AWS Guia de decisão

Escolhendo um serviço de AWS análise

Escolhendo um serviço de AWS análise: AWS Guia de decisão

Table of Contents

Escolhendo um serviço de AWS análise

Dando o primeiro passo

Finalidade	Ajude a determinar quais serviços de AWS análise são mais adequados para sua organização.
Última atualização	24 de setembro de 2025
Serviços cobertos	• Amazon Athena • AWS Clean Rooms • Amazon Data Firehose • Amazon DataZone • Amazon EMR • AWS Glue • Amazon Kinesis Data Streams • Amazon Managed Service for Apache Flink • Amazon Managed Streaming para Apache Kafka • Amazon Managed Workflows for Apache Airflow • OpenSearch Serviço Amazon • Rápido • Amazon Redshift • Amazon S3 • Amazon SageMaker

Introdução

Os dados são fundamentais para os negócios modernos. Pessoas e aplicativos precisam acessar e analisar dados com segurança, provenientes de fontes novas e diversas. O volume de dados

também está aumentando constantemente, o que pode fazer com que as organizações tenham dificuldade em capturar, armazenar e analisar todos os dados necessários.

Enfrentar esses desafios significa criar uma arquitetura de dados moderna que divide todos os seus silos de dados para análises e insights, incluindo dados de terceiros, e os torna acessíveis a todos na organização, em um só lugar, com governança. end-to-end Também é cada vez mais importante conectar seus sistemas de análise e aprendizado de máquina (ML) para permitir a análise preditiva.

Este guia de decisão ajuda você a fazer as perguntas certas para criar sua arquitetura de dados moderna em AWS serviços. Ele explica como quebrar seus silos de dados (conectando seu data lake e data warehouses), seus silos de sistema (conectando ML e análises) e seus silos de pessoal (colocando os dados nas mãos de todos em sua organização).

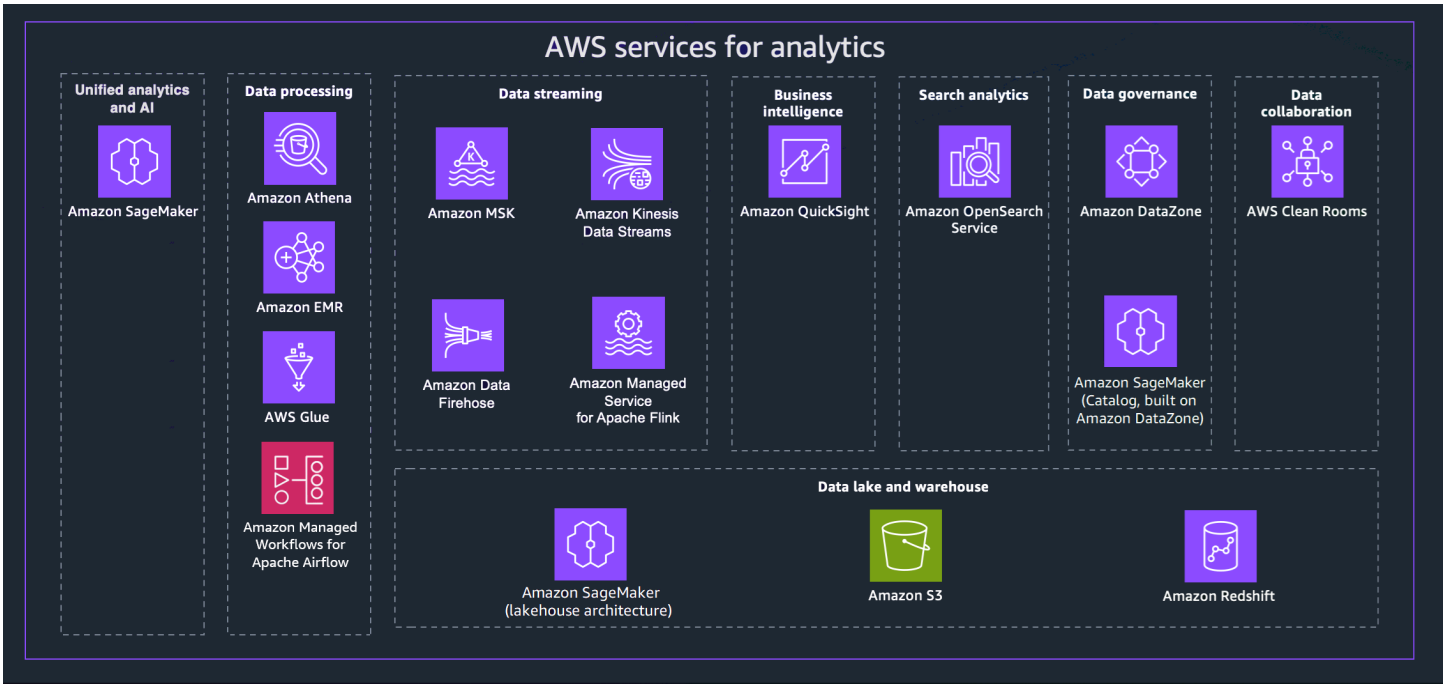
Este trecho de oito minutos é de uma apresentação de uma hora de Sirish Chandrasekaran e Rick Sears no re:Invent 2024. Ele fornece uma visão geral de como uma empresa fictícia, a Maxdome, usa a IA e a análise do SageMaker Unified Studio, parte da próxima geração da Amazon, para desbloquear insights de dados. SageMaker

Entenda os serviços de AWS análise

Uma estratégia de dados moderna é criada com um conjunto de componentes tecnológicos que ajudam você a gerenciar, acessar, analisar e agir com base nos dados. Ele também oferece várias opções para se conectar às fontes de dados. Uma estratégia de dados moderna deve capacitar suas equipes a:

- Use suas ferramentas ou técnicas preferidas
- Use inteligência artificial (IA) para ajudar a encontrar respostas para perguntas específicas sobre seus dados
- Gerencie quem tem acesso aos dados com os controles adequados de segurança e governança de dados
- Elimine os silos de dados para oferecer a você o melhor dos data lakes e dos armazenamentos de dados específicos
- Armazene qualquer quantidade de dados, a baixo custo e em formatos de dados abertos e baseados em padrões
- Conecte seus data lakes, data warehouses, bancos de dados operacionais, aplicativos e fontes de dados federadas em um todo coerente

AWS oferece uma variedade de serviços para ajudá-lo a alcançar uma estratégia de dados moderna. O diagrama a seguir mostra os AWS serviços de análise abordados neste guia. As guias a seguir fornecem detalhes adicionais.



