## Case GB - Arquitetura


1. **Camada de Ingestão:**

- Utilização do Cloud Storage como um data lake para receber dados brutos de diferentes fontes, em diferentes formatos.
- Utilização do DataFlow para a ingestão de dados, que permite a execução de pipelines de ETL em escala, com alta disponibilidade e escalabilidade.
- Utilização do Pub/Sub para receber e distribuir eventos em tempo real.

A escolha de utilizar ferramentas de integração de dados como DataFlow e Pub/Sub para a ingestão de dados é importante para garantir que os dados sejam coletados em tempo real. Essas ferramentas são altamente escaláveis e podem lidar com grandes volumes de dados de diversas fontes.

2. **Camada de Processamento:**

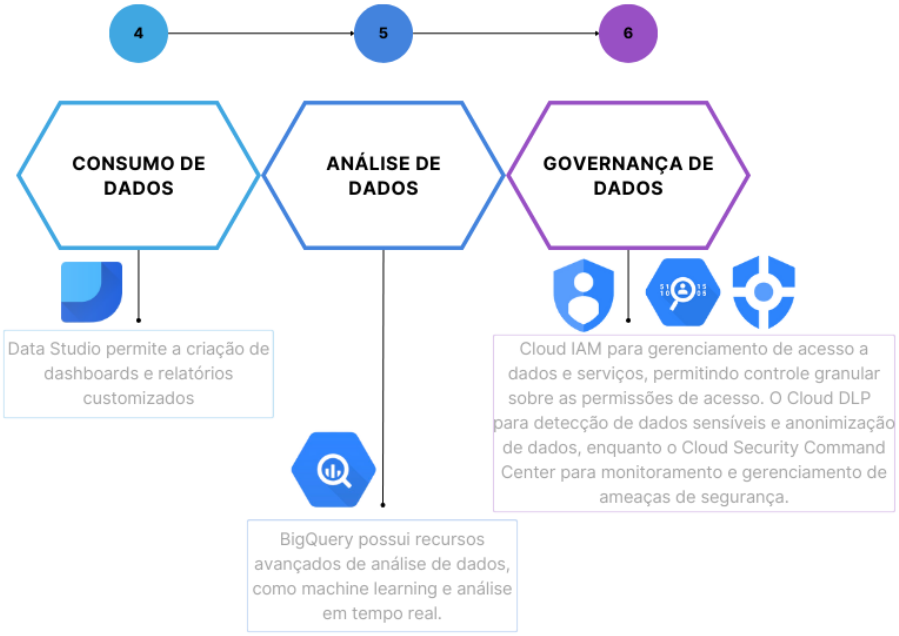
- Utilização do BigQuery como data warehouse, que oferece uma estrutura de armazenamento escalável e flexível para armazenamento e análise de dados. Além disso, o BigQuery é

capaz de lidar com cargas de trabalho analíticas complexas e executar consultas em tempo real.
- Utilização do Cloud Functions para execução de funções de processamento de dados em tempo real.
- Utilização do DataPLex para automação de processos de ETL e extração de dados do SAP Hana.

O processamento de dados é uma etapa crítica na arquitetura de dados, pois é responsável por transformar e limpar dados brutos em informações úteis para análise e consumo. Nesse sentido, a escolha de ferramentas eficientes é fundamental.
O DataFlow é uma ferramenta de processamento de dados em tempo real que permite a criação de pipelines de processamento escaláveis e com alta disponibilidade. Ela é adequada para lidar com grandes volumes de dados e permite a execução de transformações complexas. Além disso, sua integração com outras ferramentas do GCP, como o Pub/Sub, facilita a ingestão de dados em tempo real.
O Storage Procedure é uma ferramenta que permite a criação de rotinas de otimização de consultas no BigQuery, melhorando o desempenho e reduzindo o custo de processamento. Já o Cortex é uma solução de otimização de consultas que utiliza técnicas de machine learning para identificar e corrigir gargalos de desempenho.
O Cloud Functions é uma ferramenta que permite a execução de funções em tempo real, sem a necessidade de gerenciar infraestrutura. Isso é particularmente útil para processamento de eventos em tempo real, como monitoramento de transações e detecção de fraudes.
As vantagens dessas ferramentas incluem a capacidade de lidar com grandes volumes de dados em tempo real, a facilidade de criação de rotinas de otimização e a escalabilidade. No entanto, é importante ressaltar que a complexidade de configuração e gerenciamento dessas ferramentas pode ser um desafio, exigindo uma equipe técnica especializada. Além disso, é necessário monitorar cuidadosamente o custo do processamento, para evitar gastos desnecessários.

3. **Camada de Armazenamento:**
- Utilização do Cloud Storage para armazenamento de dados brutos e processados.
- Utilização do BigQuery para armazenamento de dados estruturados.

A escolha do BigQuery como armazenamento de dados é vantajosa, pois ele é uma solução totalmente gerenciada e escalável, ou seja, não requer a gestão de infraestrutura, permitindo que a equipe foque na análise de dados. O BigQuery possui também uma arquitetura de colunas, o que permite consultas rápidas e eficientes, mesmo em grandes volumes de dados. Além disso, o conceito de LakeHouse permite a inclusão de dados brutos e transformados em um mesmo ambiente, o que permite uma maior agilidade e flexibilidade na utilização dos dados.




