Case Engenheiro de dados

A empresa X foi contratada por um cliente para desenvolver uma solução de Business Intelligence que integrará dados de diversas fontes, processará esses dados e os disponibilizará para análise em Power BI. O cliente possui dados armazenados em um banco de dados SQL e em arquivos CSV, que são atualizados diariamente.

Desafio

Você deverá projetar, desenvolver e otimizar um pipeline de dados que:

1. Extraia dados atualizados diariamente de um banco de dados SQL e 1 arquivo CSV.
2. Realize as transformações necessárias nos dados.
3. Carregue os dados em um data warehouse preferencialmente em um ambiente cloud (BigQuery, Azure SQL Database, etc.).
4. Garanta que os dados estejam prontos e otimizados para análise no Power BI.
5. Implemente monitoramento e relatórios de qualidade dos dados.
6. Entregue a documentação necessária para entendimento e manutenção da solução a longo prazo

Obs. Use ferramentas que você tem hábito de usar e possui acesso gratuitamente, esse teste não deve gerar nenhum custo para ser implementado. Se necessário, faça tudo em sua máquina local, apenas mostre como seria a estrutura usando ambiente cloud.

Dados Fornecidos

1. `northwind.sql`: Dump inicial do banco de dados Northwind.
2. `order_details.csv`: Arquivo CSV com detalhes de pedidos.

Requisitos

1. **Extração de Dados**:
    - Configure a extração de dados do banco de dados Northwind a partir do dump `northwind.sql`.
    - Configure a extração de dados do arquivo `order_details.csv`.
    - Configure um processo para extrair dados atualizados diariamente do banco de dados e do arquivo CSV (considere que o arquivo CSV estará sempre no mesmo link).
2. **Transformação e Limpeza de Dados**:
    - Utilize ferramentas de ETL/ELT para transformar e limpar os dados.
    - Resolva problemas de inconsistência e qualidade dos dados.
3. **Carga de Dados**:
    - Carregue os dados transformados em um data warehouse na nuvem (BigQuery, Azure SQL Database, etc.).

    - Modele os dados de forma a otimizar o desempenho para consultas no Power BI.
4. **Integração com Power BI**:
    - Configure a conexão do Power BI com o data warehouse escolhido.
    - Crie um conjunto básico de relatórios e dashboards no Power BI para demonstrar a viabilidade da solução.
5. **Monitoramento e Qualidade de Dados**:
    - Implemente um sistema de monitoramento para acompanhar a performance do pipeline.
    - Gere relatórios de qualidade dos dados e identifique possíveis melhorias.
    - Inclua formas de corrigir erros no pipeline, garantindo a estabilidade da solução.
6. **Documentação e Colaboração**:
    - Documente todo o processo, incluindo diagramas de arquitetura, scripts utilizados e instruções para replicar o ambiente.
    - Demonstre habilidades de comunicação ao colaborar com uma equipe fictícia, fornecendo atualizações regulares sobre o progresso do projeto.

Entregáveis

1. Código-fonte do pipeline de dados, transformação e carregamento de dados.
2. Diagrama e documentação detalhada do processo.
3. Relatórios de monitoramento e qualidade dos dados.
4. Relatórios e dashboards no Power BI (apenas para validar dados).
5. Cases com entrega adequada serão convidados a fazer uma apresentação da solução, destacando desafios encontrados, soluções implementadas e possíveis melhorias.

Orientações

Use as ferramentas que você tem confiança nas suas habilidades. Priorize uma solução robusta e bem estruturada do que uma solução complexa e potencialmente instável.

Monte uma documentação que seja o suficiente para você explicar as suas decisões e que seja visualmente fácil de entender o caminho dos dados. Recomendado montar também um dicionário de dados.

Relate os desafios encontrados e o motivo das suas escolhas.