Case Técnico: Pipeline de Dados com Arquitetura Medallion

Contexto

Você foi contratado para desenvolver uma solução completa de ingestão e processamento de dados para uma empresa fictícia que deseja consolidar informações provenientes de diversas fontes em um ambiente de Data Lakehouse na GCP. O objetivo é criar um pipeline de dados que siga os princípios da arquitetura Medallion, distribuindo os dados em três camadas:

- **Camada Bronze (Dados Brutos):** Recebe os dados em seu formato original ou extraídos de uma fonte externa.
- **Camada Silver (Dados Tabulares):** Processa e organiza os dados para torná-los estruturados e prontos para análises.
- **Camada Gold (Dados Agregados):** Consolida os dados, realizando agregações e criando novas colunas derivadas para facilitar insights de negócio.

Descrição Geral do Desafio

Desenvolva um pipeline de dados utilizando os serviços nativos da GCP, que inclua os seguintes componentes:

- **Google Cloud Storage (GCS):** Criação de buckets para armazenar os dados brutos provenientes da ingestão e dados transformados.
- **Cloud Functions:** Implementação de funções para processar os dados. Essas funções devem ser responsáveis por transformar os dados brutos em um formato tabular e realizar, se possível, agregações ou construções de novas colunas.
- **BigQuery:** Armazenamento dos dados processados, permitindo consultas analíticas e a geração de insights.
- **Orquestração:** Desenvolvimento de um script Python simples que dispare e orquestre as Cloud Functions, iniciando o pipeline de forma automatizada.

A ideia é que o pipeline extraia os dados de uma fonte pública (API e dados de livre escolha), armazene-os inicialmente no GCS e, a partir daí, utilize Cloud Functions para processar esses dados, enviando-os para o BigQuery para análise. A orquestração pode ser implementada por meio de um script Python que interaja com os endpoints das Cloud Functions (por exemplo, utilizando HTTP triggers ou a API do GCP).

Requisitos Técnicos

- **Plataforma:** A solução deve ser desenvolvida na GCP.
- **Armazenamento:** Criação e gerenciamento de buckets no GCS para manter os dados brutos.
- **Processamento:** Utilização de Cloud Functions para realizar a transformação dos dados. As funções devem:
    - Ler os dados armazenados no GCS;
    - Processar e converter os dados para um formato tabular com uso de python;
    - Realizar agregações ou construir colunas derivadas com uso de SQL.
- **Armazenamento Analítico:** Inserção dos dados processados no BigQuery para consultas e análises.
- **Orquestração:** Um script Python que inicie as Cloud Functions na ordem correta, garantindo o fluxo de dados do início ao fim do pipeline.

Entregáveis

- **Código Fonte:** Link para o repositório do GitHub contendo:
    - Scripts e funções (Cloud Functions) para a ingestão, processamento e envio dos dados.
    - O script Python de orquestração.
- **Infraestrutura:** Serviços criados.
- **Documentação:** README com instruções de execução e uma breve descrição da arquitetura proposta.
- **Diagramas:** Fluxogramas ou diagramas que ilustrem a arquitetura e o fluxo dos dados.

Observações

- O candidato terá um tempo limitado de 2 dias para desenvolver a solução. O foco é demonstrar a capacidade de arquitetar e desenvolver um pipeline utilizando os serviços da GCP.
- Sinta-se à vontade para fazer adaptações que julgar necessárias, desde que os pontos essenciais (GCS, Cloud Functions, BigQuery e orquestração via Python) sejam contemplados.