Unified analytics and AI

A próxima geração da Amazon SageMaker combina recursos amplamente adotados AWS de aprendizado de máquina (ML) e análise para oferecer uma experiência integrada de análise e IA, fornecendo acesso unificado a todos os seus dados. Usando o Amazon SageMaker Unified Studio, você pode colaborar e criar mais rapidamente com AWS ferramentas familiares para desenvolvimento de modelos, desenvolvimento de aplicativos de IA generativa, processamento de dados e análise de SQL, tudo acelerado pelo Amazon Q Developer, nosso assistente generativo de IA para desenvolvimento de software. Acesse seus dados de data lakes, data warehouses ou fontes terceirizadas e federadas, com governança integrada para atender aos requisitos de segurança corporativa.

Data processing

- O Amazon Athena ajuda você a analisar dados não estruturados, semiestruturados e estruturados armazenados no Amazon S3. Entre os exemplos estão formatos de dados CSV, JSON ou colunares, como Apache Parquet e Apache ORC. Você pode usar o Athena para executar consultas ad-hoc com o ANSI SQL, sem necessidade de agregar ou carregar os dados no Athena. O Athena se integra ao Quick e a AWS Glue Data Catalog outros serviços.

AWS Você também pode analisar dados em grande escala com o Trino, sem precisar gerenciar a infraestrutura, e criar análises em tempo real usando o Apache Flink e o Apache Spark.

- O Amazon EMR é uma plataforma de cluster gerenciada que simplifica a execução de estruturas de big data, como Apache Hadoop e Apache Spark, para processar e analisar grandes quantidades de dados. AWS Ao usar essas estruturas e projetos de código aberto relacionados, é possível processar dados para finalidades de analytics e workloads de inteligência de negócios. O Amazon EMR também permite transformar e mover grandes quantidades de dados para dentro e para fora de outros bancos de dados e bancos de AWS dados, como o Amazon S3.
- Com AWS Glue, você pode descobrir e se conectar a mais de 100 fontes de dados diversas e gerenciar seus dados em um catálogo de dados centralizado. Você pode criar, executar e monitorar visualmente pipelines de ETL para carregar dados em seus data lakes. Além disso, você pode pesquisar e consultar imediatamente dados catalogados usando o Athena, o Amazon EMR e o Amazon Redshift Spectrum.
- O Amazon Managed Workflows for Apache Airflow (MWAA) é uma implementação totalmente gerenciada do Apache Airflow que facilita a criação, o agendamento e o monitoramento de fluxos de trabalho de dados na nuvem. O MWAA dimensiona automaticamente a capacidade do fluxo de trabalho para atender às suas necessidades e se integra aos AWS serviços de segurança. Você pode usar o MWAA para orquestrar fluxos de trabalho em seus serviços de análise, incluindo processamento de dados, trabalhos de ETL e pipelines de aprendizado de máquina.

Data streaming

- Com o Amazon Managed Streaming for Apache Kafka (Amazon MSK), você pode criar e executar aplicativos que usam o Apache Kafka para processar dados de streaming. O Amazon MSK fornece as operações do ambiente de gerenciamento, como as operações para criar, atualizar e excluir clusters. Ele permite usar operações do plano de dados do Apache Kafka, como aqueles para produzir e consumir dados.
- Com o Amazon Kinesis Data Streams, você pode coletar e processar grandes fluxos de registros de dados em tempo real. O tipo de dados usado pode incluir dados de log de infraestrutura de TI, logs de aplicativo, mídias sociais, feeds de dados de mercado e dados de sequência de cliques da web.
- O Amazon Data Firehose é um serviço totalmente gerenciado para fornecer dados de streaming em tempo real para destinos como Amazon S3, Amazon Redshift, OpenSearch

Amazon Service, Splunk e Apache Iceberg Tables. Você também pode enviar dados para qualquer endpoint HTTP personalizado ou endpoints HTTP de propriedade de provedores de serviços terceirizados compatíveis, incluindo Datadog, Dynatrace, MongoDB, New Relic LogicMonitor, Coralogix e Elastic.

- Com o Amazon Managed Service para Apache Flink, você pode usar Java, Scala, Python ou SQL para processar e analisar dados de streaming. Você pode criar e executar código em fontes de streaming e fontes estáticas para realizar análises de séries temporais, alimentar painéis e métricas em tempo real.

Business intelligence

O Quick oferece aos tomadores de decisão a oportunidade de explorar e interpretar informações em um ambiente visual interativo. Em um único painel de dados, o Quick pode incluir AWS dados, dados de terceiros, big data, dados de planilhas, dados de SaaS, dados B2B e muito mais. Com o Quick Q, você pode usar linguagem natural para fazer perguntas sobre seus dados e receber uma resposta. Por exemplo, “Quais são as categorias mais vendidas na Califórnia?”

Search analytics

O Amazon OpenSearch Service provisiona todos os recursos do seu OpenSearch cluster e o executa. Ele também detecta e substitui automaticamente os nós de OpenSearch serviço com falha, reduzindo a sobrecarga associada às infraestruturas autogerenciadas. Você pode usar o OpenSearch Service Direct Query para analisar dados no Amazon S3 e em outros AWS serviços.

Data governance

Com a Amazon DataZone, você pode gerenciar e controlar o acesso aos dados usando controles refinados. Esses controles ajudam a garantir o acesso com o nível certo de privilégios e contexto. A Amazon DataZone simplifica sua arquitetura integrando serviços de gerenciamento de dados, incluindo Amazon Redshift, Athena, AWS Glue Quick, fontes locais e fontes terceirizadas.

