Grupo Boticário

Case de arquitetura

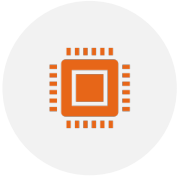
Introdução

- Permear as camadas de ingestão, processamento, armazenamento, consumo, análise, segurança e governança;
- Substituir gradativamente o cenário *on-premises* atual;
- Incorporar componentes e tecnologias que permitam a analisarmos dados em tempo real;
- Organizar e fornecer dados para diferentes fins, tais como: *Analytics*, *Data Science*, APIs e serviços para integrações com aplicações. Ressaltando que necessariamente precisaremos manter a comunicação *on-premises* x *cloud* para diversas finalidades.

Cenário atual

SAP Hana é o repositório principal de *data warehouse*;

Existem processos de ETL que fazem ingestão de dados de 50 bases transacionais;



Mais de 90% das bases são de origem transacionais de diferentes DBMSs e estão alocados em ambiente *on-premises*;

A empresa também possui algumas aplicações hospedadas em nuvens públicas como **Microsoft Azure** e **Amazon AWS**;

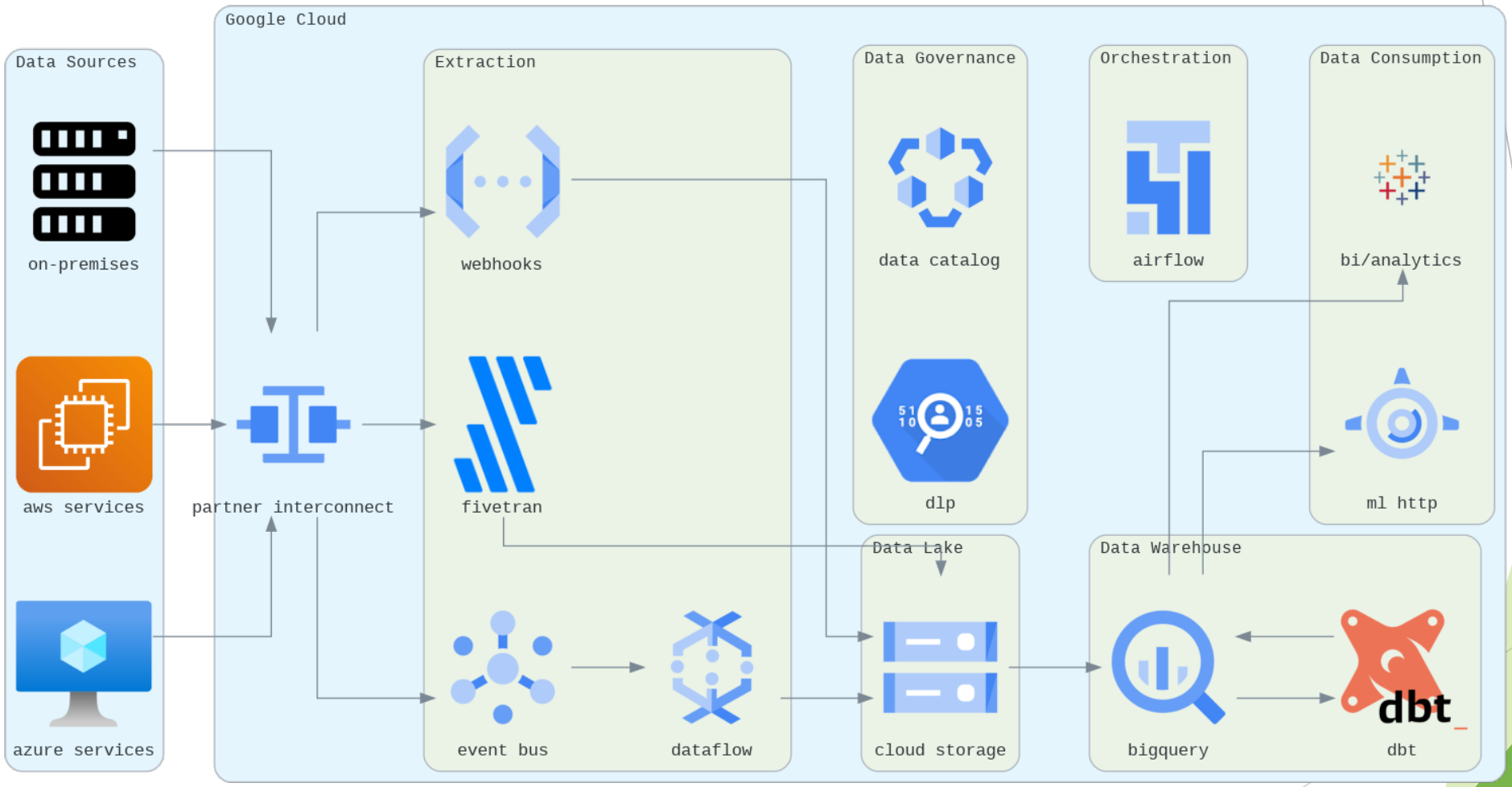
Diferentes BUs acabam utilizando diferentes ferramentas para processar, analisar e apresentar dados; e

Relativo à **governança de dados**, aspectos como acesso a dados sensíveis, catalogação e permissionamento carecem de melhorias.

Proposta

- Configuração do Cloud Interconnect para tráfego de dados entre a GCP e diferentes CSPs;
- Movimentação de dados SAP com o Fivetran;
- Migração gradual do SAP BW/4HANA (*data warehouse* atual) para o Google BigQuery (*on-demand*);
- Validação de *reports* a partir da solução com Google BigQuery;
- Entendimento e categorização dos dados com Dataplex e Cloud DLP;
- Criação de grupos de acesso no Dataplex;
- Desenvolvimento/orquestração dos fluxos no Cloud Composer; e
- Organizar diferentes estratégias de implantação dos modelos de DS.

Arquitetura final



Grupo Boticário: Architectural Diagram

Trade-offs

- Cloud Interconnect: conexão dedicada e baixa latência/custo superior à Cloud VPN
- Fivetran: versatilidade e quantidade de conectores *batch* e *streaming*/acúmulo de responsabilidades
- Dataflow: *engine* única para processos *batch* e *stream*, e suporta testes locais/BeamSQL ainda não é *production-ready*;
- Cloud BigQuery: produto gerenciado e permite *streaming inserts*/pode gerar altos custos para analisar tabelas históricas;
- Dataplex: integrado ao Google BigQuery para acesso granular/customização limitada comparado a alternativas, como Amundsen;
- Cloud DLP: integração a diferentes fontes de dados da GCP/???; e
- DBT: permite adotar boas práticas para desenvolvimento de transformações SQL/não permite *streaming*, e possui suporte limitado para Python.

Referências

- https://cloud.google.com/architecture/patterns-for-connecting-other-csps-with-gcp
- https://www.sap.com/products/technology-platform/bw4hana-data-warehousing.html
- https://blogs.sap.com/2019/05/15/sap-hana-data-warehousing-for-non-experts/
- https://fivetran.com/docs/databases/sap-erp
- https://www.linkedin.com/pulse/how-replace-sap-bw-google-bigquery-barry-kelly/
- https://cloud.google.com/blog/products/data-analytics/bigquery-connector-for-sap
- https://polleyg.dev/posts/data-engineering-tips/
- https://cloud.google.com/dataplex/docs/introduction