的模型，用于微调模型训练。

12. 对于环境变量，请指定任何特定于算法的变量及其预期值。环境变量是在 Docker 容器中设置的。
13. 对于加密，要使用自定义密钥 AWS KMS key，请选中使用自定义 KMS 密钥加密密钥复选框。
14. 对于 EC2 资源配置，请指定有关用于模型训练的计算资源的信息。
    a. 在实例类型中，选择要运行的实例类型。
    b. 在实例数中，输入实例数。
    c. 对于以 GB 为单位的卷大小，请输入 ML 存储卷大小。
15. 选择根据版本创建经过训练的模型。

API

运行增量训练作业 (API)

使用您的特定参数运行以下代码：

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.create_trained_model(
    membershipIdentifier= 'membership_id',
    configuredModelAlgorithmAssociationArn = 'arn:aws:cleanrooms-
ml:region:account:membership/membershipIdentifier/configured-model-algorithm-
association/identifier',
    name='trained_model_name',
    resourceConfig={
        'instanceType': 'ml.m5.xlarge',
        'volumeSizeInGB': 1
    },
    incrementalTrainingDataChannels=[
        {
            'trainedModelArn': trained_model_arn,
            'channelName': 'channel_name'
        },
    ]
    dataChannels=[
        {
            'mlInputChannelArn': channel_arn_1,
            'channelName': 'channel_name'
```

```
        },
        {
            'mlInputChannelArn': channel_arn_2,
            'channelName': 'channel_name'
        }
    ]
)
```

> **Note**
>
> 限制：总共最多 20 个频道（包括两个dataChannels和incrementalTrainingDataChannels）。

> **Note**
>
> 创建训练后的模型后，您无法对其进行编辑。要进行更改，请删除经过训练的模型并创建一个新模型。

在 AWS Clean Rooms ML 中使用分布式培训

先决条件：

- 可以 AWS 账户 访问的 AWS Clean Rooms
- 合作建立在 AWS Clean Rooms
- 支持分布式训练的配置模型算法
- 适用于分布式处理的大型数据集
- 在协作中创建和管理机器学习模型的适当权限
- 有足够的 Amazon EC2 配额来运行多个实例进行分布式训练

分布式训练利用许多并行运行的计算节点的强大功能来处理大量数据并有效地更新模型参数。

有关分布式训练的更多信息，请参阅 Amazon A SageMaker I 开发人员指南中的分布式训练概念。

Console

运行分布式训练作业（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面上，选择要在其中创建训练模型的协作。
4. 协作打开后，选择机器学习模型选项卡。
5. 在 “自定义 ML 模型” 下的 “经过训练的模型” 部分中，选择创建经过训练的模型。
6. 在创建训练模型页面上，为关联模型算法指定算法
7. 有关训练模型的详细信息，请输入以下内容：

   a. 在名称中，输入协作中模型的唯一名称。

   b. （可选）在描述中，输入训练模型的描述。

   c. 对于训练数据输入模式，请选择以下选项之一：

      - 如果您的数据集较小，可以容纳机器学习存储卷，并且您更喜欢使用传统的文件系统访问权限来访问训练脚本，请选择 “文件”。
      - 对于大型数据集，选择 Pip e 可以直接从 S3 流式传输数据，无需将所有内容下载到磁盘，这样可以提高训练速度并降低存储需求。
      - FastFile如果要将从 S3 进行流式传输的优势与文件系统访问相结合，尤其是在顺序读取数据或处理较少文件以缩短启动时间时，请选择此选项。
8. 要了解 ML 输入通道的详细信息，请执行以下操作：

   a. 对于 ML 输入通道，请指定为模型算法提供数据的 ML 输入通道。

      要添加其他频道，请选择添加另一个 ML 输入频道。您最多可以添加 19 个额外的 ML 输入通道。

   b. 在频道名称中，输入 ML 输入频道的名称。

   c. 对于 Amazon S3 数据分配类型，请选择以下选项之一：

      - 选择 “完全复制”，为每个训练实例提供数据集的完整副本。当您的数据集足够小以容纳内存时，或者当每个实例都需要访问所有数据时，这种方法效果最好。
      - 选择 “按 S3 密钥分片”，根据 S3 密钥将您的数据集划分到训练实例。每个实例接收大约 S3 对象总数的 1/n，其中 'n' 是实例数。这最适合您想要并行处理的大型数据集。

> **Note**
>
> 选择分布类型时，请考虑您的数据集大小和训练要求。完全复制可提供完整的数据访问权限，但需要更多存储空间，而 Sharded by S3 密钥支持对大型数据集进行分布式处理。

9. 在最长训练持续时间中，选择要训练模型的最大时间。
10. 对于超参数，请指定任何特定于算法的参数及其预期值。超参数特定于正在训练的模型，用于微调模型训练。
11. 对于环境变量，请指定任何特定于算法的变量及其预期值。环境变量是在 Docker 容器中设置的。
12. 对于加密，要使用自定义密钥 AWS KMS key，请选中使用自定义 KMS 密钥加密密钥复选框。
13. 对于 EC2 资源配置，请指定有关用于模型训练的计算资源的信息。
    a. 在实例类型中，选择要运行的实例类型。

       分布式训练支持的实例类型有：

       - ml.m5.4xlarge
       - ml.m5.12xlarge
       - ml.m5.2xlarge
       - ml.g5.12xlarge
       - ml.g5.24xlarge

    b. 在实例数中，输入实例数。
    c. 对于以 GB 为单位的卷大小，请输入 ML 存储卷大小。
14. 选择创建经过训练的模型。

API

运行分布式训练作业 (API)

使用您的特定参数运行以下代码：

```
import boto3
```

```
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.create_trained_model(
    membershipIdentifier= 'membership_id',
    configuredModelAlgorithmAssociationArn = 'arn:aws:cleanrooms-
ml:region:account:membership/membershipIdentifier/configured-model-algorithm-
association/identifier',
    name='trained_model_name',
    trainingInputMode: "File",
    resourceConfig={
        'instanceCount': "3"
        'instanceType': "ml.m5.xlarge",
        'volumeSizeInGB': 3
    },
    dataChannels=[
        {
            "mlInputChannelArn": channel_arn_1,
            "channelName": "channel_name",
            "S3DataDistributionType:" "FullyReplicated"
        }

    ]
)
```

> **Note**
>
> 创建训练后的模型后，您无法对其进行编辑。要进行更改，请删除经过训练的模型并创建一个新模型。

从 AWS Clean Rooms ML 中导出模型工件

此任务是可选的，应在您将CAN_RECEIVE_MODEL_OUTPUT成员权限分配给协作成员后完成。

模型训练完成后，训练模型的成员可以启动模型构件的导出。训练模型的成员选择谁将接收模型工件，前提是该成员可以接收结果和有效的机器学习配置。

Console

配置自定义 ML 模型算法（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面上，选择包含要导出的自定义模型的协作。
4. 协作打开后，选择 “机器学习模型” 选项卡，然后从 “自定义训练模型” 表中选择您的模型
5. 在自定义训练模型详细信息页面上，单击导出模型输出。
6. 对于导出模型输出，在导出模型输出详细信息中，输入名称和可选的描述。

   在导出给协作成员的模型输出下拉列表中选择哪个成员将接收模型工件。
7. 选择导出。

   结果将导出到机器学习配置中指定的 Amazon S3 位置的以下路径：`yourSpecifiedS3Path/collaborationIdentifier/trainedModelName/callerAccountId/jobName`。仅导出您在关联配置的模型算法时选择的要导出的文件（不超过指定的最大文件大小）。

API

配置自定义 ML 模型算法 (API)

通过运行以下代码启动模型导出：

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.start_trained_model_export_job(
    membershipIdentifier='membership_id',
    trainedModelArn='arn:aws:cleanrooms-ml:region:account:membership/
membershipIdentifier/trained-model/identifier',
    outputConfiguration={
        'member': {
            'accountId': 'model_output_receiver_account'
        }
    },
    name='export_job_name'
```

```
)
```

结果将导出到机器学习配置中指定的 Amazon S3 位置的以下路径：`yourSpecifiedS3Path/collaborationIdentifier/trainedModelName/callerAccountId/jobName`。仅导出您在关联已配置模型算法时选择的（不超过`maxSize`指定值）。`filesToExport`

在 AWS Clean Rooms ML 中对经过训练的模型运行推理

能够运行查询的成员也可以在训练作业完成后启动推理作业。他们选择要对其进行推理的推理数据集，并引用他们想要用来运行推理容器的训练模型输出。

必须向将接收推理输出的成员授予成员能力`CAN_RECEIVE_INFERENCE_OUTPUT`。

Console

创建模型推理作业（控制台）

1. 登录 AWS 管理控制台 并在 https://console.aws.amazon.com/clean room AWS Clean Rooms s 上打开控制台。
2. 在左侧导航窗格中，选择协作。
3. 在协作页面上，选择包含要创建推理作业的自定义模型的协作。
4. 协作打开后，选择 “机器学习模型” 选项卡，然后从 “自定义训练模型” 表中选择您的模型。
5. 在自定义训练模型详情页面上，单击启动推理作业。
6. 对于启动推理作业，在推理作业的详细信息中，输入名称和可选的描述。

   输入以下信息：

   - 关联模型算法-推理作业期间使用的关联模型算法。
   - ML 输入通道详情-将为该推理作业提供数据的 ML 输入通道。
   - 转换资源-用于执行推理作业转换功能的计算实例。
   - 输出配置-谁将接收推理作业输出以及输出的 MIME 类型。
   - 加密-选择自定义加密设置以指定您自己的 KMS 密钥和相关信息。否则，Clean Rooms ML 将管理加密。
   - 转换作业详细信息-推理作业的最大有效负载，以 MB 为单位。
   - 环境变量-访问推理作业容器镜像所需的任何环境变量。
7. 选择启动推理作业。

结果将导出到机器学习配置中指定的 Amazon S3 位置的以下路径：yourSpecifiedS3Path/collaborationIdentifier/trainedModelName/callerAccountId/jobName。

API

创建模型推理作业 (API)

通过运行以下代码启动推理作业：

```
import boto3
acr_ml_client= boto3.client('cleanroomsml')

acr_ml_client.start_trained_model_inference_job(
    name="inference_job",
    membershipIdentifier='membership_id',
    trainedModelArn='arn:aws:cleanrooms-ml:region:account:membership/
membershipIdentifier/trained-model/identifier',

    dataSource={
        "mlInputChannelArn": 'channel_arn_3'
    },
    resourceConfig={'instanceType': 'ml.m5.xlarge'},
    outputConfiguration={
        'accept': 'text/csv',
        'members': [
            {
                "accountId": 'member_account_id'
            }
        ]
    }
)
```

结果将导出到机器学习配置中指定的 Amazon S3 位置的以下路径：yourSpecifiedS3Path/collaborationIdentifier/trainedModelName/callerAccountId/jobName。

以培训数据提供者的身份创建 AWS Clean Rooms 机器学习模型

相似模型 是训练数据提供者的数据的模型，它允许种子数据提供者创建训练数据提供者数据的相似细分，该细分与其种子数据最相似。要创建可以在协作中使用的相似模型，您必须导入训练数据，创建相似模型，配置该相似模型，然后将其与一个协作相关联。

使用相似模型需要两方，即训练数据提供者和种子数据提供者，按顺序合作，将他们的数据整合到协作中 AWS Clean Rooms 。以下是训练数据提供者必须先完成的工作流程：

1. 训练数据提供者的数据必须存储在用户-项目交互 AWS Glue 的数据目录表中。训练数据必须至少包含用户 ID 列、交互 ID 列和时间戳列。
2. 训练数据提供者向注册训练数据 AWS Clean Rooms。
3. 训练数据提供者创建一个相似模型，可以将其与多个种子数据提供者共享。相似模型是一种深度神经网络，训练时间可能长达 24 小时。它不会自动重新训练，我们建议您每周重新训练一次。
4. 训练数据提供者配置相似模型，包括是否共享相关性指标以及输出细分的 Amazon S3 位置。训练数据提供者可以通过单个相似模型创建多个配置的相似模型。
5. 训练数据提供者将配置的受众模型关联到与某个种子数据提供者共享的协作。

训练数据提供者创建 ML 模型后，种子数据提供者可以创建和导出相似的区段。

主题

- 导入训练数据
- 创建外观相似的模型
- 配置外观相似的模型
- 关联已配置的相似模型
- 更新已配置的相似模型

导入训练数据

> **Note**
>
> 您只能提供训练数据集，以便在数据存储在 Amazon S3 中的 Clean Rooms ML 相似模型中使用。但是，您可以使用 SQL 为相似模型提供种子数据，该模型跨存储在任何支持的数据源中的数据运行。

在创建相似模型之前，必须指定包含训练数据的 AWS Glue 表。Clean Rooms ML 不存储该数据的副本，仅存储允许其访问该数据的元数据。

要在中导入训练数据 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择 AWS ML 模型。
3. 在训练数据集选项卡上，选择创建训练数据集。
4. 在创建训练数据集页面上，对于训练数据集详细信息，请输入名称，以及描述（可选）。
5. 通过从下拉列表中选择要配置的数据库和表来选择训练数据来源。

   > **Note**
   >
   > 要验证是否是正确的表，请执行以下任一操作：
   >
   > - 选择 “查看方式” AWS Glue。
   > - 打开查看架构以查看架构。

6. 对于训练详细信息，请从下拉列表中选择用户标识符列、项目标识符列和时间戳列。训练数据必须包含这三个列。您也可以选择要在训练数据中包含的任何其他列。

   时间戳列中的数据必须采用 Unix 纪元时间格式，以秒为单位。
7. （可选）如果您还有要训练的其他列，请从下拉列表中选择列名称和类型。
8. 在服务访问中，您必须指定可以访问您数据的服务角色，如果您的数据已加密，则必须提供 KMS 密钥。选择创建并使用新的服务角色，Clean Rooms ML 将自动创建服务角色并添加必要的权限策略。如果您要使用特定的服务角色，请选择使用现有服务角色，并将其输入到服务角色名称字段中。

如果您的数据已加密，请在 AWS KMS key字段中输入您的 KMS 密钥，或者单击创建 AWS KMS key以生成新的 KMS 密钥。

9. 如果要为训练数据集启用标签，请选择添加新标签，然后输入键和值对。
10. 选择创建训练数据集。

有关相应的 API 操作，请参阅CreateTrainingDataset。

创建外观相似的模型

在创建训练数据集后，您就可以创建相似模型了。您可以通过单个训练数据集创建很多相似模型。

您必须在中创建默认数据库， AWS Glue Data Catalog 或者在提供的角色中包含该`glue:createDatabase`权限。

要在中创建外观相似的模型 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择 AWS ML 模型。
3. 在相似模型选项卡上，选择创建相似模型。
4. 在创建相似模型页面上，对于相似模型详细信息，请输入名称，以及描述（可选）。
   a. 从下拉列表中选择要建模的训练数据集。

      > **Note**
      >
      > 要验证这是否是正确的训练数据集，请打开显示训练数据集详细信息以查看详细信息。
      > 要创建新训练数据集，请选择创建训练数据集。

   b. （可选）输入训练窗口。
5. 如果要为相似模型启用自定义加密设置，请选择自定义加密设置，然后输入 KMS 密钥。
6. 如果要为相似模型启用标签，请选择添加新标签，然后输入键和值对。
7. 选择创建相似模型。

> **Note**
>
> 模型训练可能需要几个小时到 2 天。

有关相应的 API 操作，请参阅CreateAudienceModel。

配置外观相似的模型

在创建相似模型后，您就可以对其进行配置以在协作中使用。您可以通过单个相似模型创建多个配置的相似模型。

要在中配置外观相似的模型 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择 AWS ML 模型。
3. 在配置的相似模型选项卡上，选择配置相似模型。
4. 在配置相似模型页面上，对于配置相似模型详细信息，请输入名称，以及描述（可选）。
   a. 从下拉列表中选择您要配置的相似模型。

      > **Note**
      >
      > 要验证这是否是正确的相似模型，请打开显示相似模型详细信息以查看详细信息。要创建新的相似模型，请选择创建相似模型。

   b. 选择您希望的最小匹配种子大小。这是种子数据提供者数据中与训练数据中的用户重叠的最小用户数。该值必须大于 0。
5. 对于与其他成员共享的指标，选择您是否希望协作中的种子数据提供者接收模型指标，包括相关性分数。
6. 对于相似区段目标位置，请输入导出相似区段的 Amazon S3 存储桶。此存储桶必须与您的其他资源位于同一区域。
7. 对于服务访问，选择将用于访问该表的现有服务角色名称。
8. 对于高级素材箱大小配置，请将受众大小类型指定为绝对数字或百分比。
9. 如果要为已配置的表资源启用标签，请选择添加新标签，然后输入键和值对。

10. 选择配置相似模型。

有关相应的 API 操作，请参阅CreateConfiguredAudienceModel。

关联已配置的相似模型

在配置相似模型后，您可以将其与一个协作相关联。

将配置的相似模型关联到 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 在具有活跃成员身份选项卡上，选择一个协作。
4. 在 “机器学习模型” 选项卡上的 “R eady-to-use 相似模型” 下，选择 “关联相似模型”。
5. 在关联配置的相似模型页面上，对于配置的相似模型关联详细信息：
   a. 输入关联的配置受众模型的名称。
   b. 输入表的描述。

      该描述有助于区分具有相似名称的其他关联的配置受众模型。
6. 对于配置的相似模型，从下拉列表中选择一个配置的相似模型。
7. 选择关联 。

有关相应的 API 操作，请参阅CreateConfiguredAudienceModelAssociation。

更新已配置的相似模型

关联已配置的相似模型后，您可以对其进行更新以更改诸如名称、要共享的指标或输出 Amazon S3 位置之类的信息。

要在中更新相关配置的相似模型 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择 AWS ML 模型。

3. 在配置的相似模型选项卡上，在Ready-to-use 相似模型下，选择已配置的相似模型，然后选择编辑。
4. 在编辑页面上，对于配置的相似模型关联详细信息：
   a. 更新名称以及（可选）描述。
   b. 从下拉列表中选择您要配置的相似模型。
   c. 选择您希望的最小匹配种子大小。这是种子数据提供者数据中与训练数据中的用户重叠的最小用户数。该值必须大于 0。
5. 对于与其他成员共享的指标，选择您是否希望协作中的种子数据提供者接收模型指标，包括相关性分数。
6. 对于相似细分目标位置，输入将相似细分导出到的 Amazon S3 存储桶。此存储桶必须与您的其他资源位于同一区域。
7. 对于服务访问，选择将用于访问该表的现有服务角色名称。
8. 对于高级桶大小配置，请选择要如何配置受众桶大小。
9. 选择保存更改。

有关相应的 API 操作，请参阅UpdateConfiguredAudienceModel。

以种子数据提供者的身份创建 AWS Clean Rooms 机器学习模型

训练数据提供者创建 ML 模型后，种子数据提供者可以创建和导出相似的区段。相似区段是训练数据的一个子集，与种子数据最为相似。

这是种子数据提供者必须完成的工作流程：

1. 种子数据提供者的数据可以存储在 Amazon S3 存储桶中，也可以来自查询结果。
2. 种子数据提供者开启与训练数据提供者共享的协作。
3. 种子数据提供者从协作页面的“Clean Rooms ML”选项卡中创建一个相似细分。
4. 种子数据提供者可以评估相关性指标（如果已共享），并导出相似细分以在 AWS Clean Rooms外部使用。

主题

创建长相相似的区段

> **Note**
>
> 您只能提供训练数据集，以便在数据存储在 Amazon S3 中的 Clean Rooms ML 相似模型中使用。但是，您可以使用 SQL 为相似模型提供种子数据，该模型跨存储在任何支持的数据源中的数据运行。

相似细分是与种子数据最相似的训练数据子集。

要在中创建相似的区段 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。

3. 在具有活跃成员身份选项卡上，选择一个协作。
4. 在 “机器学习模型” 选项卡上，选择 “创建相似区段”。
5. 在 “创建相似区段” 页面上，对于关联的配置相似模型，选择要用于此相似区段的关联配置相似模型。
6. 对于相似细分详细信息，输入名称以及（可选）描述。
7. 对于种子配置文件，请选择一个选项，然后采取建议的操作来选择种子方法。

Option	推荐操作
Amazon S3 路径	1. 选择 Amazon S3 位置。 2. （可选）选择在输出中包含种子配置文件。
SQL 查询	编写 SQL 查询并将其结果用作种子数据。
分析模板	从下拉列表中选择一个分析模板，并使用分析模板创建的结果。

8. 选择创建此数据源时要使用的工作器类型。默认的工作器类型为 CR.1X。指定要使用的员工人数。默认为工作人员编号为 16。要指定火花属性，请执行以下操作：

   a. 展开 Spark 属性。
   b. 选择 “添加 Spark 属性”。
   c. 在 Spark 属性对话框中，从下拉列表中选择一个属性名称并输入值。

   下表提供了每个属性的定义。

   有关 Spark 属性的更多信息，请参阅 Apache Spark 文档中的 Spark 属性。

属性名称	说明	默认值
Spark.task.maxFail	控制任务在失败之前可以连续失败多少次。需要一个大于或等于 1 的值。允许的重试次数等于该值减去 1。如果任何尝试成功，则失败计数将重置。不同任务的失败不会累积到这个极限。	4

属性名称	说明	默认值
spark.sql.files。maxPartitionBytes	设置从基于文件的源（例如 Parquet、JSON 和 ORC）读取数据时要打包到单个分区的最大字节数。	128MB
spark.hadoop.fs.s3.maxRetries	设置 Amazon S3 文件操作的最大重试次数。	
spark.network.	设置所有网络交互的默认超时时间。如果未配置，则覆盖以下超时设置： • Spark.storage。 blockManagerHeartbeatTimeoutMs • shuffle.io.connectionTimeout • Spark.rpc.askTimeout • spark.rpc.lookupTim	120
spark.rdd.com	指定是否使用 spark.io.compression.codec 压缩序列化的 RDD 分区。适用于 Java 和 Scala 中的 StorageLevel .MEMORY_ONLY_SER，或 Python 中的.MEMORY_ONLY。 StorageLevel减少存储空间，但需要额外的 CPU 处理时间。	FALSE
Spark.shuffle.spill.compress	指定是否使用 spark.io.compression.codec 压缩随机播放数据。	TRUE
spark.sql.自适应。advisoryPartitionSizeInBytes	当 spark.sql.adaptive.enabled 为真时，设置自适应优化期间洗牌分区的目标大小（以字节为单位）。控制合并小分区或拆分倾斜分区时的分区大小。	（spark.sql.adaptive.shuffle 的值。targetPostShuffleInputSize)
spark.sql.自适应。autoBroadcastJoin阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。仅适用于自适应框架。使用与 spark.sql 相同的默认值。 autoBroadcastJoin阈值。设置为 -1 可禁用广播。	（无）

属性名称	说明	默认值
Spark.sql.adaptive .coalescePartition s.enabled	指定是否根据 spark.sql.adaptive 合并连续的洗牌分区。 advisoryPartitionSizeInByte s 以优化任务规模。需要 spark.sql.adaptive .enabled 才为真。	TRUE
Spark.sql.adaptive .coalescePartive.coal initialPartitionNum	定义合并前随机分区的初始数量。需要同时启用 spark.sql.adaptive.enabled 和 spark.sql .adaptive.coalescePartitions.enabled 才能成真。默认为 spark.sql.shuffle.partitions 的值。	（无）
Spark.sql.adaptive .coalescePartive.coal minPartitionSize	设置合并后的随机分区的最小大小，以防止自适应优化期间分区变得太小。	1 MB
Spark.sql.adaptive .coalescePartition s.parallelism Firs	指定是否根据集群并行度而不是 spark.sql .adaptive 来计算分区大小。 advisoryP artitionSizeInBytes 在分区合并期间。生成的分区大小小于配置的目标大小，以最大限度地提高并行度。我们建议在繁忙的群集上将其设置为 false，以通过防止过多的小任务来提高资源利用率。	TRUE
sql.adaptive.enabled	指定是否启用自适应查询执行，以便在查询执行期间根据准确的运行时统计数据重新优化查询计划。	TRUE
spark.sql.自适应。 forceOptimizeSkewe d加入	指定是否强制启用， OptimizeSkewedJoin 即使它引入了额外的随机播放。	FALSE
spark.sql.自适应。 localShuffleReader .enabled	指定在不需要随机分区时（例如从排序合并联接转换为广播哈希联接之后）是否使用本地随机播放读取器。需要 spark.sql.adaptive .enabled 才为真。	TRUE

属性名称	说明	默认值
spark.sql.自适应。maxShuffledHashJoinLocalMapThreshold	设置用于构建本地哈希映射的最大分区大小（以字节为单位）。在以下情况下，优先考虑洗牌后的哈希联接而不是排序合并联接： • 此值等于或超过 spark.sql.adaptive。advisoryPartitionSizeInBytes • 所有分区大小均在此限制范围内 覆盖 spark.sql.join。 preferSortMerge加入设置。	0 字节
spark.sql.自适应。optimizeSkewsInRebalancePartitions. 已启用	指定是否通过基于 spark.sql.adaptive 将倾斜的随机分区拆分为较小的分区来优化这些分区。 advisoryPartitionSizeInBytes。需要 spark.sql.adaptive.enabled 才为真。	TRUE
spark.sql.自适应。rebalancePartitionsSmallPartitionFactor	定义拆分期间合并分区的大小阈值系数。小于此因子的分区乘以 spark.sql.adaptive。advisoryPartitionSizeInBytes 已合并。	0.2
Spark.sql.adaptive.skewjoin.enable	指定是否通过拆分和可选复制倾斜的分区来处理洗牌联接中的数据倾斜。适用于排序合并和洗牌哈希联接。需要 spark.sql.adaptive.enabled 才为真。	TRUE
Spark.sql.adaptive.skewJoin。skewedPartitionFactor	确定决定分区偏斜的大小系数。当分区的大小超过两个分区时，分区就会出现偏差： • 该因子乘以分区大小中位数 • spark.sql.adaptive.skewJoin 的值。skewedPartitionThresholdInBytes	5

