Explicação da arquitetura

terça-feira, 15 de abril de 2025 17:12

A arquitetura proposta utiliza o **AWS DMS** para realizar CDC no banco MySQL; os dados são armazenados em **camadas no Amazon S3** (landing zone, bronze, silver e gold) para suportar uma arquitetura de lakehouse com separação de responsabilidades; o processamento é feito com **AWS Glue em PySpark**, escolhida pela escalabilidade serverless e integração nativa com o Glue Data Catalog; a **catalogação dos metadados** permite consultas SQL via Athena; por fim, a **governança de acesso por usuário é garantida com AWS Lake Formation e IAM**, assegurando controle fino de permissões por database, tabela e coluna.

1. Captura de Dados: MySQL + AWS DMS (CDC)

- **Tecnologia**: AWS DMS
- **Objetivo:** Realizar **CDC (Change Data Capture)** com base em logs binários do MySQL.
- **Justificativa técnica:**
  - Permite ingestão **incremental** de dados.
  - Reduz custo e tempo comparado à extração completa (full load).
  - É **gerenciado pela AWS**, com suporte nativo a S3 como destino.

2. Armazenamento em Lakehouse no S3

- **Tecnologia:** Amazon S3
- **Camadas:** Landing zone, Bronze , Silver, Gold
- **Justificativa técnica:**
  - Alta durabilidade e escalabilidade de armazenamento.
  - Organização em camadas facilita **tracing, governança e reprocessamento**.

3. Processamento com AWS Glue (PySpark)

- **Tecnologia:** AWS Glue + Apache Spark
- **Tarefas:** Limpeza, normalização, enriquecimento, deduplicação.
- **Justificativa técnica:**
  - Engine escalável, serverless, com auto-provisionamento.
  - Integração nativa com S3, Glue Catalog, Lake Formation e Delta Lake.
  - Permite desenvolvimento em PySpark com suporte a Delta Lake para ACID.

4. Catalogação com AWS Glue Data Catalog

- **Tecnologia:** AWS Glue Crawler + Glue Catalog
- **Objetivo:** Tornar dados consultáveis via SQL e acessíveis por ferramentas como Athena.
- **Justificativa técnica:**
  - Centraliza e versiona metadados.
  - Suporte automático a particionamento, schema evolution e integração com Lake Formation.

5. Governança e Segurança com Lake Formation e IAM

- **Tecnologias:** AWS Lake Formation + IAM Policies
- **Controles aplicados:**
  - Permissões por **database, tabela e coluna**.
  - Restrições por **grupos de usuários** ou tags.
  - Logs e trilha de auditoria via CloudTrail.
- **Justificativa técnica:**
  - Cumpre exigências de LGPD, ISO, SOX e outras regulamentações.
  - Define perfis como exemplo:
    - Cientistas de dados: acesso à camada Silver
    - Time de Negocio: acesso apenas à Gold
    - Engenheiros: permissão total

6. Consumo Analítico

- **Ferramentas:** Athena, Jupyter

- **Justificativa técnica:**
    - Oferece **consultas serverless** e visuais sobre dados.
    - Compatibilidade com BI tradicional e notebooks Python/SQL.
    - Permite democratização dos dados com controle de acesso.