AWS 决策指南

选择 AWS 分析服务

选择 AWS 分析服务: AWS 决策指南

Table of Contents

选择 AWS 分析服务

迈出第一步

目的	帮助确定哪些 AWS 分析服务最适合您的组织。
上次更新	2025 年 9 月 24 日
承保服务	• Amazon Athena • AWS Clean Rooms • Amazon Data Firehose • Amazon DataZone • Amazon EMR • AWS Glue • Amazon Kinesis Data Streams • 适用于 Apache Flink 的亚马逊托管服务 • Amazon Managed Streaming for Apache Kafka • Amazon Managed Workflows for Apache Airflow • 亚马逊 OpenSearch 服务 • 快点 • Amazon Redshift • Amazon S3 • Amazon SageMaker

简介

数据是现代业务的基础。人员和应用程序需要安全地访问和分析来自各种新来源的数据。数据量也在不断增加，这可能导致组织难以捕获、存储和分析所有必要的数据。

应对这些挑战意味着要构建一个现代化的数据架构，打破所有用于分析和洞察的数据孤岛（包括第三方数据），并通过治理让组织中的每个人都能在一个地方访问这些数据。 end-to-end连接分析和机器学习 (ML) 系统以实现预测性分析也变得越来越重要。

本决策指南可帮助您提出正确的问题，以便在 AWS 服务上构建现代数据架构。它解释了如何打破数据孤岛（通过连接数据湖和数据仓库）、系统孤岛（通过连接机器学习和分析）和人员孤岛（通过将数据交到组织中的每个人手中）。

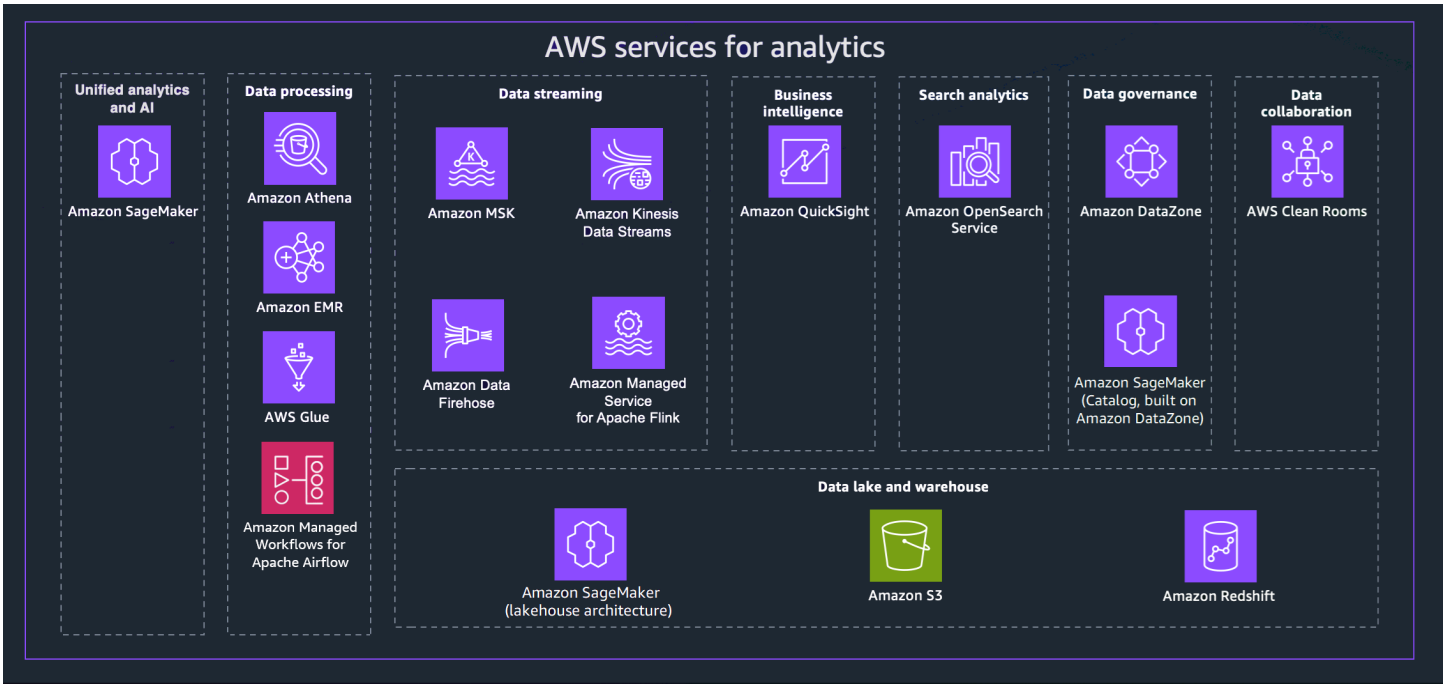
这段八分钟的片段摘自 Sirish Chandrasekaran 和 Rick Sears 在 re: Invent 2024 上发表的一小时演讲。它概述了虚构的公司Maxdome如何使用 SageMaker Unified Studio人工智能和分析（下一代亚马逊 SageMaker的一部分）来解锁数据见解。