属性名称	说明	默认值
Spark.sql.adaptive .skewJoin。 skewedPartitionThr esholdInBytes	设置用于识别偏斜分区的大小阈值（以字节为单位）。当分区的大小超过两个分区时，分区就会出现偏差： • 这个门槛 • 分区大小中位数乘以 spark.sql.adaptive .skewJoin。 skewedPartitionFactor 我们建议将此值设置为大于 spark.sql .adaptive。 advisoryPartitionSizeInBytes。	256MB
spark.sql。 autoBroadcastJoin阈值	设置在联接期间向工作节点广播的最大表大小（以字节为单位）。设置为 -1 可禁用广播。	10MB
sql.broadcastTimeo ut	控制广播加入期间广播操作的超时时间（以秒为单位）。	300 秒
spark.sql.cbo.enab led	指定是否为计划统计数据估算启用基于成本的优化 (CBO)。	FALSE
spark.sql.cbo.join reorder.dp.star.Filter	指定是否在基于开销的联接枚举期间应用星型联接过滤器启发式算法。	FALSE
spark.sql.cbo.join reorder.dp.Thresh	设置动态规划算法中允许的最大连接节点数。	12
Spark.sql.cbo.join reorder.enabled	指定是否在基于成本的优化 (CBO) 中启用联接重新排序。	FALSE
Spark.sql.cbo.plan stats.enabled	指定在逻辑计划生成期间是否从目录中提取行数和列统计信息。	FALSE
spark.sql.cbo。 starSchemaDetectio n	指定是否启用基于星型架构检测的联接重新排序。	FALSE

属性名称	说明	默认值
spark.sql.files。maxPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最大数量。当初始计数超过此值时，重新缩放分区。这是建议的目标，而不是保证的上限。	（无）
spark.sql.files。maxRecordsPer文件	设置写入单个文件的最大记录数。如果设置为零或负值，则不适用任何限制。	0
spark.sql.files。minPartitionNum	为基于文件的源（Parquet、JSON 和 ORC）设置拆分文件分区的目标最小数量。默认为 spark.sql。 leafNodeDefault并行性。这是建议的目标，而不是保证的上限。	（无）
spark.sql。inMemoryColumnarStorage.batchSize	控制列式缓存的批次大小。增加大小可以提高内存利用率和压缩率，但会增加 out-of-memory出错的风险。	10000
spark.sql。inMemoryColumnar存储. 已压缩	指定是否根据数据统计信息自动为列选择压缩编解码器。	TRUE
spark.sql。inMemoryColumnar存储。 enableVectorizedReader	指定是否为列式缓存启用矢量化读取。	TRUE
Spark.sql.legacy。allowHashOnMapType	指定是否允许对地图类型数据结构进行哈希操作。此传统设置保持了与旧版 Spark 地图类型处理的兼容性。	
Spark.sql.legacy。allowNegativeScaleOfDecimal	指定是否允许在十进制类型定义中使用负比例值。此传统设置保持了与支持负十进制小数位数的旧 Spark 版本的兼容性。	

属性名称	说明	默认值
Spark.sql.legacy。castCompl exTypesToString. 已启用	指定是否启用将复杂类型转换为字符串的传统行为。保持与旧版 Spark 的类型转换规则的兼容性。	
Spark.sql.legacy。charVarcharAs字符串	指定是否将 CHAR 和 VARCHAR 类型视为字符串类型。此传统设置提供了与旧版 Spark 的字符串类型处理的兼容性。	
Spark.sql.legacy。createEmptyCollectionUsingStringType	指定是否使用字符串类型元素创建空集合。此传统设置保持了与旧版 Spark 的集合初始化行为的兼容性。	
Spark.sql.legacy。exponentLiteralAsdecimal. 已启用	指定是否将指数文字解释为十进制类型。此传统设置保持了与旧版 Spark 的数字文字处理的兼容性。	
spark.sql.legacy.json。 allowEmptyString.enabled	指定是否允许在 JSON 处理中使用空字符串。此传统设置保持了与旧版 Spark 的 JSON 解析行为的兼容性。	
spark.sql.legacy.parquet.int96RebaseModelRead	指定在读取 Parquet 文件时是否使用传统 INT96 的时间戳变基模式。此传统设置保持了与旧版 Spark 的时间戳处理的兼容性。	
Spark.sql.legacy。timeParserPolicy	控制时间解析行为以实现向后兼容。此传统设置决定了如何从字符串中解析时间戳和日期。	
Spark.sql.legacy.typeCorcion。datetimeToString.enabled	指定在将日期时间值转换为字符串时是否启用传统类型强制行为。保持与旧版 Spark 版本的日期时间转换规则的兼容性。	
spark.sql。maxSinglePartition字节	以字节为单位设置最大分区大小。规划器为较大的分区引入了洗牌操作以提高并行度。	128m

属性名称	说明	默认值
Spark.sql.metadata Cache TTLSeconds	控制元数据缓存的 time-to-live (TTL)。适用于分区文件元数据和会话目录缓存。需要： • 大于零的正值 • Spark.sql.catalog实现设置为蜂巢 • spark.sql.hive。 filesourcePartitio nFileCacheSize 大于零 • spark.sql.hive。 manageFilesourcePa rtitions 设置为 true	-1000 毫秒
火花.sql.optimizer。 collapseProjectAlw ays内联	指定是否折叠相邻的投影和行内表达式，即使这会导致重复。	FALSE
火花.sql.optimizer。 dynamicPartitionPr uning.enabled	指定是否为用作联接键的分区列生成谓词。	TRUE
火花.sql.optimizer。 enableCsvExpressio n优化	指定是否通过从 from_csv 操作中删除不必要的列来优化 SQL 优化器中的 CSV 表达式。	TRUE
火花.sql.optimizer。 enableJsonExpressi on优化	通过以下方式指定是否优化 SQL 优化器中的 JSON 表达式： • 从 from_json 操作中删除不必要的列 • 简化 from_json 和 to_json 的组合 • 优化 named_struct 操作	TRUE
spark.sql.Optimize r.ExcludedR	定义要禁用的优化器规则，由逗号分隔的规则名称标识。某些规则无法禁用，因为它们是正确性所必需的。优化器会记录哪些规则已成功禁用。	（无）

属性名称	说明	默认值
spark.sql.Optimizer.runtime.bloomFilapplicationSideScanSizeThreshold	设置在应用程序端注入 Bloom 过滤器所需的最小聚合扫描大小（以字节为单位）。	10GB
spark.sql.Optimizer.runtime.bloomFilcreationSideThreshold	定义在创建端注入 Bloom 滤镜的最大大小阈值。	10MB
Spark.sql.Optimizer.runtime.bloomFilter.enable	指定当随机连接的一侧具有选择性谓词时，是否插入布隆过滤器以减少随机播放数据。	TRUE
spark.sql.Optimizer.runtime.bloomFilexpectedNumItems	定义运行时 Bloom 过滤器中预期项目的默认数量。	1000000
spark.sql.Optimizer.runtime.bloomFilmaxNumBits	设置运行时 Bloom 过滤器中允许的最大位数。	67108864
spark.sql.Optimizer.runtime.bloomFilmaxNumItems	设置运行时 Bloom 过滤器中允许的最大预期项目数。	4000000
spark.sql.Optimizer.runtime.bloomFilter.number.	限制每次查询允许的非 DPP 运行时过滤器的最大数量，以防止驱动程序 out-of-memory出错。	10
Spark.sql.Optimizer.runtime.bloomfilter.numbit	定义运行时 Bloom 过滤器中使用的默认位数。	8388608

属性名称	说明	默认值
Spark.sql.optimizer.runtime rowlevelOperationGroup过滤器. 已启用	指定是否为行级操作启用运行时组筛选。允许数据源： • 使用数据源筛选器修剪整组数据（例如文件或分区） • 执行运行时查询以识别匹配的记录 • 丢弃不必要的组以避免昂贵的重写 限制： • 并非所有表达式都可以转换为数据源筛选器 • 有些表达式需要 Spark 求值（例如子查询）	TRUE
Spark.sql.Optimizer.runtimeFilter semiJoinReduction.enabled	指定当随机连接的一侧具有选择性谓词时，是否插入半联接以减少随机播放数据。	FALSE
spark.sql.parquet.AgregatePus	指定是否将聚合向下推送到 Parquet 进行优化。支持： • 布尔型、整数、浮点型和日期类型的最小值和最大值 • 所有数据类型的计数 如果任何 Parquet 文件页脚中缺少统计信息，则会抛出异常。	FALSE
sql.parquet。columnarReaderBatch大小	控制每个 Parquet 矢量化阅读器批次中的行数。选择一个平衡性能开销和内存使用量的值，以防止 out-of-memory出错。	4096

属性名称	说明	默认值
Spark.sql.session.time	定义会话时区，用于处理字符串文字中的时间戳和 Java 对象转换。接受： • 以地区为基础 IDs 的 area/city 格式（例如美国/洛杉矶） • 区域偏移量采用 (+/-) HH、(+/-) HH: mm 或 (+/-) HH: mm: ss 格式（例如 -08 或 + 01:00） • UTC 或 Z 作为 + 00:00 的别名	（当地时区的值）
spark.sql.shuffle.part	设置联接或聚合期间用于数据洗牌的默认分区数。无法在结构化流式查询从同一检查点位置重新启动之间进行修改。	200
spark.sql。shuffledHashJoin因子	定义用于确定 shuffle 哈希加入资格的乘法系数。当小边数据大小乘以此系数小于大边数据大小时，将选择随机哈希联接。	3
火花.sql.sources。parallelPartitionDiscovery. 阈值	使用基于文件的源（Parquet、JSON 和 ORC）设置驱动端文件列表的最大路径数。如果在分区发现期间超出限制，则使用单独的 Spark 分布式作业列出文件。	32
spark.sql.statistics.histics.h	指定是否在列统计数据计算期间生成等高直方图以提高估计精度。除了基本列统计数据所需的扫描之外，还需要进行额外的表扫描。	FALSE

9. 对于服务访问，选择将用于访问该表的现有服务角色名称。
10. 如果要为训练数据集启用标签，请选择添加新标签，然后输入键和值对。
11. 选择创建相似细分。

有关相应的 API 操作，请参阅StartAudienceGenerationJob。

导出相似的区段

在创建相似细分后，您可以将该数据导出到一个 Amazon S3 存储桶。

要在中导出相似的区段 AWS Clean Rooms

1. 登录 AWS 管理控制台 并使用您的AWS Clean Rooms 主机打开主机 AWS 账户 （如果您尚未这样做）。
2. 在左侧导航窗格中，选择协作。
3. 在具有活跃成员身份选项卡上，选择一个协作。
4. 在 “机器学习模型” 选项卡上，选择一个相似的区段，然后选择导出。
5. 对于导出相似模型，为导出相似模型详细信息输入名称和可选描述。
6. 对于细分大小，选择导出的细分所需的大小。
7. 选择导出。

有关相应的 API 操作，请参阅StartAudienceExportJob。

故障排除 AWS Clean Rooms

本节介绍使用时可能出现的一些常见问题 AWS Clean Rooms 以及如何解决这些问题。

问题

查询所引用的一个或多个表不能由其关联的服务角色访问。 table/role 所有者必须向服务角色授予对表的访问权限。

- 验证服务角色的权限是否已按要求设置。有关更多信息，请参阅设置 AWS Clean Rooms。

其中一个底层数据集的文件格式不受支持。

- 确保您的数据集采用支持的文件格式之一：
  - Parquet
  - RCFile
  - TextFile
  - SequenceFile
  - RegexSerde
  - OpenCSV
  - AVRO
  - JSON

  有关更多信息，请参阅 的数据格式 AWS Clean Rooms。

使用 Clean Rooms 加密计算时，查询结果不如预期。

如果您使用 Clean Rooms 加密计算 (C3R)，请验证您的查询是否正确使用了加密列：

- sealed 列仅用于 SELECT 子句。
- fingerprint 列仅用于 JOIN 子句（以及某些条件下的 GROUP BY 子句）。
- 只有在协作设置要求的情况下，才 JOINing 具有相同名称的 fingerprint 列。

有关更多信息，请参阅the section called “加密计算”和the section called “列类型”。

AWS Clean Rooms Spark SQL：缺少分区

中的所有分区还 AWS Glue Data Catalog 必须在 S3 中包含数据。引擎使用 Spark 设置 `spark.sql.files.ignoreMissingFiles=False`

有关更多信息，请参阅 https://spark.apache.org/docs/latest/sql-data-sources-generic-options.html#ignore-missing-files。

如果遇到此错误，您将收到以下错误消息：`"Missing partition data: One of the configured tables is partitioned and one or more of the partitions does not have data".`

将您在 Amazon S3 中的数据与表中列出的分区 AWS Glue Data Catalog 进行比较。删除 S3 中没有相应数据的分区。

安全性 AWS Clean Rooms

云安全 AWS 是重中之重。作为 AWS 客户，您可以受益于专为满足大多数安全敏感型组织的要求而构建的数据中心和网络架构。

安全是双方共同承担 AWS 的责任。责任共担模式将此描述为云的安全性和云中的安全性：

- 云安全 — AWS 负责保护在 AWS 云中运行 AWS 服务的基础架构。 AWS 还为您提供可以安全使用的服务。作为AWS 合规计划合规计划合规计划合的一部分，第三方审计师定期测试和验证我们安全的有效性。要了解适用于的合规计划 AWS Clean Rooms，请参阅按合规计划划分的 AWS 范围内服务 AWS 按合规计划。
- 云端安全-您的责任由您使用的 AWS 服务决定。您还需要对其他因素负责，包括您的数据的敏感性、您公司的要求以及适用的法律法规。

本文档可帮助您了解在使用时如何应用分担责任模型 AWS Clean Rooms。它向您展示了如何进行配置 AWS Clean Rooms 以满足您的安全和合规性目标。您还将学习如何使用其他 AWS 服务来帮助您监控和保护您的 AWS Clean Rooms 资源。

内容

中的数据保护 AWS Clean Rooms

分 AWS 担责任模型适用于中的数据保护 AWS Clean Rooms。如本模型所述 AWS ，负责保护运行所有内容的全球基础架构 AWS Cloud。您负责维护对托管在此基础结构上的内容的控制。您还负责您所使用的 AWS 服务 的安全配置和管理任务。有关数据隐私的更多信息，请参阅数据隐私常见问

题。有关欧洲数据保护的信息，请参阅 AWS Security Blog 上的 AWS Shared Responsibility Model and GDPR 博客文章。

出于数据保护目的，我们建议您保护 AWS 账户 凭证并使用 AWS IAM Identity Center 或 AWS Identity and Access Management (IAM) 设置个人用户。这样，每个用户只获得履行其工作职责所需的权限。还建议您通过以下方式保护数据：

- 对每个账户使用多重身份验证（MFA）。
- 用于 SSL/TLS 与 AWS 资源通信。我们要求使用 TLS 1.2，建议使用 TLS 1.3。
- 使用设置 API 和用户活动日志 AWS CloudTrail。有关使用 CloudTrail 跟踪捕获 AWS 活动的信息，请参阅AWS CloudTrail 用户指南中的使用跟 CloudTrail 踪。
- 使用 AWS 加密解决方案以及其中的所有默认安全控件 AWS 服务。
- 使用高级托管安全服务（例如 Amazon Macie），它有助于发现和保护存储在 Amazon S3 中的敏感数据。
- 如果您在 AWS 通过命令行界面或 API 进行访问时需要经过 FIPS 140-3 验证的加密模块，请使用 FIPS 端点。有关可用的 FIPS 端点的更多信息，请参阅《美国联邦信息处理标准（FIPS）第 140-3 版》https://aws.amazon.com/compliance/fips/。

强烈建议您切勿将机密信息或敏感信息（如您客户的电子邮件地址）放入标签或自由格式文本字段（如名称字段）。这包括您使用控制台、API AWS Clean Rooms 或以其他 AWS 服务 方式使用控制台 AWS CLI、API 或时 AWS SDKs。在用于名称的标签或自由格式文本字段中输入的任何数据都可能会用于计费或诊断日志。如果您向外部服务器提供 URL，强烈建议您不要在网址中包含凭证信息来验证对该服务器的请求。

静态加密

AWS Clean Rooms 始终对所有静态服务元数据进行加密，无需任何其他配置。使用时会自动进行加密 AWS Clean Rooms。

Clean Rooms ML 对存储在服务中的所有静态数据进行加密。 AWS KMS如果您选择提供自己的 KMS 密钥，则相似模型和相似细分生成作业内容将使用您的 KMS 密钥进行静态加密。

使用 AWS Clean Rooms 自定义 ML 模型时，该服务会对所有静态存储的数据进行加密。 AWS KMS AWS Clean Rooms 支持使用您创建、拥有和管理的对称客户托管密钥来加密静态数据。如果未指定客户管理的密钥，则默认使用 AWS 拥有的密钥 这些密钥。

AWS Clean Rooms 使用授权和密钥策略来访问客户托管的密钥。您可以随时撤销授予访问权限，或删除服务对客户托管密钥的访问权限。如果这样做，将 AWS Clean Rooms 无法访问由客户托管密钥加

密的任何数据，这会影响依赖该数据的操作。例如，如果您尝试从 AWS Clean Rooms 无法访问的加密机器学习输入通道创建经过训练的模型，则该操作将返回`ValidationException`错误。

> **Note**
>
> 您可以利用 Amazon S3 中的加密选项来保护静态数据。
> 有关更多信息，请参阅《Amazon S3 用户指南》中的指定 Amazon S3 加密。

在中使用 ID 映射表时 AWS Clean Rooms，该服务会对所有静态存储的数据进行加密。 AWS KMS如果您选择提供自己的 KMS 密钥，则会通过 KMS 密钥对您的 ID 映射表中的内容进行静态加密 AWS Entity Resolution 数据匹配服务。有关使用 ID 映射工作流程进行加密所需的权限的更多详细信息，请参阅《AWS Entity Resolution 数据匹配服务 用户指南》中的为 AWS Entity Resolution 数据匹配服务创建工作流程作业角色。

传输中加密

AWS Clean Rooms 使用传输层安全 (TLS) 对传输过程进行加密。与的 AWS Clean Rooms 通信始终通过 HTTPS 完成，因此无论数据存储在 Amazon S3、Amazon Athena 还是 Snowflake 中，您的数据在传输过程中始终处于加密状态。这包括使用 Clean Rooms ML 时的所有传输中数据。

加密底层数据

有关如何解密您的底层数据的更多信息，请参阅加密计算 Clean Rooms。

密钥策略

密钥策略控制对客户自主管理型密钥的访问。每个客户托管式密钥必须只有一个密钥策略，其中包含确定谁可以使用密钥以及如何使用密钥的声明。创建客户托管式密钥时，可以指定密钥策略。有关更多信息，请参阅《 AWS Key Management Service 开发人员指南》中的管理客户托管密钥的访问权限。

要将客户托管密钥用于您的 AWS Clean Rooms 自定义 ML 模型，必须在密钥策略中允许以下 API 操作：

- `kms:DescribeKey`— 提供客户管理的密钥详细信息 AWS Clean Rooms 以允许验证密钥。
- `kms:Decrypt`— 提供访问权限 AWS Clean Rooms 以解密加密数据并将其用于相关作业。
- `kms:CreateGrant`-Clean Rooms ML 通过为 Amazon ECR 创建补助金来加密 Amazon ECR 中静态的训练和推理图像。要了解更多信息，请参阅 Amazon ECR 中的静态加密。Clean Rooms ML 还使用 SageMaker Amazon AI 来运行训练和推理作业，并为 SageMaker AI 创建补助金，用于加密附

加到实例的 Amazon EBS 卷以及 Amazon S3 中的输出数据。要了解更多信息，请参阅在 Amazon A SageMaker I 中使用加密保护静态数据。

- `kms:GenerateDataKey`-Clean Rooms ML 使用服务器端加密对存储在 Amazon S3 中的静态数据进行加密。 AWS KMS keys要了解更多信息，请参阅在 A mazon S3 中使用服务器端加密 AWS KMS keys (SSE-KMS)。

以下是您可以为以下资源添加 AWS Clean Rooms 的策略声明示例：

带有合成数据的 ML 输入通道

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
        "Sid": "Allow access to principals authorized to use AWS Clean Rooms ML",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::666666666666:role/ExampleRole"
        },
        "Action": [
            "kms:GenerateDataKey",
            "kms:CreateGrant",
            "kms:Decrypt"
        ],
        "Resource": "*",
        "Condition": {
            "StringEquals": {
                "kms:ViaService": "cleanrooms-ml.us-east-1.amazonaws.com"
            },
                "ForAllValues:StringEquals": {
                    "kms:GrantOperations": [
                        "Decrypt",
                        "Encrypt",
                        "GenerateDataKeyWithoutPlaintext",
                        "ReEncryptFrom",
                        "ReEncryptTo",
                        "CreateGrant",
                        "DescribeKey",
                        "RetireGrant",
                        "GenerateDataKey"
                    ]
                },
```

```
                "BoolIfExists": {
                    "kms:GrantIsForAWSResource": true
        }
            }
        },
        {
            "Sid": "Allow describe key for principals authorized to use AWS Clean Rooms
ML",
            "Effect": "Allow",
            "Principal": {
                "AWS": "arn:aws:iam::444455556666:role/ExampleRole"
            },
            "Action": [
                "kms:DescribeKey"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "kms:ViaService": "cleanrooms-ml.us-east-1.amazonaws.com"
                }
            }
        },
        {
            "Sid": "Allow grant operations for AWS Clean Rooms ML service principal",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": [
                "kms:GenerateDataKey",
                "kms:CreateGrant",
                "kms:Decrypt"
            ],
            "Resource": "*",
            "Condition": {
                "ForAllValues:StringEquals": {
                    "kms:GrantOperations": [
                        "Decrypt",
                        "Encrypt",
                        "GenerateDataKeyWithoutPlaintext",
                        "ReEncryptFrom",
                        "ReEncryptTo",
                        "CreateGrant",
                        "DescribeKey",
```

```
                    "RetireGrant",
                    "GenerateDataKey"
                ]
            }
        }
    },
    {
        "Sid": "Allow describe key for AWS Clean Rooms ML service principal",
        "Effect": "Allow",
        "Principal": {
            "Service": "cleanrooms-ml.amazonaws.com"
        },
        "Action": [
            "kms:DescribeKey"
        ],
        "Resource": "*"
    }
  ]
}
```

没有合成数据的 ML 输入通道

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
        "Sid": "Allow access to principals authorized to use AWS Clean Rooms ML",
        "Effect": "Allow",
        "Principal": {
            "AWS": "arn:aws:iam::666666666666:role/ExampleRole"
        },
        "Action": [
            "kms:DescribeKey",
            "kms:GenerateDataKey",
            "kms:Decrypt"
        ],
        "Resource": "*",
        "Condition": {
            "StringEquals": {
                "kms:ViaService": "cleanrooms-ml.us-east-1.amazonaws.com"
            }
```

```
        }
    },
    {
        "Sid": "Allow access to AWS Clean Rooms ML service principal",
        "Effect": "Allow",
        "Principal": {
            "Service": "cleanrooms-ml.amazonaws.com"
        },
        "Action": [
            "kms:DescribeKey",
            "kms:GenerateDataKey",
            "kms:Decrypt"
        ],
        "Resource": "*"
    }
  ]
}
```

经过训练的模型作业或经过训练的模型推理作业

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "Allow grant operations for principals authorized to use AWS
 Clean Rooms ML",
            "Effect": "Allow",
            "Principal": {
                "AWS": "arn:aws:iam::666666666666:role/ExampleRole"
            },
            "Action": [
                "kms:GenerateDataKey",
                "kms:CreateGrant",
                "kms:Decrypt"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "kms:ViaService": "cleanrooms-ml.us-east-1.amazonaws.com"
                },
```

```
                "ForAllValues:StringEquals": {
                    "kms:GrantOperations": [
                        "Decrypt",
                        "Encrypt",
                        "GenerateDataKeyWithoutPlaintext",
                        "ReEncryptFrom",
                        "ReEncryptTo",
                        "CreateGrant",
                        "DescribeKey",
                        "RetireGrant",
                        "GenerateDataKey"
                    ]
                },
                "BoolIfExists": {
                    "kms:GrantIsForAWSResource": true
                }
            }
        },
        {
            "Sid": "Allow describe key for principals authorized to use AWS Clean
Rooms ML",
            "Effect": "Allow",
            "Principal": {
                "AWS": "arn:aws:iam::444455556666:role/ExampleRole"
            },
            "Action": [
                "kms:DescribeKey"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "kms:ViaService": "cleanrooms-ml.us-east-1.amazonaws.com"
                }
            }
        },
        {
            "Sid": "Allow grant operations for AWS Clean Rooms ML service
principal",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": [
                "kms:GenerateDataKey",
```

