

PROJETO SALES TECH

PARTICIPANTES:

Augusto Batista – RA: 10444612

Daniela Alexandra – RA: 10444894

Luan Ferrazzo – RA: 10397276

Objetivo:

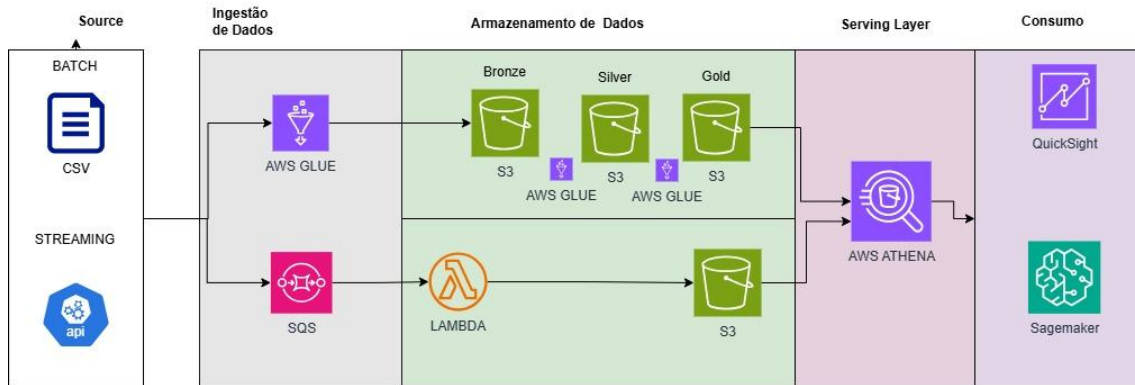
Criar um fluxo que extraia dados transacionais de vendas de produtos de tecnologia e computação, e transformá-los para atender as camadas RAW, SILVER e GOLD em cloud privada - AWS a fim de promover melhorias na escalabilidade, automação e estudos de vendas e promoções. O projeto será com processamento de **rotina Batch** devido a tarefas que não demandam resposta rápida e que envolvem grandes volumes de dados estáticos, como relatórios financeiros, geralmente em intervalos programados (diário, semanal).

Características:

- Trabalha com **lotes grandes de dados**.
- Ideal para **tarefas não interativas** e que podem ser realizadas em horários fora de pico (ex.: durante a madrugada).
- Alta **latência**, já que o processamento só começa quando todos os dados estão disponíveis.
- Usado em sistemas como **ETL** (Extract, Transform, Load), relatórios gerenciais.

1. Etapa de criação da arquitetura

ARQUITETURA:



DESCRIÇÃO: A divisão segue os módulos apresentados: **Fonte de Dados (Source)**, **Ingestão de Dados**, **Armazenamento de Dados**, **Serving Layer** e **Consumo**.

1. Fonte de Dados (Source)

- **Batch (CSV):**
 - Arquivos CSV são carregados em lotes de uma fonte externa (ex.: sistemas legados, exportações de banco de dados).
 - Estes arquivos entram no pipeline por meio do AWS Glue.
- **Streaming (API):**
 - Dados contínuos e em tempo real provenientes de APIs ou serviços externos.
 - Estes dados são direcionados ao **Amazon SQS** para processamento assíncrono.

2. Ingestão de Dados

- **AWS Glue:**
 - Para dados em lote, o AWS Glue processa os arquivos CSV carregados.
 - Realiza a extração, transformação e carregamento (ETL), preparando os dados para armazenamento na camada "Bronze" do **Amazon S3**.
- **Amazon SQS e AWS Lambda:**
 - Para dados em streaming:

- O **Amazon SQS** (Simple Queue Service) atua como um buffer, garantindo que as mensagens sejam entregues com resiliência e desacoplamento.
 - O **AWS Lambda** consome as mensagens da fila e executa transformações ou validações antes de armazenar os dados no **Amazon S3**.
-

3. Armazenamento de Dados

O armazenamento segue o modelo de **Data Lake** em camadas no **Amazon S3**:

- **Bronze:**
 - Contém os dados brutos, diretamente após a ingestão, sem processamento significativo.
 - Utilizado para auditoria ou caso seja necessário reprocessar dados.
 - **Silver:**
 - Dados parcialmente processados, com transformações iniciais e validações aplicadas.
 - Esta camada é gerenciada pelo **AWS Glue**, que executa jobs de ETL.
 - **Gold:**
 - Dados altamente processados e prontos para consumo analítico.
 - Otimizado para consultas por ferramentas como **Amazon Athena**.
-

4. Serving Layer

- **Amazon Athena:**
 - Serviço de consulta que permite acessar os dados armazenados no **Amazon S3** usando SQL.
 - É integrado às camadas de dados (Silver e Gold), facilitando a análise sem necessidade de carregar dados para outro ambiente.
-

5. Consumo

- **Amazon QuickSight:**
 - Ferramenta de BI (Business Intelligence) para visualização de dados.
 - Conecta-se ao **Amazon Athena** para criar relatórios e dashboards interativos.

- **Amazon SageMaker:**
 - Serviço de aprendizado de máquina usado para criar, treinar e implantar modelos de machine learning.
 - Consome dados das camadas Gold (via Athena ou diretamente do S3).

Resumo dos Serviços Utilizados

1. **AWS Glue:**
 - **O que é:** Serviço de ETL gerenciado para descoberta, preparação e transformação de dados.
 - **Benefícios:** Automatiza pipelines de dados, suporta vários formatos e integra-se nativamente ao S3.
 2. **Amazon SQS:**
 - **O que é:** Serviço de fila de mensagens para comunicação assíncrona.
 - **Benefícios:** Alta resiliência, escalabilidade e desacoplamento de sistemas.
 3. **AWS Lambda:**
 - **O que é:** Serviço de computação serverless para execução de código baseado em eventos.
 - **Benefícios:** Escala automaticamente, paga apenas pelo tempo de execução.
 4. **Amazon S3:**
 - **O que é:** Serviço de armazenamento escalável e durável.
 - **Benefícios:** Ideal para Data Lakes, baixo custo e integração nativa com outros serviços AWS.
 5. **Amazon Athena:**
 - **O que é:** Serviço de consulta SQL diretamente no S3.
 - **Benefícios:** Sem necessidade de configurar servidores, ótimo para análises ad-hoc.
 6. **Amazon QuickSight:**
 - **O que é:** Serviço de BI para visualização e análise de dados.
 - **Benefícios:** Criação rápida de dashboards, integração com Athena e outros serviços AWS.
 7. **Amazon SageMaker:**
 - **O que é:** Plataforma para machine learning, abrangendo desde o treinamento até a implantação de modelos.
 - **Benefícios:** Automatiza processos de ML, escalabilidade e integração com dados do S3.
-

Benefícios da Arquitetura

1. **Escalabilidade:** Todos os serviços são altamente escaláveis, permitindo lidar com volumes crescentes de dados.
2. **Custo-Eficiência:** Cobra-se apenas pelo uso real de serviços como Lambda, Athena e S3.
3. **Resiliência:** O uso de SQS e S3 garante alta durabilidade e disponibilidade.
4. **Flexibilidade:** Capacidade de atender a diferentes fontes (batch e streaming) e casos de uso (análises e machine learning).
5. **Integração:** Todos os serviços se integram nativamente, reduzindo complexidade e custos operacionais.