Data collaboration

AWS Clean Roomsé um espaço de trabalho de colaboração seguro onde você pode analisar conjuntos de dados coletivos sem fornecer acesso aos dados brutos. Você pode colaborar com outras empresas escolhendo os parceiros com os quais deseja colaborar, selecionando seus conjuntos de dados e configurando controles de aprimoramento de privacidade para esses parceiros. Quando você executa consultas, AWS Clean Rooms lê dados do local original desses dados e aplica regras de análise integradas para ajudá-lo a manter o controle sobre esses dados.

Data lake and data warehouse

- A próxima geração da Amazon SageMaker é totalmente compatível com o Apache Iceberg, permitindo que você unifique dados nos data lakes do Amazon Simple Storage Service (Amazon S3) e nos armazéns de dados do Amazon Redshift. Isso permite criar aplicativos de análise, IA e aprendizado de máquina (ML) em uma única cópia de dados. Por meio de integrações sem ETL, você pode transmitir dados de fontes operacionais quase em tempo real, executar consultas federadas em várias fontes e acessar dados com ferramentas compatíveis com o Apache Iceberg. Você pode proteger seus dados definindo permissões refinadas que são aplicadas em todas as suas ferramentas e mecanismos de análise e ML.
- O Amazon S3 pode armazenar e proteger praticamente qualquer quantidade e tipo de dados, que você pode usar para sua base de data lake. O Amazon S3 fornece recursos de gerenciamento para que você possa otimizar, organizar e configurar o acesso aos seus dados para atender aos seus requisitos específicos de negócios, organizacionais e de compatibilidade. As tabelas Amazon S3 fornecem armazenamento S3 otimizado para cargas de trabalho de análise. Usando instruções SQL padrão, você pode consultar suas tabelas com mecanismos de consulta compatíveis com o Iceberg, como Athena, Amazon Redshift e Apache Spark.
- O Amazon Redshift é um serviço de armazém de dados totalmente gerenciado em escala de petabytes. O Amazon Redshift pode ser conectado a um data lakehouse na SageMaker Amazon, permitindo que você use seus poderosos recursos analíticos de SQL em seus dados unificados nos armazéns de dados do Amazon Redshift e nos data lakes do Amazon S3. Você também pode usar o Amazon Q no Amazon Redshift, o que simplifica a criação de SQL por meio da linguagem natural.

Considere os critérios para serviços de AWS análise

Há muitos motivos para desenvolver a análise de dados AWS. Talvez você precise apoiar um projeto novo ou piloto como primeira etapa em sua jornada de migração para a nuvem. Como alternativa, você pode migrar uma carga de trabalho existente com o mínimo de interrupção possível. Seja qual for seu objetivo, as considerações a seguir podem ser úteis para fazer sua escolha.

Assess data sources and data types

Analise as fontes de dados e os tipos de dados disponíveis para obter uma compreensão abrangente da diversidade, frequência e qualidade dos dados. Entenda todos os possíveis desafios no processamento e análise dos dados. Essa análise é crucial porque:

- As fontes de dados são diversas e vêm de vários sistemas, aplicativos, dispositivos e plataformas externas.
- As fontes de dados têm estrutura, formato e frequência exclusivos de atualizações de dados. A análise dessas fontes ajuda a identificar métodos e tecnologias de coleta de dados adequados.
- A análise de tipos de dados, como dados estruturados, semiestruturados e não estruturados, determina as abordagens apropriadas de processamento e armazenamento de dados.
- A análise de fontes e tipos de dados facilita a avaliação da qualidade dos dados e ajuda a prever possíveis problemas de qualidade dos dados — valores ausentes, inconsistências ou imprecisões.

Data processing requirements

Determine os requisitos de processamento de dados sobre como os dados são ingeridos, transformados, limpos e preparados para análise. As principais considerações incluem:

- Transformação de dados: determine as transformações específicas necessárias para tornar os dados brutos adequados para análise. Isso envolve tarefas como agregação, normalização, filtragem e enriquecimento de dados.
- Limpeza de dados: avalie a qualidade dos dados e defina processos para lidar com dados ausentes, imprecisos ou inconsistentes. Implemente técnicas de limpeza de dados para garantir dados de alta qualidade para obter insights confiáveis.
- Frequência de processamento: determine se o processamento em tempo real, quase em tempo real ou em lote é necessário com base nas necessidades analíticas. O processamento em tempo real permite insights imediatos, enquanto o processamento em lote pode ser suficiente para análises periódicas.
- Escalabilidade e taxa de transferência: avalie os requisitos de escalabilidade para lidar com volumes de dados, velocidade de processamento e número de solicitações de dados simultâneas. Certifique-se de que a abordagem de processamento escolhida possa acomodar o crescimento futuro.
- Latência: considere a latência aceitável para o processamento de dados e o tempo que leva desde a ingestão dos dados até os resultados da análise. Isso é particularmente importante para análises em tempo real ou urgentes.

Storage requirements

Determine as necessidades de armazenamento determinando como e onde os dados são armazenados em todo o pipeline de análise. Considerações importantes incluem:

- Volume de dados: avalie a quantidade de dados gerados e coletados e estime o crescimento futuro dos dados para planejar a capacidade de armazenamento suficiente.
- Retenção de dados: defina a duração pela qual os dados devem ser retidos para fins de análise histórica ou conformidade. Determine as políticas de retenção de dados apropriadas.
- Padrões de acesso aos dados: entenda como os dados serão acessados e consultados para escolher a solução de armazenamento mais adequada. Considere as operações de leitura e gravação, a frequência de acesso aos dados e a localidade dos dados.
- Segurança de dados: priorize a segurança dos dados avaliando as opções de criptografia, controles de acesso e mecanismos de proteção de dados para proteger informações confidenciais.
- Otimização de custos: otimize os custos de armazenamento selecionando as soluções de armazenamento mais econômicas com base nos padrões de acesso e uso de dados.
- Integração com serviços de análise: garanta uma integração perfeita entre a solução de armazenamento escolhida e as ferramentas de processamento e análise de dados em andamento.