```
                "kms:CreateGrant",
                "kms:Decrypt"
            ],
            "Resource": "*",
            "Condition": {
                "ForAllValues:StringEquals": {
                    "kms:GrantOperations": [
                        "Decrypt",
                        "Encrypt",
                        "GenerateDataKeyWithoutPlaintext",
                        "ReEncryptFrom",
                        "ReEncryptTo",
                        "CreateGrant",
                        "DescribeKey",
                        "RetireGrant",
                        "GenerateDataKey"
                    ]
                }
            }
        },
        {
            "Sid": "Allow describe key for AWS Clean Rooms ML service principal",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms-ml.amazonaws.com"
            },
            "Action": [
                "kms:DescribeKey"
            ],
            "Resource": "*"
        }
    ]
}
```

Clean Rooms ML 不支持在客户托管密钥策略中指定服务加密上下文或源上下文。客户可以在中看到服务内部使用的加密上下文 CloudTrail。

将服务相关角色用于 AWS Clean Rooms

AWS Clean Rooms 使用 AWS Identity and Access Management (IAM) 服务相关角色。服务相关角色是一种与之直接关联的 IAM 角色的独特类型。 AWS Clean Rooms服务相关角色由服务预定义 AWS Clean Rooms ，包括该服务代表您调用其他 AWS 服务所需的所有权限。

服务相关角色使设置变得 AWS Clean Rooms 更加容易，因为您不必手动添加必要的权限。 AWS Clean Rooms 定义其服务相关角色的权限，除非另有定义，否则 AWS Clean Rooms 只能担任其角色。定义的权限包括信任策略和权限策略，以及不能附加到任何其他 IAM 实体的权限策略。

只有在首先删除相关资源后，您才能删除服务关联角色。这样可以保护您的 AWS Clean Rooms 资源，因为您不会无意中删除访问资源的权限。

有关支持服务相关角色的其他服务的信息，请参阅与 IAM 配合使用的AWS 服务，并在服务相关角色列中查找标有 “是” 的服务。选择是和链接，查看该服务的服务关联角色文档。

的服务相关角色权限 AWS Clean Rooms

AWS Clean Rooms 使用名为 R AWSServiceRoleForAWSCleanoo ms 的服务相关角色向您的 AWS 账户发布与 Clean Room CloudWatch s 相关的指标。

R AWSService RoleFor AWSClean ooms 服务相关角色信任以下服务来代替该角色：

- `cleanrooms.amazonaws.com`

名为的角色权限策略 AWSCleanRoomsServiceRolePolicy AWS Clean Rooms 允许对指定资源完成以下操作：

- 操作：`all AWS resources, restricted to the AWS Clean Rooms namespace` 上的 `cloudwatch:PutMetricData`

您必须配置使用户、组或角色能够创建、编辑或删除服务相关角色的权限。有关更多信息，请参阅《IAM 用户指南》中的服务相关角色权限。

为创建服务相关角色 AWS Clean Rooms

您可以使用 IAM 控制台在 R AWSServiceRoleForAWSCleanoo ms 用例中创建服务相关角色。在 AWS CLI 或 AWS API 中，使用服务名称创建服务相关角色。`cleanrooms.amazonaws.com`有关更

多信息，请参阅 IAM 用户指南 中的创建服务相关角色。如果您删除了此服务相关角色，可以使用同样的过程再次创建角色。

编辑的服务相关角色 AWS Clean Rooms

AWS Clean Rooms 不允许您编辑 Rooms AWSService RoleFor AWSClean 服务相关角色。创建服务关联角色后，您将无法更改角色的名称，因为可能有多种实体引用该角色。但是可以使用 IAM 编辑角色描述。有关更多信息，请参阅《IAM 用户指南》中的编辑服务关联角色。

删除的服务相关角色 AWS Clean Rooms

如果不再需要使用某个需要服务关联角色的功能或服务，我们建议您删除该角色。这样就没有未被主动监控或维护的未使用实体。但是，必须先清除服务相关角色的资源，然后才能手动删除它。

要删除聊天AWSServiceRoleForAWSClean室，您必须先删除其中的所有协作和成员资格。 AWS 账户

> **Note**
>
> 如果您尝试删除资源时 AWS Clean Rooms 服务正在使用该角色，则删除可能会失败。如果发生这种情况，请等待几分钟后重试。

使用 IAM 手动删除服务关联角色

使用 IAM 控制台 AWS CLI、或 AWS API 删除 R AWSService RoleFor AWSClean ooms 服务相关角色。有关更多信息，请参阅《IAM 用户指南》中的删除服务关联角色。

AWS Clean Rooms 服务相关角色支持的区域

AWS Clean Rooms 支持在提供服务的所有区域中使用服务相关角色。有关更多信息，请参阅AWS 区域和端点。

数据保留在 AWS Clean Rooms

查询完成后，任何临时读入 AWS Clean Rooms 协作的数据都会被删除。

在您创建相似模型时，Clean Rooms ML 读取您的训练数据，将其转换为适合我们的机器学习模型的格式，并将训练的模型参数存储在 Clean Rooms ML 中。Clean Rooms ML 不会保留您的训练数据的副本。 AWS Clean Rooms 查询运行后，SQL 查询不会保留您的任何数据。然后，Clean Rooms ML 使

用训练的模型总结您的所有用户的行为。只要您的相似模型处于活跃状态，Clean Rooms ML 就会为您的数据中的每个用户存储用户级数据集。

当您启动相似区段生成作业时，Clean Rooms ML 会读取种子数据，从关联的相似模型中读取行为摘要，然后创建存储在服务中的相似区段。 AWS Clean Rooms Clean Rooms ML 不会保留您的种子数据的副本。只要作业处于活跃状态，Clean Rooms ML 就会存储作业的用户级输出。

如果您的种子数据来自 SQL 查询，则该查询的输出仅在作业持续期间存储在服务中。查询结果会进行静态和传输中加密。

如果要删除相似模型或相似细分生成作业数据，请使用 API 将其删除。Clean Rooms ML 异步删除与模型或作业关联的所有数据。在该过程完成后，Clean Rooms ML 删除模型或作业的元数据，而不再在 API 中显示这些数据。Clean Rooms ML 会将删除的数据保留 3 天以防进行灾难恢复。在 API 中不再显示作业或模型并且经过 3 天后，将永久删除与模型或作业关联的所有数据。

中数据协作的最佳实践 AWS Clean Rooms

本主题介绍在 AWS Clean Rooms中开展数据协作的最佳实践。

AWS Clean Rooms 遵循AWS 分担责任模型。 AWS Clean Rooms 提供了分析规则，您可以配置这些规则以增强在协作中保护敏感数据的能力。您在中配置的分析规则 AWS Clean Rooms 将强制执行您配置的限制（查询控件和查询输出控件）。您负责确定限制并相应地配置分析规则。

数据协作可能涉及的不仅仅是您的使用。 AWS Clean Rooms为了帮助您最大限度地发挥数据协作的优势，我们建议您在使用分析规则时执行以下最佳实践 AWS Clean Rooms ，特别是分析规则。

主题

- 最佳实践 AWS Clean Rooms
- 在中使用分析规则的最佳实践 AWS Clean Rooms

最佳实践 AWS Clean Rooms

您负责评估每个数据协作的风险，并将其与您的隐私要求（例如外部和内部合规性计划和策略）进行比较。我们建议您在使用时采取其他措施 AWS Clean Rooms。这些操作可能有助于进一步管理风险，并有助于防范第三方试图重新识别您的数据（例如，差异攻击或侧信道攻击）。

例如，考虑对您的其他协作者进行尽职调查，并在进行协作之前 与他们签订法律协议。要监控数据的使用情况，还要考虑在使用 AWS Clean Rooms时采用其他审计机制。

在中使用分析规则的最佳实践 AWS Clean Rooms

中的分析规则 AWS Clean Rooms 允许您通过在已配置的表上设置查询控件来限制可以运行的查询。例如，您可以设置查询控制，以确定如何联接配置表以及可以选择哪些列。您还可以通过设置查询结果控制（例如输出行的聚合阈值）来限制查询输出。该服务拒绝任何查询，并删除不符合成员在查询中配置表上设置的分析规则的行。

对于在配置表上使用分析规则，我们推荐以下 10 种最佳实践：

- 为不同的查询使用案例（例如受众规划或归因）创建单独的配置表。您可以使用同一底层 AWS Glue 表创建多个配置表。
- 在分析规则中指定协作中查询所必需的列（例如维度列、列表列、联接列）。这可能有助于降低差异攻击的风险或使其他成员能够对您的数据进行逆向工程。使用允许列表列功能记下将来可能要设置为可查询的其他列。要自定义可用于特定协作的列，请使用相同的基础表创建其他已配置 AWS Glue 表。
- 在分析规则中指定协作中分析所必需的函数。这有助于降低因罕见的函数错误而带来的风险，这些错误可能会显示单个数据点的信息。要自定义可用于特定协作的函数，请使用同一底层 AWS Glue 表创建其他配置表。
- 对行级值敏感的任何列添加聚合约束。这包括您的配置表中的列，这些列也存在于其他协作成员的表中，并且有分析规则作为聚合约束。这也包括您的配置表中不可查询的列，即配置表中有但不在分析规则中的列。聚合约束可以帮助降低将查询结果与协作之外的数据关联起来的风险。
- 创建测试协作和分析规则，以测试使用指定分析规则创建的限制。
- 查看协作者配置表和成员对配置表的分析规则，以检查它们是否符合协作商定的内容。这可以帮助降低其他成员设计自己的数据以运行未商定的查询所带来的风险。
- 查看提供的示例查询（仅限控制台），该查询在设置分析规则后在配置表上启用。

> **Note**
>
> 除了提供的示例查询外，还可以根据分析规则和其他协作成员表和分析规则进行其他查询。

- 您可以为协作中的配置表添加或更新分析规则。完成后，请查看与配置表关联的所有协作及其产生的影响。这有助于确保任何协作都不会使用过时的分析规则。
- 审核协作中运行的查询，检查查询是否与协作中商定的使用案例或查询相匹配。（打开查询日志记录功能后，可在查询日志中查看查询）。这可以帮助降低成员运行未商定的分析和潜在攻击（例如侧信道攻击）带来的风险。

- 审核协作成员分析规则和查询中使用的配置表列，检查它们是否与协作中商定的内容相匹配。（打开该功能后，可在查询日志中查看查询。）这可以帮助降低其他成员设计自己的数据以进行未商定的查询所带来的风险。

Identity and Access Management AWS Clean Rooms

AWS Identity and Access Management (IAM) AWS 服务 可帮助管理员安全地控制对 AWS 资源的访问权限。IAM 管理员控制谁可以进行身份验证（登录）和授权（拥有权限）使用 AWS Clean Rooms 资源。您可以使用 IAM AWS 服务 ，无需支付额外费用。

主题

受众

您的使用方式 AWS Identity and Access Management (IAM) 因您的角色而异：

- 服务用户：如果您无法访问功能，请从管理员处请求权限（请参阅对 AWS Clean Rooms 身份和访问进行故障排除）
- 服务管理员：确定用户访问权限并提交权限请求（请参阅如何 AWS Clean Rooms 与 IAM 配合使用）
- IAM 管理员：编写用于管理访问权限的策略（请参阅基于身份的策略示例 AWS Clean Rooms）

使用身份进行身份验证

身份验证是您 AWS 使用身份凭证登录的方式。您必须以 IAM 用户身份或通过担 AWS 账户根用户任 IAM 角色进行身份验证（登录 AWS）。

您可以使用通过身份源提供的凭据以 AWS 联合身份登录。 AWS IAM Identity Center （IAM Identity Center）用户或贵公司的单点登录身份验证就是联合身份的示例。当您以联合身份登录时，您的管理员以前使用 IAM 角色设置了身份联合验证。当你使用联合访问 AWS 时，你就是在间接扮演一个角色。

根据您的用户类型，您可以登录 AWS 管理控制台 或 AWS 访问门户。有关登录的更多信息 AWS，请参阅《AWS 登录 用户指南》中的如何登录到您 AWS 账户的。

如果您 AWS 以编程方式访问，则会 AWS 提供软件开发套件 (SDK) 和命令行接口 (CLI)，以便使用您的凭据对请求进行加密签名。如果您不使用 AWS 工具，则必须自己签署请求。有关使用推荐的方法自行对请求签名的更多信息，请参阅《AWS 一般参考》中的 签名版本 4 签名流程。

无论使用何种身份验证方法，您都可能需要提供其他安全信息。例如， AWS 建议您使用多重身份验证 (MFA) 来提高账户的安全性。要了解更多信息，请参阅《AWS IAM Identity Center 用户指南》中的多重身份验证和《IAM 用户指南》中的在 AWS中使用多重身份验证（MFA）。

AWS 账户 root 用户

创建时 AWS 账户，首先要有一个登录身份，该身份可以完全访问账户中的所有资源 AWS 服务 和资源。此身份称为 AWS 账户 根用户，使用您创建账户时所用的电子邮件地址和密码登录，即可获得该身份。强烈建议您不要使用根用户执行日常任务。保护好根用户凭证，并使用这些凭证来执行仅根用户可以执行的任务。有关需要您以根用户身份登录的任务的完整列表，请参阅《AWS 一般参考》中的 AWS 账户根用户 凭证和 IAM 身份。

联合身份

作为最佳实践，要求人类用户使用与身份提供商的联合身份验证才能 AWS 服务 使用临时证书进行访问。

联合身份是指来自您的企业目录、Web 身份提供商的用户 Directory Service ，或者 AWS 服务 使用来自身份源的凭据进行访问的用户。联合身份代入可提供临时凭证的角色。

要集中管理访问权限，建议使用。 AWS IAM Identity Center有关更多信息，请参阅《AWS IAM Identity Center 用户指南》中的什么是 IAM Identity Center？。

IAM 用户和群组

IAM 用户是对某个人员或应用程序具有特定权限的一个身份。建议使用临时凭证，而非具有长期凭证的 IAM 用户。有关更多信息，请参阅 IAM 用户指南中的要求人类用户使用身份提供商的联合身份验证才能 AWS 使用临时证书进行访问。

IAM 组指定一组 IAM 用户，便于更轻松地对大量用户进行权限管理。有关更多信息，请参阅《IAM 用户指南》中的 IAM 用户使用案例。

IAM 角色

IAM 角色是具有特定权限的身份，可提供临时凭证。您可以通过从用户切换到 IAM 角色（控制台）或调用 AWS CLI 或 AWS API 操作来代入角色。有关更多信息，请参阅《IAM 用户指南》中的担任角色的方法。

IAM 角色对于联合用户访问、临时 IAM 用户权限、跨账户访问、跨服务访问以及在 Amazon EC2 上运行的应用程序非常有用。有关更多信息，请参阅《IAM 用户指南》中的 IAM 中的跨账户资源访问。

使用策略管理访问

您可以 AWS 通过创建策略并将其附加到 AWS 身份或资源来控制中的访问权限。策略是其中的一个对象 AWS ，当与身份或资源关联时，它会定义其权限。 AWS 在委托人（用户、root 用户或角色会话）发出请求时评估这些策略。策略中的权限确定是允许还是拒绝请求。大多数策略都以 JSON 文档的 AWS 形式存储在中。有关 JSON 策略文档的结构和内容的更多信息，请参阅 IAM 用户指南中的 JSON 策略概览。

管理员可以使用 AWS JSON 策略来指定谁有权访问什么。也就是说，哪个主体 可以对什么资源 执行操作，以及在什么条件 下执行。

每个 IAM 实体（用户或角色）最初没有任何权限。原定设置情况下，用户什么都不能做，甚至不能更改他们自己的密码。要为用户授予执行某些操作的权限，管理员必须将权限策略附加到用户。或者，管理员可以将用户添加到具有预期权限的组中。当管理员为某个组授予访问权限时，该组内的全部用户都会获得这些访问权限。

IAM 策略定义操作的权限，无关乎您使用哪种方法执行操作。例如，假设您有一个允许 `iam:GetRole` 操作的策略。拥有该策略的用户可以从 AWS 管理控制台 AWS CLI、或 AWS API 获取角色信息。

基于身份的策略

基于身份的策略是可附加到身份（如 IAM 用户、用户组或角色）的 JSON 权限策略文档。这些策略控制用户和角色可在何种条件下对哪些资源执行哪些操作。要了解如何创建基于身份的策略，请参阅《IAM 用户指南》中的使用客户管理型策略定义自定义 IAM 权限。

基于身份的策略可以进一步归类为内联策略或托管式策略。内联策略直接嵌入单个用户、组或角色中。托管策略是可以附加到 AWS 账户中的多个用户、组和角色的独立策略。托管策略包括 AWS 托管策略和客户托管策略。要了解如何在托管式策略和内联策略之间进行选择，请参阅 IAM 用户指南中的在托管式策略与内联策略之间进行选择。

基于资源的策略

基于资源的策略是附加到资源的 JSON 策略文档。基于资源的策略的示例包括 IAM 角色信任策略和 Amazon S3 存储桶策略。在支持基于资源的策略的服务中，服务管理员可以使用它们来控制对特定资源的访问。对于在其中附加策略的资源，策略定义指定主体可以对该资源执行哪些操作以及在什么条件下执行。您必须在基于资源的策略中指定主体。委托人可以包括账户、用户、角色、联合用户或 AWS 服务。

基于资源的策略是位于该服务中的内联策略。您不能在基于资源的策略中使用 IAM 中的 AWS 托管策略。

其他策略类型

AWS 支持其他不太常见的策略类型。这些策略类型可以设置更常用的策略类型向您授予的最大权限。

- 权限边界：权限边界是一个高级特征，用于设置基于身份的策略可以为 IAM 实体（IAM 用户或角色）授予的最大权限。您可为实体设置权限边界。这些结果权限是实体的基于身份的策略及其权限边界的交集。在 `Principal` 中指定用户或角色的基于资源的策略不受权限边界限制。任一项策略中的显式拒绝将覆盖允许。有关权限边界的更多信息，请参阅IAM 用户指南中的 IAM 实体的权限边界。
- 服务控制策略 (SCPs) — SCPs 是 JSON 策略，用于指定中组织或组织单位 (OU) 的最大权限 AWS Organizations。 AWS Organizations 是一项用于对您的企业拥有的多 AWS 账户 项进行分组和集中管理的服务。如果您启用组织中的所有功能，则可以将服务控制策略 (SCPs) 应用于您的任何或所有帐户。SCP 限制成员账户中的实体（包括每个 AWS 账户根用户实体）的权限。有关 Organizations 和的更多信息 SCPs，请参阅《AWS Organizations 用户指南》中的 SCPs工作原理。
- 会话策略：会话策略是当您以编程方式为角色或联合用户创建临时会话时作为参数传递的高级策略。结果会话的权限是用户或角色的基于身份的策略和会话策略的交集。权限也可以来自基于资源的策略。任一项策略中的显式拒绝将覆盖允许。有关更多信息，请参阅 IAM 用户指南中的会话策略。

多个策略类型

当多个类型的策略应用于一个请求时，生成的权限更加复杂和难以理解。要了解在涉及多种策略类型时如何 AWS 确定是否允许请求，请参阅 IAM 用户指南中的策略评估逻辑。

如何 AWS Clean Rooms 与 IAM 配合使用

在使用 IAM 管理访问权限之前 AWS Clean Rooms，请先了解哪些可用的 IAM 功能 AWS Clean Rooms。

您可以搭配使用的 IAM 功能 AWS Clean Rooms

IAM 功能	AWS Clean Rooms 支持
基于身份的策略	是
基于资源的策略	部分
策略操作	是
策略资源	是
策略条件键（特定于服务）	部分
ACLs	否
ABAC（策略中的标签）	是
临时凭证	是
转发访问会话（FAS）	是
服务角色	是
服务关联角色	否

要全面了解大多数 IAM 功能的使用 AWS 服务 方式 AWS Clean Rooms 和其他功能，请参阅AWS 服务 IAM 用户指南中的与 IA M 配合使用的内容。

基于身份的策略 AWS Clean Rooms

支持基于身份的策略：是

基于身份的策略是可附加到身份（如 IAM 用户、用户组或角色）的 JSON 权限策略文档。这些策略控制用户和角色可在何种条件下对哪些资源执行哪些操作。要了解如何创建基于身份的策略，请参阅《IAM 用户指南》中的使用客户管理型策略定义自定义 IAM 权限。

通过使用 IAM 基于身份的策略，您可以指定允许或拒绝的操作和资源以及允许或拒绝操作的条件。要了解可在 JSON 策略中使用的所有元素，请参阅《IAM 用户指南》中的 IAM JSON 策略元素引用。

基于身份的策略示例 AWS Clean Rooms

要查看 AWS Clean Rooms 基于身份的策略的示例，请参阅。基于身份的策略示例 AWS Clean Rooms

内部基于资源的政策 AWS Clean Rooms

支持基于资源的策略：部分支持

基于资源的策略是附加到资源的 JSON 策略文档。基于资源的策略的示例包括 IAM 角色信任策略和 Amazon S3 存储桶策略。在支持基于资源的策略的服务中，服务管理员可以使用它们来控制对特定资源的访问。对于在其中附加策略的资源，策略定义指定主体可以对该资源执行哪些操作以及在什么条件下执行。您必须在基于资源的策略中指定主体。委托人可以包括账户、用户、角色、联合用户或 AWS 服务。

要启用跨账户访问，您可以将整个账户或其他账户中的 IAM 实体指定为基于资源的策略中的主体。有关更多信息，请参阅《IAM 用户指南》中的 IAM 中的跨账户资源访问。

该 AWS Clean Rooms 服务仅支持一种基于资源的策略，称为配置的相似模型托管资源策略，该策略附加到已配置的相似模型上。此策略定义了哪些主体可以对配置的相似模型执行操作。

要了解如何将基于资源的策略附加到配置的相似模型，请参阅AWS Clean Rooms ML 的 IAM 行为。

的政策行动 AWS Clean Rooms

支持策略操作：是

管理员可以使用 AWS JSON 策略来指定谁有权访问什么。也就是说，哪个主体可以对什么资源执行操作，以及在什么条件下执行。

JSON 策略的 `Action` 元素描述可用于在策略中允许或拒绝访问的操作。在策略中包含操作以授予执行关联操作的权限。

要查看 AWS Clean Rooms 操作列表，请参阅《服务授权参考》 AWS Clean Rooms中定义的操作。

正在执行的策略操作在操作前 AWS Clean Rooms 使用以下前缀。

```
cleanrooms
```

要在单个语句中指定多项操作，请使用逗号将它们隔开。

```
"Action": [
      "cleanrooms:action1",
      "cleanrooms:action2"
         ]
```

要查看 AWS Clean Rooms 基于身份的策略的示例，请参阅。基于身份的策略示例 AWS Clean Rooms

的政策资源 AWS Clean Rooms

支持策略资源：是

管理员可以使用 AWS JSON 策略来指定谁有权访问什么。也就是说，哪个主体可以对什么资源执行操作，以及在什么条件下执行。

`Resource` JSON 策略元素指定要向其应用操作的一个或多个对象。作为最佳实践，请使用其 Amazon 资源名称（ARN）指定资源。对于不支持资源级权限的操作，请使用通配符 (*) 指示语句应用于所有资源。