了解 AWS 分析服务

现代数据策略由一组技术构建块构建，可帮助您管理、访问、分析和处理数据。它还为您提供了多个连接数据源的选项。现代数据策略应使您的团队能够：

- 使用你喜欢的工具或技术
- 使用人工智能 (AI) 来帮助寻找有关数据的特定问题的答案
- 通过适当的安全和数据治理控制来管理谁有权访问数据
- 打破数据孤岛，让你充分利用数据湖和专门构建的数据存储
- 以开放的、基于标准的数据格式以低成本存储任意数量的数据
- 将您的数据湖、数据仓库、操作数据库、应用程序和联合数据源连接成一个连贯的整体

AWS 提供各种服务来帮助您实现现代数据策略。下图描述了本指南涵盖的分析 AWS 服务。随后的选项卡提供了更多详细信息。



Unified analytics and AI

下一代 Amazon SageMaker 结合了广泛采用的 AWS 机器学习 (ML) 和分析功能，可提供分析和 AI 的集成体验，提供对所有数据的统一访问。使用 Amazon SageMaker Unified Studio，您可以使用熟悉的模型开发、生成式人工智能应用程序开发、数据处理和 SQL 分析 AWS 工具更快地进行协作和构建，所有这些工具均由我们用于软件开发的生成人工智能助手 Amazon Q Developer 加速。通过内置监管功能，访问来自数据湖、数据仓库或第三方和联合来源的数据，以满足企业安全要求。

Data processing

- Amazon A thena 可帮助您分析存储在 Amazon S3 中的非结构化、半结构化和结构化数据。示例包括 CSV、JSON 或列式数据格式，如 Apache Parquet 和 Apache ORC。您可以使用 ANSI SQL 通过 Athena 运行临时查询，而无需将数据聚合或加载到 Athena 中。Athena 与 Quick AWS Glue Data Catalog、和其他服务集成。 AWS您还可以使用 Trino 大规模分析数据，无需管理基础架构，并使用 Apache Flink 和 Apache Spark 进行实时分析。
- Amazon EMR 是一个托管集群平台，可简化大数据框架（例如 Apache Hadoop 和 Apache Spark）的运行， AWS 以处理和分析大量数据。使用这些框架和相关的开源项目，您可以处理用于分析目的的数据和业务情报工作负载。Amazon EMR 还允许您在其他数据存储和数据库（例如 Amazon S3）中转换和移出大量 AWS 数据。
- 借AWS Glue助，您可以发现并连接到 100 多个不同的数据源，并在集中式数据目录中管理您的数据。您可以直观地创建、运行和监控 ETL 管道，将数据加载到数据湖中。此外，您还可以使用 Athena、亚马逊 EMR 和 Amazon Redshift Spectrum 立即搜索和查询编目数据。

- 适用于 Apache Airflow 的亚马逊托管工作流程 (MWAA) 是 Apache Airflow 的完全托管实施，它使在云中创建、安排和监控数据工作流程变得更加容易。MWAA 会自动扩展工作流程容量以满足您的需求，并与 AWS 安全服务集成。您可以使用 MWAA 来协调分析服务中的工作流程，包括数据处理、ETL 作业和机器学习管道。

Data streaming

- 借助适用于 Apache Kafka 的亚马逊托管流媒体（亚马逊 MSK），您可以构建和运行使用 Apache Kafka 处理流数据的应用程序。Amazon MSK 提供控制面板操作，例如，用于创建、更新和删除集群的操作。它允许您使用 Apache Kafka 数据层面操作，例如，用于生成和使用数据的操作。
- 借助 Amazon Kinesis D ata Streams，您可以实时收集和处理大量数据记录流。使用的数据类型可以包括 IT 基础设施日志数据、应用程序日志、社交媒体、市场数据源和 Web 点击流数据。
- Amazon Data Firehose 是一项完全托管的服务，用于向亚马逊 S3、亚马逊 Redshift、OpenSearch 亚马逊服务、Splunk 和 Apache Iceberg Tables 等目的地提供实时流数据。您还可以将数据发送到受支持的第三方服务提供商拥有的任何自定义 HTTP 端点或 HTTP 终端节点，包括 Datadog、Dynatrace、 LogicMonitor MongoDB、New Relic、Coralogix 和 Elastic。
- 借助适用于 Apache Flink 的亚马逊托管服务，你可以使用 Java、Scala、Python 或 SQL 来处理和分析流数据。您可以针对流媒体源和静态源编写和运行代码，以执行时间序列分析、提供实时仪表板和指标。

Business intelligence

Quick 让决策者有机会在交互式视觉环境中探索和解释信息。在单个数据仪表板中，Quick 可以包含 AWS 数据、第三方数据、大数据、电子表格数据、SaaS 数据、B2B 数据等。使用 Quick Q，您可以使用自然语言询问有关您的数据的问题并获得回复。例如，“加州最畅销的类别有哪些？”