4. Camada de Consumo:

- Utilização do Tableau para visualização de dados e análise de dados.
- Utilização do Jupyter Notebook para análise e experimentação de dados.
- Utilização do Qlik Sense para análise de dados e criação de painéis.

Para o consumo de dados, a arquitetura escolheu ferramentas que permitem a análise e a disponibilização dos dados para diferentes usuários e fins. O Tableau é uma ferramenta amplamente utilizada para criação de dashboards e relatórios personalizados, permitindo que diferentes usuários da empresa possam analisar os dados de acordo com suas necessidades.
O Amundsen também pode ser usado, por sua vez, é uma ferramenta de catalogação de dados que permite a organização dos dados para fácil acesso, permitindo aos usuários localizar e utilizar os dados de forma mais eficiente.
A criação de APIs e serviços para integração com outras aplicações também é importante para permitir o acesso aos dados de diferentes formas. Essas APIs e serviços podem ser utilizados para integração com outras ferramentas de análise de dados ou para fornecer acesso aos dados para outras aplicações internas da empresa.
Uma vantagem do uso do Tableau é a sua facilidade de uso e de criação de dashboards, permitindo que diferentes usuários da empresa possam criar relatórios personalizados de acordo com suas necessidades sem a necessidade de conhecimento técnico avançado. Já o Amundsen permite a organização e a catalogação dos dados, o que facilita a sua localização e utilização pelos usuários.
Uma desvantagem pode ser o custo associado ao licenciamento das ferramentas de análise, como o Tableau. Além disso, a criação de APIs e serviços pode demandar tempo e recursos adicionais para desenvolvimento e manutenção.

5. Camada de Análise de dados:
- A arquitetura utilizará o BigQuery, que possui recursos avançados de análise de dados, como machine learning e análise em tempo real.

O GCP possui outras ferramentas de análise de dados que podem ser utilizadas em conjunto com o BigQuery, como o AI Platform e o Data Studio. O AI Platform permite a criação de modelos de machine learning para análise preditiva e classificação de dados, enquanto o Data Studio é uma ferramenta de visualização de dados que permite a criação de dashboards interativos e relatórios personalizados. A utilização dessas ferramentas em conjunto com o BigQuery permite uma análise mais completa e aprofundada dos dados, possibilitando insights valiosos para o negócio.

6. Camada de Segurança e Governança:
- Utilização do Google Cloud IAM para gerenciamento de acesso a recursos.
- Utilização do Cloud Data Loss Prevention para proteção de dados sensíveis.
- Utilização do Cloud Security Command Center para monitoramento de segurança e conformidade.
- Utilização do Amundsen para catalogação e descoberta de dados.
- Utilização do Alvin para linhagem de dados.

A arquitetura também inclui a definição de padrões de nomenclatura e documentação dos dados, com o objetivo de manter a qualidade e a consistência dos dados armazenados. A utilização do Google Cloud Data Loss Prevention (DLP) também será uma medida importante para garantir a proteção de dados sensíveis.
As vantagens da implementação de uma governança de dados efetiva incluem a melhoria da qualidade dos dados, o aumento da confiança nas informações utilizadas nas tomadas de decisão, a garantia de conformidade com regulamentações e políticas internas, além da redução do risco de vazamento de dados sensíveis.
No entanto, a implementação de governança de dados também apresenta desvantagens e riscos, como o aumento da complexidade da arquitetura, o tempo e o custo necessários para a implementação, além da necessidade de mudanças culturais dentro da empresa para garantir a adesão às políticas e padrões definidos.


7. Integração com Aplicações On-Premises:

- Utilização do SLT como uma ferramenta de ETL do SAP para realizar a replicação de dados do SAP Hana para o BigQuery.
- Utilização do BigQuery Connector e Cortex para integração do BigQuery com outras fontes de dados.
- Utilização do TDD e CICD para orquestração de processos de integração de dados e deploy de aplicações.

A comunicação entre ambientes on-premises e a nuvem do GCP é essencial para garantir a integração e consistência dos dados em toda a empresa. Nesse sentido, a utilização de ferramentas de orquestração como o Airflow Compose permite a criação de fluxos de dados automatizados, garantindo que os dados sejam transferidos de forma rápida e confiável.
Além disso, a utilização do SLT como uma ferramenta de ETL do SAP para realizar a replicação de dados do SAP Hana para o BigQuery é uma escolha inteligente, pois permite que os dados de um dos principais sistemas da empresa sejam integrados à nuvem do GCP de forma segura e confiável. O uso do BigQuery Connector e Cortex para integração do BigQuery com outras fontes de dados também é importante, pois permite que os dados sejam integrados de várias fontes, garantindo que as informações sejam centralizadas em um único ambiente.
No entanto, a comunicação on-premises x cloud também apresenta alguns riscos, como possíveis interrupções no serviço devido a falhas na conexão ou segurança de dados comprometida durante a transferência. Por isso, é importante estabelecer políticas de segurança e controle de acesso, utilizando o TDD e CICD, para minimizar esses riscos.

Com essa arquitetura de referência, a empresa será capaz de integrar seus dados de diferentes fontes e transformá-los em informações úteis para diferentes fins, tais como Analytics, Data Science, API's e serviços para integrações com aplicações. A substituição gradativa do cenário on-premises atual para a nuvem pública do GCP irá trazer escalabilidade, disponibilidade e segurança para o gerenciamento de dados da empresa. Além disso, a arquitetura permitirá a empresa analisar dados em tempo real, permitindo uma tomada de decisão mais rápida e precisa.