```
"Resource": "*"
```

要查看 AWS Clean Rooms 资源类型及其列表 ARNs，请参阅《服务授权参考》 AWS Clean Rooms中定义的资源。要了解可以在哪些操作中指定每个资源的 ARN，请参阅 AWS Clean Rooms定义的操作。

要查看 AWS Clean Rooms 基于身份的策略的示例，请参阅。基于身份的策略示例 AWS Clean Rooms

的策略条件密钥 AWS Clean Rooms

支持特定于服务的策略条件键：部分

管理员可以使用 AWS JSON 策略来指定谁有权访问什么。也就是说，哪个主体可以对什么资源执行操作，以及在什么条件下执行。

`Condition` 元素根据定义的条件指定语句何时执行。您可以创建使用条件运算符（例如，等于或小于）的条件表达式，以使策略中的条件与请求中的值相匹配。要查看所有 AWS 全局条件键，请参阅 IAM 用户指南中的AWS 全局条件上下文密钥。

要了解 AWS Clean Rooms ML 如何使用策略条件密钥，请参阅AWS Clean Rooms ML 的 IAM 行为。

ACLs in AWS Clean Rooms

支持 ACLs：否

访问控制列表 (ACLs) 控制哪些委托人（账户成员、用户或角色）有权访问资源。 ACLs 与基于资源的策略类似，尽管它们不使用 JSON 策略文档格式。

ABAC with AWS Clean Rooms

支持 ABAC（策略中的标签）：是

基于属性的访问权限控制（ABAC）是一种授权策略，该策略基于称为标签的属性来定义权限。您可以将标签附加到 IAM 实体和 AWS 资源，然后设计 ABAC 策略以允许在委托人的标签与资源上的标签匹配时进行操作。

要基于标签控制访问，您需要使用 `aws:ResourceTag/`*`key-name`*`aws:RequestTag/`*`key-name`* 或 `aws:TagKeys` 条件键在策略的条件元素中提供标签信息。

如果某个服务对于每种资源类型都支持所有这三个条件键，则对于该服务，该值为是。如果某个服务仅对于部分资源类型支持所有这三个条件键，则该值为部分。

有关 ABAC 的更多信息，请参阅《IAM 用户指南》中的使用 ABAC 授权定义权限。要查看设置 ABAC 步骤的教程，请参阅《IAM 用户指南》中的使用基于属性的访问权限控制（ABAC）。

将临时凭证与配合使用 AWS Clean Rooms

支持临时凭证：是

临时证书提供对 AWS 资源的短期访问权限，并且是在您使用联合身份或切换角色时自动创建的。AWS 建议您动态生成临时证书，而不是使用长期访问密钥。有关更多信息，请参阅《IAM 用户指南》中的 IAM 中的临时安全凭证和使用 IAM 的。AWS 服务

转发访问会话 AWS Clean Rooms

支持转发访问会话（FAS）：是

转发访问会话 (FAS) 使用调用主体的权限 AWS 服务，再加上 AWS 服务 向下游服务发出请求的请求。有关发出 FAS 请求时的策略详情，请参阅转发访问会话。

的服务角色 AWS Clean Rooms

支持服务角色：是

服务角色是由一项服务担任、代表您执行操作的 IAM 角色。IAM 管理员可以在 IAM 中创建、修改和删除服务角色。有关更多信息，请参阅《IAM 用户指南》中的创建向 AWS 服务委派权限的角色。

> **Warning**
>
> 更改服务角色的权限可能会中断 AWS Clean Rooms 功能。只有在 AWS Clean Rooms 提供操作指导时才编辑服务角色。

的服务相关角色 AWS Clean Rooms

支持服务相关角色：否

服务相关角色是一种与服务相关联的 AWS 服务服务角色。服务可以代入代表您执行操作的角色。服务相关角色出现在您的中 AWS 账户 ，并且归服务所有。IAM 管理员可以查看但不能编辑服务关联角色的权限。

有关创建或管理服务相关角色的详细信息，请参阅能够与 IAM 搭配使用的AWS 服务。在表中查找服务相关角色列中包含 Yes 的表。选择是链接以查看该服务的服务相关角色文档。

基于身份的策略示例 AWS Clean Rooms

默认情况下，用户和角色没有创建或修改 AWS Clean Rooms 资源的权限。要授予用户对所需资源执行操作的权限，IAM 管理员可以创建 IAM 策略。

要了解如何使用这些示例 JSON 策略文档创建基于 IAM 身份的策略，请参阅《IAM 用户指南》中的创建 IAM 策略（控制台）。

有关由 AWS Clean Rooms定义的操作和资源类型（包括每种资源类型的格式）的详细信息，请参阅《服务授权参考》 AWS Clean Rooms中的操作、资源和条件密钥。 ARNs

主题

策略最佳实践

基于身份的策略决定了某人是否可以在您的账户中创建、访问或删除 AWS Clean Rooms 资源。这些操作可能会使 AWS 账户产生成本。创建或编辑基于身份的策略时，请遵循以下指南和建议：

- 开始使用 AWS 托管策略并转向最低权限权限 — 要开始向用户和工作负载授予权限，请使用为许多常见用例授予权限的AWS 托管策略。它们在你的版本中可用 AWS 账户。我们建议您通过定义针对您的用例的 AWS 客户托管策略来进一步减少权限。有关更多信息，请参阅《IAM 用户指南》中的 AWS 托管策略或工作职能的AWS 托管策略。
- 应用最低权限：在使用 IAM 策略设置权限时，请仅授予执行任务所需的权限。为此，您可以定义在特定条件下可以对特定资源执行的操作，也称为最低权限许可。有关使用 IAM 应用权限的更多信息，请参阅《IAM 用户指南》中的 IAM 中的策略和权限。
- 使用 IAM 策略中的条件进一步限制访问权限：您可以向策略添加条件来限制对操作和资源的访问。例如，您可以编写策略条件来指定必须使用 SSL 发送所有请求。如果服务操作是通过特定的方式使用的，则也可以使用条件来授予对服务操作的访问权限 AWS 服务，例如 CloudFormation。有关更多信息，请参阅《IAM 用户指南》中的 IAM JSON 策略元素：条件。
- 使用 IAM Access Analyzer 验证您的 IAM 策略，以确保权限的安全性和功能性：IAM Access Analyzer 会验证新策略和现有策略，以确保策略符合 IAM 策略语言（JSON）和 IAM 最佳实践。IAM Access Analyzer 提供 100 多项策略检查和可操作的建议，以帮助您制定安全且功能性强的策略。有关更多信息，请参阅《IAM 用户指南》中的使用 IAM Access Analyzer 验证策略。
- 需要多重身份验证 (MFA)-如果 AWS 账户您的场景需要 IAM 用户或根用户，请启用 MFA 以提高安全性。若要在调用 API 操作时需要 MFA，请将 MFA 条件添加到您的策略中。有关更多信息，请参阅《IAM 用户指南》中的使用 MFA 保护 API 访问。

有关 IAM 中的最佳实操的更多信息，请参阅《IAM 用户指南》中的 IAM 中的安全最佳实践。

使用控制 AWS Clean Rooms 台

要访问 AWS Clean Rooms 控制台，您必须拥有一组最低权限。这些权限必须允许您列出和查看有关您的 AWS Clean Rooms 资源的详细信息 AWS 账户。如果创建比必需的最低权限更为严格的基于身份的策略，对于附加了该策略的实体（用户或角色），控制台将无法按预期正常运行。

对于仅调用 AWS CLI 或 AWS API 的用户，您无需为其设置最低控制台权限。相反，只允许访问与其尝试执行的 API 操作相匹配的操作。

为确保用户和角色仍然可以使用 AWS Clean Rooms 控制台，还需要将 AWS Clean Rooms *FullAccess*或*ReadOnly* AWS 托管策略附加到实体。有关更多信息，请参阅《IAM 用户指南》中的为用户添加权限。

允许用户查看他们自己的权限

该示例说明了您如何创建策略，以允许 IAM 用户查看附加到其用户身份的内联和托管式策略。此策略包括在控制台上或使用 AWS CLI 或 AWS API 以编程方式完成此操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ViewOwnUserInfo",
            "Effect": "Allow",
            "Action": [
                "iam:GetUserPolicy",
                "iam:ListGroupsForUser",
                "iam:ListAttachedUserPolicies",
                "iam:ListUserPolicies",
                "iam:GetUser"
            ],
            "Resource": ["arn:aws:iam::*:user/${aws:username}"]
        },
        {
            "Sid": "NavigateInConsole",
            "Effect": "Allow",
            "Action": [
                "iam:GetGroupPolicy",
                "iam:GetPolicyVersion",
                "iam:GetPolicy",
                "iam:ListAttachedGroupPolicies",
                "iam:ListGroupPolicies",
```

```
                "iam:ListPolicyVersions",
                "iam:ListPolicies",
                "iam:ListUsers"
            ],
            "Resource": "*"
        }
    ]
}
```

AWS 的托管策略 AWS Clean Rooms

AWS 托管策略是由创建和管理的独立策略 AWS。 AWS 托管策略旨在为许多常见用例提供权限，以便您可以开始为用户、组和角色分配权限。

请记住， AWS 托管策略可能不会为您的特定用例授予最低权限权限，因为它们可供所有 AWS 客户使用。我们建议通过定义特定于使用案例的客户管理型策略来进一步减少权限。

您无法更改 AWS 托管策略中定义的权限。如果 AWS 更新 AWS 托管策略中定义的权限，则更新会影响该策略所关联的所有委托人身份（用户、组和角色）。 AWS 最有可能在启动新的 API 或现有服务可以使用新 AWS 服务 的 API 操作时更新 AWS 托管策略。

有关更多信息，请参阅《IAM 用户指南》中的 AWS 托管式策略。

AWS 托管策略：**AWSCleanRoomsReadOnlyAccess**

您可以将 AWSCleanRoomsReadOnlyAccess 附加到 IAM 主体。

该策略授予 AWSCleanRoomsReadOnlyAccess 协作中的资源和元数据的只读权限。

权限详细信息

该策略包含以下权限：

- CleanRoomsRead - 允许主体对服务进行只读访问。
- ConsoleDisplayTables— 允许委托人对在控制台上显示有关基础 AWS Glue 表的数据所需的 AWS Glue 元数据的只读访问权限。
- ConsoleLogSummaryQueryLogs - 允许主体查看查询日志。
- ConsoleLogSummaryObtainLogs - 允许主体检索日志结果。

有关策略详细信息的 JSON 列表，请参阅AWS 托管策略参考指南AWSCleanRoomsReadOnlyAccess中的。

AWS 托管策略：AWSCleanRoomsFullAccess

您可以将 AWSCleanRoomsFullAccess 附加到 IAM 主体。

此策略授予管理权限，允许对 AWS Clean Rooms 协作中的资源和元数据进行完全访问（读取、写入和更新）。此策略包括执行查询的权限。

权限详细信息

该策略包含以下权限：

- CleanRoomsAccess— 授予对所有资源执行所有操作的完全访问权限 AWS Clean Rooms。
- PassServiceRole - 仅授予将服务角色传递给名称中带有“cleanrooms”的服务（PassedToService 条件）的访问权限。
- ListRolesToPickServiceRole— 允许委托人列出其所有角色以便在使用 AWS Clean Rooms时选择服务角色。
- GetRoleAndListRolePoliciesToInspectServiceRole - 允许主体在 IAM 中查看服务角色和相应的策略。
- ListPoliciesToInspectServiceRolePolicy - 允许主体在 IAM 中查看服务角色和相应的策略。
- GetPolicyToInspectServiceRolePolicy - 允许主体在 IAM 中查看服务角色和相应的策略。
- ConsoleDisplayTables— 允许委托人对在控制台上显示有关基础 AWS Glue 表的数据所需的 AWS Glue 元数据的只读访问权限。
- ConsolePickQueryResultsBucketListAll - 允许主体从查询结果写入的所有可用 S3 存储桶的列表中选择一个 Amazon S3 存储桶。
- SetQueryResultsBucket - 允许主体选择查询结果写入的 S3 存储桶。
- ConsoleDisplayQueryResults - 允许主体向客户显示从 S3 存储桶读取的查询结果。
- WriteQueryResults - 允许主体将查询结果写入客户拥有的 S3 存储桶。
- EstablishLogDeliveries— 允许委托人将查询日志传送到客户的 Amazon Lo CloudWatch gs 日志组。
- SetupLogGroupsDescribe— 允许委托人使用 Amazon Logs CloudWatch 日志组的创建流程。
- SetupLogGroupsCreate— 允许委托人创建 Amazon CloudWatch 日志组。

- SetupLogGroupsResourcePolicy— 允许委托人在 Amazon Logs CloudWatch 日志组上设置资源策略。
- ConsoleLogSummaryQueryLogs - 允许主体查看查询日志。
- ConsoleLogSummaryObtainLogs - 允许主体检索日志结果。

有关策略详细信息的 JSON 列表，请参阅AWS 托管策略参考指南AWSCleanRoomsFullAccess中的。

AWS 托管策略：AWSCleanRoomsFullAccessNoQuerying

您可以将 AWSCleanRoomsFullAccessNoQuerying 附加到 IAM principals。

此策略授予管理权限，允许对 AWS Clean Rooms 协作中的资源和元数据进行完全访问（读取、写入和更新）。此策略不包括执行查询的权限。

权限详细信息

该策略包含以下权限：

- CleanRoomsAccess— 授予对所有资源的所有操作的完全访问权限 AWS Clean Rooms，协作中查询除外。
- CleanRoomsNoQuerying - 明确拒绝 StartProtectedQuery 和 UpdateProtectedQuery，阻止查询。
- PassServiceRole - 仅授予将服务角色传递给名称中带有“cleanrooms”的服务（PassedToService 条件）的访问权限。
- ListRolesToPickServiceRole— 允许委托人列出其所有角色以便在使用 AWS Clean Rooms时选择服务角色。
- GetRoleAndListRolePoliciesToInspectServiceRole - 允许主体在 IAM 中查看服务角色和相应的策略。
- ListPoliciesToInspectServiceRolePolicy - 允许主体在 IAM 中查看服务角色和相应的策略。
- GetPolicyToInspectServiceRolePolicy - 允许主体在 IAM 中查看服务角色和相应的策略。
- ConsoleDisplayTables— 允许委托人对在控制台上显示有关基础 AWS Glue 表的数据所需的 AWS Glue 元数据的只读访问权限。
- EstablishLogDeliveries— 允许委托人将查询日志传送到客户的 Amazon Lo CloudWatch gs 日志组。
- SetupLogGroupsDescribe— 允许委托人使用 Amazon Logs CloudWatch 日志组的创建流程。

- SetupLogGroupsCreate— 允许委托人创建 Amazon CloudWatch 日志组。
- SetupLogGroupsResourcePolicy— 允许委托人在 Amazon Logs CloudWatch 日志组上设置资源策略。
- ConsoleLogSummaryQueryLogs - 允许主体查看查询日志。
- ConsoleLogSummaryObtainLogs - 允许主体检索日志结果。
- cleanrooms - 管理 AWS Clean Rooms 服务中的协作、分析模板、配置表、成员身份和关联资源。执行各种操作，例如创建、更新、删除、列出和检索有关这些资源的信息。
- iam— 将名称包含 “cleanrooms” 的服务角色传递给 AWS Clean Rooms 服务。列出角色、策略，并检查服务角色和与 AWS Clean Rooms 服务相关的策略。
- glue— 从中检索有关数据库、表、分区和架构的信息。 AWS Glue这是 AWS Clean Rooms 服务显示底层数据源并与之交互所必需的。
- logs— 管理日志传送、日志组和 CloudWatch 日志资源策略。查询和检索与 AWS Clean Rooms 服务相关的日志。对于在服务中进行监控、审计和故障排除，必须具备这些权限。

该策略还明确拒绝 cleanrooms:StartProtectedQuery 和 cleanrooms:UpdateProtectedQuery 操作，以防用户直接执行或更新受保护的查询，这些操作应当通过 AWS Clean Rooms 受控机制完成。

有关策略详细信息的 JSON 列表，请参阅AWS 托管策略参考指南AWSCleanRoomsFullAccessNoQuerying中的。

AWS 托管策略：AWSCleanRoomsMLReadOnlyAccess

您可以将 AWSCleanRoomsMLReadOnlyAccess 附加到 IAM 主体。

该策略授予 AWSCleanRoomsMLReadOnlyAccess 协作中的资源和元数据的只读权限。

该策略包含以下权限：

- CleanRoomsConsoleNavigation— 授予查看 AWS Clean Rooms 控制台屏幕的权限。
- CleanRoomsMLRead - 允许 Clean Rooms ML 对服务进行只读访问。
- PassCleanRoomsResources— 授予传递指定 AWS Clean Rooms 资源的访问权限。

有关策略详细信息的 JSON 列表，请参阅AWSClean《AWS 托管策略参考指南》MLReadOnlyAccess中的 Rooms。

AWS 托管策略：**AWSCleanRoomsMLFullAccess**

您可以将 AWSCleanRoomsMLFullAcces 附加到 IAM 主体。该策略授予管理权限，以允许对 Clean Rooms ML 所需的资源和元数据进行完全访问（读取、写入和更新）。

权限详细信息

该策略包含以下权限：

- CleanRoomsMLFullAccess - 授予所有 Clean Rooms ML 操作的访问权限。
- PassServiceRole - 仅授予将服务角色传递给名称中带有“cleanrooms-ml”的服务（PassedToService 条件）的访问权限。
- CleanRoomsConsoleNavigation— 授予查看 AWS Clean Rooms 控制台屏幕的权限。
- CollaborationMembershipCheck— 当您在协作中启动受众生成（相似区段）工作时，Clean Rooms ML 服务会调用ListMembers以检查协作是否有效，来电者是否为活跃成员，配置的受众模型所有者是否为活跃成员。始终需要该权限；仅控制台用户需要控制台导航 SID。
- PassCleanRoomsResources— 授予传递指定 AWS Clean Rooms 资源的访问权限。
- AssociateModels - 允许主体将 Clean Rooms ML 模型与您的协作相关联。
- TagAssociations - 允许主体将标签添加到相似模型和协作之间的关联中。
- ListRolesToPickServiceRole— 允许委托人列出其所有角色以便在使用 AWS Clean Rooms时选择服务角色。
- GetRoleAndListRolePoliciesToInspectServiceRole - 允许主体在 IAM 中查看服务角色和相应的策略。
- ListPoliciesToInspectServiceRolePolicy - 允许主体在 IAM 中查看服务角色和相应的策略。
- GetPolicyToInspectServiceRolePolicy - 允许主体在 IAM 中查看服务角色和相应的策略。
- ConsoleDisplayTables— 允许委托人对在控制台上显示有关基础 AWS Glue 表的数据所需的 AWS Glue 元数据的只读访问权限。
- ConsolePickOutputBucket - 允许主体为配置的受众模型输出选择 Amazon S3 存储桶。
- ConsolePickS3Location - 允许主体为配置的受众模型输出选择存储桶中的位置。
- ConsoleDescribeECRRepositories— 允许委托人描述 Amazon ECR 存储库和映像。

有关策略详细信息的 JSON 列表，请参阅《AWS 托管策略参考指南》中的AWSClean房间MLFull访问权限。

AWS Clean Rooms AWS 托管策略的更新

查看 AWS Clean Rooms 自该服务开始跟踪这些更改以来 AWS 托管策略更新的详细信息。要获得有关此页面变更的自动提醒，请订阅 “ AWS Clean Rooms 文档历史记录” 页面上的 RSS feed。

更改	描述	日期
AWSCleanRoomsFullAccessNoQuerying— 更新现有政策	将洁净室:UpdateConfiguredTableAllowedColumns 和洁净室:UpdateConfiguredTableReference 添加到。CleanRoomsAccess	2025 年 7 月 29 日
AWSCleanRoomsMLReadOnlyAccess - 对现有策略的更新	已将 PassCleanRoomsResources 添加到 AWSCleanRoomsMLReadOnlyAccess。已将 PassCleanRoomsResources 和 ConsoleDescribeECRRepositories 添加到 AWSCleanRoomsMLFullAccess。	2025 年 1 月 10 日
AWSCleanRoomsFullAccessNoQuerying - 对现有策略的更新	已将 cleanrooms:BatchGetSchemaAnalysisRule 添加到 CleanRoomsAccess。	2024 年 5 月 13 日
AWSCleanRoomsFullAccess - 对现有策略的更新	在此策略中将 AWSCleanRoomsFullAccess 中的语句 ID 从 ConsolePickQueryResultsBucket 更新为 SetQueryResultsBucket，以更好地表示权限，因为无论使用控制台还是不使用控制台，都需要这些权限来设置查询结果存储桶。	2024 年 3 月 21 日
AWSCleanRoomsMLReadOnlyAccess - 新策略 AWSCleanRoomsMLFullAccess - 新策略	添加AWSCleanRoomsMLReadOnlyAccess并AWSCleanRoomsMLFullAccess支持 AWS Clean Rooms ML。	2023 年 11 月 29 日
AWSCleanRoomsFullAccessNoQuerying - 对现有策略的更新	向 CleanRoomsAccess 添加了 cleanrooms:CreateAnalysisTe	2023 年 7 月 31 日

更改	描述	日期
	mplate、cleanrooms:GetAnalysisTemplate、cleanrooms:UpdateAnalysisTemplate、 cleanrooms:DeleteAnalysisTemplate、cleanrooms:ListAnalysisTemplates、cleanrooms:GetCollaborationAnalysisTemplate、cleanrooms:BatchGetCollaborationAnalysisTemplate 和 cleanrooms:ListCollaborationAnalysisTemplates，以启用新的分析模板特征。	
AWSCleanRoomsFullAccessNoQuerying - 对现有策略的更新	向 CleanRoomsAccess 添加了 cleanrooms:ListTagsForResource、cleanrooms:UntagResource 和 cleanrooms:TagResource，以启用资源标记。	2023 年 3 月 21 日
AWS Clean Rooms 已开始跟踪更改	AWS Clean Rooms 开始跟踪其 AWS 托管策略的更改。	2023 年 1 月 12 日

对 AWS Clean Rooms 身份和访问进行故障排除

使用以下信息来帮助您诊断和修复在使用 AWS Clean Rooms 和 IAM 时可能遇到的常见问题。

主题

- 我无权在以下位置执行操作 AWS Clean Rooms
- 我无权执行 iam：PassRole
- 我想允许我以外的人 AWS 账户 访问我的 AWS Clean Rooms 资源

我无权在以下位置执行操作 AWS Clean Rooms

如果您收到错误提示，指明您无权执行某个操作，则必须更新策略以允许执行该操作。

当 mateojackson IAM 用户尝试使用控制台查看有关虚构 *my-example-widget* 资源的详细信息，但不拥有虚构 cleanrooms:*GetWidget* 权限时，会发生以下示例错误。

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform:
 cleanrooms:GetWidget on resource: my-example-widget
```

在此情况下，Mateo 的策略必须更新以允许其使用 cleanrooms:*GetWidget* 操作访问 *my-example-widget* 资源。

如果您需要帮助，请联系您的 AWS 管理员。您的管理员是提供登录凭证的人。

我无权执行 iam：PassRole

如果您收到一个错误，表明您无权执行 iam:PassRole 操作，则必须更新策略以允许您将角色传递给。 AWS Clean Rooms

有些 AWS 服务 允许您将现有角色传递给该服务，而不是创建新的服务角色或服务相关角色。为此，您必须具有将角色传递到服务的权限。

当名为 marymajor 的 IAM 用户尝试使用控制台在 AWS Clean Rooms中执行操作时，会发生以下示例错误。但是，服务必须具有服务角色所授予的权限才可执行此操作。Mary 不具有将角色传递到服务的权限。

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
 iam:PassRole
```

在这种情况下，必须更新 Mary 的策略以允许她执行 iam:PassRole 操作。

如果您需要帮助，请联系您的 AWS 管理员。您的管理员是提供登录凭证的人。

我想允许我以外的人 AWS 账户 访问我的 AWS Clean Rooms 资源

您可以创建一个角色，以便其他账户中的用户或您组织外的人员可以使用该角色来访问您的资源。您可以指定谁值得信赖，可以担任角色。

要了解更多信息，请参阅以下内容：

- 要了解是否 AWS Clean Rooms 支持这些功能，请参阅如何 AWS Clean Rooms 与 IAM 配合使用。
- 要了解如何提供对您拥有的资源的访问权限 AWS 账户 ，请参阅 IAM 用户指南中的向您拥有 AWS 账户 的另一个 IAM 用户提供访问权限。

- 要了解如何向第三方提供对您的资源的访问权限 AWS 账户，请参阅 IAM 用户指南中的向第三方提供访问权限。 AWS 账户
- 要了解如何通过身份联合验证提供访问权限，请参阅《IAM 用户指南》中的为经过外部身份验证的用户（联合身份验证）提供访问权限。
- 要了解使用角色和基于资源的策略进行跨账户存取之间的差别，请参阅《IAM 用户指南》中的 IAM 角色与基于资源的策略有何不同。

防止跨服务混淆代理

混淆代理问题是一个安全性问题，即不具有某操作执行权限的实体可能会迫使具有更高权限的实体执行该操作。在中 AWS，跨服务模仿可能会导致混乱的副手问题。一个服务（呼叫服务）调用另一项服务（所谓的服务）时，可能会发生跨服务模拟。可以操纵调用服务，使用其权限以在其他情况下该服务不应有访问权限的方式对另一个客户的资源进行操作。为防止这种情况， AWS 提供可帮助您保护所有服务的数据的工具，而这些服务中的服务主体有权限访问账户中的资源。

我们建议在资源策略中使用 `aws:SourceArn` 全局条件上下文键，以限制 AWSClean Rooms 授予其他服务对资源的权限。如果您只希望将一个资源与跨服务访问相关联，请使用。`aws:SourceArn`

防范混淆代理问题最有效的方法是使用 `aws:SourceArn` 全局条件上下文键和资源的完整 ARN。在中 AWSClean Rooms，您还必须与`sts:ExternalId`条件键进行比较。

`aws:SourceArn` 的值必须设置为所担任角色的成员身份的 ARN。

以下示例演示如何使用 AWSClean Rooms 中的 `aws:SourceArn` 全局条件上下文键来防范混淆代理问题。

> **Note**
>
> 示例策略适用于 AWS Clean Rooms 用于访问已配置表的数据和元数据的服务角色的信任策略。
> 的值*<query-runner-membership-id>*需要设置为查询运行器的成员资格 ID。
> 协作的所有成员都可以查看已配置的表格元数据，因此每个成员资格 ARN 都必须包含在成员资格列表中。 ARNs

> **Note**
>
> 通过 AWS Clean Rooms 控制台创建服务角色时，默认情况下，协作的所有当前成员都将包含在混乱的副手条件中。
> 如果您要向已配置与其关联的表格的协作中添加新成员，请务必使用新成员的成员资格 ARN 更新服务角色的混淆副手条件。
> 如果您在添加新成员后没有更新服务角色混乱的副手状况，则该新成员将无法访问使用 AWS Clean Rooms 该角色检索到的信息。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowIfExternalIdMatches",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringLike": {
                    "sts:ExternalId": "arn:aws:*:us-east-1:*:dbuser:*/<query-
runner-membership-id>*"
                }
            }
        },
        {
            "Sid": "AllowIfSourceArnMatches",
            "Effect": "Allow",
            "Principal": {
                "Service": "cleanrooms.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "ForAnyValue:ArnEquals": {
                    "aws:SourceArn": [
                        "arn:aws:cleanrooms:us-
east-1:111122223333:membership/<member-1-membership-id>",
```

```
                    "arn:aws:cleanrooms:us-
east-1:444455556666:membership/<member-2-membership-id>"
                ]
            }
        }
    }
]
}
```

AWS Clean Rooms ML 的 IAM 行为

跨账户作业

Clean Rooms ML AWS 账户 允许一个人在其帐户中安全地访问另一个人创建的某些资源 AWS 账户。当 AWS 账户 A 中的客户端调用 AWS 账户 B 拥有`StartAudienceGenerationJob`的`ConfiguredAudienceModel`资源时，Clean Rooms ML 会 ARNs 为该任务创建两个资源。一个 ARN 在 AWS 账户 A 中，另一个 ARN 在 B 中 AWS 账户 除了 ARNs 它们之外，它们是相同的 AWS 账户。

Clean Rooms ML ARNs 为任务创建了两个，以确保两个账户都可以将自己的 IAM 策略应用于任务。例如，两个账户都可以使用基于标签的访问控制并应用其 AWS 组织的策略。作业处理来自两个账户的数据，因此，两个账户都可以删除作业及其关联数据。两个账户都不能阻止另一个账户删除作业。

只能执行一个作业，两个账户可以在调用 `ListAudienceGenerationJobs` 时看到该作业。两个账户都可以使用自己 AWS 账户 的 `Export` APIs ID 的 ARN 调用`Delete`、和。`Get`

使用带有另一 AWS 账户 个 ID 的 ARN 时，两者都 AWS 账户 无法访问任务。

作业的名称在 AWS 账户中必须是唯一的。 AWS 账户 B 中的名字是*$accountA-$name*。在 B 中查看作业时 AWS 账户 ，A 选择的名称以 AWS 账户 A 为前缀。 AWS 账户

为了使跨账户`StartAudienceGenerationJob`成功， AWS 账户 B 必须使用类似于以下示例的资源策略允许对 AWS 账户 B 中的新任务和 AWS 账户 B `ConfiguredAudienceModel` 中的新任务执行该操作：

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
```

```
        {
            "Sid": "Clean-Rooms-CAMA-ID",
            "Effect": "Allow",
            "Principal": {
                "AWS": [
                    "111122223333"
                ]
            },
            "Action": [
                "cleanrooms-ml:StartAudienceGenerationJob"
            ],
            "Resource": [
                "arn:aws:cleanrooms-ml:us-east-1:444455556666:configured-
audience-model/id",
                "arn:aws:cleanrooms-ml:us-east-1:444455556666:audience-
generation-job/*"
            ],
            "Condition":{"StringEquals":{"cleanrooms-ml:CollaborationId":"UUID"}}
        }
    ]
}
```