Search analytics

Amazon S OpenSearch ervic e 会为您的 OpenSearch 集群配置所有资源并启动它。它还可以自动检测和替换出现故障的 OpenSearch 服务节点，从而减少与自我管理基础架构相关的开销。您可以使用 OpenSearch 服务直接查询来分析 Amazon S3 和其他 AWS 服务中的数据。

Data governance

借助 Amazon DataZone，您可以使用精细的控制来管理和控制对数据的访问。这些控件有助于确保使用适当级别的权限和上下文进行访问。亚马逊通过集成数据管理服务来 DataZone 简化您的架构，包括Amazon Redshift、Athena、Quick AWS Glue、本地来源和第三方来源。

Data collaboration

AWS Clean Rooms是一个安全的协作工作空间，您可以在其中分析集体数据集，而无需提供对原始数据的访问权限。您可以通过选择要与之合作的合作伙伴、选择他们的数据集以及为这些合作伙伴配置隐私增强控制来与其他公司合作。运行查询时，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。

Data lake and data warehouse

- 下一代亚马逊 SageMaker与Apache Iceberg完全兼容，允许您统一亚马逊简单存储服务（Amazon S3）数据湖和亚马逊Redshift数据仓库中的数据。这样就可以在单个数据副本上构建分析、AI 和机器学习 (ML) 应用程序。通过零 ETL 集成，您可以近乎实时地流式传输来自运营源的数据，跨多个来源运行联合查询，并使用兼容 Apache Iceberg 的工具访问数据。您可以通过定义在所有分析、机器学习工具和引擎中强制执行的精细权限来保护您的数据。
- Amazon S3 几乎可以存储和保护任何数量和类型的数据，您可以将这些数据用作数据湖基础。Amazon S3 提供了管理功能，使您可以优化、组织和配置对数据的访问，以满足您的特定业务、组织和合规性要求。Amazon S3 表提供针对分析工作负载进行了优化的 S3 存储。使用标准 SQL 语句，您可以使用支持 Iceberg 的查询引擎（例如 Athena、Amazon Redshift 和 Apache Spark）来查询表。
- Amazon Redshift 是一项完全托管的 PB 级数据仓库服务。Amazon Redshift 可以连接到亚马逊的数据湖库 SageMaker，允许您使用其强大的 SQL 分析功能处理跨亚马逊 Redshift 数据仓库和亚马逊 S3 数据湖的统一数据。你也可以在 Amazon Redshift 中使用 Amazon Q，它通过自然语言简化了 SQL 的创作。

考虑 AWS 分析服务的标准

建立数据分析的原因有很多 AWS。作为云迁移之旅的第一步，您可能需要支持绿地项目或试点项目。或者，您可能在迁移现有工作负载时尽可能减少中断。无论您的目标是什么，以下注意事项都可能有助于您做出选择。

Assess data sources and data types

分析可用的数据源和数据类型，全面了解数据的多样性、频率和质量。了解在处理和分析数据时可能遇到的任何挑战。这种分析至关重要，因为：

- 数据源多种多样，来自不同的系统、应用程序、设备和外部平台。
- 数据源具有独特的结构、格式和数据更新频率。分析这些来源有助于确定合适的数据收集方法和技术。

- 分析数据类型（例如结构化、半结构化和非结构化数据）可确定适当的数据处理和存储方法。
- 分析数据源和类型有助于进行数据质量评估，帮助您预测潜在的数据质量问题——缺失值、不一致或不准确。

Data processing requirements

确定数据处理要求，以了解如何摄取、转换、清理和准备分析数据。关键考虑因素包括：

- 数据转换：确定使原始数据适合分析所需的特定转换。这涉及诸如数据聚合、标准化、筛选和丰富之类的任务。
- 数据清理：评估数据质量并定义处理缺失、不准确或不一致数据的流程。实施数据清理技术，确保提供高质量的数据，从而获得可靠的见解。
- 处理频率：根据分析需求确定是需要实时、近实时还是批处理。实时处理可以立即获得见解，而批处理可能足以进行定期分析。
- 可扩展性和吞吐量：评估处理数据量、处理速度和并发数据请求数量的可扩展性要求。确保所选的处理方法能够适应未来的增长。
- 延迟：考虑数据处理的可接受延迟，以及从数据摄取到分析结果所花费的时间。这对于实时分析或时间敏感型分析尤其重要。

Storage requirements

通过确定数据在整个分析管道中的存储方式和位置来确定存储需求。重要的注意事项包括：

- 数据量：评估生成和收集的数据量，并估计 future 的数据增长，以规划足够的存储容量。
- 数据保留：定义出于历史分析或合规目的应保留数据的期限。确定适当的数据保留政策。
- 数据访问模式：了解如何访问和查询数据，以选择最合适的存储解决方案。考虑读写操作、数据访问频率和数据局部性。
- 数据安全：通过评估加密选项、访问控制和数据保护机制来保护敏感信息，从而优先考虑数据安全。
- 成本优化：根据数据访问模式和使用情况选择最具成本效益的存储解决方案，从而优化存储成本。
- 与分析服务集成：确保所选存储解决方案与正在开发的数据处理和分析工具之间的无缝集成。

