Vantagens da Arquitetura Medalhão

Gosto muito do modelo medalhão, é o que eu gosto de trabalhar.

Bronze: Onde é armazenado os dados de origem externa, não sofrem alterações e correspondem ao formato e estrutura de onde vieram. Focando na captura do dado e o fornecimento de um histórico e reprocessamento de dados se necessário.

Silver: Derivada da camada Bronze, os dados são correspondidos, mesclados, conformados e limpos para que possa ter uma visão mais personalizada para a empresa, mas sem trabalhar tanto os dados ainda. Quase sempre já reunindo dados de diferentes fontes e fornecendo estrutura para análises e até mesmo ML.

Geralmente serve para outras pessoas usarem como referência se quiserem fazer uma análise personalizada.

Gold: Geralmente direcionadas a um projeto específico e prontas para consumo. Onde se aplicam transformações e regras nos dados para cada necessidade de um projeto.

As vantagens dessa arquitetura:

Modelo de dados simples

Fácil de entender e implementar

Habilita ETL incremental

Pode recriar suas tabelas a partir de dados brutos a qualquer momento

Transações ACID, viagem no tempo

Controle de Duplicação e Falhas

Checkpointing: salva o ponto em que o streaming parou. Se der falha, ele volta desse ponto, sem repetir tudo.

Watermark: controla janelas de tempo para lidar com eventos atrasados, evitando reprocessar o mesmo dado.

Deduplicação: na camada Silver, usamos o ID da transação ou cliente para garantir que cada registro apareça só uma vez.

MERGE/UPSERT: quando um registro muda (ex.: status de transação), usamos merge para atualizar sem duplicar.

Jobs com parâmetros: passamos `execution_date` e `refresh_rate` nos jobs, assim é possível repetir execuções ou ajustar a frequência.

3. Delta vs Iceberg vs Parquet

Parquet: É apenas um formato, sem “framework”. Bom para armazenagem de dados frios. Ou análises pequenas. Não tem, evolução de esquema, time travel e ACID.

Delta Lake: Muito otimizado para Spark, ideal para ETL Spark e arquitetura medalhão, utiliza como base Parquet e inclui metadados. Já tem ACID, time travel e evolução parcial de esquema.

Iceberg: Projetado para múltiplas plataformas, utiliza como base Parquet e inclui metadados. Também tem ACID, time travel e evolução robusta de esquema.