Note

此 AWS Clean Rooms 机器学习资源策略引用了两种不同的策略 AWS 账户 IDs 来支持跨账户受众生成：

- 111122223333-该账户包含授权开始受众生成工作的委托人（用户、角色或服务）。此账户启动机器学习处理工作流程。
- 444455556666-这是拥有 AWS Clean Rooms 机器学习资源（配置的受众模型和受众生成作业）的账户。此账户托管 ML 模型并管理任务执行。

其他配置说明：

- 报表 ID (Sid)：`CAMA-ID`替换为您的实际 AWS Clean Rooms 受众模型应用程序 (CAMA) 标识符，以使政策声明易于识别。
- 资源 IDs：*id*替换为您配置的受众模型的实际 ID 和*UUID*您的特定协作 ID。
- 条件：该`cleanrooms-ml:CollaborationId`条件可确保受众生成作业只能在指定的 AWS Clean Rooms 协作环境中启动，从而提供了额外的安全边界。

这种跨账户配置支持这样的场景：一个组织管理机器学习模型和基础架构，同时允许授权合作伙伴在其合作协议的范围内启动受众生成流程。

如果您使用 AWS Clean Rooms ML API 创建manageResourcePolicies设置为 true 的配置相似模型，则会为您 AWS Clean Rooms 创建此策略。

此外， AWS 账户 A 中来电者的身份策略需要获得StartAudienceGenerationJob许可arn:aws:cleanrooms-ml:us-west-1:AccountA:audience-generation-job/*。因此，有三个 IAM 资源可供操作StartAudienceGenerationJob： AWS 账户 A 作业、 AWS 账户 B 作业和 AWS 账户 B ConfiguredAudienceModel。

Warning

启动 AWS 账户 该作业的用户会收到有关该作业的 AWS CloudTrail 审核日志事件。拥有 ConfiguredAudienceModel 的 AWS 账户 不会收到 AWS CloudTrail 审核日志事件。

标记作业

在您设置 CreateConfiguredAudienceModel 的 childResourceTagOnCreatePolicy=FROM_PARENT_RESOURCE 参数时，您的账户中通过该配置的相似模型创建的所有相似细分生成作业默认具有与配置的相似模型相同的标签。配置的相似模型是父模型，相似细分生成作业是子模型。

如果您在自己的账户中创建作业，作业的请求标签将覆盖父标签。其他账户创建的作业绝不会在您的账户中创建标签。如果您设置 childResourceTagOnCreatePolicy=FROM_PARENT_RESOURCE 并且另一个账户创建作业，则作业具有两个副本。您的账户中的副本具有父资源标签，作业提交者账户中的副本具有来自请求的标签。

验证协作者

向 AWS Clean Rooms 协作中的其他成员授予权限时，资源策略应包含条件键cleanrooms-ml:CollaborationId。这会强制要求collaborationId参数包含在StartAudienceGenerationJob请求中。在请求中包含 collaborationId 参数时，Clean Rooms ML 验证协作是否存在，作业提交者是否为协作的活跃成员，以及配置的相似模型所有者是否为协作的活跃成员。

AWS Clean Rooms 管理您配置的相似模型资源策略（manageResourcePolicies参数在CreateConfiguredAudienceModelAssociation 请求TRUE中）时，将在资源策略中设置此条件密钥。因此，必须在 in collaborationId 中指定StartAudienceGenerationJob。

跨账户访问

只能跨账户调用 StartAudienceGenerationJob。所有其他 Clean Rooms ML APIs 只能与您自己账户中的资源一起使用。这可确保您的训练数据、相似模型配置和其他信息保持私密。

Clean Rooms ML 永远不会透露 Amazon S3 或各个账户 AWS Glue 的位置。训练数据位置、配置的相似模型输出位置和相似细分生成作业种子位置绝不会在账户之间可见。除非在协作中启用了查询日志记录，否则种子数据是否来自某个 SQL 查询以及查询本身在账户中都不可见。如果您获取 (Get) 另一个账户提交的受众生成作业，该服务不会显示种子位置。

洁净室机器学习自定义模型的 IAM 行为

跨账户作业

Clean Rooms ML 允许另一个人在其帐户中安全地访问与一个人 AWS 账户 创建的协作关联的某些资源 AWS 账户。 AWS 账户 A 中具有成员运行查询能力的客户可以调用CreateTrainedModelCreateMLInputChannel、或StartTrainedModelInferenceJob对协作中其他成员拥有的ConfiguredModelAlgorithmAssociation资源进行调用，前提ConfiguredModelAlgorithmAssociation是使用创建的自定义分析规则允许CreateConfiguredTableAnalysisRule。

此外，协作中的任何活跃成员都可以通过DeleteTrainedModelOutput和删除与训练模型或机器学习输入通道关联的数据DeleteMLInputChannelData APIs。

跨账户访问

Clean Rooms ML 允许用户通过GetCollaboration和检索有关其他账户创建的资源的元数据ListCollaboration APIs。Clean Rooms ML 不会向其他账户透露 KMS 密钥 ARNs、标签、环境变量或超参数（用于TrainedModel操作）。

成员资格和协作访问权限

在 Clean Rooms ML 自定义模型的上下文中访问成员资格和协作资源时，用户的身份策略需要操作权限 cleanrooms:PassMembershipcleanrooms:PassCollaboration，或两者兼而有之。所有 APIs 接受的人都membershipId需要cleanrooms:PassMembership许可，而所有 APIs 接受的人

都collaborationId需要cleanrooms:PassCollaboration许可。提供了一个角色的身份策略示例，该角色可以在成员身份 ID 的上下文GetCollaborationTrainedModel中调用，并且可以在协作 ID 的上下文中进行调用。createTrainedModel

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Sid": "AllowCleanroomsMLActions",
            "Effect": "Allow",
            "Action": [
                "cleanrooms:PassCollaboration",
                "cleanrooms:PassMembership"
            ],
            "Resource": [
                "*"
            ]
        },
        {
            "Sid": "AllowMembershipAccess",
            "Effect": "Allow",
            "Action": [
                "cleanrooms:GetMembership"
            ],
            "Resource": [
                "arn:aws:cleanrooms:us-east-1:111122223333:membership/memberId"
            ]
        },
        {
            "Sid": "AllowCollaborationAccess",
            "Effect": "Allow",
            "Action": [
                "cleanrooms:GetCollaboration"
            ],
            "Resource": [
                "arn:aws:cleanrooms:us-
east-1:111122223333:collaboration/collaborationId"
            ]
        }
    ]
```

```
}
```

合规性验证 AWS Clean Rooms

要了解是否属于特定合规计划的范围，请参阅AWS 服务 “按合规计划划分的范围” ”，然后选择您感兴趣的合规计划。 AWS 服务 有关一般信息，请参阅AWS 合规计划AWS。

您可以使用下载第三方审计报告 AWS Artifact。有关更多信息，请参阅中的 “下载报告” 中的 “ AWS Artifact。

您在使用 AWS 服务 时的合规责任取决于您的数据的敏感性、贵公司的合规目标以及适用的法律和法规。有关您在使用时的合规责任的更多信息 AWS 服务，请参阅AWS 安全文档。

韧性在 AWS Clean Rooms

AWS 全球基础设施是围绕 AWS 区域和可用区构建的。各区域提供多个在物理上独立且隔离的可用区，这些可用区通过延迟低、吞吐量高且冗余性高的网络连接在一起。利用可用区，您可以设计和操作在可用区之间无中断地自动实现故障转移的应用程序和数据库。与传统的单个或多个数据中心基础设施相比，可用区具有更高的可用性、容错能力和可扩展性。

有关 AWS 区域和可用区的更多信息，请参阅AWS 全球基础设施。

中的基础设施安全 AWS Clean Rooms

作为一项托管服务 AWS Clean Rooms ，受 AWS 全球网络安全的保护。有关 AWS 安全服务以及如何 AWS 保护基础设施的信息，请参阅AWS 云安全。要使用基础设施安全的最佳实践来设计您的 AWS 环境，请参阅 S AWS ecurity Pillar Well-Architected Fram ework 中的基础设施保护。

您可以使用 AWS 已发布的 API 调用 AWS Clean Rooms 通过网络进行访问。客户端必须支持以下内容：

- 传输层安全性协议（TLS）。我们要求使用 TLS 1.2，建议使用 TLS 1.3。
- 具有完全向前保密（PFS）的密码套件，例如 DHE（临时 Diffie-Hellman）或 ECDHE（临时椭圆曲线 Diffie-Hellman）。大多数现代系统（如 Java 7 及更高版本）都支持这些模式。

网络安全

在查询执行期间从 S3 存储桶 AWS Clean Rooms 读取数据时，与 Amazon S3 AWS Clean Rooms 之间的流量将通过 AWS 私有网络安全路由。飞行中的流量使用亚马逊签名版本 4 协议 (SIGv4) 进行签名，并使用 HTTPS 进行加密。此流量根据您为配置表设置的 IAM 服务角色进行授权。

您可以 AWS Clean Rooms 通过终端节点以编程方式连接到。有关服务端点的列表，请参阅《AWS 一般参考》中的 AWS Clean Rooms 端点和配额。

所有服务端点都只支持 HTTPS。如果您想从 VPC 连接但又不想连接互联 AWS Clean Rooms 网，则可以使用亚马逊虚拟私有云 (VPC) 终端节点。有关更多信息，请参阅AWS PrivateLink 指南 AWS PrivateLink中的通过访问 AWS 服务。

您可以为您的 IAM 委托人分配 IAM 策略，这些委托人使用 aws: SourceVpce 上下文密钥来限制您的 IAM 委托人只能通过 VPC 终端节点进行调用，而不能 AWS Clean Rooms 通过互联网进行调用。

使用接口端点进行访问 AWS Clean Rooms 或 AWS Clean Rooms ML (AWS PrivateLink)

您可以使用 AWS PrivateLink 在您的虚拟私有云 (VPC) 和/ AWS Clean Rooms 或 AWS Clean Rooms ML 之间创建私有连接。您可以像在 VPC 中一样访问 AWS Clean Rooms 或 AWS Clean Rooms ML，无需使用互联网网关、NAT 设备、VPN 连接或 Direct Connect 连接。VPC 中的实例不需要公有 IP 地址即可访问 AWS Clean Rooms。

您可以通过创建由 AWS PrivateLink提供支持的接口端点来建立此私有连接。我们将在您为接口端点启用的每个子网中创建一个端点网络接口。这些是请求者托管的网络接口，用作发往 AWS Clean Rooms 的流量的入口点。

有关更多信息，请参阅《AWS PrivateLink 指南》中的通过 AWS PrivateLink访问 AWS 服务。

的注意事项 AWS Clean Rooms

在为设置接口端点之前 AWS Clean Rooms，请查看AWS PrivateLink 指南中的注意事项。

AWS Clean Rooms 而且 AWS Clean Rooms ML 支持通过接口端点调用其所有 API 操作。

AWS Clean Rooms 或 AWS Clean Rooms ML 不支持 VPC 终端节点策略。默认情况下，允许通过接口端点对 AWS Clean Rooms 和 AWS Clean Rooms ML 进行完全访问。或者，您可以将安全组与端点

网络接口关联，以控制通过接口终端节点 AWS Clean Rooms 传入或 AWS Clean Rooms 机器学习的流量。

为创建接口终端节点 AWS Clean Rooms

您可以使用 Amazon VPC 控制台 AWS Clean Rooms 或 AWS Command Line Interface (AWS CLI) 为或 AWS Clean Rooms ML 创建接口终端节点。有关更多信息，请参阅《AWS PrivateLink 指南》中的创建接口端点。

AWS Clean Rooms 使用以下服务名称创建接口终端节点。

```
com.amazonaws.region.cleanrooms
```

使用以下服务名称为 AWS Clean Rooms ML 创建接口终端节点。

```
com.amazonaws.region.cleanrooms-ml
```

如果为接口端点启用私有 DNS，则可使用其默认区域 DNS 名称向 AWS Clean Rooms 发出 API 请求。例如 `cleanrooms-ml.us-east-1.amazonaws.com`。

监控 AWS Clean Rooms

监控是维护和其他 AWS 解决方案的可靠性、可用性和性能的重要组成部分。 AWS Clean Rooms AWS 提供以下监控工具 AWS Clean Rooms，供您监视、报告问题并在适当时自动采取措施：

- Amazon CloudWatch Logs 允许您监控、存储和访问来自 Amazon EC2 实例和其他来源的日志文件。 AWS CloudTrail Amazon Log CloudWatch s 可以监控日志文件中的信息，并在达到特定阈值时通知您。您还可以在高持久性存储中检索您的日志数据。有关更多信息，请参阅 Amazon CloudWatch 日志用户指南。

  Clean Rooms ML 允许跨账户作业，以便执行某些 API 操作。启动 AWS 账户 该作业的将收到该作业的 AWS CloudTrail 审核日志事件。有关更多信息，请参阅 AWS Clean Rooms ML 的 IAM 行为。
- AWS CloudTrail捕获由您或代表您发起的 API 调用和相关事件， AWS 账户 并将日志文件传输到您指定的 Amazon S3 存储桶。您可以识别哪些用户和帐户拨打了电话 AWS、发出呼叫的源 IP 地址以及呼叫发生的时间。有关更多信息，请参阅 AWS CloudTrail 《用户指南》。
- Amazon EventBridge 是一项无服务器事件总线服务，可以轻松地将您的应用程序与来自各种来源的数据连接起来。 EventBridge 提供来自您自己的应用程序、 Software-as-a-Service (SaaS) 应用程序和 AWS 服务的实时数据流，并将这些数据路由到 Lambda 等目标。这使您能够监控服务中发生的事件，并构建事件驱动的架构。有关更多信息，请参阅《亚马逊 EventBridge 用户指南》和《亚马逊 EventBridge活动参考》。
- AWS 账单与成本管理是一项服务，其提供的功能可帮助您支付账单和优化成本。Amazon Web Services 根据使用情况向您的账户开具账单，确保您只需为实际用量付费。 AWS 还允许您对资源应用成本分配标签，以跟踪您的 AWS 成本并对其进行分类。例如，在中 AWS Clean Rooms，您可以将标签应用于协作成员资格，以跟踪每次协作的成本。有关更多信息，请参阅 AWS Billing 用户指南。

分析登录 AWS Clean Rooms

分析日志是中的一项功能 AWS Clean Rooms。当您创建协作并打开分析日志时，成员可以将查询中的相关日志或任务日志存储在 Amazon L CloudWatch ogs 中。

通过查询日志和作业日志，成员可以确定查询是否符合分析规则并符合协作协议。此外，查询日志有助于支持审计。

在 AWS Clean Rooms 控制台中启用 Analysis 日志记录选项后，查询日志包括以下内容：

- `analysisRule` — 已配置表的分析规则。

- analysisTemplateArn — 已运行的分析模板（根据分析规则显示）。
- collaborationId — 运行查询的协作的唯一标识符。
- configuredTableID — 查询中引用的已配置表的唯一标识符。
- directQueryAnalysisRulePolicy.custom.allowedAnalysis — 允许在配置表上运行的分析模板（根据分析规则显示）。
- directQueryAnalysisRulePolicy.v1.custom.allowedAnalysisProviders — 允许创建查询的查询提供者（根据分析规则显示）。
- errorCode - 查询未能正确执行时的错误代码。
- errorMessage - 查询未能正确执行时的错误消息。
- eventID — 查询运行的唯一标识符。2023 年 8 月 31 日之后，唯一标识符与 protectedQueryID 相同。
- eventTimestamp — 查询运行时间。
- parameters.parametervalue — 参数值（根据查询文本显示）。
- queryText — 查询运行的 SQL 定义。如果有参数，则会将其标记为 :parametervalue。
- queryValidationErrors — 查询验证时的查询错误。
- schemaName — 查询中引用的已配置表关联的名称。
- status – 查询的执行状态。

接收查询和作业日志

您无需在之外执行任何操作 AWS Clean Rooms 即可设置查询日志和作业日志。 AWS Clean Rooms 在每个协作成员创建成员资格后，为协作创建日志组。

可以查询的成员、可以运行查询和作业的成员、可以接收结果的成员，以及在查询中引用配置表的成员，都将收到查询日志或作业日志。

可以查询的成员、可以接收结果的成员将收到查询中引用的每个已配置表的查询日志。如果他们不拥有配置表，将无法查看配置表 ID (configuredTableID)。

可以运行查询和作业的成员以及可以接收结果的成员将收到作业中引用的每个已配置表的作业日志。如果他们不拥有配置表，将无法查看配置表 ID (configuredTableID)。

如果成员在查询中引用了多个配置表关联，则他们将收到每个配置表的查询日志。

如果成员在作业中引用了多个已配置的表关联，则他们将收到每个已配置表的作业日志。

将为包含 AWS Clean Rooms中不支持和支持的 SQL 的查询创建日志。有关详细信息，请参阅 AWS Clean Rooms SQL 参考。

当查询或作业引用与协作无关的已配置表时，也会创建日志。

日志可能包含有关不正确的 SQL 的信息。

查询和作业日志会显示查询的状态，但不报告查询输出是否已送达。他们确认查询或任务是由可以查询的成员提交的。查询日志还确认查询中包含支持的 SQL， AWS Clean Rooms 并引用了与协作关联的已配置表。

Example

例如，如果在 AWS Clean Rooms 验证查询是否符合分析规则之后并在查询处理期间取消查询，则不会生成日志。

如果删除日志组，则必须使用相同的日志组名称（协作的协作 ID）手动重新创建日志组。或者，您也可以在成员身份中禁用和启用日志记录。

有关如何开启分析日志记录的更多信息，请参阅创建协作。

有关 Amazon CloudWatch 日志的更多信息，请参阅亚马逊 CloudWatch 日志用户指南。

查询和作业日志的推荐操作

我们建议成员定期采取以下行动：

- 要验证查询和作业是否与为协作商定的用例或查询相匹配，请查看协作中运行的查询和作业。

  有关如何查看最近查询的更多信息，请参阅查看最近的查询。

  有关如何查看最近作业的更多信息，请参阅查看最近的工作。
- 要验证配置表列是否与协作商定的内容相匹配，请查看协作成员分析规则和查询中使用的配置表列。

  有关如何查看已配置列的更多信息，请参阅查看表和分析规则。

带有 in 的 CloudWatch 详细监控 AWS Clean Rooms

详细监控是中的一项可选功能。AWS Clean Rooms当您创建协作并开启详细监控时，可以运行查询并配置付款人的成员可以选择在其 CloudWatch 账户中接收详细的可观察性指标。

通过详细监控，成员可以监控查询性能，跟踪资源利用率，并创建仪表板以获取运营见解。这些指标有助于容量规划、成本分析和性能问题故障排除。

> **Note**
>
> 指标可用性因配置的表数据源而异：
>
> - 对于引用以 Athena 作为数据源的配置表的分析，`QueryFileScanned`和指标始终返`BytesRead`回 0。
> - 对于引用以 Snowflake 作为数据源的已配置表的分析，`BytesRead`、`QueryFileScanned`和`RecordsRead`指标始终返回 0。
> - 用于引用具有不同数据源的多个已配置表（例如，Snowflake 和 S3 支持的AWS Glue 表），`BytesRead`并且仅`QueryFileScanned`返回支持 S3 的表的指标的分析。AWS Glue `RecordsRead`返回支持 S3 的表和 Athena AWS Glue 表的指标。

指标类型

AWS Clean Rooms将两类指标发布到 CloudWatch：

汇总的详细监控指标

这些指标提供了跨分析的汇总级可见性：

`AnalysisRuntime`

显示分析的总持续时间（以毫秒为单位）。

`CRPUConsumed`

用于特定分析的 CRPU（洁净室处理单元）。有关 CRPU 计算的更多信息，请参阅AWS Clean Rooms定价。

`ConcurrentVCPUs`

在给定时间点使用并发 vCPU（与服务配额有关）。

`AnalysesCount`

在一段时间内对指定维度运行的分析次数。

`ConcurrentQueries`

给定时间点的并发查询数（与服务配额有关）。

查询级别的详细监控指标

这些指标提供了对单个查询执行的详细见解：

`BytesRead`

从源表中读取的用于分析的总数据（以字节为单位）。

`BytesWritten`

为分析写入的总数据（以字节为单位）。

`RecordsRead`

从源表中读取的用于分析的记录总数。

`RecordsWritten`

分析结果中返回的记录总数。

`QueryFileScanned`

为分析而扫描的数据文件总数。

`Memory utilization`

用于分析的内存资源。

`vCPU utilization`

用于分析的 vCPU 资源。

`Disk utilization`

用于分析的存储资源。

指标维度

指标按以下维度组织：

- 分析状态：已完成
- 分析类型：sparkSQL
- 成员资格：会员 ID

- 查询：个人查询 ID（用于查询级别的指标）
- 账户/地区：您的AWS 账户和地区

所有指标都发布到中的AWS/CleanRooms命名空间 CloudWatch。

谁可以访问指标

详细的监控指标可用于：

- 查询运行器：可以在协作中运行查询的成员
- 配置的付款人：指定支付查询执行费用的会员

仅提供数据的成员无法访问详细的监控指标。

定价

AWS Clean Rooms不收取向发布指标的费用 CloudWatch。但是，指标存储、仪表板使用情况和警报将 CloudWatch收取标准费用。有关更多信息，请参阅 Amazon CloudWatch 定价。

使用记录 AWS Clean Rooms API 调用 AWS CloudTrail

AWS Clean Rooms 与 AWS CloudTrail一项服务集成，该服务提供用户、角色或角色所执行操作 AWS 服务 的记录 AWS Clean Rooms。 CloudTrail 将所有 API 调用捕获 AWS Clean Rooms 为事件。捕获的调用包括来自 AWS Clean Rooms 控制台的调用和对 AWS Clean Rooms API 操作的代码调用。如果您创建了跟踪，则可以允许将 CloudTrail事件持续传输到 Amazon S3 存储桶，包括的事件 AWS Clean Rooms。如果您未配置跟踪，您仍然可以在 CloudTrail 控制台的 “事件历史记录” 中查看最新的事件。使用收集的信息 CloudTrail，您可以确定向哪个请求发出 AWS Clean Rooms、发出请求的 IP 地址、谁发出了请求、何时发出请求以及其他详细信息。

要了解更多信息 CloudTrail，请参阅《AWS CloudTrail 用户指南》。

AWS Clean Rooms 信息在 CloudTrail

CloudTrail 在您创建账户 AWS 账户 时已在您的账户上启用。当活动发生在中时 AWS Clean Rooms，该活动会与其他 CloudTrail 事件一起记录在 AWS 服务 事件历史记录中。您可以在 AWS 账户中查看、搜索和下载最新事件。有关更多信息，请参阅使用事件历史记录查看 CloudTrail 事件。

要持续记录您的 AWS 账户事件（包括的事件） AWS Clean Rooms，请创建跟踪。跟踪允许 CloudTrail 将日志文件传输到 Amazon S3 存储桶。默认情况下，在控制台中创建跟踪记录时，此跟踪记录应用于所有 AWS 区域。跟踪记录 AWS 分区中所有区域的事件，并将日志文件传送到您指定的 Amazon S3 存储桶。此外，您可以配置其他 AWS 服务 ，以进一步分析和处理 CloudTrail 日志中收集的事件数据。有关更多信息，请参阅下列内容：

- 创建跟踪记录概述
- CloudTrail 支持的服务和集成
- 配置 Amazon SNS 通知 CloudTrail
- 接收来自多个区域的 CloudTrail 日志文件
- 接收来自多个账户的 CloudTrail 日志文件

所有 AWS Clean Rooms 操作均由《API 参考》记录 CloudTrail 并记录在《AWS Clean Rooms API 参考》中。

每个事件或日志条目都包含有关生成请求的人员信息。身份信息有助于您确定以下内容：

- 请求是使用根用户凭证还是 IAM 用户凭证发出的。
- 请求是使用角色还是联合用户的临时安全凭证发出的。
- 请求是否由其他 AWS 服务发出。

有关更多信息，请参阅 CloudTrail userIdentity 元素。

了解 AWS Clean Rooms 日志文件条目

跟踪是一种配置，允许将事件作为日志文件传输到您指定的 Amazon S3 存储桶。 CloudTrail 日志文件包含一个或多个日志条目。事件代表来自任何来源的单个请求，包括有关请求的操作、操作的日期和时间、请求参数等的信息。 CloudTrail 日志文件不是公共 API 调用的有序堆栈跟踪，因此它们不会按任何特定的顺序出现。

示例 AWS Clean Rooms CloudTrail 事件

以下示例演示了以下 CloudTrail 各项的事件：

主题

- StartProtectedQuery（成功）

- StartProtectedQuery（失败）

StartProtectedQuery（成功）


