Teste técnico - Engenheiro de Dados Pleno

Proposta de Arquitetura de Pipeline de Dados na Google Cloud Platform

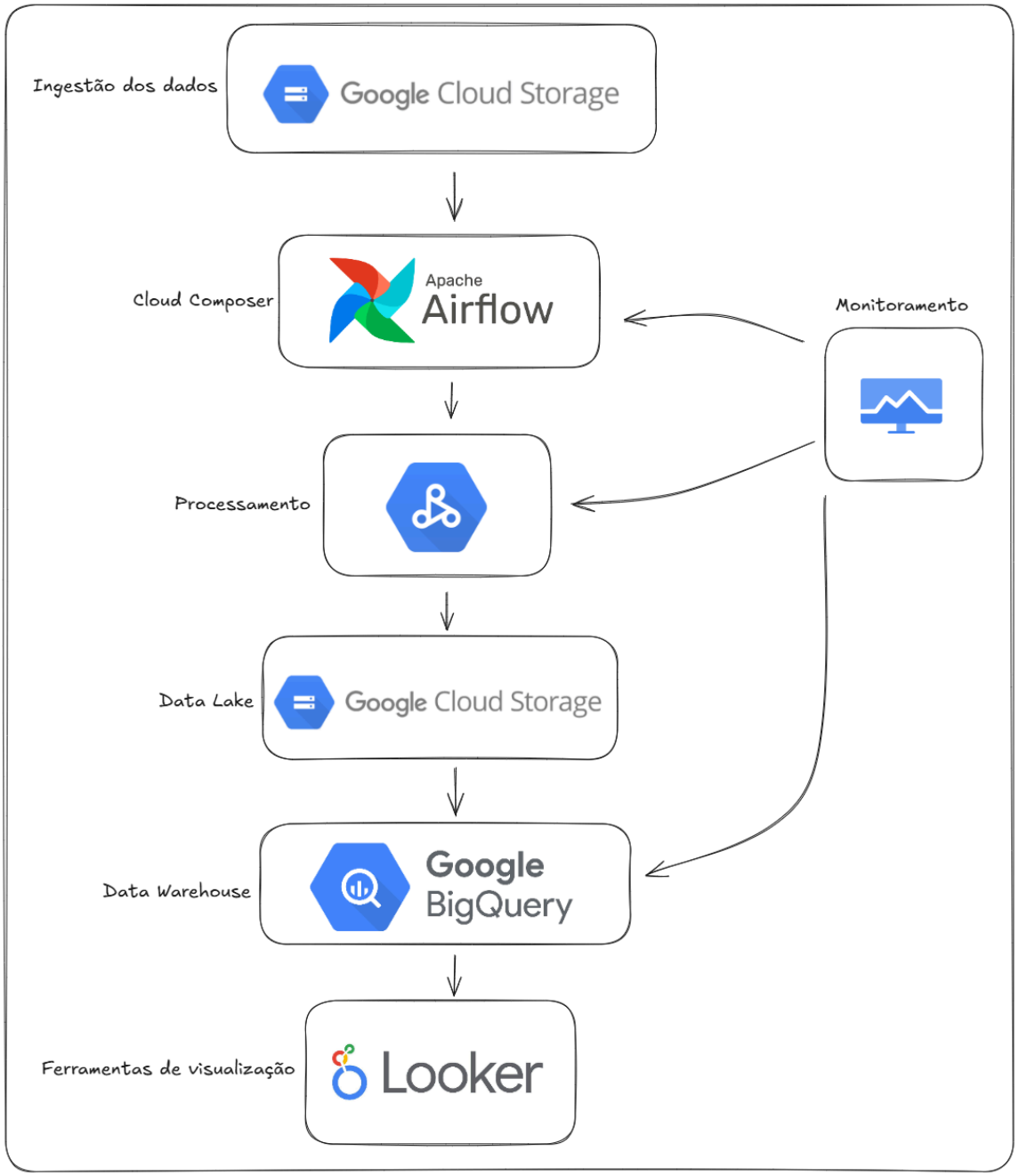
Henrique Albuquerque Ferreira - 30/09/2024

Índice

1. Design do pipeline

1.1. Visão geral

O pipeline de dados foi projetado para coletar, transformar e analisar dados de vendas de forma eficiente e escalável. A arquitetura proposta utiliza o Google Cloud Platform (GCP) para implementar um fluxo de trabalho robusto, permitindo uma tomada de decisão estratégica baseada em dados.