Types of data

在决定使用用于收集和摄取数据的分析服务时，请考虑与组织需求和目标相关的各种类型的数据。您可能需要考虑的常见数据类型包括：

- 交易数据：包括有关个人互动或交易的信息，例如客户购买、财务交易、在线订单和用户活动日志。
- 基于文件的数据：指存储在文件中的结构化或非结构化数据，例如日志文件、电子表格、文档、图像、音频文件和视频文件。分析服务应支持摄取不同的文件格式。
- 事件数据：捕获重大事件或事件，例如用户操作、系统事件、计算机事件或业务事件。事件可以包括为上游或下游处理而捕获的任何高速到达的数据。

Operational considerations

运营责任由您共同承担 AWS，不同现代化级别的责任分工各不相同。您可以选择在自己的分析基础架构上自行管理自己的分析基础架构， AWS 也可以利用众多的无服务器分析服务来减轻基础设施管理负担。

自我管理选项使用户可以更好地控制基础架构和配置，但需要更多的操作工作。

无服务器选项消除了大部分运营负担，提供了自动可扩展性、高可用性和强大的安全功能，使用户可以将更多精力集中在构建分析解决方案和推动见解上，而不是管理基础设施和运营任务。考虑一下无服务器分析解决方案的以下好处：

- 基础架构抽象：无服务器服务将基础设施管理抽象化，使用户无需进行配置、扩展和维护任务。AWS 处理这些操作方面，从而减少了管理开销。
- 自动扩展和性能：无服务器服务根据工作负载需求自动扩展资源，无需人工干预即可确保最佳性能。
- 高可用性和灾难恢复： AWS 为无服务器服务提供高可用性。 AWS 管理数据冗余、复制和灾难恢复，以增强数据的可用性和可靠性。
- 安全性与合规性： AWS 管理无服务器服务的安全措施、数据加密和合规性，遵守行业标准和最佳实践。
- 监控和日志记录： AWS 为无服务器服务提供内置监控、日志和警报功能。用户可以通过 Amazon 访问详细的指标和日志 CloudWatch。

Type of workload

在构建现代分析管道时，决定要支持的工作负载类型对于有效满足不同的分析需求至关重要。每种工作负载类型需要考虑的关键决策点包括：

Batch 工作负载

- 数据量和频率：Batch 处理适用于定期更新的大量数据。
- 数据延迟：与实时处理相比，批处理可能会在提供见解方面带来一些延迟。

交互式分析

- 数据查询的复杂性：交互式分析需要低延迟响应以获得快速反馈。
- 数据可视化：评估对交互式数据可视化工具的需求，以使业务用户能够直观地探索数据。

流媒体工作负载

- 数据速度和量：流式处理工作负载需要实时处理才能处理高速数据。
- 数据窗口：为流式数据定义数据窗口和基于时间的聚合，以提取相关见解。

Type of analysis needed

明确定义业务目标和您要从分析中获得的见解。不同类型的分析有不同的用途。例如：

- 描述性分析非常适合获取历史概览
- 诊断分析有助于了解过去事件背后的原因
- 预测分析预测 future 的结果
- 规范性分析为最佳行动提供建议

将您的业务目标与相关类型的分析相匹配。以下是一些关键决策标准，可帮助您选择正确的分析类型：

- 数据可用性和质量：描述性和诊断分析依赖于历史数据，而预测和规范性分析需要足够的历史数据和高质量的数据来构建准确的模型。
- 数据量和复杂性：预测性和规范性分析需要大量的数据处理和计算资源。确保您的基础架构和工具能够处理数据量和复杂性。

- 决策复杂性：如果决策涉及多个变量、约束条件和目标，则规范性分析可能更适合指导最佳行动。
- 风险承受能力：规范性分析可能会提供建议，但会带来相关的不确定性。确保决策者了解与分析结果相关的风险。

Evaluate scalability and performance

评估架构的可扩展性和性能需求。设计必须处理不断增长的数据量、用户需求和分析工作负载。需要考虑的关键决策因素包括：

- 数据量和增长：评估当前的数据量并预测未来的增长。
- 数据速度和实时要求：确定是需要实时还是近乎实时地处理和分析数据。
- 数据处理复杂性：分析数据处理和分析任务的复杂性。对于计算密集型任务，Amazon EMR 等服务为大数据处理提供了可扩展的托管环境。
- 并发和用户负载：考虑系统的并发用户数量和用户负载级别。
- 自动扩展功能：考虑提供自动缩放功能的服务，允许资源根据需求自动向上或向下扩展。这确保了高效的资源利用率和成本优化。
- 地理分布：如果您的数据架构需要分布在多个区域或地点，请考虑采用全球复制和低延迟数据访问的服务。
- 性价比权衡：在性能需求和成本考虑之间取得平衡。高性能的服务可能要付出更高的成本。
- 服务级别协议 (SLAs)：检查 AWS 服务 SLAs提供的服务，确保它们符合您的可扩展性和性能预期。

