# Estruturas de Dados eficazes em Data Lakehouses
## Uma opinião pessoal

Aviso Legal – todas as opiniões aqui emitidas são de cunho pessoal com base na experiência adquirida, e, portanto, sujeitas e divergências de opinião. Todos os comentários são sempre bem-vindos.

Resumo

Muito das estruturas de dados usadas no Data Lakehouses deriva em parte das teorias de Ralph Kimball que se utiliza a snapshots e fact tables. Dentro de padrões de desempenho os snapshots são uma alternativa que cria volume excessivo de dados, incorrendo em maiores tempos de processamento e custos.

O uso de modelos canônicos ao contrário, simplifica e reduz os volumes de dados aumentando as velocidades de processamento em mais de 45%. Nesse artigo eu descrevo as vantagens de uso dos modelos canônicos em comparação às visões snapshot.

A arquitetura de Data Lakehouse (aberta)

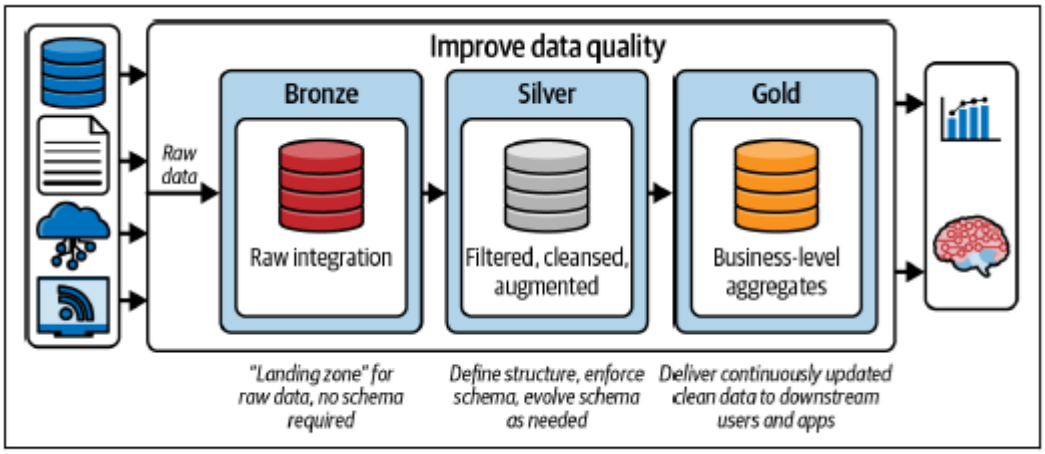



*Figure 1-8. Data lakehouse solution architecture*

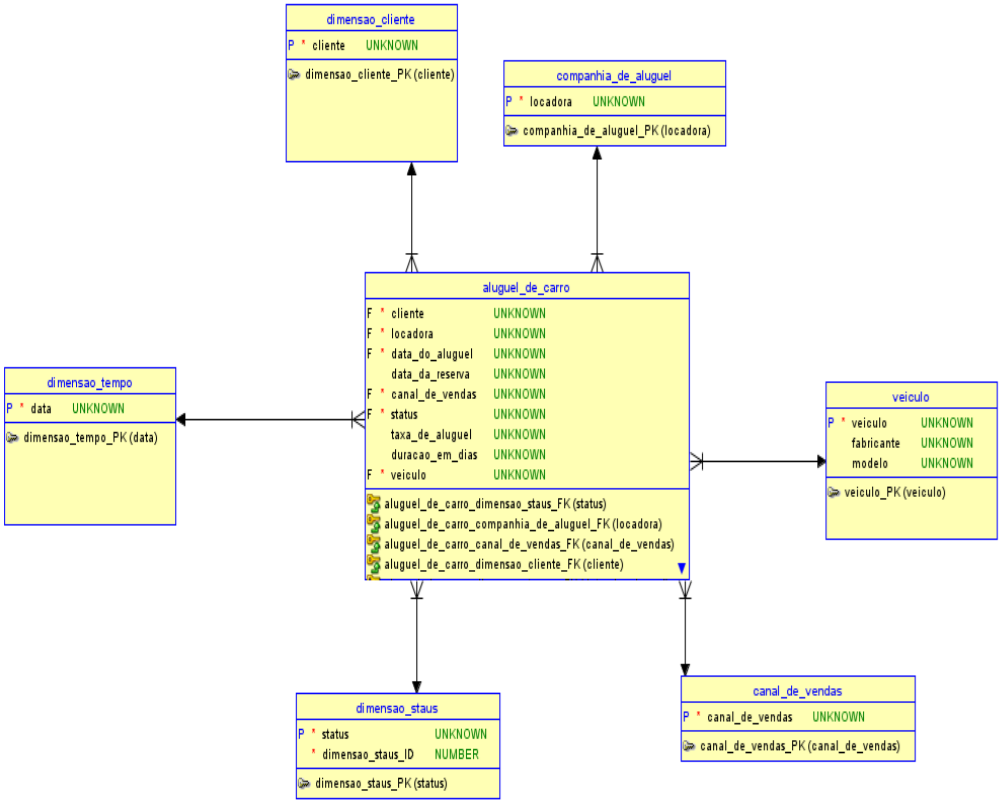
© Delta Lake Up and Running by Bennie Haelen and Dan Davis