Types of data

Ao decidir sobre serviços de análise para coleta e ingestão de dados, considere vários tipos de dados que são relevantes para as necessidades e objetivos da sua organização. Os tipos comuns de dados que talvez você precise considerar incluem:

- Dados transacionais: incluem informações sobre interações ou transações individuais, como compras de clientes, transações financeiras, pedidos on-line e registros de atividades do usuário.
- Dados baseados em arquivos: referem-se a dados estruturados ou não estruturados que são armazenados em arquivos, como arquivos de log, planilhas, documentos, imagens, arquivos de áudio e arquivos de vídeo. Os serviços de análise devem oferecer suporte à ingestão de diferentes formatos de arquivo.
- Dados de eventos: captura ocorrências ou incidentes significativos, como ações do usuário, eventos do sistema, eventos de máquinas ou eventos de negócios. Os eventos podem

incluir qualquer dado que esteja chegando em alta velocidade e que seja capturado para processamento contínuo ou posterior.

Operational considerations

A responsabilidade operacional é compartilhada entre você e AWS, com a divisão de responsabilidade variando em diferentes níveis de modernização. Você tem a opção de autogerenciar sua infraestrutura de análise AWS ou aproveitar os vários serviços de análise sem servidor para reduzir sua carga de gerenciamento de infraestrutura.

As opções autogerenciadas concedem aos usuários maior controle sobre a infraestrutura e as configurações, mas exigem mais esforço operacional.

As opções sem servidor eliminam grande parte da carga operacional, fornecendo escalabilidade automática, alta disponibilidade e recursos de segurança robustos, permitindo que os usuários se concentrem mais na criação de soluções analíticas e na geração de insights, em vez de gerenciar tarefas operacionais e de infraestrutura. Considere estes benefícios das soluções de análise sem servidor:

- Abstração da infraestrutura: os serviços sem servidor abstraem o gerenciamento da infraestrutura, dispensando os usuários das tarefas de provisionamento, dimensionamento e manutenção. AWS lida com esses aspectos operacionais, reduzindo a sobrecarga de gerenciamento.
- Escalabilidade e desempenho automáticos: os serviços sem servidor escalam automaticamente os recursos com base nas demandas da carga de trabalho, garantindo um desempenho ideal sem intervenção manual.
- Alta disponibilidade e recuperação de desastres: AWS fornece alta disponibilidade para serviços sem servidor. AWS gerencia a redundância de dados, a replicação e a recuperação de desastres para melhorar a disponibilidade e a confiabilidade dos dados.
- Segurança e conformidade: AWS gerencia medidas de segurança, criptografia de dados e conformidade para serviços sem servidor, aderindo aos padrões e melhores práticas do setor.
- Monitoramento e registro: AWS oferece recursos integrados de monitoramento, registro e alerta para serviços sem servidor. Os usuários podem acessar métricas e registros detalhados por meio da Amazon CloudWatch.

Type of workload

Ao criar um pipeline de análise moderno, decidir sobre os tipos de carga de trabalho a serem suportados é crucial para atender às diferentes necessidades analíticas de forma eficaz. Os principais pontos de decisão a serem considerados para cada tipo de carga de trabalho incluem:

Carga de trabalho em lote

- Volume e frequência de dados: o processamento em lote é adequado para grandes volumes de dados com atualizações periódicas.
- Latência de dados: o processamento em lote pode causar algum atraso na entrega de insights em comparação com o processamento em tempo real.

Análise interativa

- Complexidade da consulta de dados: a análise interativa requer respostas de baixa latência para um feedback rápido.
- Visualização de dados: avalie a necessidade de ferramentas interativas de visualização de dados para permitir que os usuários corporativos explorem os dados visualmente.

Cargas de trabalho de streaming

- Velocidade e volume de dados: as cargas de trabalho de streaming exigem processamento em tempo real para lidar com dados de alta velocidade.
- Janela de dados: defina janelas de dados e agregações baseadas em tempo para dados de streaming para extrair insights relevantes.

Type of analysis needed

Defina claramente os objetivos de negócios e os insights que você pretende obter das análises. Diferentes tipos de análise têm finalidades diferentes. Por exemplo:

- A análise descritiva é ideal para obter uma visão geral histórica
- A análise de diagnóstico ajuda a entender os motivos por trás de eventos passados
- A análise preditiva prevê resultados futuros
- A análise prescritiva fornece recomendações para ações ideais

Combine suas metas de negócios com os tipos relevantes de análise. Aqui estão alguns dos principais critérios de decisão para ajudar você a escolher os tipos certos de análise:

- Disponibilidade e qualidade dos dados: a análise descritiva e diagnóstica depende de dados históricos, enquanto a análise preditiva e prescritiva exige dados históricos suficientes e dados de alta qualidade para criar modelos precisos.
- Volume e complexidade dos dados: a análise preditiva e prescritiva exige recursos computacionais e de processamento de dados substanciais. Garanta que sua infraestrutura e suas ferramentas possam lidar com o volume e a complexidade dos dados.
- Complexidade da decisão: se as decisões envolverem várias variáveis, restrições e objetivos, a análise prescritiva pode ser mais adequada para orientar as ações ideais.
- Tolerância ao risco: a análise prescritiva pode fornecer recomendações, mas vem com incertezas associadas. Garanta que os tomadores de decisão entendam os riscos associados aos resultados da análise.

Evaluate scalability and performance

Avalie as necessidades de escalabilidade e desempenho da arquitetura. O design deve lidar com volumes crescentes de dados, demandas de usuários e cargas de trabalho analíticas. Os principais fatores de decisão a serem considerados incluem:

- Volume e crescimento de dados: avalie o volume de dados atual e antecipe o crescimento futuro.
- Velocidade dos dados e requisitos em tempo real: determine se os dados precisam ser processados e analisados em tempo real ou quase em tempo real.
- Complexidade do processamento de dados: analise a complexidade de suas tarefas de processamento e análise de dados. Para tarefas computacionalmente intensivas, serviços como o Amazon EMR fornecem um ambiente escalável e gerenciado para processamento de big data.
- Concorrência e carga do usuário: considere o número de usuários simultâneos e o nível de carga do usuário no sistema.
- Recursos de escalonamento automático: considere serviços que oferecem recursos de escalonamento automático, permitindo que os recursos aumentem ou diminuam automaticamente com base na demanda. Isso garante a utilização eficiente dos recursos e a otimização de custos.