Data governance

数据治理是您需要实施的一组流程、策略和控制措施，以确保数据资产的有效管理、质量、安全性和合规性。需要考虑的关键决策点包括：

- 数据保留政策：根据监管要求和业务需求定义数据保留政策，并建立在不再需要时安全处置数据的流程。
- 审计跟踪和日志：决定用于监控数据访问和使用情况的日志和审计机制。实施全面的审计跟踪，以跟踪数据更改、访问尝试和用户活动，以实现合规性和安全监控。
- 合规性要求：了解适用于您的组织的特定行业和地理数据合规性法规。确保数据架构符合这些法规和指导方针。
- 数据分类：根据数据的敏感度对数据进行分类，并为每个数据类别定义适当的安全控制措施。

- 灾难恢复和业务连续性：制定灾难恢复和业务连续性计划，以确保发生意外事件或系统故障时数据的可用性和弹性。
- 第三方数据共享：如果与第三方实体共享数据，请实施安全的数据共享协议和协议，以保护数据机密性并防止数据滥用。

Security

分析管道中数据的安全性涉及在管道的每个阶段保护数据，以确保其机密性、完整性和可用性。需要考虑的关键决策点包括：

- 访问控制和授权：实施强大的身份验证和授权协议，确保只有经过授权的用户才能访问特定的数据资源。
- 数据加密：为存储在数据库、数据湖中以及在架构不同组件之间移动数据期间的数据选择适当的加密方法。
- 数据屏蔽和匿名化：考虑是否需要进行数据屏蔽或匿名化以保护敏感数据，例如 PII 或敏感业务数据，同时允许某些分析过程继续进行。
- 安全的数据集成：建立安全的数据集成实践，确保数据在架构的不同组件之间安全流动，避免在数据移动过程中出现数据泄露或未经授权的访问。
- 网络隔离：考虑支持 Amazon VPC 终端节点的服务，以避免将资源暴露给公共互联网。

Plan for integration and data flows

定义分析管道各个组件之间的集成点和数据流，以确保无缝的数据流和互操作性。需要考虑的关键决策点包括：

- 数据源集成：确定要从中收集数据的数据源，例如数据库、应用程序、文件或外部 APIs。确定数据提取方法（批量、实时、基于事件），以便以最小的延迟高效地将数据引入管道。
- 数据转换：确定准备数据以供分析所需的转换。决定在数据流经管道时清理、聚合、标准化或丰富数据的工具和流程。
- 数据移动架构：为管道组件之间的数据移动选择合适的架构。根据实时要求和数据量，考虑批处理、流处理或两者的组合。
- 数据复制和同步：决定数据复制和同步机制，以 up-to-date跨所有组件保留数据。根据数据新鲜度要求，考虑实时复制解决方案或定期数据同步。
- 数据质量和验证：实施数据质量检查和验证步骤，以确保数据在流经管道时的完整性。决定当数据未通过验证时要采取的操作，例如警报或错误处理。

- 数据安全和加密：确定在传输期间和静态数据将如何得到保护。根据数据敏感度考虑所需的安全级别，决定在整个管道中保护敏感数据的加密方法。
- 可扩展性和弹性：确保数据流设计允许横向可扩展性，并且可以处理增加的数据量和流量。

Architect for cost optimization

在此基础上构建分析渠道 AWS 可提供各种成本优化机会。为确保成本效益，请考虑以下策略：

- 资源大小和选择：根据实际工作负载要求调整资源规模。选择与工作负载性能需求相匹配的 AWS 服务和实例类型，同时避免过度配置。
- 自动扩展：为遇到不同工作负载的服务实现自动缩放。Auto-Scaling 可根据需求动态调整实例数量，从而在流量较低的时段降低成本。
- 竞价型实例：将 Amazon EC2 竞价型实例用于非关键和容错工作负载。与按需实例相比，竞价型实例可以显著降低成本。
- 预留实例：考虑购买 AWS 预留实例，与按需定价相比，为使用量可预测的稳定工作负载节省大量成本。
- 数据存储分层：根据数据访问频率使用不同的存储类别，从而优化数据存储成本。
- 数据生命周期策略：制定数据生命周期策略，根据数据的使用年限和使用模式自动移动或删除数据。这有助于管理存储成本并使数据存储与其价值保持一致。

选择 AWS 分析服务

既然您已经知道了评估分析需求的标准，就可以选择适合您的组织需求的 AWS 分析服务了。下表将服务集与通用功能和业务目标对齐。

类别	它针对什么进行了优化？	Services
统一分析和 AI	分析和 AI 开发 针对使用单一开发环境（Amazon SageMaker Unified Studio）访问数据、分析和 AI 功能进行了优化。	Amazon SageMaker
数据处理	交互式分析	Amazon Athena