```
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "AssumedRole",
        "principalId": "EXAMPLE_PRINCIPAL_ID",
        "arn": "arn:aws:sts::123456789012:assumed-role/query-runner/jdoe",
        "accountId": "123456789012",
        "accessKeyId": "EXAMPLE_KEY_ID",
        "sessionContext": {
            "sessionIssuer": {
                "type": "Role",
                "principalId": "EXAMPLE_PRINCIPAL_ID",
                "arn": "arn:aws:iam::123456789012:role/query-runner",
                "accountId": "123456789012",
                "userName": "query-runner"
            },
            "webIdFederationData": {},
            "attributes": {
                "creationDate": "2023-04-07T19:34:32Z",
                "mfaAuthenticated": "false"
            }
        }
    },
    "eventTime": "2023-04-07T19:53:32Z",
    "eventSource": "cleanrooms.amazonaws.com",
    "eventName": "StartProtectedQuery",
    "awsRegion": "us-east-2",
    "sourceIPAddress": "203.0.113.1",
    "userAgent": "aws-internal/3",
    "requestParameters": {
        "resultConfiguration": {
            "outputConfiguration": {
                "s3": {
                    "resultFormat": "CSV",
                    "bucket": "cleanrooms-queryresults-jdoe-test",
                    "keyPrefix": "test"
                }
            }
        },
```

```
        "sqlParameters": "***",
        "membershipIdentifier": "a1b2c3d4-5678-90ab-cdef-EXAMPLE11111",
        "type": "SQL"
    },
    "responseElements": {
        "Access-Control-Expose-Headers": "x-amzn-RequestId,x-amzn-ErrorType,x-amzn-
ErrorMessage,Date",
        "protectedQuery": {
            "createTime": 1680897212.279,
            "id": "f5988bf1-771a-4141-82a8-26fcc4e41c9f",
            "membershipArn": "arn:aws:cleanrooms:us-east-2:123456789012:membership/
a1b2c3d4-5678-90ab-cdef-EXAMPLE11111",
            "membershipId": "a1b2c3d4-5678-90ab-cdef-EXAMPLE11111",
            "resultConfiguration": {
                "outputConfiguration": {
                    "s3": {
                        "bucket": "cleanrooms-queryresults-jdoe-test",
                        "keyPrefix": "test",
                        "resultFormat": "CSV"
                    }
                }
            },
            "sqlParameters": "***",
            "status": "SUBMITTED"
        }
    },
    "requestID": "7464211b-2277-4b55-9723-fb4f259aefd2",
    "eventID": "f7610f5e-74b9-420f-ae43-206571ebcbf7",
    "readOnly": false,
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "123456789012",
    "eventCategory": "Management"
}
```

StartProtectedQuery（失败）

```
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "AssumedRole",
        "principalId": "EXAMPLE_PRINCIPAL_ID",
        "arn": "arn:aws:sts::123456789012:assumed-role/query-runner/jdoe",
```

```
        "accountId": "123456789012",
        "accessKeyId": "EXAMPLE_KEY_ID",
        "sessionContext": {
            "sessionIssuer": {
                "type": "Role",
                "principalId": "EXAMPLE_PRINCIPAL_ID",
                "arn": "arn:aws:iam::123456789012:role/query-runner",
                "accountId": "123456789012",
                "userName": "query-runner"
            },
            "webIdFederationData": {},
            "attributes": {
                "creationDate": "2023-04-07T19:34:32Z",
                "mfaAuthenticated": "false"
            }
        }
    },
    "eventTime": "2023-04-07T19:47:27Z",
    "eventSource": "cleanrooms.amazonaws.com",
    "eventName": "StartProtectedQuery",
    "awsRegion": "us-east-2",
    "sourceIPAddress": "203.0.113.1",
    "userAgent": "aws-internal/3",
    "errorCode": "ValidationException",
    "requestParameters": {
        "resultConfiguration": {
            "outputConfiguration": {
                "s3": {
                    "resultFormat": "CSV",
                    "bucket": "cleanrooms-queryresults-jdoe-test",
                    "keyPrefix": "test"
                }
            }
        },
        "sqlParameters": "***",
        "membershipIdentifier": "a1b2c3d4-5678-90ab-cdef-EXAMPLE11111",
        "type": "SQL"
    },
    "responseElements": {
        "Access-Control-Expose-Headers": "x-amzn-RequestId,x-amzn-ErrorType,x-amzn-
ErrorMessage,Date",
        "message": "Column(s) [identifier] is not allowed in select"
    },
    "requestID": "e29f9f74-8299-4a83-9d18-5ddce7302f07",
```

```
    "eventID": "c8ee3498-8e4e-44b5-87e4-ab9477e56eb5",
    "readOnly": false,
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "123456789012",
    "eventCategory": "Management"
}
```

使用 Ama AWS Clean Rooms zon 集成到事件驱动的应用程序中 EventBridge

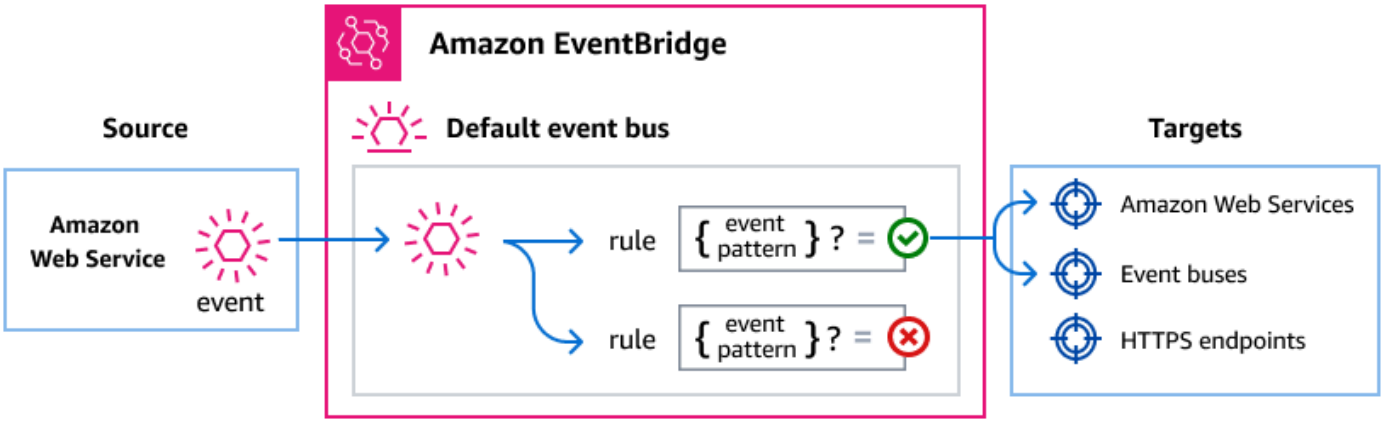
您可以将事件驱动的应用程序 (EDAs) 整合 AWS Clean Rooms 到事件驱动的应用程序中，这些应用程序使用中发生的事件在 AWS Clean Rooms 应用程序组件之间进行通信并启动下游进程。为此，您可以使用 Amazon EventBridge 将事件从其他软件组件路由 AWS Clean Rooms 到其他软件组件。Amazon EventBridge 是一项无服务器服务，它使用事件将应用程序组件连接在一起，这样您就可以更轻松地将 AWS 服务集成 AWS Clean Rooms 到事件驱动的架构中，而无需额外的代码和操作。

事件驱动型架构是一种构建松耦合软件系统的风格，这些系统通过发出和响应事件 来协同工作。在此模型中，事件表示资源或环境的变化。

以下是 EventBridge 使用方法 AWS Clean Rooms：

与许多 AWS 服务一样， AWS Clean Rooms 生成事件并将其发送到 EventBridge 默认事件总线。事件总线是接收事件并将其路由到您所指定的目的地或目标的路由器。目标可以包括其他 AWS 服务、定制应用程序和 SaaS 合作伙伴应用程序。

EventBridge 根据您在事件总线上创建的规则对事件进行路由。对于每条规则，您可以指定筛选条件或事件模式，以便仅选择所需的事件。每当向事件总线发送事件时，都要将其 EventBridge 与每条规则进行比较。如果事件符合规则，则将事件 EventBridge 路由到指定的目标。



例如，假设每次在账户中创建新 AWS Clean Rooms 协作时，您都想知道。您可以在默认事件总线上创建规则。在规则中，您将创建一个事件模式，该模式指定来自 AWS Clean Rooms 的事件命名**`Collaboration Created`**。每次 EventBridge 收到与这些属性匹配的事件时，它都会将该事件路由到指定的工作流程。

AWS Clean Rooms 事件

AWS 服务可以将事件直接发送到 EventBridge 默认事件总线。此外，还 AWS CloudTrail 会向发送源自多个 AWS 服务的事件。 EventBridge 这些事件可能包括 API 调用、控制台登录和操作、服务事件和 Ins CloudTrail ights。有关更多信息，请参阅《EventBridge 用户指南》 AWS CloudTrail中的通过交付的AWS 服务事件。

有关发送到 AWS Clean Rooms 的事件的完整列表 EventBridge，请参阅《EventBridge事件参考》中的 AWS Clean Rooms 主题。

事件详细信息类型	说明
分析模板已创建	创建分析模板时，系统会通知分析模板所有者和协作的所有活跃成员。
分析模板已更新	当分析模板更新时，系统会通知分析模板所有者和所有可以查看更新的协作活跃成员。
分析模板已删除	删除分析模板时，系统会通知分析模板所有者和协作的所有活跃成员。
协作已创建	创建协作时，协作所有者会收到通知。

事件详细信息类型	说明
协作已更新	协作更新时，协作所有者和所有能查看更新的协作活跃成员都会收到通知。
已创建协作变更请求	创建协作变更请求时，协作所有者和协作的所有活跃成员都会收到通知。
协作变更请求已批准	协作变更请求获得批准后，协作所有者和协作的所有活跃成员都会收到通知。
协作变更请求已取消	取消协作变更请求时，协作所有者和协作的所有活跃成员都会收到通知。
协作变更请求已提交	提交协作变更请求时，协作所有者和协作的所有活跃成员都会收到通知。
已配置的表关联已创建	创建配置表关联时，会通知配置的表关联所有者和协作的所有活动成员。
已配置的表关联已更新	当配置的表关联更新时，会通知配置的表关联所有者和所有可以查看更新的协作活动成员。
已配置的表关联已删除	当已配置的表关联被删除时，会通知配置的表关联所有者和协作的所有活动成员。
已配置的表关联分析规则已创建	创建配置的表关联分析规则时，会通知配置的表关联分析规则所有者和协作的所有活动成员。
已配置的表关联分析规则已更新	当配置的表关联分析规则更新时，会通知配置的表关联分析规则所有者和所有可以查看更新的协作活动成员。
已配置的表关联分析规则已删除	删除已配置表关联分析规则时，会通知已配置的表关联分析规则所有者和协作的所有活动成员。
已创建 ID 映射表	创建 ID 映射表时，会通知 ID 映射表所有者和协作的所有活动成员。
身份映射表已更新	当 ID 映射表更新时，会通知 ID 映射表所有者和协作中所有可以查看更新的活跃成员。

事件详细信息类型	说明
ID 映射表已删除	删除 ID 映射表时，会通知 ID 映射表所有者和协作的所有活动成员。
ID 命名空间关联已创建	创建 ID 命名空间关联时，会通知 ID 命名空间协会所有者和协作的所有活跃成员。
Id 命名空间关联已更新	当 ID 命名空间关联更新时，会通知 ID 命名空间协会所有者和所有可以查看更新的协作活跃成员。
ID 命名空间关联已删除	删除 ID 命名空间关联时，会通知 ID 命名空间协会所有者和协作的所有活跃成员。
受邀合作	当受邀成员被邀请加入协作时，他们会收到通知。
已创建会员资格	创建成员资格时，会通知成员资格所有者和协作的所有活跃成员。
成员资格已更新	当成员资格更新时，会通知成员资格所有者，除非成员资格已从协作中删除，在这种情况下，协作的所有活跃成员都会收到通知。
已删除会员资格	删除成员资格时，会通知成员资格所有者和协作的所有活跃成员。
受保护的 Job 已提交	提交受保护作业时，会通知受保护作业的作业运行者、作业付款人和结果接收者。
受保护的 Job 已启动	当受保护的作业启动时，会通知受保护作业的作业运行者、作业付款人和结果接收者。
受保护的 Job 取消	当受保护的作业取消时，会通知受保护作业的作业运行者、作业付款人和结果接收者。
受保护的 Job 已取消	当受保护的作业被取消时，会通知受保护作业的作业运行者、作业付款人和结果接收者。
受保护的 Job 已成功	当受保护的作业成功时，会通知受保护作业的作业运行者、作业付款人和结果接收者。

事件详细信息类型	说明
受保护的 Job 失败	当受保护的作业失败时，会通知受保护作业的作业运行者、作业付款人和结果接收者。
受保护的查询已提交	提交受保护查询时，会通知受保护查询的查询运行器、查询付款人和结果接收者。
受保护的查询已启动	受保护查询启动时，会通知受保护查询的查询运行器、查询付款人和结果接收者。
受保护的查询取消	当受保护查询取消时，会通知受保护查询的查询运行器、查询付款人和结果接收者。
受保护的查询已取消	当受保护的查询被取消时，会通知受保护查询的查询运行器、查询付款人和结果接收者。
受保护的查询成功	受保护查询成功后，会通知受保护查询的查询运行器、查询付款人和结果接收者。
受保护的查询失败	受保护查询失败时，会通知受保护查询的查询运行器、查询付款人和结果接收者。
受保护的查询已超时	当受保护查询超时时，会通知受保护查询的查询运行器、查询付款人和结果接收者。

使用路由 AWS Clean Rooms 事件 EventBridge

要将 AWS Clean Rooms 事件 EventBridge 路由到目标，必须创建规则。每条规则都包含一个事件模式，该模式与事件总线上接收到的每个事件进行 EventBridge 匹配。如果事件数据与指定的事件模式匹配，则会将该事件 EventBridge 路由到规则的目标。

有关创建事件总线规则的全面说明，请参阅《EventBridge 用户指南》中的创建对事件作出反应的规则。

创建与事件匹配 AWS Clean Rooms 的事件模式

每个事件模式是一个 JSON 对象，其中包含：

- （可选）：用于标识发送事件的服务的source属性。对于 AWS Clean Rooms 事件，来源是aws.cleanrooms。
- （可选）：包含要匹配的事件名称数组的 detail-type 属性。
- （可选）：包含要匹配的其他事件数据的 detail 属性。

例如，以下事件模式与从中删除协作的所有 Id Namespace Association Updated 事件相匹配 AWS Clean Rooms：

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Namespace Association Updated"],
  "detail": {
    "status": ["COLLABORATION_DELETED"]
  }
}
```

有关写入事件模式的更多信息，请参阅《EventBridge 用户指南》中的事件模式。

AWS Clean Rooms 事件详情参考

来自 AWS 服务的所有事件都有一组公共字段，其中包含有关事件的元数据，例如作为事件来源的 AWS 服务、事件的生成时间、事件发生的账户和区域等。有关这些常规字段的定义，请参阅 Amazon Events 参考中的 EventBridge 事件结构。

此外，每个事件都有一个 detail 字段，其中包该特定事件专有的数据。下面的参考定义了各种 AWS Clean Rooms 事件的详细信息字段。

使用 EventBridge 来选择和管理 AWS Clean Rooms 事件时，记住以下几点很有用：

- 来自的所有事件的source字段均设置 AWS Clean Rooms 为aws.cleanrooms。
- detail-type 字段指定事件类型。

  例如 Collaboration Created。
- detail 字段包含该特定事件专有的数据。

有关构建使规则匹配 AWS Clean Rooms 事件的事件模式的信息，请参阅 Amazon EventBridge 用户指南中的事件模式。

分析模板已创建事件

以下是 Analysis Template Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Analysis Template Created"]
}
```

分析模板已更新事件

以下是 Analysis Template Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Analysis Template Updated"]
}
```

分析模板已删除事件

以下是 Analysis Template Deleted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Analysis Template Deleted"]
}
```

协作创建的事件

以下是 Collaboration Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Collaboration Created"]
}
```

协作已更新活动

以下是 Collaboration Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Collaboration Updated"]
}
```

协作变更请求已创建事件

以下是 Collaboration Change Request Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Collaboration Change Request Created"]
}
```

协作变更请求已批准事件

以下是 Collaboration Change Request Approved 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Collaboration Change Request Approved"]
}
```

协作变更请求已取消活动

以下是 Collaboration Change Request Cancelled 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Collaboration Change Request Cancelled"]
}
```

协作变更请求已提交事件

以下是 Collaboration Change Request Committed 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Collaboration Change Request Committed"]
}
```

已配置的表关联已创建事件

以下是 Configured Table Association Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Configured Table Association Created"]
}
```

已配置的表关联已更新事件

以下是 Configured Table Association Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Configured Table Association Updated"]
}
```

已配置的表关联已删除事件

以下是 Configured Table Association Deleted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Configured Table Association Deleted"]
}
```

已配置的表关联分析规则已创建事件

以下是 Configured Table Association Analysis Rule Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Configured Table Association Analysis Rule Created"]
}
```

已配置的表关联分析规则已更新事件

以下是 Configured Table Association Analysis Rule Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Configured Table Association Analysis Rule Updated"]
}
```

已配置的表关联分析规则已删除事件

以下是 Configured Table Association Analysis Rule Deleted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Configured Table Association Analysis Rule Deleted"]
}
```

Id 映射表已创建事件

以下是 Id Mapping Table Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Mapping Table Created"]
}
```

ID 映射表已更新事件

以下是 Id Mapping Table Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Mapping Table Updated"]
}
```

ID 映射表已删除事件

以下是 Id Mapping Table Deleted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Mapping Table Deleted"]
```

```
}
```

ID 命名空间关联已创建事件

以下是 Id Namespace Association Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Namespace Association Created"]
}
```

ID 命名空间关联已更新事件

以下是 Id Namespace Association Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Namespace Association Updated"]
}
```

> **Note**
>
> 删除协作后，将在detail.status字段的 “ID 命名空间关联已更新” 事件中捕获该事件。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Namespace Association Updated"],
  "detail": {
    "status": ["COLLABORATION_DELETED"]
}
```

ID 命名空间关联已删除事件

以下是 Id Namespace Association Deleted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Id Namespace Association Deleted"]
```

```
}
```

受邀参加合作活动

以下是 Invited To Collaboration 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Invited To Collaboration"]
}
```

创建会员资格的活动

以下是 Membership Created 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Membership Created"]
}
```

会员资格更新活动

以下是 Membership Updated 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Membership Updated"]
}
```

会员资格删除活动

以下是 Membership Deleted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Membership Deleted"]
}
```

受保护的 Job 已提交事件

以下是 Protected Job Submitted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Job Submitted"]
}
```

受保护的 Job 已启动事件

以下是 Protected Job Started 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Job Started"]
}
```

受保护的 Job 取消事件

以下是 Protected Job Cancelling 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Job Cancelling"]
}
```

受保护的 Job 已取消事件

以下是 Protected Job Cancelled 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Job Cancelled"]
}
```

受保护的 Job 成功事件

以下是 Protected Job Succeeded 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Job Succeeded"]
}
```

受保护的 Job 失败事件

以下是 Protected Job Failed 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Job Failed"]
}
```

受保护的查询已提交事件

以下是 Protected Query Submitted 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Submitted"]
}
```

“受保护的查询已启动” 事件

以下是 Protected Query Started 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Started"]
}
```

受保护的查询取消事件

以下是 Protected Query Cancelling 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Cancelling"]
}
```

受保护的查询已取消事件

以下是 Protected Query Cancelled 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Cancelled"]
}
```

受保护的查询成功事件

以下是 Protected Query Succeeded 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Succeeded"]
}
```

受保护的查询失败事件

以下是 Protected Query Failed 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Failed"]
}
```

受保护的查询超时事件

以下是 Protected Query Timed Out 事件的详细信息字段。

```
{
  "source": ["aws.cleanrooms"],
  "detail-type": ["Protected Query Timed Out"]
}
```

成本分配标记

AWS Clean Rooms 支持使用成本分配标签来跟踪您的 AWS 成本。您可以在 AWS 账单与成本管理 控制面板上激活这些标签。 AWS 使用标签对您的成本进行分类，并向您提供每月成本分配报告。用户定义的成本分配标签可以应用于 AWS Clean Rooms 资源，以帮助在整个协作中跟踪和分配成本。

用于成本分配的可标记资源

下表列出了会产生费用的 AWS Clean Rooms 资源，这些资源可以用用户定义的成本分配标签进行标记，以跟踪和组织成本。

计费资源	带标签的资源
SQL 查询	成员身份
PySpark Job	成员身份
机器学习训练 Job	成员身份
机器学习推理 Job	成员身份
合成数据生成	成员身份
相似模型训练	长相模特
相似区段导出	相似区段

使用创建 AWS Clean Rooms 资源 AWS CloudFormation

AWS Clean Rooms 与一项服务集成 AWS CloudFormation，该服务可帮助您对 AWS 资源进行建模和设置。通过这种集成，您可以花费更少的时间来创建和管理您的资源和基础设施。您可以创建一个描述所需所有 AWS 资源的模板，并为您预 CloudFormation 置和配置这些资源。资源的示例包括协作、配置表、配置表关联和成员身份。

使用时 CloudFormation，您可以重复使用模板来一致且重复地设置 AWS Clean Rooms 资源。一次描述您的资源，然后在多个 AWS 账户 和（或）中一遍又一遍地配置相同的资源 AWS 区域。

AWS Clean Rooms 和 CloudFormation 模板

要为和相关服务配置 AWS Clean Rooms 和配置资源，必须了解CloudFormation 模板。模板是 JSON 或 YAML 格式的文本文件。这些模板描述了您要在 CloudFormation 堆栈中配置的资源。如果你不熟悉 JSON 或 YAML，可以使用 D CloudFormation esigner 来帮助你开始使用 CloudFormation 模板。有关更多信息，请参阅《AWS CloudFormation 用户指南》中的什么是 CloudFormation Designer？。

AWS Clean Rooms 支持在中创建协作、配置的表、配置的表关联和成员资格。 CloudFormation有关更多信息，包括用于协作的 JSON 和 YAML 模板、配置的表、配置的表关联和成员资格的示例，请参阅用户指南中的和 AWS Clean RoomsAW S Clean Rooms ML 资源类型参考。AWS CloudFormation

了解更多关于 CloudFormation

要了解更多信息 CloudFormation，请参阅以下资源：

- AWS CloudFormation
- AWS CloudFormation 用户指南
- CloudFormation API 引用
- 《AWS CloudFormation 命令行界面用户指南》

的配额 AWS Clean Rooms

您的每个配额 AWS 账户 都有默认配额，以前称为限制 AWS 服务。除非另有说明，否则每个配额都特定于 AWS 区域。您可以请求增加某些配额，但其他一些配额无法增加。

要查看的配额 AWS Clean Rooms，请打开 S ervice Quotas 控制台。在导航窗格中，选择 AWS 服务，然后选择 AWS Clean Rooms。

要请求提高配额，请参阅《Service Quotas 用户指南》中的请求提高配额。如果配额在服务配额中尚不可用，请使用服务限制提高表单。

主题

AWS Clean Rooms 配额

您的 AWS 账户 配额与以下有关 AWS Clean Rooms。

Name	默认值	可调整	说明
分析规则大小	每个受支持的区域：100 KB	否	分析规则的最大 JSON 大小
每个成员资格的分析模板数	每个受支持的区域：25 个	是	每个成员资格的最大分析模板数
每个账户创建的协作数	每个受支持的区域：10 个	是	每个账户的最大协作数
每个配置允许列表的列数	每个受支持的区域：100 个	是	每个已配置表可以列入许可名单的最大列数
每个账户的并发 PySpark 作业 vCPU 使用率	us-east-1：1,024	是	每个账户所有并发运行 PySpark 的作业的最大 vCPU 总使用率

Name	默认值	可调整	说明
	其他所有支持的区域：512		
每个账户的并发 PySpark 任务	us-east-1：5 个 每个其他支持的区域：2 个	是	每个账户同时运行的最大 PySpark 任务数
每个账户的并发 SQL 查询数量	us-east-1：5 个 每个其他支持的区域：2 个	是	每个账户并发运行的 SQL 查询数量上限
每个账户的并发 SQL 查询 vCPU 使用量	每个受支持的区域：512 个	是	每个账户所有并发运行的 SQL 查询的 vCPU 总使用量上限
每个成员的并发进行中作业数量	每个受支持的区域：1 个	否	每个成员的并发进行中作业数量上限
每个成员资格正在进行的并发查询数	每个受支持的区域：5 个	否	每个成员资格正在进行的最大并发查询数
为每个成员配置的受众模型关联数量	每个受支持的区域：5 个	是	为每个成员配置的受众模型关联数量上限
每个账户的配置表数	每个支持的区域：250 个	是	每个账户创建的最大配置表数
每个受保护的查询的配置表数	每个支持的区域：15 个	是	受保护查询中的最大已配置表数
每个成员的 ID 映射表数量	每个受支持的区域：5 个	是	每个成员的 ID 映射表数量上限

Name	默认值	可调整	说明
每个成员的 ID 命名空间关联数	每个受支持的区域：10 个	是	每个成员的 ID 命名空间关联数量上限
每次协作邀请的成员数	每个受支持的区域：5 个	是	每次协作邀请的最大成员人数
每个账户的成员数	每个受支持的区域：100 个	是	每个账户的最大成员资格数
每次合作的访问预算的隐私预算模板数量	每个受支持的区域：25 个	是	每次合作的访问预算的隐私预算模板数量上限
查询文本长度（使用差别隐私）	每个受支持的区域：8 KB	否	使用差别隐私的 SQL 查询语句的文本长度上限
Clean Rooms SQL 分析引擎上的查询文本长度	每个支持的区域：90 KB	否	Clean Rooms SQL 分析引擎上 SQL 查询语句的文本长度上限
Spark 分析引擎上的查询文本长度	每个受支持的区域：500 KB	否	Spark SQL 分析引擎上 SQL 查询语句的文本长度上限
BatchGetCollaborationAnalysisTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 BatchGetCollaborationAnalysisTemplate API 调用的最大次数
BatchGetSchema 请求率	每个受支持的区域：每秒 5 个	是	每秒 BatchGetSchema API 调用的最大次数
BatchGetSchemaAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 BatchGetSchemaAnalysisRule API 调用的最大次数