Trata-se de uma arquitetura de três camadas, em diversos níveis de agregação usando essencialmente o formato PARQUET; os dados provêm de diversas origens, são capturados na camada RAW, convertidos numa segunda camada e

finalmente expostos como agregados de negócios (KPIs). Se a proposta de Ralph Kimball for seguida à risca teremos os dois seguintes níveis de informação a partir de tabelas brutas.

1. Snapshots(Camada Prata)

Tabelas cumulativas que carregam os valores das transações individuais de cada objeto de dados. Em suma, os snapshots são visões resumidas dos registros operacionais e carregam consigo um alto volume de registros.

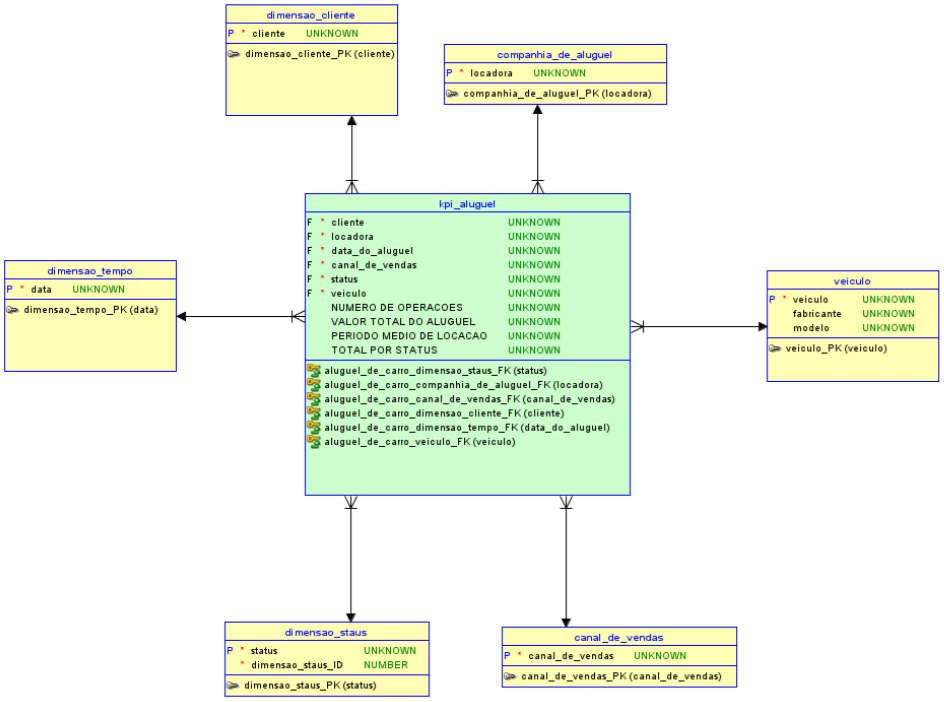



Esta estrutura, no entanto, apresenta problemas, já que deriva de uma tabela de dados atômicos, o que é uma repetição dos dados dos sistemas comerciais, antes da camada de dados brutos.

Outro ponto de atenção se refere ao fato de que muitas tabelas sem fato são repetidas no modelo (em todo ou em parte), o que implica em inconsistência de dados, justo o que tentamos evitar.

2. Tabelas fato e KPIs

Usam as mesmas estruturas dos snapshots, mas aqui ao invés de valores brutos temos os KPIs que são os valores calculados em função de agregação de valores das operações.