- Distribuição geográfica: considere serviços com replicação global e acesso a dados de baixa latência se sua arquitetura de dados precisar ser distribuída em várias regiões ou locais.
- Compensação entre custo e desempenho: equilibre as necessidades de desempenho com as considerações de custo. Serviços com alto desempenho podem ter um custo maior.
- Acordos de nível de serviço (SLAs): verifique os AWS serviços SLAs fornecidos para garantir que eles atendam às suas expectativas de escalabilidade e desempenho.

Data governance

A governança de dados é o conjunto de processos, políticas e controles que você precisa implementar para garantir gerenciamento, qualidade, segurança e conformidade eficazes de seus ativos de dados. Os principais pontos de decisão a serem considerados incluem:

- Políticas de retenção de dados: defina políticas de retenção de dados com base nos requisitos normativos e nas necessidades comerciais e estabeleça processos para descarte seguro de dados quando não forem mais necessários.
- Trilha de auditoria e registro: decida sobre os mecanismos de registro e auditoria para monitorar o acesso e o uso dos dados. Implemente trilhas de auditoria abrangentes para rastrear alterações de dados, tentativas de acesso e atividades do usuário para monitoramento de conformidade e segurança.
- Requisitos de conformidade: entenda os regulamentos de conformidade de dados geográficos e específicos do setor que se aplicam à sua organização. Certifique-se de que a arquitetura de dados esteja alinhada com esses regulamentos e diretrizes.
- Classificação de dados: classifique os dados com base em sua sensibilidade e defina os controles de segurança apropriados para cada classe de dados.
- Recuperação de desastres e continuidade dos negócios: Planeje a recuperação de desastres e a continuidade dos negócios para garantir a disponibilidade e a resiliência dos dados em caso de eventos inesperados ou falhas no sistema.
- Compartilhamento de dados com terceiros: se estiver compartilhando dados com entidades terceirizadas, implemente protocolos e acordos seguros de compartilhamento de dados para proteger a confidencialidade dos dados e evitar o uso indevido dos dados.

Security

A segurança dos dados no pipeline de análise envolve a proteção dos dados em todas as etapas do pipeline para garantir sua confidencialidade, integridade e disponibilidade. Os principais pontos de decisão a serem considerados incluem:

- Controle e autorização de acesso: implemente protocolos robustos de autenticação e autorização para garantir que somente usuários autorizados possam acessar recursos de dados específicos.
- Criptografia de dados: escolha métodos de criptografia apropriados para dados armazenados em bancos de dados, lagos de dados e durante a movimentação de dados entre diferentes componentes da arquitetura.
- Mascaramento e anonimização de dados: considere a necessidade de mascaramento ou anonimização de dados para proteger dados confidenciais, como PII ou dados comerciais confidenciais, permitindo que determinados processos analíticos continuem.
- Integração segura de dados: estabeleça práticas seguras de integração de dados para garantir que os dados fluam com segurança entre os diferentes componentes da arquitetura, evitando vazamentos de dados ou acesso não autorizado durante a movimentação dos dados.
- Isolamento de rede: considere serviços que ofereçam suporte aos Amazon VPC Endpoints para evitar a exposição de recursos à Internet pública.

Plan for integration and data flows

Defina os pontos de integração e os fluxos de dados entre os vários componentes do pipeline de análise para garantir o fluxo de dados e a interoperabilidade contínuos. Os principais pontos de decisão a serem considerados incluem:

- Integração da fonte de dados: identifique as fontes de dados das quais os dados serão coletados, como bancos de dados, aplicativos, arquivos ou externos APIs. Escolha os métodos de ingestão de dados (em lote, em tempo real, com base em eventos) para levar os dados para o pipeline de forma eficiente e com latência mínima.
- Transformação de dados: determine as transformações necessárias para preparar os dados para análise. Escolha as ferramentas e os processos para limpar, agregar, normalizar ou enriquecer os dados à medida que eles se movem pelo pipeline.
- Arquitetura de movimentação de dados: escolha a arquitetura apropriada para movimentação de dados entre os componentes do pipeline. Considere o processamento em lote, o

processamento em fluxo ou uma combinação de ambos com base nos requisitos em tempo real e no volume de dados.

- Replicação e sincronização de dados: escolha os mecanismos de replicação e sincronização de dados para manter os dados up-to-date em todos os componentes. Considere soluções de replicação em tempo real ou sincronizações periódicas de dados, dependendo dos requisitos de atualização dos dados.
- Qualidade e validação de dados: implemente verificações de qualidade de dados e etapas de validação para garantir a integridade dos dados à medida que eles se movem pelo pipeline. Decida as ações a serem tomadas quando os dados falharem na validação, como alertas ou tratamento de erros.
- Segurança e criptografia de dados: determine como os dados serão protegidos durante o trânsito e em repouso. Decida sobre os métodos de criptografia para proteger dados confidenciais em todo o pipeline, considerando o nível de segurança necessário com base na confidencialidade dos dados.
- Escalabilidade e resiliência: garanta que o design do fluxo de dados permita escalabilidade horizontal e possa lidar com maiores volumes de dados e tráfego.

Architect for cost optimization

A criação de seu pipeline de análise AWS oferece várias oportunidades de otimização de custos. Para garantir a eficiência de custos, considere as seguintes estratégias:

- Dimensionamento e seleção de recursos: dimensione corretamente seus recursos com base nos requisitos reais da carga de trabalho. Escolha AWS serviços e tipos de instância que atendam às necessidades de desempenho das cargas de trabalho, evitando o provisionamento excessivo.
- Escalonamento automático: implemente o escalonamento automático para serviços que experimentam cargas de trabalho variadas. O escalonamento automático ajusta dinamicamente o número de instâncias com base na demanda, reduzindo os custos durante períodos de baixo tráfego.
- Instâncias spot: use as instâncias spot do Amazon EC2 para cargas de trabalho não críticas e tolerantes a falhas. As instâncias spot podem reduzir significativamente os custos em comparação com as instâncias sob demanda.
- Instâncias reservadas: considere comprar instâncias AWS reservadas para obter economias de custo significativas em relação aos preços sob demanda para cargas de trabalho estáveis com uso previsível.