1.2. Camadas de dados

O pipeline é organizado em três camadas principais:

- Camada Bronze: Armazena os dados brutos coletados de vendas.
- Camada Silver: Contém dados limpos e transformados, prontos para análise.
- Camada Gold: Armazena dados agregados e prontos para relatórios.

1.3. Serviços utilizados

Os principais serviços utilizados na arquitetura incluem:

- Google Cloud Storage (GCS): Para armazenamento das camadas de dados.
- Google Cloud Dataproc: Para processamento e transformação dos dados.
- Google BigQuery: Para análise de dados e criação de relatórios.
- Google Cloud Composer: Para orquestração do pipeline.

2. Extração e carregamento dos dados

2.1. Ingestão

Os dados são automaticamente inseridos no GCS em sua forma bruta. Por exemplo, um sistema SAP dentro da empresa pode ter uma rotina que envia os dados diretamente para o GCS, garantindo a consistência e a organização do carregamento dos dados. Essa abordagem elimina a necessidade de intervenções manuais na inserção de novos dados.

2.2. Organização no Cloud Storage

Os dados são armazenados em diretórios específicos no GCS, organizados de acordo com as camadas de dados:

- gs://seu-bucket/bronze/vendas/ para dados brutos.
- gs://seu-bucket/silver/vendas/ para dados limpos.
- gs://seu-bucket/gold/vendas_agregadas/ para dados agregados.

3. Transformação dos dados

3.1. Transformação com PySpark

Utilizando o Dataproc e PySpark, os dados são limpos e transformados na camada Silver. Exemplos de transformação incluem remoção de duplicatas e conversão de tipos de dados.

3.2. Uso do formato Parquet

Os dados são armazenados no formato Parquet, que permite uma compressão eficiente e otimiza a leitura e a consulta de dados no BigQuery.

3.3. Armazenamento em BigQuery

Os dados processados na camada Gold são carregados no BigQuery para análise. Isso permite consultas rápidas e a criação de relatórios.

4. Orquestração do pipeline

4.1. Cloud Composer (Airflow)

O Cloud Composer é utilizado para orquestrar todo o pipeline. Ele permite programar, monitorar e gerenciar tarefas de forma eficiente, garantindo que cada etapa do processo ocorra na sequência correta. O Airflow é essencial para a automação de workflows complexos, facilitando a execução programada de tarefas de transformação e carga. Além disso, ele proporciona visibilidade sobre o estado do pipeline, permitindo que a equipe monitore a execução das DAGs.

5. Segurança e governança de dados

5.1. Controle de acesso

Implementamos políticas de controle de acesso no GCS e no BigQuery para garantir que apenas usuários autorizados tenham acesso aos dados.

5.2. Auditoria

Registros de auditoria são mantidos para monitorar quem acessou ou modificou os dados, garantindo a conformidade e a segurança.

6. Monitoramento e observabilidade

6.1. Cloud Logging

Utilizamos o Cloud Logging para registrar atividades e erros no pipeline, permitindo a detecção de problemas em tempo real.

6.2. Cloud Monitoring

O Cloud Monitoring é configurado para acompanhar a performance do pipeline, garantindo que os jobs sejam executados conforme esperado.

7. Conclusão

A arquitetura proposta para o pipeline de dados no Google Cloud Platform é robusta e escalável, permitindo a coleta, transformação e análise de dados de

vendas de forma eficiente. As camadas de dados (Bronze, Silver, Gold) organizam o fluxo de trabalho, e a utilização do GCS para armazenamento proporciona eficiência em custo e gestão de dados. Com o Cloud Composer, garantimos a orquestração do pipeline, enquanto monitoramento e segurança são implementados para assegurar a integridade dos dados.