类别	它针对什么进行了优化？	Services
	针对执行实时数据分析和探索进行了优化，允许用户以交互方式查询和可视化数据。	
	大数据处理 针对处理、移动和转换大量数据进行了优化。	Amazon EMR
	数据目录 经过优化，可提供有关可用数据、其结构、特征和关系的详细信息。	AWS Glue
	工作流程编排 针对使用 Apache Airflow 创建、计划和监控数据工作流进行了优化，以协调分析流程和 ETL 作业。	亚马逊 MWAA
数据流	Apache Kafka 对流数据的处理 针对使用 Apache Kafka 数据平面操作和运行 Apache Kafka 的开源版本进行了优化。	Amazon MSK
	实时处理 针对快速持续的数据采集和聚合进行了优化，包括 IT 基础架构日志数据、应用程序日志、社交媒体、市场数据源和网络点击流数据。	Amazon Kinesis Data Streams

类别	它针对什么进行了优化？	Services
	实时流媒体数据交付 经过优化，可将实时流数据传输到目的地，例如亚马逊 S3、Amazon Redshift、Service、Splunk、Apache Iceberg Tables 以及受支持的第三方 OpenSearch 服务提供商拥有的任何自定义 HTTP 终端节点或 HTTP 终端节点。	Amazon Data Firehose
	构建 Apache Flink 应用程序 针对使用 Java、Scala、Python 或 SQL 处理和分析流数据进行了优化。	适用于 Apache Flink 的亚马逊托管服务
商业智能	仪表板和可视化 经过优化，可直观地呈现复杂的数据集，并提供数据的自然语言查询。	快点
搜索分析	托管 OpenSearch 集群 针对日志分析、实时应用程序监控和点击流分析进行了优化。	亚马逊 OpenSearch 服务
数据治理	管理数据访问权限 针对在数据的整个生命周期中设置适当的管理、可用性、可用性、完整性和安全性进行了优化。	Amazon DataZone

类别	它针对什么进行了优化？	Services
数据协作	安全的数据清洁室 经过优化，无需共享原始底层数据，即可与其他公司协作。	AWS Clean Rooms
数据湖和仓库	统一访问数据湖和数据仓库 基于湖仓架构构建，可进行优化，以统一对 Amazon S3 数据湖、Amazon Redshift 数据仓库、操作数据库以及第三方和联合数据源的数据访问。	Amazon SageMaker
	数据湖的对象存储 经过优化，可提供具有几乎无限可扩展性和高耐久性的数据湖基础。	Amazon S3
	数据仓库 针对集中存储、组织和检索来自组织内各种来源的大量结构化数据（有时甚至是半结构化数据）进行了优化。	Amazon Redshift

使用 AWS 分析服务

现在，您应该清楚地了解自己的业务目标，以及开始构建数据管道时要摄取和分析的数据量和速度。

为了探索如何使用每种可用服务并了解有关这些服务的更多信息，我们提供了探索每项服务如何运作的途径。以下部分提供了指向深入文档、动手教程和资源的链接，可帮助您从基本用法入门到更高级的深度探索。

Amazon Athena

- 亚马逊 Athena 入门

了解如何使用 Amazon Athena 查询数据，并根据存储在 Amazon S3 中的示例数据创建表、查询表以及检查查询结果。

开始阅读本教程

- 开始在 Athena 上使用 Apache Spark

  使用 Athena 控制台中简化的笔记本体验，使用 Python 或 Athena 笔记本开发 Apache Spark 应用程序。 APIs

  开始阅读本教程

- 使用 Amazon 湖屋架构对 Athena 联合查询进行分类和管理 SageMaker

  了解如何通过亚马逊的数据湖库连接、管理和对存储在 Amazon Redshift、DynamoDB 和 Snowflake 中的数据进行联合查询。 SageMaker

  阅读博客

- 使用 Athena 分析 Amazon S3 中的数据

  探索如何在 Elastic Load Balancers 的日志上使用 Athena，这些日志以预定义格式生成为文本文件。我们将向您展示如何创建表、以 Athena 使用的格式对数据进行分区、将其转换为 Parquet 以及比较查询性能。