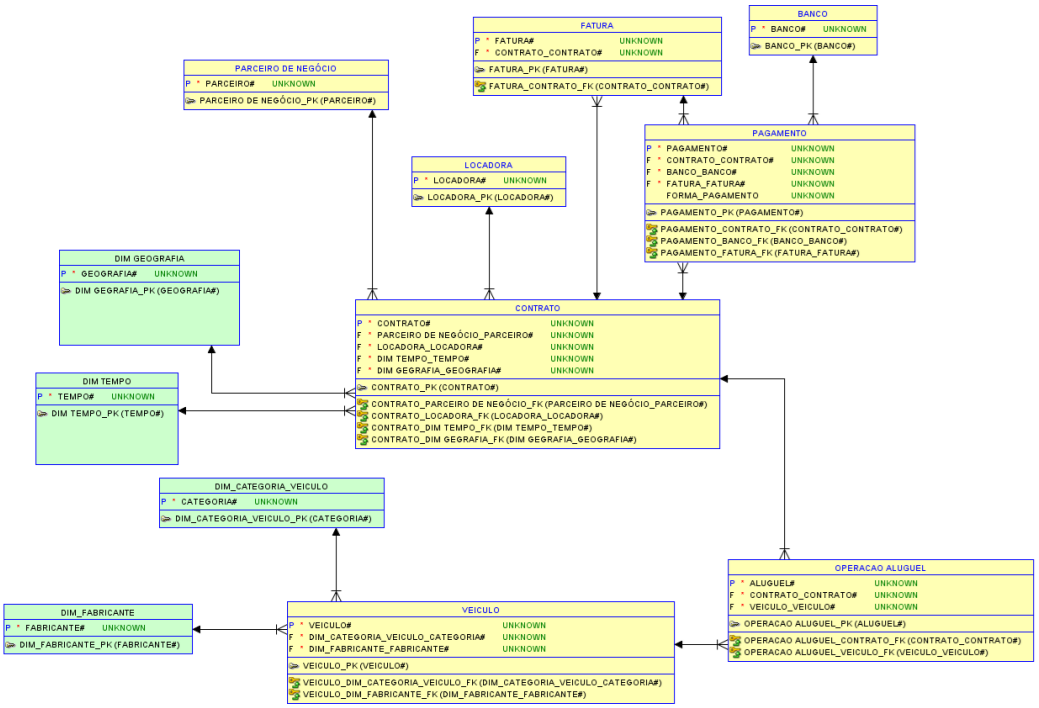

Aqui temos as dimensões e os fatos cumulativos. Esse é o tipo de dado que vai para o nível OURO de um Data Lakehouse.

Nas tabelas-fato incluímos todos os indicadores relevantes ao modelo de decisão e por uma questão de maior eficácia pode conter um ou muitos indicadores pelas mesmas dimensões.

Uma coisa é certa: caso não consigamos um conjunto de snapshots únicos, a consistência dos dados fica sacrificada, por implicar num novo processo de harmonização e consistência aumentando o consumo de recursos de processamento.

3. Modelos Canônicos

Diferente dos snapshots, um modelo canônico não possui entidades repetidas num modelo de dados; cada modelo representa as entidades de negócio de forma inequívoca, de tal maneira que os dados representam plenamente o conceito de “visão única da verdade”.




*Um exemplo de modelo canônico de uma locadora de automóveis, onde a unidade principal é o contrato de locação (mensal ou por período).*

Um modelo canônico armazena as transações unificadas de todos os sistemas como um conjunto de entidades, permitindo infinitas combinações de dados e extrações de tabelas-fato sem a necessidade de extrações complexas, já que a fonte de dados é única.

Nos snapshots, precisamos antes ter a certeza de que não há duplicidades ou inconsistências, enquanto nos modelos canônicos todo o processo de validação já foi realizado e temos apenas uma fonte de dados.

Conclusão

Devemos orientar nossos modelos à criação de um modelo canônico na camada SILVER, e fazer deste tanto a base de importação de dados brutos como a base de criação das tabelas-fato que vão conter um ou mais indicadores. A ideia do modelo único é a de trazer consistência e ao mesmo tempo flexibilidade aos Data Lakehouses, reduzindo o número de processos envolvidos e assegurando maior qualidade da informação.