

Arquitetura Medalhão

DATA LAKEHOUSE

A **arquitetura medalhão** é um padrão de design de dados usado para organizar logicamente os dados do lakehouse, que visa melhorar de forma incremental e progressiva a estrutura e a qualidade dos dados à medida que fluem pelas três camadas da arquitetura (tabelas Bronze ⇒ Prata ⇒ Ouro).

As arquiteturas medalhão também são conhecidas como arquitetura “multi-hop”.



Arquitetura Medalhão

LANDING

- É uma camada provisória, responsável pela primeira ingestão dos dados no data lake.
- Formato original dos arquivos em seus sistema de origens (CSV, JSON, XML, etc).
- Geralmente não guarda histórico (mas pode ser utilizada como histórico).
- Normalmente usada como backup para recriar um conjunto de dados.
- Ingestão full ou incremental.
- Utilizada apenas por engenheiros de dados.

Arquitetura Medalhão

BRONZE

- Cópia da camada anterior (landing).
- Mantem histórico completo dos dados brutos (ainda não processados).
- Formato delta (ACID, metadados, etc).
- Dados imutáveis (apenas leitura).
- Normalmente particionados por data (ano, mês, dia).
- Podem incluir metadados adicionais (colunas), como nomes de arquivos de origem, data e hora em que os dados foram processados, etc.
- Utilizada apenas por engenheiros de dados.

Arquitetura Medalhão

SILVER

- Formato delta (ACID, metadados, etc).
- Geralmente alinhados e organizados com o sistema de origem.
- Padronização e regras de qualidade (nomenclatura dos campos - ex: inglês para português, minúsculo x maiúsculo, retirada de abreviações - cd_cliente = CODIGO_CLIENTE).
- Remoção de dados duplicados, trata dados ausentes, padroniza vazios ou inválidos.
- Aqui pode ocorrer pequenas agregações* e união de tabelas do mesmo assunto.
- Incremental. Normalmente usa técnicas de SCD 2.
- Utilizada por engenheiros de dados e engenheiros de ML e IA (consultas Ad Hoc).
- Podem incluir metadados adicionais (colunas), como nomes de arquivos de origem, data e hora em que os dados foram processados, etc.

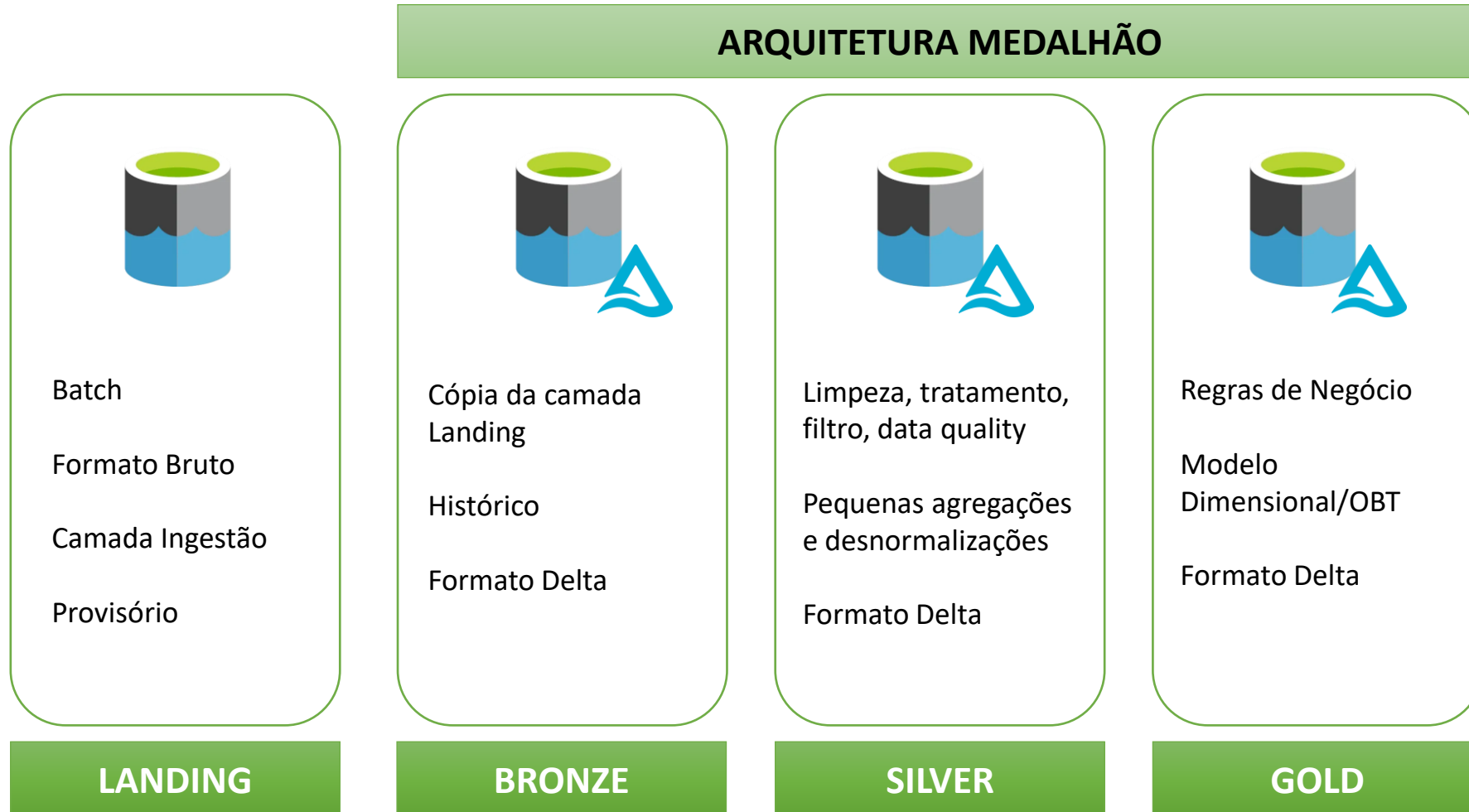
Arquitetura Medalhão

GOLD

- Formato delta (ACID, metadados, etc).
- Dados transformados de acordo com as regras de negócio para consumo ou casos de uso.
- Modelo de dados otimizado para leitura (Dimensional, OBT, etc).
- Controle de dados históricos são aplicados apenas para o conjunto de dados dos casos de uso de acordo com suas regras, tendo em mente que esta camada pode ser uma seleção ou agregação de dados encontrados na camada Prata.
- Os dados são altamente governados e bem documentados.

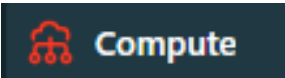
Arquitetura Medalhão

DATA LAKEHOUSE



Atividade Prática

ARQUITETURA MEDALHAO COM DATABRICKS COMMUNITY E AZURE DATA LAKE STORAGE

1. Logar no Databricks Community - <https://community.cloud.databricks.com/>
2. Criar um cluster no Databricks. 
3. Criar um Azure Data Lake Storage Gen2 através do Terraform e Azure CLI
<https://github.com/jlsilva01/adls-azure/>
4. Configurar o Databricks para acessar o ADLS criado (SAS ou Access Key).
5. Importar os arquivos CSV no landing e movê-los para o bronze, silver e gold no modelo dimensional (já feito em aula).