Name	默认值	可调整	说明
CreateAnalysisTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateAnalysisTemplate API 调用的最大次数
CreateCollaboration 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateCollaboration API 调用的最大次数
CreateCollaborationChangeRequest 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateCollaborationChangeRequest API 调用的最大次数
CreateConfiguredAudienceModelAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateConfiguredAudienceModelAssociation API 调用的最大次数
CreateConfiguredTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateConfiguredTable API 调用的最大次数
CreateConfiguredTableAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateConfiguredTableAnalysisRule API 调用的最大次数
CreateConfiguredTableAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateConfiguredTableAssociation API 调用的最大次数
CreateConfiguredTableAssociationAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateConfiguredTableAssociationAnalysisRule API 调用的最大次数
CreateIdMappingTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateIdMappingTable API 调用的最大次数

Name	默认值	可调整	说明
CreateIdNamespaceAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateIdNamespaceAssociation API 调用的最大次数
CreateMembership 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreateMembership API 调用的最大次数
CreatePrivacyBudgetTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 CreatePrivacyBudgetTemplate API 调用的最大次数
DeleteAnalysisTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteAnalysisTemplate API 调用的最大次数
DeleteCollaboration 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteCollaboration API 调用的最大次数
DeleteConfiguredAudienceModelAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteConfiguredAudienceModelAssociation API 调用的最大次数
DeleteConfiguredTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteConfiguredTable API 调用的最大次数
DeleteConfiguredTableAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteConfiguredTableAnalysisRule API 调用的最大次数
DeleteConfiguredTableAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteConfiguredTableAssociation API 调用的最大次数

Name	默认值	可调整	说明
DeleteConfiguredTableAssociationAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteConfiguredTableAssociationAnalysisRule API 调用的最大次数
DeleteIdMappingTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteIdMappingTable API 调用的最大次数
DeleteIdNamespaceAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteIdNamespaceAssociation API 调用的最大次数
DeleteMember 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteMember API 调用的最大次数
DeleteMembership 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeleteMembership API 调用的最大次数
DeletePrivacyBudgetTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 DeletePrivacyBudgetTemplate API 调用的最大次数
GetAnalysisTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetAnalysisTemplate API 调用的最大次数
GetCollaboration 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetCollaboration API 调用的最大次数
GetCollaborationAnalysisTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetCollaborationAnalysisTemplate API 调用的最大次数
GetCollaborationChangeRequest 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetCollaborationChangeRequest API 调用的最大次数

Name	默认值	可调整	说明
GetCollaborationConfiguredAudienceModelAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetCollaborationConfiguredAudienceModelAssociation API 调用的最大次数
GetCollaborationIdNamespaceAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetCollaborationIdNamespaceAssociation API 调用的最大次数
GetCollaborationPrivacyBudgetTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetCollaborationPrivacyBudgetTemplate API 调用的最大次数
GetConfiguredAudienceModelAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetConfiguredAudienceModelAssociation API 调用的最大次数
GetConfiguredTable 请求率	每个受支持的区域：每秒 20 个	是	每秒 GetConfiguredTable API 调用的最大次数
GetConfiguredTableAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetConfiguredTableAnalysisRule API 调用的最大次数
GetConfiguredTableAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetConfiguredTableAssociation API 调用的最大次数
GetConfiguredTableAssociationAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetConfiguredTableAssociationAnalysisRule API 调用的最大次数
GetIdMappingTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetIdMappingTable API 调用的最大次数

Name	默认值	可调整	说明
GetIdNamespaceAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetIdNamespaceAssociation API 调用的最大次数
GetMembership 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetMembership API 调用的最大次数
GetPrivacyBudgetTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetPrivacyBudgetTemplate API 调用的最大次数
GetProtectedJob 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetProtectedJob API 调用的最大次数
GetProtectedQuery 请求率	每个受支持的区域：每秒 20 个	是	每秒 GetProtectedQuery API 调用的最大次数
GetSchema 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetSchema API 调用的最大次数
GetSchemaAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 GetSchemaAnalysisRule API 调用的最大次数
ListAnalysisTemplates 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListAnalysisTemplates API 调用的最大次数
ListCollaborationAnalysisTemplates 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborationAnalysisTemplates API 调用的最大次数
ListCollaborationChangeRequests 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborationChangeRequests API 调用的最大次数

Name	默认值	可调整	说明
ListCollaborationConfiguredAudienceModelAssociations 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborationConfiguredAudienceModelAssociations API 调用的最大次数
ListCollaborationIdNamespaceAssociations 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborationIdNamespaceAssociations API 调用的最大次数
ListCollaborationPrivacyBudgetTemplates 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborationPrivacyBudgetTemplates API 调用的最大次数
ListCollaborationPrivacyBudgets 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborationPrivacyBudgets API 调用的最大次数
ListCollaborations 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListCollaborations API 调用的最大次数
ListConfiguredAudienceModelAssociations 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListConfiguredAudienceModelAssociations API 调用的最大次数
ListConfiguredTableAssociations 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListConfiguredTableAssociations API 调用的最大次数
ListConfiguredTables 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListConfiguredTables API 调用的最大次数
ListIdMappingTables 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListIdMappingTables API 调用的最大次数

Name	默认值	可调整	说明
ListIdNamespaceAssociations 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListIdNamespaceAssociations API 调用的最大次数
ListMembers 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListMembers API 调用的最大次数
ListMemberships 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListMemberships API 调用的最大次数
ListPrivacyBudgetTemplates 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListPrivacyBudgetTemplates API 调用的最大次数
ListPrivacyBudgets 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListPrivacyBudgets API 调用的最大次数
ListProtectedJobs 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListProtectedJobs API 调用的最大次数
ListProtectedQueries 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListProtectedQueries API 调用的最大次数
ListSchemas 请求率	每个受支持的区域：5 个	是	每秒 ListSchemas API 调用的最大次数
ListTagsForResource 请求率	每个受支持的区域：每秒 5 个	是	每秒 ListTagsForResource API 调用的最大次数
PopulateIdMappingTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 PopulateIdMappingTable API 调用的最大次数
PreviewPrivacyImpact 请求率	每个受支持的区域：每秒 5 个	是	每秒 PreviewPrivacyImpact API 调用的最大次数

Name	默认值	可调整	说明
StartProtectedJob 请求率	每个受支持的区域：每秒 5 个	是	每秒 StartProtectedJob API 调用的最大次数
StartProtectedQuery 请求率	每个受支持的区域：每秒 5 个	是	每秒 StartProtectedQuery API 调用的最大次数
TagResource 请求率	每个受支持的区域：每秒 5 个	是	每秒 TagResource API 调用的最大次数
UntagResource 请求率	每个受支持的区域：每秒 5 个	是	每秒 UntagResource API 调用的最大次数
UpdateAnalysisTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateAnalysisTemplate API 调用的最大次数
UpdateCollaboration 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateCollaboration API 调用的最大次数
UpdateCollaborationChangeRequest 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateCollaborationChangeRequest API 调用的最大次数
UpdateConfiguredAudienceModelAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateConfiguredAudienceModelAssociation API 调用的最大次数
UpdateConfiguredTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateConfiguredTable API 调用的最大次数
UpdateConfiguredTableAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateConfiguredTableAnalysisRule API 调用的最大次数

Name	默认值	可调整	说明
UpdateConfiguredTableAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateConfiguredTableAssociation API 调用的最大次数
UpdateConfiguredTableAssociationAnalysisRule 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateConfiguredTableAssociationAnalysisRule API 调用的最大次数
UpdateIdMappingTable 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateIdMappingTable API 调用的最大次数
UpdateIdNamespaceAssociation 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateIdNamespaceAssociation API 调用的最大次数
UpdatePrivacyBudgetTemplate 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdatePrivacyBudgetTemplate API 调用的最大次数
UpdateProtectedJob 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateProtectedJob API 调用的最大次数
UpdateProtectedQuery 请求率	每个受支持的区域：每秒 5 个	是	每秒 UpdateProtectedQuery API 调用的最大次数
每个成员资格的表关联数	每个受支持的区域：25 个	是	每个成员的最大表关联数

AWS Clean Rooms 资源参数限制

资源	默认值	说明
查询文本长度	90 KB	SQL 查询语句的最大文本长度
查询文本长度（使用差别隐私）	90 KB	使用差别隐私的 SQL 查询语句的文本长度上限
查询运行时间	24 小时	查询在超时前运行的最长持续时间

AWS 无尘室机器学习配额

您 AWS 账户 拥有 AWS Clean Rooms ML 的默认配额（以前称为限制）。

要查看 AWS Clean Rooms ML 的服务配额，请执行以下操作之一：

- 按照 Service Quotas 用户指南中的查看服务配额中的步骤进行操作，然后选择 AWS Clean Rooms ML 作为服务。
- 请参阅中的 A WS Clean Rooms 机器学习服务配额Amazon Web Services 一般参考。

为了保持服务的性能并确保适当使用 AWS Clean Rooms ML，分配给账户的默认配额可能会根据地区因素、付款历史记录、欺诈性使用、配额增加请求的 and/or 批准进行更新。

Name	默认值	可调整	说明
每个受众生成作业的活跃受众导出作业数	每个受支持的区域：25 个	否	一个受众生成作业的活跃受众导出作业数上限
每个成员的活动配置模型算法关联数量	每个受支持的区域：1,000 个	是	每个成员的活动配置模型算法关联数量上限
每个成员的活动配置模型算法数量	每个受支持的区域：1,000 个	是	每个成员的活动配置模型算法数量上限

Name	默认值	可调整	说明
每个成员的活动自定义模型输入通道数量	每个受支持的区域：100 个	是	每个成员的活动自定义模型输入通道数量上限
每个账户的活动训练实例数量上限	每个受支持的区域：10 个	是	每个账户可用于创建训练模型的活跃训练实例数量上限。
每个训练模型的最大active/pending/in进度训练模型版本数	每个受支持的区域：100 个	是	每个训练模型可以创建的最大active/pending/in进度训练模型版本数。
生成合成数据的最大输入列数	每个受支持的区域：1000 个	否	生成合成数据的最大输入列数
生成合成数据的最大输入行数	每个支持的区域：2,500,000	否	生成合成数据的最大输入行数
每个账户的 ml.c4.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c4.2xlarge 训练实例数量上限。
每个账户的 ml.c4.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c4.4xlarge 训练实例数量上限。
每个账户的 ml.c4.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c4.8xlarge 训练实例数量上限。
每个账户的 ml.c4.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c4.xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.c5.18xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5.18xlarge 训练实例数量上限。
每个账户的 ml.c5.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5.2xlarge 训练实例数量上限。
每个账户的 ml.c5.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5.4xlarge 训练实例数量上限。
每个账户的 ml.c5.9xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5.9xlarge 训练实例数量上限。
每个账户的 ml.c5.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5.xlarge 训练实例数量上限。
每个账户的 ml.c5n.18xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5n.18xlarge 训练实例数量上限。
每个账户的 ml.c5n.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5n.2xlarge 训练实例数量上限。
每个账户的 ml.c5n.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5n.4xlarge 训练实例数量上限。
每个账户的 ml.c5n.9xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5n.9xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.c5n.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c5n.xlarge 训练实例数量上限。
每个账户的 ml.c6i.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.12xlarge 训练实例数量上限。
每个账户的 ml.c6i.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.16xlarge 训练实例数量上限。
每个账户的 ml.c6i.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.24xlarge 训练实例数量上限。
每个账户的 ml.c6i.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.2xlarge 训练实例数量上限。
每个账户的 ml.c6i.32xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.32xlarge 训练实例数量上限。
每个账户的 ml.c6i.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.4xlarge 训练实例数量上限。
每个账户的 ml.c6i.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.8xlarge 训练实例数量上限。
每个账户的 ml.c6i.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c6i.xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.c7i.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.12xlarge 训练实例数量上限。
每个账户的 ml.c7i.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.16xlarge 训练实例数量上限。
每个账户的 ml.c7i.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.24xlarge 训练实例数量上限。
每个账户的 ml.c7i.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.2xlarge 训练实例数量上限。
每个账户的 ml.c7i.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.48xlarge 训练实例数量上限。
每个账户的 ml.c7i.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.4xlarge 训练实例数量上限。
每个账户的 ml.c7i.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.8xlarge 训练实例数量上限。
每个账户的 ml.c7i.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.large 训练实例数量上限。
每个账户的 ml.c7i.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.c7i.xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.g4dn.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g4dn.12xlarge 训练实例数量上限。
每个账户的 ml.g4dn.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g4dn.16xlarge 训练实例数量上限。
每个账户的 ml.g4dn.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g4dn.2xlarge 训练实例数量上限。
每个账户的 ml.g4dn.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g4dn.4xlarge 训练实例数量上限。
每个账户的 ml.g4dn.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g4dn.8xlarge 训练实例数量上限。
每个账户的 ml.g4dn.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g4dn.xlarge 训练实例数量上限。
每个账户的 ml.g5.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.12xlarge 训练实例数量上限。
每个账户的 ml.g5.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.16xlarge 训练实例数量上限。
每个账户的 ml.g5.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.24xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.g5.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.2xlarge 训练实例数量上限。
每个账户的 ml.g5.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.48xlarge 训练实例数量上限。
每个账户的 ml.g5.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.4xlarge 训练实例数量上限。
每个账户的 ml.g5.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.8xlarge 训练实例数量上限。
每个账户的 ml.g5.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g5.xlarge 训练实例数量上限。
每个账户的 ml.g6.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.12xlarge 训练实例数量上限。
每个账户的 ml.g6.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.16xlarge 训练实例数量上限。
每个账户的 ml.g6.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.24xlarge 训练实例数量上限。
每个账户的 ml.g6.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.2xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.g6.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.48xlarge 训练实例数量上限。
每个账户的 ml.g6.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.4xlarge 训练实例数量上限。
每个账户的 ml.g6.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.8xlarge 训练实例数量上限。
每个账户的 ml.g6.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6.xlarge 训练实例数量上限。
每个账户的 ml.g6e.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.12xlarge 训练实例数量上限。
每个账户的 ml.g6e.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.16xlarge 训练实例数量上限。
每个账户的 ml.g6e.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.24xlarge 训练实例数量上限。
每个账户的 ml.g6e.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.2xlarge 训练实例数量上限。
每个账户的 ml.g6e.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.48xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.g6e.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.4xlarge 训练实例数量上限。
每个账户的 ml.g6e.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.8xlarge 训练实例数量上限。
每个账户的 ml.g6e.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.g6e.xlarge 训练实例数量上限。
每个账户的 ml.m4.10xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m4.10xlarge 训练实例数量上限。
每个账户的 ml.m4.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m4.16xlarge 训练实例数量上限。
每个账户的 ml.m4.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m4.2xlarge 训练实例数量上限。
每个账户的 ml.m4.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m4.4xlarge 训练实例数量上限。
每个账户的 ml.m4.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m4.xlarge 训练实例数量上限。
每个账户的 ml.m5.12xlarge 训练实例数量上限	每个受支持的区域：3 个	是	每个账户可用于创建训练模型的 ml.m5.12xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.m5.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m5.24xlarge 训练实例数量上限。
每个账户的 ml.m5.2xlarge 训练实例数量上限	每个受支持的区域：3 个	是	每个账户可用于创建训练模型的 ml.m5.2xlarge 训练实例数量上限。
每个账户的 ml.m5.4xlarge 训练实例数量上限	每个受支持的区域：3 个	是	每个账户可用于创建训练模型的 ml.m5.4xlarge 训练实例数量上限。
每个账户的 ml.m5.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m5.large 训练实例数量上限。
每个账户的 ml.m5.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m5.xlarge 训练实例数量上限。
每个账户的 ml.m6i.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.12xlarge 训练实例数量上限。
每个账户的 ml.m6i.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.16xlarge 训练实例数量上限。
每个账户的 ml.m6i.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.24xlarge 训练实例数量上限。
每个账户的 ml.m6i.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.2xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.m6i.32xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.32xlarge 训练实例数量上限。
每个账户的 ml.m6i.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.4xlarge 训练实例数量上限。
每个账户的 ml.m6i.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.8xlarge 训练实例数量上限。
每个账户的 ml.m6i.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.large 训练实例数量上限。
每个账户的 ml.m6i.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m6i.xlarge 训练实例数量上限。
每个账户的 ml.m7i.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.12xlarge 训练实例数量上限。
每个账户的 ml.m7i.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.16xlarge 训练实例数量上限。
每个账户的 ml.m7i.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.24xlarge 训练实例数量上限。
每个账户的 ml.m7i.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.2xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.m7i.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.48xlarge 训练实例数量上限。
每个账户的 ml.m7i.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.4xlarge 训练实例数量上限。
每个账户的 ml.m7i.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.8xlarge 训练实例数量上限。
每个账户的 ml.m7i.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.large 训练实例数量上限。
每个账户的 ml.m7i.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.m7i.xlarge 训练实例数量上限。
每个账户的 ml.p2.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p2.16xlarge 训练实例数量上限。
每个账户的 ml.p2.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p2.8xlarge 训练实例数量上限。
每个账户的 ml.p2.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p2.xlarge 训练实例数量上限。
每个账户的 ml.p4d.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p4d.24xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.p4de.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p4de.24xlarge 训练实例数量上限。
每个账户的 ml.p5.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p5.48xlarge 训练实例数量上限。
每个账户的 ml.p5en.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.p5en.48xlarge 训练实例数量上限。
每个账户的 ml.r5.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.12xlarge 训练实例数量上限。
每个账户的 ml.r5.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.16xlarge 训练实例数量上限。
每个账户的 ml.r5.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.24xlarge 训练实例数量上限。
每个账户的 ml.r5.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.2xlarge 训练实例数量上限。
每个账户的 ml.r5.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.4xlarge 训练实例数量上限。
每个账户的 ml.r5.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.8xlarge 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.r5.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.large 训练实例数量上限。
每个账户的 ml.r5.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5.xlarge 训练实例数量上限。
每个账户的 ml.r5d.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.12xlarge 训练实例数量上限。
每个账户的 ml.r5d.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.16xlarge 训练实例数量上限。
每个账户的 ml.r5d.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.24xlarge 训练实例数量上限。
每个账户的 ml.r5d.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.2xlarge 训练实例数量上限。
每个账户的 ml.r5d.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.4xlarge 训练实例数量上限。
每个账户的 ml.r5d.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.8xlarge 训练实例数量上限。
每个账户的 ml.r5d.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.large 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.r5d.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r5d.xlarge 训练实例数量上限。
每个账户的 ml.r7i.12xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.12xlarge 训练实例数量上限。
每个账户的 ml.r7i.16xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.16xlarge 训练实例数量上限。
每个账户的 ml.r7i.24xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.24xlarge 训练实例数量上限。
每个账户的 ml.r7i.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.2xlarge 训练实例数量上限。
每个账户的 ml.r7i.48xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.48xlarge 训练实例数量上限。
每个账户的 ml.r7i.4xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.4xlarge 训练实例数量上限。
每个账户的 ml.r7i.8xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.8xlarge 训练实例数量上限。
每个账户的 ml.r7i.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.large 训练实例数量上限。

Name	默认值	可调整	说明
每个账户的 ml.r7i.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.r7i.xlarge 训练实例数量上限。
每个账户的 ml.t3.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.t3.2xlarge 训练实例数量上限。
每个账户的 ml.t3.large 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.t3.large 训练实例数量上限。
每个账户的 ml.t3.medium 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.t3.medium 训练实例数量上限。
每个账户的 ml.t3.xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.t3.xlarge 训练实例数量上限。
每个账户的 ml.trn1.2xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.trn1.2xlarge 训练实例数量上限。
每个账户的 ml.trn1.32xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.trn1.32xlarge 训练实例数量上限。
每个账户的 ml.trn1n.32xlarge 训练实例数量上限	每个受支持的区域：0 个	是	每个账户可用于创建训练模型的 ml.trn1n.32xlarge 训练实例数量上限。
类别预测列中的最大唯一值	每个受支持的区域：100 个	否	类别预测列中的最大唯一值

Name	默认值	可调整	说明
生成合成数据所需的最小输入列数	每个受支持的区域：5 个	否	生成合成数据所需的最小输入列数
生成合成数据所需的最小输入行数	每个受支持的区域：1,500 个	否	生成合成数据所需的最小输入行数
每位客户待处理/进行中的受众导出作业数	每个受支持的区域：20 个	否	每位客户待处理/进行中的受众导出作业数上限
每位客户待处理/进行中的受众生成作业数	每个受支持的区域：10 个	是	每位客户待处理/进行中的受众生成作业数上限
每位客户待处理/进行中的受众模型数	每个受支持的区域：2 个	是	每位客户待处理/进行中的受众模型训练作业数上限
每个账户的待处理/进行中自定义模型推理作业数量	每个受支持的区域：10 个	是	每个账户的待处理/进行中自定义模型推理作业数量上限
每个成员的待处理/进行中自定义模型推理作业数量	每个受支持的区域：5 个	是	每个成员的待处理/进行中自定义模型推理作业数量上限
每个账户的待处理/进行中自定义模型训练作业数量	每个受支持的区域：10 个	是	每个账户的待处理/进行中自定义模型训练作业数量上限
每个成员的待处理/进行中自定义模型训练作业数量	每个受支持的区域：5 个	是	每个成员的待处理/进行中自定义模型训练作业数量上限
每个账户的待处理/正在进行合成 MLIC 生成作业	每个受支持的区域：2 个	是	每个账户的待处理/正在进行的合成 MLIC 生成任务的最大数量

AWS Clean Rooms ML 还有下面列出的其他配额

Clean Rooms ML 配额

资源	默认值	说明
自定义建模推理作业的最大持续时间	25 个小时	
最大交互次数	200 亿	训练数据中允许的最大交互次数。对于较大的输入，将缩减采样。
最小交互次数	100 万	
用于相似模型训练的最大不同用户数	1 亿	如果包含更多用户，则仅使用前 1 亿个用户（按交互次数排名）。
用于相似模型训练的最小不同用户数	100000	
导出相似细分（受众）作业的用户数下限	10000	
用于模型训练的最大不同项目数。	100 万	您最多可以包含 5000 万个项目，但仅使用最常用的 100 万个项目。
训练数据集中的最大特征列数。	10	
每位用户的不同项目数下限	2	AWS Clean Rooms ML 要求每行或每用户都有两个或更多项目，包括重复的项目。
种子受众的大小上限	500,000	
种子受众的大小下限	500	训练数据提供者最低可以将此值设置为 25。

资源	默认值	说明
APIs	每个客户	
活跃训练数据集总数	500	
活跃相似模型（受众模型）总数	500	
配置的活跃相似模型（受众模型）总数	10000	
完成的相似细分（受众）生成作业总数	无限制	
完成的导出相似细分（受众）作业总数	无限制	
相似模型（受众模型）生成作业的最长持续时间	1 天（24 小时）	
相似细分（受众）生成作业的最长持续时间	10 小时	提供种子后，Clean Rooms ML 最多需要 10 个小时来生成相似细分。如果您使用 SQL 查询作为种子数据，则除了需要 10 个小时来生成相似细分外，最多可能还需要 12 个小时来运行查询。
细分（受众）大小区间的最小百分比	1%	
细分（受众）大小区间的最大百分比	20%	
细分（受众）大小区间的最小绝对大小	不同用户数量的 1%	

资源	默认值	说明
细分（受众）大小区间的最大绝对大小	不同用户数量的 20%	

Clean Rooms ML API 限制配额

AWS 账户 每个终端节点配额中每个账户的每秒 API 交易量 (TPS) 如下。

资源	速率限制	说明
CreateAudienceModel 请求速率	1 TPS 速率，3 TPS 突增	每秒的最大 CreateAudienceModel API 调用数
CreateConfiguredAudienceModel 请求速率	10 TPS	每秒的最大 CreateConfiguredAudienceModel API 调用数
CreateTrainingDataset 请求速率	10 TPS	每秒的最大 CreateTrainingDataset API 调用数
DeleteAudienceGenerationJob 请求速率	2 TPS 速率，10 TPS 突增	每秒的最大 DeleteAudienceGenerationJob API 调用数
DeleteAudienceModel 请求速率	2 TPS 速率，10 TPS 突增	每秒的最大 DeleteAudienceModel API 调用数
DeleteConfiguredAudienceModel 请求速率	10 TPS	每秒的最大 DeleteConfiguredAudienceModel API 调用数
DeleteConfiguredAudienceModelPolicy 请求速率	25 TPS	每秒的最大 DeleteConfiguredAudienceModelPolicy API 调用数
DeleteTrainingDataset 请求速率	10 TPS	每秒的最大 DeleteTrainingDataset API 调用数

资源	速率限制	说明
GetAudienceGenerationJob 请求速率	50 TPS	每秒的最大 GetAudienceGenerationJob API 调用数
GetAudienceModel 请求速率	50 TPS	每秒的最大 GetAudienceModel API 调用数
GetConfiguredAudienceModel 请求速率	50 TPS	每秒的最大 GetConfiguredAudienceModel API 调用数
GetConfiguredAudienceModelPolicy 请求速率	50 TPS	每秒的最大 GetConfiguredAudienceModelPolicy API 调用数
GetTrainingDataset 请求速率	50 TPS	每秒的最大 GetTrainingDataset API 调用数
ListAudienceExportJobs 请求速率	50 TPS	每秒的最大 ListAudienceExportJobs API 调用数
ListAudienceGenerationJobs 请求速率	50 TPS	每秒的最大 ListAudienceGenerationJobs API 调用数
ListAudienceModels 请求速率	50 TPS	每秒的最大 ListAudienceModels API 调用数
ListConfiguredAudienceModels 请求速率	50 TPS	每秒的最大 ListConfiguredAudienceModels API 调用数
ListTagsForResource 请求速率	50 TPS	每秒的最大 ListTagsForResource API 调用数