  阅读博客文章

AWS Clean Rooms

- 设置 AWS Clean Rooms

  了解如何在您的 AWS 账户 AWS Clean Rooms 中进行设置。

  阅读指南

- 开启 AWS 实体解析功能， AWS Clean Rooms 无需共享底层数据，即可解锁跨多方数据集的数据见解

  学习如何使用准备和匹配来帮助改善与协作者的数据匹配。

  阅读博客文章

- 差异隐私如何在不泄露个人层面的数据的情况下帮助解锁见解

了解 AWS Clean Rooms 差异隐私如何简化差异隐私的应用并帮助保护用户的隐私。

阅读博客

Amazon Data Firehose

- 教程：从控制台创建 Firehose 直播

  了解如何使用 AWS 管理控制台 或 AWS SDK 创建到您所选目的地的 Firehose 直播。

  阅读指南

- 向 Firehose 直播发送数据

  了解如何使用不同的数据源向你的 Firehose 直播发送数据。

  阅读指南

- 在 Firehose 中转换源数据

  了解如何调用 Lambda 函数来转换传入的源数据并将转换后的数据传送到目的地。

  阅读指南

Amazon DataZone

- 开始使用亚马逊 DataZone

  学习如何创建 Amazon DataZone 根域、获取数据门户 URL、演练面向数据创建者和数据使用者的基本亚马逊 DataZone 工作流程。

  开始阅读本教程

- 宣布数据谱系将在下一代亚马逊 SageMaker 和亚马逊中全面推出 DataZone

  了解亚马逊如何 DataZone 使用自动世系捕获来专注于自动收集和映射来自 AWS Glue Amazon Redshift 的血统信息。

  阅读博客

Amazon EMR

- 亚马逊 EMR 入门

  学习如何使用 Spark 启动示例集群，以及如何运行存储在 Amazon S3 存储桶中的简单 PySpark 脚本。

  开始阅读本教程

- 开始在亚马逊 EKS 上使用亚马逊 EMR

  我们将向您展示如何通过在虚拟集群上部署 Spark 应用程序来开始在 Amazon EKS 上使用 Amazon EMR。

  浏览指南

- 开始使用 EMR Serverless

  探索 Amazon EMR Serverless 如何提供无服务器运行时环境，简化使用最新开源框架的分析应用程序的操作。

  开始阅读本教程

AWS Glue

- 入门 AWS Glue DataBrew

  学习如何创建您的第一个 DataBrew 项目。您可以加载数据集示例，在该数据集上运行转换，构建用于捕获这些转换的配方，然后运行作业以将转换后的数据写入 Amazon S3。

  开始阅读本教程

- 使用转换数据 AWS Glue DataBrew

  Learn about AWS Glue DataBrew，这是一款可视化数据准备工具，可让数据分析师和数据科学家轻松清理和标准化数据，为分析和机器学习做好准备。学习如何使用 AWS Glue DataBrew构建 ETL 流程。

  开始使用实验室

- AWS Glue DataBrew 沉浸日

  探索 AWS Glue DataBrew 如何使用清理和标准化数据，以便进行分析和机器学习。

从研讨会开始吧

- 开始使用 AWS Glue Data Catalog

  了解如何创建您的第一个存储桶 AWS Glue Data Catalog，该存储桶使用 Amazon S3 存储桶作为数据源。

  开始阅读本教程

- 中的数据目录和爬虫 AWS Glue

  了解如何使用数据目录中的信息来创建和监控 ETL 作业。

  浏览指南

Amazon Kinesis Data Streams

- 亚马逊 Kinesis Data Streams 入门教程

  学习如何处理和分析实时股票数据。

  开始使用教程

- 使用 Amazon Kinesis Data Streams 进行实时分析的架构模式，第 1 部分

  了解两个用例的常见架构模式：时间序列数据分析和事件驱动的微服务。

  阅读博客

- 使用 Amazon Kinesis Data Streams 进行实时分析的架构模式，第 2 部分

  使用 Kinesis Data Streams 在三种场景中了解人工智能应用：实时生成商业智能、实时推荐系统以及物联网数据流和推理。

  阅读博客

Amazon Managed Service for Apache Flink

- 什么是 Apache Flink 的亚马逊托管服务？

  了解适用于 Apache Flink 的亚马逊托管服务的基本概念。

  浏览指南

- 适用于 Apache Flink 研讨会的亚马逊托管服务

  在本次研讨会中，您将学习如何使用适用于 Apache Flink 的亚马逊托管服务部署、操作和扩展 Flink 应用程序。

  参加虚拟研讨会

Amazon MSK

- 亚马逊 MSK 入门

  学习如何创建 Amazon MSK 集群、生成和使用数据，以及如何使用指标监控集群的运行状况。

  开始使用该指南

- 亚马逊 MSK 研讨会

  通过这个动手实践的 Amazon MSK 研讨会深入了解。

  从研讨会开始吧

Amazon MWAA

- 亚马逊 MWAA 入门

  了解如何创建您的第一个 MWAA 环境、将 DAG 上传到 Amazon S3 以及如何运行您的第一个工作流程。

  开始阅读本教程

- 使用 Amazon MWAA 构建数据管道

  学习如何构建 end-to-end数据管道来编排 Glue、EMR 和 Redshift 等其他 AWS 分析服务。这篇博客文章探讨了一种简化的、以配置为导向的方法，该方法使用 MWAA 和 Cosmos 编排 dbt Core 作业，任务在 Amazon Redshift 上进行转换。

  阅读博客文章

- 亚马逊 MWAA 研讨会

  探索动手练习，学习如何部署、配置和使用 Amazon MWAA 进行数据工作流程编排。

  从研讨会开始吧

- 亚马逊 MWAA 的最佳实践

  了解在分析工作流程中使用 Amazon MWAA 的架构模式和最佳实践。

  阅读指南

OpenSearch Service

- OpenSearch 服务入门

  了解如何使用 Amazon OpenSearch 服务创建和配置测试域。

  开始阅读本教程

- 使用 OpenSearch 服务和仪表板可视化客户支持电话 OpenSearch

  了解以下情况的完整演练：企业接到一定数量的客户支持电话并希望对其进行分析。每个呼叫的主题是什么？ 多少是正面的？ 多少是负面的？ 经理如何搜索或查看这些呼叫的脚本？