- Classificação por níveis de armazenamento de dados: otimize os custos de armazenamento de dados usando diferentes classes de armazenamento com base na frequência de acesso aos dados.
- Políticas de ciclo de vida de dados: estabeleça políticas de ciclo de vida de dados para mover ou excluir dados automaticamente com base em sua idade e padrões de uso. Isso ajuda a gerenciar os custos de armazenamento e mantém o armazenamento de dados alinhado com seu valor.

Escolha serviços AWS de análise

Agora que você conhece os critérios para avaliar suas necessidades de análise, você está pronto para escolher quais serviços de AWS análise são adequados para suas necessidades organizacionais. A tabela a seguir alinha conjuntos de serviços com recursos e metas comerciais comuns.

Categorias	Para que é otimizado?	Services
Análise unificada e IA	Análise e desenvolvimento de IA Otimizado para usar um único ambiente de desenvolvimento, o Amazon SageMaker Unified Studio, para acessar dados, análises e recursos de IA.	Amazon SageMaker
Processamento de dados	Análise interativa Otimizado para realizar análise e exploração de dados em tempo real, o que permite que os usuários consultem e visualizem dados de forma interativa.	Amazon Athena
	Processamento de big data	Amazon EMR

Categorias	Para que é otimizado?	Services
	Otimizado para processar, mover e transformar grandes quantidades de dados.	
	Catálogo de dados Otimizado para fornecer informações detalhadas sobre os dados disponíveis, sua estrutura, características e relacionamentos.	AWS Glue
	Orquestração do fluxo de trabalho Otimizado para criar, programar e monitorar fluxos de trabalho de dados usando o Apache Airflow para coordenar processos de análise e tarefas de ETL.	Amazon MAA
Streaming de dados	Processamento de dados de streaming pelo Apache Kafka Otimizado para usar operações de plano de dados do Apache Kafka e executar versões de código aberto do Apache Kafka.	Amazon MSK

Categorias	Para que é otimizado?	Services
	Processamento em tempo real Otimizado para ingestão e agregação de dados rápidas e contínuas, incluindo dados de registro da infraestrutura de TI, registros de aplicativos, mídias sociais, feeds de dados de mercado e dados de fluxo de cliques na web.	Amazon Kinesis Data Streams
	Entrega de dados em tempo real Otimizado para fornecer dados de streaming em tempo real para destinos como Amazon S3, Amazon Redshift, Service, Splunk OpenSearch , Apache Iceberg Tables e qualquer endpoint HTTP personalizado ou endpoints HTTP de propriedade de provedores de serviços terceirizados compatíveis.	Amazon Data Firehose
	Criando aplicativos Apache Flink Otimizado para usar Java, Scala, Python ou SQL para processar e analisar dados de streaming.	Amazon Managed Service for Apache Flink

Categorias	Para que é otimizado?	Services
Inteligência de negócios	Painéis e visualizações Otimizado para represent ar visualmente conjuntos de dados complexos e fornecer consultas em linguagem natural de seus dados.	Rápido
Análise de pesquisa	OpenSearch Clusters gerenciados Otimizado para análise de registros, monitoramento de aplicativos em tempo real e análise de fluxo de cliques.	OpenSearch Serviço Amazon
Governança de dados	Gerenciando o acesso aos dados Otimizado para configurar o gerenciamento, a disponibi lidade, a usabilidade, a integridade e a segurança adequados dos dados em todo o ciclo de vida.	Amazon DataZone
Colaboração de dados	Salas limpas de dados seguras Otimizado para colaborar com outras empresas sem compartilhar dados subjacent es brutos.	AWS Clean Rooms

Categorias	Para que é otimizado?	Services
Lago de dados e armazém	Acesso unificado a data lakes e data warehouses Construído em uma arquitetura lakehouse para otimizar a unificação do acesso aos dados em lagos de dados do Amazon S3, armazéns de dados do Amazon Redshift, bancos de dados operacionais e fontes de dados federadas e terceirizadas.	Amazon SageMaker
	Armazenamento de objetos para data lakes Otimizado para fornecer uma base de data lake com escalabilidade praticamente ilimitada e alta durabilidade.	Amazon S3
	Armazenamento de dados Otimizado para armazenar, organizar e recuperar centralmente grandes volumes de dados estruturados e, às vezes, semiestruturados de várias fontes dentro de uma organização.	Amazon Redshift

Use serviços AWS de análise

Agora você deve ter uma compreensão clara dos seus objetivos de negócios e do volume e da velocidade dos dados que você ingerirá e analisará para começar a criar seus pipelines de dados.

Para explorar como usar e aprender mais sobre cada um dos serviços disponíveis, fornecemos um caminho para explorar como cada um dos serviços funciona. As seções a seguir fornecem links para documentação detalhada, tutoriais práticos e recursos para você começar do uso básico aos mergulhos mais avançados.

Amazon Athena

- Comece a usar o Amazon Athena

  Aprenda a usar o Amazon Athena para consultar dados e criar uma tabela com base em dados de amostra armazenados no Amazon S3, consultar a tabela e verificar os resultados da consulta.

  Comece com o tutorial

- Comece a usar o Apache Spark no Athena

  Use a experiência simplificada de notebook no console Athena para desenvolver aplicativos Apache Spark usando Python ou notebook Athena. APIs

  Comece com o tutorial

- Catalogue e controle as consultas federadas do Athena com a arquitetura Amazon Lakehouse SageMaker

  Saiba como se conectar, governar e executar consultas federadas em dados armazenados no Amazon Redshift, no DynamoDB e no Snowflake por meio do data lakehouse na Amazon. SageMaker

  Leia o blog

- Análise de dados no Amazon S3 usando o Athena

  Explore como usar o Athena em registros do Elastic Load Balancers, gerados como arquivos de texto em um formato predefinido. Mostramos como criar uma tabela, particionar os dados em um formato usado pelo Athena, convertê-los em Parquet e comparar o desempenho das consultas.

  Leia a postagem do blog

AWS Clean Rooms

- Configurar AWS Clean Rooms

  Saiba como configurar AWS Clean Rooms sua AWS conta.