资源	速率限制	说明
ListTrainingDatasets 请求速率	50 TPS	每秒的最大 ListTrainingDatasets API 调用数
PutConfiguredAudienceModelPolicy 请求速率	25 TPS	每秒的最大 PutConfiguredAudienceModelPolicy API 调用数
StartAudienceExportJob 请求速率	1 TPS 速率，3 TPS 突增	每秒的最大 StartAudienceExportJob API 调用数
StartAudienceGenerationJob 请求速率	1 TPS 速率，5 TPS 突增	每秒的最大 StartAudienceGenerationJob API 调用数
TagResource 请求速率	10 TPS	每秒的最大 TagResource API 调用数
UntagResource 请求速率	50 TPS	每秒的最大 UntagResource API 调用数
UpdateConfiguredAudienceModel 请求速率	10 TPS	每秒的最大 UpdateConfiguredAudienceModel API 调用数
CreateConfiguredModelAlgorithm 请求速率	10 TPS	每秒 CreateConfiguredModelAlgorithm API 调用的最大次数。
CreateConfiguredModelAlgorithmAssociation 请求速率	10 TPS	每秒 CreateConfiguredModelAlgorithmAssociaton API 调用的最大次数。
PutMLConfiguration 请求速率	10 TPS	每秒 PutMLConfiguration API 调用的最大次数。

资源	速率限制	说明
CreateTrainedModel 请求速率	1 TPS 速率，3 TPS 突增	每秒 CreateTrainedModel API 调用的最大次数。
StartTrainedModelExportJob 请求速率	10 TPS	每秒 StartTrainedModelExportJob API 调用的最大次数。
StartTrainedModelInferenceJob 请求速率	1 TPS 速率，3 TPS 速率	每秒 StartTrainedModelInferenceJob API 调用的最大次数。
GetConfiguredModelAlgorithm 请求率	50 TPS	每秒 GetConfiguredModelAlgorithm API 调用的最大次数。
GetConfiguredModelAlgorithmAssociation 请求率	50 TPS	每秒 GetConfiguredModelAlgorithmAssociaton API 调用的最大次数。
GetTrainedModel 请求速率	50 TPS	每秒 GetTrainedModel API 调用的最大次数。
GetMLConfiguration 请求速率	50 TPS	每秒 GetMLConfiguration API 调用的最大次数。
GetTrainedModelInferenceJob 请求速率	50 TPS	每秒 GetTrainedModelInferenceJob API 调用的最大次数。
ListConfiguredModelAlgorithm 请求速率	50 TPS	每秒 ListConfiguredModelAlgorithm API 调用的最大次数。
ListConfiguredModelAlgorithmAssociations 请求速率	50 TPS	每秒 ListConfiguredModelAlgorithmAssociations API 调用的最大次数。

资源	速率限制	说明
ListTrainedModels 请求速率	50 TPS	每秒 ListTrainedModels API 调用的最大次数。
ListCollaborationTrainedModelExportJobs 请求速率	50 TPS	每秒 ListCollaborationTrainedModelExportJobs API 调用的最大次数。
ListCollaborationTrainedModelInferenceJobs 请求速率	50 TPS	每秒 ListCollaborationTrainedModelInferenceJobs API 调用的最大次数。
DeleteConfiguredModelAlgorithm 请求速率	2 TPS 速率，10 TPS 突增	每秒 DeleteConfiguredModelAlgorithm API 调用的最大次数。
DeleteConfiguredModelAlgorithmAssociation 请求速率	2 TPS 速率，10 TPS 突增	每秒 DeleteConfiguredModelAlgorithmAssociaton API 请求的最大数量。
DeleteMLConfiguration 请求速率	2 TPS 速率，10 TPS 突增	每秒 DeleteMLConfiguration API 请求的最大数量。
DeleteTrainedModelOutput 请求速率	2 TPS 速率，10 TPS 突增	每秒 DeleteTrainedModelOutput API 请求的最大数量。

《 AWS Clean Rooms 用户指南》的文档历史记录

下表描述了的文档版本 AWS Clean Rooms。

如需有关此文档的更新通知，您可以订阅 RSS 源。要订阅 RSS 更新，您必须为当前使用的浏览器启用 RSS 插件。

变更	说明	日期
更新了成本分配标签部分的文档	为成本分配相关的可标记资源添加了文档 AWS 账单与成本管理 并创建了章节。	2026年2月20日
支持 cleanrooms-ml spark 计算配置	现在，客户可以通过为 SQL 查询配置支持的 Spark 属性来自定义 Spark 运行时行为并优化 SQL 性能。此功能现已可用于创造受众就业机会和创建MLInput频道。	2026年1月5日
Support 支持详细监控 CloudWatch	客户现在可以选择使用详细监控 CloudWatch 指标对协作进行运营监控，包括查询性能和资源利用率。	2026年1月2日
Support 对协作变更请求 EventBridge 通知事件的支持	为协作变更请求生命周期事件添加了新 EventBridge 事件。	2025 年 12 月 19 日
Support 支持更改请求以管理协作成员资格和设置	协作创建者现在可以提交更改请求以添加新成员、更新现有成员权限和修改自动批准设置。所有协作成员都必须批准变更请求才能生效。	2025 年 12 月 18 日
AWS Clean Rooms SQL 分析引擎的终止支持时间表	对 AWS Clean Rooms SQL 分析引擎的支持将于 2025 年 12 月 17 日结束。客户现在只能使用 Spark SQL 分析引擎。	2025 年 12 月 17 日

Support 支持其他 EventBridge 通知事件	为分析模板、配置的表关联、配置的表关联分析规则、ID 映射表、ID 命名空间关联、受保护的作业和受保护的查询添加了新 EventBridge 事件。	2025 年 12 月 15 日
Support 支持隐私增强型合成数据集生成，用于训练自定义机器学习模型	客户现在可以生成增强隐私的合成数据集来训练自定义机器学习模型，同时保护敏感数据。	2025 年 11 月 30 日
Support 支持在 SQL 查询中配置的 Spark 属性	现在，客户可以通过为 SQL 查询配置支持的 Spark 属性来自定义 Spark 运行时行为。	2025 年 10 月 30 日
Support 支持跨区域协作	AWS Clean Rooms 现在支持 AWS Glue 和 Athena 数据表的跨区域协作，允许客户处理存储 AWS 区域 在一个数据表中的数据，并将结果交付给所有协作成员批准的区域。	2025 年 10 月 3 日
支持数据访问预算	客户现在可以设置和管理数据访问预算，以控制表的使用情况。	2025 年 10 月 1 日
填充 ID 映射表 — 更新	现在，客户可以在基于规则的 ID 映射工作流程中使用增量处理来更高效地处理大型数据集。客户还可以从 ID 映射工作流程中删除记录，以帮助遵守数据管理法规。	2025 年 9 月 23 日
Support 支持 PySpark 作业的可变工作负载大小	客户现在可以选择和调整其计算规模，以更有效地运行 PySpark 作业。	2025 年 9 月 4 日

Support 支持向现有协作中添加新成员	客户现在可以向现有协作中添加新成员。此功能包括配置自动批准设置和阻止成员资格申请的选项。	2025 年 9 月 3 日
增加了对配置错误摘要和编辑自定义模型日志的支持	客户现在可以在配置的模型算法关联中为自定义模型日志配置错误摘要和修改。	2025 年 9 月 3 日
支持 Snowflake 身份验证选项	现在，客户在存储 Snowflake 凭据以进行连接时，可以将密钥对身份验证与 Snowflake PEM 私钥一起使用。 AWS Secrets Manager AWS Clean Rooms	2025 年 8 月 29 日
为 PySpark 分析模板添加了错误配置	客户现在可以在 PySpark 分析模板中配置基本或详细的错误消息级别，以管理错误可见性和故障排除功能。	2025 年 8 月 19 日
Support EventBridge	客户现在可以使用 Amazon EventBridge 将事件从其他软件组件路由 AWS Clean Rooms 到其他软件组件。	2025 年 7 月 31 日
更新现有策略	`AWSCleanRoomsFullAccessNoQuerying` 托管策略中添加了以下新权限：`cleanrooms:UpdateConfiguredTableAllowedColumns` 和 `cleanrooms:UpdateConfiguredTableReference` 。	2025 年 7 月 29 日
增强了已配置表格的编辑功能	客户现在可以修改已配置表的允许列和表引用字段。	2025 年 7 月 29 日

ML 输入通道支持 Parquet 文件类型	客户现在可以选择 Parquet 文件类型作为 ML 输入通道可以使用的数据格式。	2025 年 7 月 17 日
在 Spark SQL 分析引擎上支持差异隐私	客户现在可以在 Spark SQL 分析引擎中使用差异隐私。此外，使用差异隐私的 SQL 查询语句的最大文本长度现在为 90 KB。	2025 年 7 月 16 日
AWS Clean Rooms SQL 分析引擎的终止支持时间表	对使用 Clean Rooms SQL 分析引擎的新协作的支持将于 2025 年 7 月 16 日结束。自 2025 年 7 月 17 日起，您无法使用传统的 Clean Rooms SQL 分析引擎创建新的协作。 AWS Clean Rooms 将于 2025 年 12 月 17 日终止对 Clean Rooms SQL 分析引擎的支持。	2025 年 7 月 16 日
Support 支持增量训练和分布式训练	客户现在可以将增量训练和分布式训练用于其自定义模型。	2025 年 7 月 1 日
Clean Rooms SQL 分析引擎的支持终止时间表	从即日起至2025年7月16日，使用传统Clean Rooms SQL 分析引擎的新协作需要提高限制。从 2025 年 7 月 17 日起，您无法使用此引擎创建新的协作关系。 AWS Clean Rooms 将于 2025 年 12 月 17 日终止对 Clean Rooms SQL 分析引擎的支持。	2025 年 5 月 22 日
更新 PySpark 分析模板	客户现在可以用任何有效的 Python 文件名命名用户脚本文件。	2025 年 5 月 15 日

Support 支持多个结果接收器	客户现在可以选择多个成员来接收来自单个查询的查询结果。	2025 年 4 月 30 日
支持将协作迁移到 Spark SQL	AWS Clean Rooms 除了自定义分析规则外，SQL 现在还支持聚合和列表分析规则。此外，客户可以更新现有协作以使用支持 Spark SQL 的 Spark 分析引擎。	2025 年 4 月 2 日
为 PySpark 工作提供支持	现在，客户可以使用经批准的分析模板通过运行作业来 PySpark 分析数据。	2025 年 3 月 18 日
更新现有策略	在 `AWSCleanRoomsMLReadOnlyAccess` 托管策略添加了以下新权限：`PassCleanRoomsResources` 。`AWSCleanRoomsMLFullAccess` 托管策略中添加了以下新权限：`PassCleanRoomsResources` 和`ConsoleDescribeECRRepositories` 。	2025 年 1 月 10 日
Support 支持多个计算工作者	现在，在创建相似区段时，客户可以指定要配置哪种类型的计算工作线程以及要配置多少计算工作线程。	2024 年 12 月 17 日
Support 支持多个数据源和云	客户现在可以使用多个数据源和云（例如 Amazon Athena 和 Snowflake）与合作伙伴的数据集进行协作。	2024 年 12 月 1 日

Clean Rooms ML 自定义建模现已推出	客户现在可以在协作中使用自己的自定义 ML 模型。	2024 年 11 月 7 日
新的分析引擎	拥有大型数据集的客户现在可以使用 Spark SQL 分析引擎支持的 SQL 函数运行复杂的查询。	2024 年 10 月 29 日
增强隐私保护，创建相似受众，选择多个结果接收者	您可以保护自己的数据，同时还可以通过其他分析和协作分析规则支持复杂的激活查询。您可以通过 SQL 查询或分析模板创建相似受众模型。您可以选择多个成员来接收结果。	2024 年 7 月 24 日
中的实体解析 AWS Clean Rooms	使用 AWS Entity Resolution 数据匹配服务 中 AWS Clean Rooms，您可以在两个 ID 命名空间之间创建 ID 映射表，以跨不同的身份空间查询事件数据。	2024 年 7 月 23 日
更新现有策略	在 `AWSCleanRoomsFullAccessNoQuerying` 托管策略添加了以下新权限：`cleanrooms:BatchGetSchemaAnalysisRule` 。	2024 年 5 月 13 日
AWS Clean Rooms ML 现已完全可用	AWS Clean Rooms ML 提供了一种增强隐私的方法，供双方识别其数据中的相似用户，而无需彼此共享数据。	2024 年 4 月 3 日

更新现有策略	AWSCleanRoomsFullAccess 托管策略中的语句 ID 已从 ConsolePickQueryResultsBucket 更新为 SetQueryResultsBucket，以更好地表示权限。	2024 年 3 月 21 日
AWS Clean Rooms ML 的新托管政策	添加了两个新的托管式策略：AWSCleanRoomsMLReadOnlyAccess 和 AWSCleanRoomsMLFullAccess 。	2023 年 11 月 29 日
AWS Clean Rooms ML（预览版）	AWS Clean Rooms ML 提供了一种增强隐私的方法，供双方识别其数据中的相似用户，而无需彼此共享数据。	2023 年 11 月 29 日
AWS Clean Rooms 差异隐私（预览版）	客户现在可以使用 AWS Clean Rooms 差分隐私来帮助保护其用户的隐私。	2023 年 11 月 29 日
付款配置	协作创建者现在可以配置可以运行查询的成员或协作中的其他成员，以收取查询计算费用。	2023 年 11 月 14 日
查询运行时间 - 更新	超时前运行查询的最长时间从 4 小时更新为 12 小时。	2023 年 10 月 6 日

CloudFormation 资源-更新	AWS Clean Rooms 添加了以下新资源：AWS::CleanRooms::Membership ProtectedQueryOutputConfiguration AWS::CleanRooms::Membership ProtectedQueryResultConfiguration 、和AWS::CleanRooms::Membership ProtectedQueryS3OutputConfiguration 。	2023 年 9 月 7 日
CloudFormation 资源-更新	AWS Clean Rooms 添加了以下新资源：AWS::CleanRooms::AnalysisTemplate 和AWS::CleanRooms::ConfiguredTable AnalysisRuleCustom 。	2023 年 8 月 31 日
成员能力分开	协作创建者现在可以指定一名成员负责查询，另一名成员负责接收结果。这样，协作创建者就能确保可以查询的成员无法访问查询结果。	2023 年 8 月 30 日
AWS Clean Rooms 术语表	仅限文档的更新以添加术语表。 AWS Clean Rooms	2023 年 8 月 30 日
对 Apache Iceberg 表的支持（预览版）	AWS Clean Rooms 现在支持Apache Iceberg表格（预览）。	2023 年 8 月 25 日
配额更新	更新了配额部分，以反映每个账户成员身份的新默认配额。	2023 年 8 月 9 日

对现有策略的更新	在 AWSCleanRoomsFullAccessNoQuerying 托管式策略中新增了以下新权限：cleanrooms:CreateAnalysisTemplate 、cleanrooms:GetAnalysisTemplate 、cleanrooms:UpdateAnalysisTemplate 、 cleanrooms:DeleteAnalysisTemplate 、cleanrooms:ListAnalysisTemplates cleanrooms:GetCollaborationAnalysisTemplate 、cleanrooms:BatchGetCollaborationAnalysisTemplate 和 cleanrooms:ListCollaborationAnalysisTemplates 。	2023 年 7 月 31 日
分析模板和自定义分析规则	AWS Clean Rooms 现在支持分析模板和自定义分析规则。分析模板使协作者能够构建或导入自己的自定义 SQL 查询，以便在协作中使用。使用自定义分析规则，表所有者可以批准对其配置表进行自定义 SQL 查询。	2023 年 7 月 31 日
分析规则支持 OR 逻辑条件	AWS Clean Rooms 分析规则现在支持子JOIN句中的OR逻辑条件。	2023 年 6 月 29 日

CloudFormation 整合	AWS Clean Rooms 现在与集成 CloudFormation。	2023 年 6 月 15 日
分析构建器	现在，能够查询和接收结果的成员可以使用分析构建器用户界面对某些表运行查询，而无需编写 SQL 代码。	2023 年 6 月 15 日
SQL 函数	仅限文档的更新，阐明支持的 SQL 函数。	2023 年 5 月 5 日
故障排查	仅限文档的更新，为常见问题添加了“疑难解答”一节。	2023 年 4 月 27 日
支持的数据类型 AWS Clean Rooms	仅限文档的更新以添加列出支持 AWS Glue Data Catalog 的数据类型的新部分。	2023 年 4 月 26 日
AWS CloudTrail 事件示例	仅限文档的更新，添加了 StartProtectedQuery（成功）和StartProtectedQuery（失败）CloudTrail 的事件示例。	2023 年 4 月 20 日
对现有策略的更新	在 `AWSCleanRoomsFullAccessNoQuerying` 托管式策略中新增了以下新权限：`cleanrooms:ListTagsForResource` 、`cleanrooms:UntagResource` 和 `cleanrooms:TagResource` 。有关更多信息，请参阅 AWS 托管式策略。	2023 年 3 月 21 日
正式发布	AWS Clean Rooms 现已正式上市。	2023 年 3 月 21 日

预览版	《AWS Clean Rooms 用户指南》的预览版	2023 年 1 月 12 日

AWS Clean Rooms 词汇表

请查阅此词汇表，熟悉 AWS Clean Rooms所用的术语。

聚合分析规则

查询限制，允许使用 COUNT、SUM 或 AVG 函数沿可选维度进行聚合分析的查询。这些查询不会泄露行级信息。

支持活动规划、媒体覆盖面、频率和换算测量值等使用案例。

其他类型的分析规则包括自定义和列表。

分析规则

授权特定类型查询的查询限制。

分析规则类型决定了可以在配置表上运行哪种分析。每种类型都有预定义的查询结构。您可以通过查询控制来控制如何在结构中使用表列。

分析规则类型包括聚合、列表和自定义。

分析模板

特定于协作的预先批准的查询，可以重复使用。

支持的格式：适用于 Spark 的 SQL 代码或 Python 代码。

如果使用 SQL，则分析模板可以包含字面值通常可能出现在 SQL 查询中的任何地方。有关支持的参数类型的更多信息，请参阅《AWS Clean Rooms SQL 参考》中的数据类型。

分析模板仅适用于自定义分析规则。

C3R 加密客户端

Clean Rooms 计算加密 (C3R) 加密客户端。

C3R 是一个具有命令行界面的客户端加密 SDK，用于加密和解密数据。

cleartext 列

在 JOIN 或 SELECT SQL 构造中未受加密保护的列。

cleartext 列可以用于 SQL 查询的任何部分。

协作

一种安全的逻辑边界， AWS Clean Rooms 成员可以在其中对已配置的表执行 SQL 查询。

协作由协作创建者创建。

只有受邀参与协作的成员才能加入协作。

一个协作只能有一个成员可以查询数据，也可以有一个成员可以运行查询和作业。

一个协作只能有一个成员可以接收结果。

协作只能让一个成员支付查询计算费用，或者让一个成员支付查询和作业计算成本。

所有成员在加入协作之前都可以看到协作的受邀参与者列表。

协作创建者

创建协作的成员。

每个协作只有一个协作创建者。

只有协作创建者才能从协作中删除成员或删除协作。

配置表

每个已配置的表都表示对中已配置为在 AWS Glue Data Catalog 中使用的现有表的引用 AWS Clean Rooms。配置表包含用于确定如何使用数据的分析规则。

目前， AWS Clean Rooms 支持关联存储在亚马逊简单存储服务 (Amazon S3) 中的数据，这些数据是通过编目的。 AWS Glue

有关的更多信息 AWS Glue，请参阅《AWS Glue 开发人员指南》。

配置表可以与一个或多个协作关联。

> **Note**
>
> AWS Clean Rooms 目前不支持注册到的 Amazon S3 存储桶位置 AWS Lake Formation。

自定义分析规则

查询限制，允许一组特定的预先批准的查询（分析模板），或者允许一组特定的账户来提供使用您的数据的查询或作业。

支持首触归因、增量分析和受众发现分析等使用案例。

支持差别隐私。

其他分析规则类型包括聚合和列表。

解密

将加密数据转换回其原始形式的过程。只有获得密钥，才能进行解密。

差别隐私

一种在数学上非常严格的技术，可以保护协作数据以防止可以接收结果的成员了解特定个人的数据。

加密

使用称为密钥的机密值将数据编码成看似随机的形式的过程。如果无法访问密钥，就无法确定原始明文。

指纹列

在 JOIN SQL 构造中未受加密保护的列。

ID 映射工作流程方法

您希望如何执行 ID 映射。

有两种 ID 映射工作流程方法：

- 基于规则的 ID 映射 - 通过该方法，可以在 ID 映射工作流程中，使用匹配规则将第一方数据从源转换为目标。
- 提供商服务 ID 映射 - 通过该方法，可以在 ID 映射工作流程中，使用提供商服务将第三方编码数据从源转换为目标。

  AWS Clean Rooms 目前支持 LiveRamp 作为基于提供商服务的身份映射工作流程方法。您必须订阅 LiveRamp 直通 AWS Data Exchange 才能使用此方法。有关更多信息，请参阅《AWS Entity Resolution 数据匹配服务 用户指南》中的在 AWS Data Exchange上订阅提供商服务。

ID 映射表

中的一种资源 AWS Clean Rooms ，可在协作中启用第一方匹配规则或多方身份转码。

ID 配置表是对 AWS Glue Data Catalog中现有表的引用。它包含一个 ID 映射表分析规则，用于确定如何在 AWS Clean Rooms中查询数据。ID 映射表可以与一个或多个协作关联。

ID 映射表分析规则

一种由 AWS Clean Rooms 管理的分析规则，用于联接不同的身份数据以方便查询。它会自动添加到 ID 映射表中，并且无法编辑。它会继承协作中其他分析规则的行为，前提是这些分析规则是同构分析规则。

ID 映射工作流程

一种数据处理作业，它根据指定的 ID 映射工作流程方法将数据从源映射到目标。它会生成一个 ID 映射表。

ID 命名空间

中的一种资源 AWS Clean Rooms ，其中包含解释多个数据集 AWS 账户 以及如何在 ID 映射工作流程中使用这些数据集的元数据。

ID 命名空间关联

ID 命名空间资源的关联，有助于您发现其 ID 映射工作流程中的输入。

任务

一种使用一组支持的函数、类和变量在协作中访问和分析已配置表的方法。

AWS Clean Rooms 目前支持该 PySpark 作业类型。

AWS Clean Rooms 目前支持使用 PySpark 分析模板运行作业。

列表分析规则

查询限制，允许对该表和可查询成员表之间的重叠情况输出行级属性分析的查询。

支持扩充以及受众拓展或抑制等使用案例。

其他分析规则类型包括聚合和自定义。

长相模特

训练数据提供者的数据模型，它允许种子数据提供者创建与其种子数据最为相似的训练数据提供者的数据段。

相似区段

与种子数据最为相似的训练数据子集。

成员

作为协作参与者的 AWS 客户。

使用 AWS 账户识别成员身份。

所有成员都可以贡献数据。

可以查询的成员

可以在协作中查询数据的成员。

每个协作中只有一个成员可以查询，而且该成员是不可变的。

管理用户可以使用 AWS Identity and Access Management (IAM) 权限来控制其哪些 IAM 委托人（例如用户或角色）可以查询协作中的数据。有关更多信息，请参阅 创建服务角色以从 Amazon S3 读取数据。

可以运行查询和作业的成员

可以对协作中的数据运行查询和作业的成员。

每次协作只有一个成员可以运行查询和作业，而且该成员是不可变的。

管理用户可以使用 AWS Identity and Access Management (IAM) 权限来控制其哪些 IAM 委托人（例如用户或角色）可以在协作中运行查询和作业。有关更多信息，请参阅 创建服务角色以从 Amazon S3 读取数据。

可以接收结果的成员

可以接收查询结果的成员。能够接收结果的成员可以指定 Amazon S3 目标的查询结果设置和查询结果格式（CSV 或 Parquet）。

对于使用 Spark 分析引擎进行分析，可以接收结果的成员还会指定文件应输出为单个文件还是多个文件。

可以有多个成员在协作中收到结果。

支付查询计算费用的成员

负责支付查询计算费用的成员。

只有一个成员负责支付每个协作的查询计算费用，而且该成员是不可变的。

如果协作创建者未将任何人指定为支付查询计算费用的成员，则可以查询的成员为默认付款人。

支付查询计算费用的成员会收到协作中已运行的查询的账单。

为查询和作业计算费用付费的会员

负责支付查询和作业计算费用的成员。

只有一个成员负责支付每次协作的查询和作业计算费用，而且该成员是不可变的。

如果协作创建者未将任何人指定为支付查询和作业计算费用的成员，则可以查询的成员为默认付款人。

为查询和任务计算费用付费的成员会收到协作中运行的查询的账单。

成员身份

成员加入协作时创建的资源。

成员关联到协作的所有资源都是成员身份的一部分，或与成员身份相关联。

只有拥有该成员身份的成员才能在该成员身份中添加、删除或编辑资源。

密封列

在 SELECT SQL 构造中未受加密保护的列。

种子数据

种子数据提供者的数据，用于创建相似区段。种子数据可以直接提供，也可以来自 AWS Clean Rooms 查询结果。相似细分输出是训练数据中与种子用户最相似的一组用户。

Spark 分析引擎

中的一个分析选项 AWS Clean Rooms ，使客户能够使用 Apache Spark SQL 函数对存储在 Amazon S3、Amazon Athena 或 Snowflake 中的大型数据集运行复杂查询。它还支持中的 PySpark 分析 AWS Clean Rooms。

当你使用 CreateCollaborationAPI 创建协作时，Spark 分析引擎的价值是SPARK。

Query

一种使用一组支持的函数、类和变量在协作中访问和分析已配置表的方法。

AWS Clean Rooms 目前支持 SQL 查询语言。

AWS Clean Rooms 目前支持运行直接 SQL 查询或使用 SQL 分析模板运行查询。

本文属于机器翻译版本。若本译文内容与英语原文存在差异，则一律以英文原文为准。