  开始阅读本教程

- Amazon OpenSearch 无服务器研讨会入门

  了解如何在 AWS 控制台中设置新的 Amazon OpenSearch 无服务器域。探索不同类型的可用搜索查询，设计引人注目的可视化效果，并了解如何根据分配的用户权限保护域名和文档。

  从研讨会开始吧

- 成本优化的矢量数据库：Amazon OpenSearch 服务量化技术简介

  了解 S OpenSearch ervice 如何支持标量和乘积量化技术，以优化内存使用并降低运营成本。

  阅读博客文章

Quick

- 快速数据分析入门

  学习如何创建您的第一个分析。使用样本数据创建简单或更高级的分析。或者，您可以连接到自己的数据来创建分析。

  浏览指南

- 使用 Quick 进行可视化

通过探索商业智能 (BI) 和数据可视化的技术方面 AWS。了解如何将仪表板嵌入到应用程序和网站中，以及如何安全地管理访问权限和权限。

开始学习这门课程

- 快速研讨会

  通过研讨会抢先开启你的快速旅程

  从研讨会开始吧

Amazon Redshift

- 亚马逊 Redshift Serverless 入门

  了解 Amazon Redshift Serverless 创建无服务器资源、连接到 Amazon Redshift Serverless、加载示例数据，然后对数据进行查询的基本流程。

  浏览指南

- 亚马逊 Redshift 深度潜水研讨会

  探索一系列练习，帮助用户开始使用 Amazon Redshift 平台。

  从研讨会开始吧

Amazon S3

- 亚马逊 S3 入门

  学习如何创建您的第一个 DataBrew 项目。您可以加载数据集示例，在该数据集上运行转换，构建用于捕获这些转换的配方，然后运行作业以将转换后的数据写入 Amazon S3。

  开始使用该指南

Amazon SageMaker

- 入门 SageMaker

  学习如何创建项目、添加成员以及如何使用示例 JupyterLab 笔记本开始构建。

阅读指南

- 推出下一代 Amazon SageMaker：所有数据、分析和 AI 的中心

  学习如何开始数据处理、模型开发和生成式 AI 应用程序开发。

  阅读博客

- 什么是 SageMaker 统一工作室？

  了解 SageMaker Unified Studio 的功能以及如何在使用亚马逊时访问这些功能 SageMaker。

  阅读指南

- 亚马逊湖畔建筑入门 SageMaker

  了解如何在 Amazon 中创建项目以及如何浏览、上传和查询业务用例的数据 SageMaker。

  阅读指南

- Amazon 湖畔架构中的数据连接 SageMaker

  了解 Lakehouse 架构如何提供统一的方法来管理跨 AWS 服务和企业应用程序的数据连接。

  阅读指南

- 使用湖屋架构对Athena联合查询进行分类和管理 SageMaker

  了解如何为您的亚马逊项目连接、管理和对存储在 Amazon Redshift、DynamoDB 和 Snowflake 中的数据进行联合查询。 SageMaker

  阅读博客

探索使用 AWS 分析服务的方法

Editable architecture diagrams

参考架构图

浏览架构图，帮助您开发、扩展和测试分析解决方案 AWS。

探索分析参考架构

Ready-to-use code

精选解决方案

使用 Apache Druid 进行可扩展分析 AWS

部署 AWS构建的代码，以帮助您在经济高效、高可用性、弹性和容错性的托管环境上设置 AWS、操作和管理 Apache Druid。

探索此解决方案

AWS 解决方案

探索由构建的预配置、可部署的解决方案及其实施指南。AWS

探索所有 AWS 安全、身份和治理解决方案

Documentation

分析白皮书

浏览白皮书，了解有关选择、实施和使用最适合您组织的分析服务的更多见解和最佳实践。

浏览分析白皮书

AWS 大数据博客

浏览针对特定大数据用例的博客文章。

浏览 AWS 大数据博客

文档历史记录

下表描述了本决策指南的重要更改。要获取有关本指南更新的通知，您可以订阅 RSS feed。

变更	说明	日期
re: Invent 更新	更新了决策指南中的链接，并添加了适用于 Apache Airflow 的亚马逊托管工作流程。	2025年9月24日
re: Invent 更新	更新了决策指南中对亚马逊 SageMaker、亚马逊 SageMaker Unified Studio（不再预览版）和亚马逊 SageMaker Lakehouse 的引用。	2025年9月9日
re: Invent 更新	添加了 SageMaker 人工智能统一工作室和 AWS Clean Rooms. 自始至终更新了文档，增加了新的 AI 特性和功能。	2025 年 2 月 20 日
初次发布	指南首次出版。	2023 年 11 月 17 日

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。