  Leia o guia

- Desbloqueie insights de dados em conjuntos de dados de várias partes usando o AWS Entity Resolution on AWS Clean Rooms sem compartilhar dados subjacentes

  Saiba como usar a preparação e a correspondência para ajudar a melhorar a correspondência de dados com os colaboradores.

  Leia a postagem do blog

- Como a privacidade diferencial ajuda a desbloquear insights sem revelar dados em nível individual

  Saiba como a Privacidade AWS Clean Rooms Diferencial simplifica a aplicação da privacidade diferencial e ajuda a proteger a privacidade de seus usuários.

  Leia o blog

Amazon Data Firehose

- Tutorial: Criar um stream do Firehose a partir do console

  Saiba como usar o Console de gerenciamento da AWS ou um AWS SDK para criar um stream do Firehose para o destino escolhido.

  Leia o guia

- Enviar dados para um stream do Firehose

  Saiba como usar fontes de dados diferentes para enviar dados para seu stream do Firehose.

  Leia o guia

- Transforme os dados de origem no Firehose

  Saiba como invocar sua função Lambda para transformar os dados de origem recebidos e entregar os dados transformados aos destinos.

Leia o guia

Amazon DataZone

- Começando com a Amazon DataZone

  Aprenda a criar o domínio DataZone raiz da Amazon, obter a URL do portal de dados e percorrer os DataZone fluxos de trabalho básicos da Amazon para produtores e consumidores de dados.

  Comece com o tutorial

- Anunciando a disponibilidade geral da linhagem de dados na próxima geração da Amazon e da Amazon SageMaker DataZone

  Saiba como a Amazon DataZone usa a captura automatizada de linhagem para se concentrar na coleta e no mapeamento automático de informações de linhagem do Amazon AWS Glue Redshift.

  Leia o blog

Amazon EMR

- Comece a usar o Amazon EMR

  Saiba como iniciar um cluster de amostra usando o Spark e como executar um PySpark script simples armazenado em um bucket do Amazon S3.

  Comece com o tutorial

- Comece a usar o Amazon EMR no Amazon EKS

  Mostramos como começar a usar o Amazon EMR no Amazon EKS implantando um aplicativo Spark em um cluster virtual.

  Explore o guia

- Comece a usar o EMR Serverless

Explore como o Amazon EMR Serverless fornece um ambiente de execução sem servidor que simplifica a operação de aplicativos de análise que usam as estruturas de código aberto mais recentes.

Comece com o tutorial

AWS Glue

- Começando com AWS Glue DataBrew

  Saiba como criar seu primeiro DataBrew projeto. Você carrega um conjunto de dados de amostra, executa transformações nesse conjunto de dados, cria uma receita para capturar essas transformações e executa um trabalho para gravar os dados transformados no Amazon S3.

  Comece com o tutorial

- Transforme dados com AWS Glue DataBrew

  Conheça AWS Glue DataBrew uma ferramenta visual de preparação de dados que facilita a limpeza e a normalização dos dados por analistas e cientistas de dados para prepará-los para análises e aprendizado de máquina. Aprenda a construir um processo de ETL usando o. AWS Glue DataBrew

  Comece a usar o laboratório

- AWS Glue DataBrew dia de imersão

  Explore como usar AWS Glue DataBrew para limpar e normalizar dados para análise e aprendizado de máquina.

  Comece com o workshop

- Começando com o AWS Glue Data Catalog

  Saiba como criar seu primeiro AWS Glue Data Catalog, que usa um bucket do Amazon S3 como fonte de dados.

  Comece com o tutorial

- Catálogo de dados e rastreadores em AWS Glue

Descubra como você pode usar as informações do Catálogo de Dados para criar e monitorar suas tarefas de ETL.

Explore o guia

Amazon Kinesis Data Streams

- Tutoriais de introdução ao Amazon Kinesis Data Streams

  Saiba como processar e analisar dados de estoque em tempo real.

  Comece com os tutoriais
- Padrões arquitetônicos para análises em tempo real usando o Amazon Kinesis Data Streams, parte 1

  Saiba mais sobre padrões de arquitetura comuns de dois casos de uso: análise de dados de séries temporais e microsserviços orientados a eventos.

  Leia o blog
- Padrões arquitetônicos para análises em tempo real usando o Amazon Kinesis Data Streams, parte 2

  Saiba mais sobre os aplicativos de IA com o Kinesis Data Streams em três cenários: inteligência comercial generativa em tempo real, sistemas de recomendação em tempo real e streaming e inferência de dados da Internet das Coisas.

  Leia o blog

Amazon Managed Service for Apache Flink

- O que é o Amazon Managed Service para Apache Flink?

  Entenda os conceitos fundamentais do Amazon Managed Service para Apache Flink.

  Explore o guia
- Workshop do Amazon Managed Service para Apache Flink

  Neste workshop, você aprenderá a implantar, operar e escalar um aplicativo Flink com o Amazon Managed Service para Apache Flink.

Participe do workshop virtual

Amazon MSK

- Conceitos básicos do Amazon MSK

  Saiba como criar um cluster Amazon MSK, produzir e consumir dados e monitorar a integridade do seu cluster usando métricas.

  Comece com o guia
- Workshop sobre Amazon MSK

  Aprofunde-se com este workshop prático sobre o Amazon MSK.

  Comece com o workshop

Amazon MWAA

- Comece a usar o Amazon MWAA

  Aprenda a criar seu primeiro ambiente MWAA, carregar um DAG para o Amazon S3 e executar seu primeiro fluxo de trabalho.

  Comece com o tutorial
- Criação de pipelines de dados com o Amazon MWAA

  Saiba como criar pipelines de end-to-end dados que orquestram outros serviços de AWS análise, como Glue, EMR e Redshift. Esta postagem do blog explora uma abordagem simplificada e orientada por configuração para orquestrar trabalhos do dbt Core usando MWAA e Cosmos, com trabalhos executando transformações no Amazon Redshift.

  Leia a postagem do blog
- Workshop da Amazon MWAA

  Explore exercícios práticos para aprender como implantar, configurar e usar o Amazon MWAA para orquestração do fluxo de trabalho de dados.

  Comece com o workshop

- Melhores práticas para o Amazon MWAA

  Aprenda os padrões de arquitetura e as melhores práticas para usar o Amazon MWAA em seus fluxos de trabalho de análise.

  Leia o guia

OpenSearch Service

- Começando com o OpenSearch serviço

  Saiba como usar o Amazon OpenSearch Service para criar e configurar um domínio de teste.

  Comece com o tutorial

- Visualizando chamadas de suporte ao cliente com OpenSearch serviços e painéis OpenSearch

  Descubra um passo a passo completo da seguinte situação: uma empresa recebe um certo número de chamadas de suporte ao cliente e deseja analisá-las. O que é o assunto de cada chamada? Quantas eram positivas? Quantas eram negativas? Como os gerentes podem pesquisar ou revisar as transcrições dessas chamadas?

  Comece com o tutorial

- Workshop sobre como começar a usar o Amazon OpenSearch Serverless

  Saiba como configurar um novo domínio Amazon OpenSearch Serverless no AWS console. Explore os diferentes tipos de consultas de pesquisa disponíveis, crie visualizações atraentes e saiba como você pode proteger seu domínio e documentos com base nos privilégios de usuário atribuídos.

  Comece com o workshop

- Banco de dados vetorial com custo otimizado: introdução às técnicas OpenSearch de quantização do Amazon Service

  Saiba como o OpenSearch Service oferece suporte a técnicas de quantização escalar e de produtos para otimizar o uso da memória e reduzir os custos operacionais.

  Leia a postagem do blog

Quick

- Introdução à análise rápida de dados

  Saiba como criar sua primeira análise. Use dados de amostra para criar uma análise simples ou mais avançada. Você também pode se conectar aos seus próprios dados para criar uma análise.

  Explore o guia
- Visualizando com o Quick

  Descubra o lado técnico da inteligência de negócios (BI) e da visualização de dados com o. AWS Saiba como incorporar painéis em aplicativos e sites e gerenciar com segurança o acesso e as permissões.

  Comece com o curso
- Workshops rápidos

  Comece sua jornada rápida com workshops

  Comece com os workshops

Amazon Redshift

- Comece a usar o Amazon Redshift Serverless

  Entenda o fluxo básico do Amazon Redshift Serverless para criar recursos sem servidor, conectar-se ao Amazon Redshift Serverless, carregar dados de amostra e, em seguida, executar consultas nos dados.

  Explore o guia
- Workshop de aprofundamento do Amazon Redshift

  Explore uma série de exercícios que ajudam os usuários a começar a usar a plataforma Amazon Redshift.

  Comece com o workshop

Amazon S3

- Começando a usar o Amazon S3

  Saiba como criar seu primeiro DataBrew projeto. Você carrega um conjunto de dados de amostra, executa transformações nesse conjunto de dados, cria uma receita para capturar essas transformações e executa um trabalho para gravar os dados transformados no Amazon S3.

  Comece com o guia

Amazon SageMaker

- Começando com SageMaker

  Saiba como criar um projeto, adicionar membros e usar o JupyterLab caderno de amostra para começar a criar.

  Leia o guia
- Apresentando a próxima geração da Amazon SageMaker: o centro de todos os seus dados, análises e inteligência artificial

  Saiba como começar a usar processamento de dados, desenvolvimento de modelos e desenvolvimento generativo de aplicativos de IA.

  Leia o blog
- O que é o SageMaker Unified Studio?

  Saiba mais sobre os recursos do SageMaker Unified Studio e como acessá-los ao usar a Amazon SageMaker.

  Leia o guia
- Começando com a arquitetura de lakehouse da Amazon SageMaker

  Saiba como criar um projeto e navegar, carregar e consultar dados para seus casos de uso comercial na Amazon SageMaker.

  Leia o guia
- Conexões de dados na arquitetura lakehouse da Amazon SageMaker

Saiba como a arquitetura lakehouse fornece uma abordagem unificada para gerenciar conexões de dados entre AWS serviços e aplicativos corporativos.

Leia o guia

- Catalogue e controle as consultas federadas do Athena com a arquitetura lakehouse SageMaker

  Saiba como se conectar, governar e executar consultas federadas em dados armazenados no Amazon Redshift, DynamoDB e Snowflake para seus projetos da Amazon. SageMaker

  Leia o blog

Explore maneiras de usar os serviços de AWS análise

Editable architecture diagrams

Diagramas de arquitetura de referência

Explore diagramas de arquitetura para ajudá-lo a desenvolver, escalar e testar suas soluções de análise. AWS

Explore as arquiteturas de referência de análise

Ready-to-use code

Solução em destaque	AWS Soluções
Análise escalável usando o Apache Druid em AWS Implante um código AWS criado para ajudá-lo a configurar, operar e gerenciar o Apache Druid on AWS, um ambiente de hospedage m econômico, altamente disponível, resiliente e tolerante a falhas.	Explore soluções pré-confi guradas e implantáveis e seus guias de implement ação, criados por. AWS Explore todas as soluções AWS de segurança, identidade e governança

Explore esta solução

Documentation

Documentos técnicos de análise

Explore os whitepapers para obter mais informações e melhores práticas sobre como escolher, implementar e usar os serviços de análise que melhor se adequam à sua organização.

Explore os whitepapers de análise

AWS Blog de Big Data

Explore postagens de blog que abordam casos específicos de uso de big data.

Explore o blog sobre AWS Big Data

Histórico do documento

A tabela a seguir descreve as mudanças importantes nesse guia de decisão. Para receber notificações sobre atualizações deste guia, você pode assinar um feed RSS.

Alteração	Descrição	Data
Atualizações do re:Invent	Links atualizados em todo o guia de decisão e adição de Amazon Managed Workflows para Apache Airflow.	24 de setembro de 2025
Atualizações do re:Invent	Referências atualizadas para Amazon SageMaker, Amazon SageMaker Unified Studio (não mais versão prévia) e Amazon SageMaker Lakehouse em todo o guia de decisão.	9 de setembro de 2025
Atualizações do re:Invent	Foi adicionado o SageMaker AI Unified Studio AWS Clean Rooms e. Documento atualizado por toda parte com novos recursos e capacidades de IA.	20 de fevereiro de 2025
Publicação inicial	Guia publicado pela primeira vez.	17 de novembro de 2023

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.