
Newton
Quem se prepara, não para.

Business Intelligence

4º período

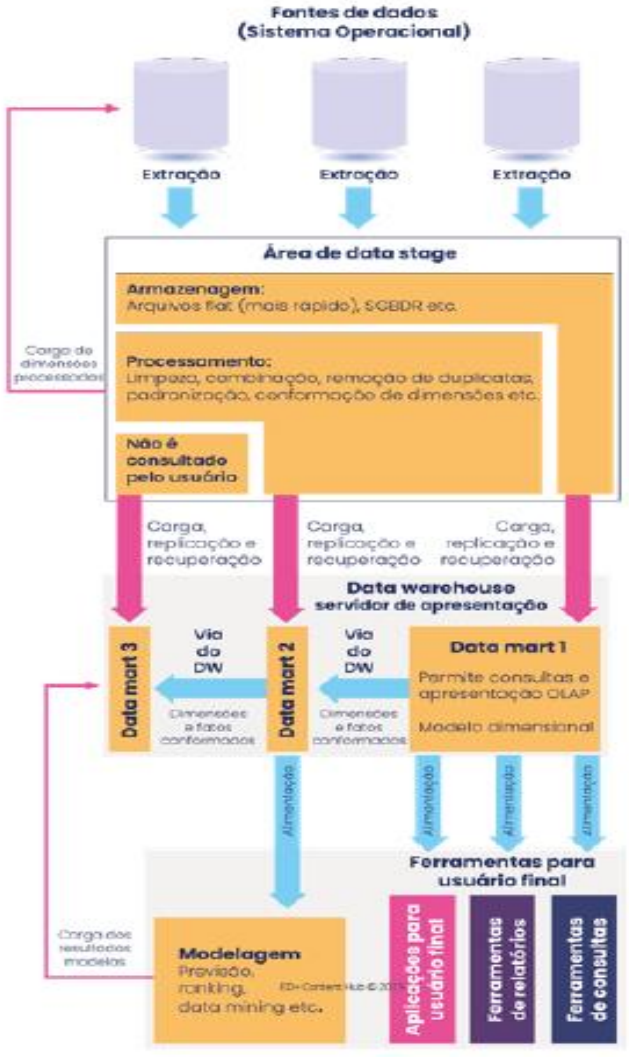
Professora: Michelle Hanne

Newton
Quem se prepara, não para.

Arquitetura de Business Intelligence

Área de Stage

A área de stage ou stage area é um armazenamento temporário de dados do data warehouse. Como também é uma área de armazenamento que será usada por um conjunto de processos comumente referido como ETL (Extract - Transformation - Load). A área de stage está entre os Sistemas Operacionais de origem (fontes dos dados) e a área de apresentação de dados.



Área ODS (Operational Data Store)

Newton
Quem se prepara, não para.

A área ODS é onde os dados são organizados, armazenados e disponibilizados para **consultas diretas por usuários**, desenvolvedores de relatórios e outros aplicativos analíticos. Representa um armazenamento intermediário dos dados antes de sua atualização no data warehouse, ou seja, o ODS é um repositório que armazena apenas as informações correntes, antes de serem carregadas para o DW.

Normalmente um ODS é implementado quando existe a necessidade de analisar informações do dia a dia.

Área ODS (Operational Data Store)

Newton
Quem se prepara, não para.

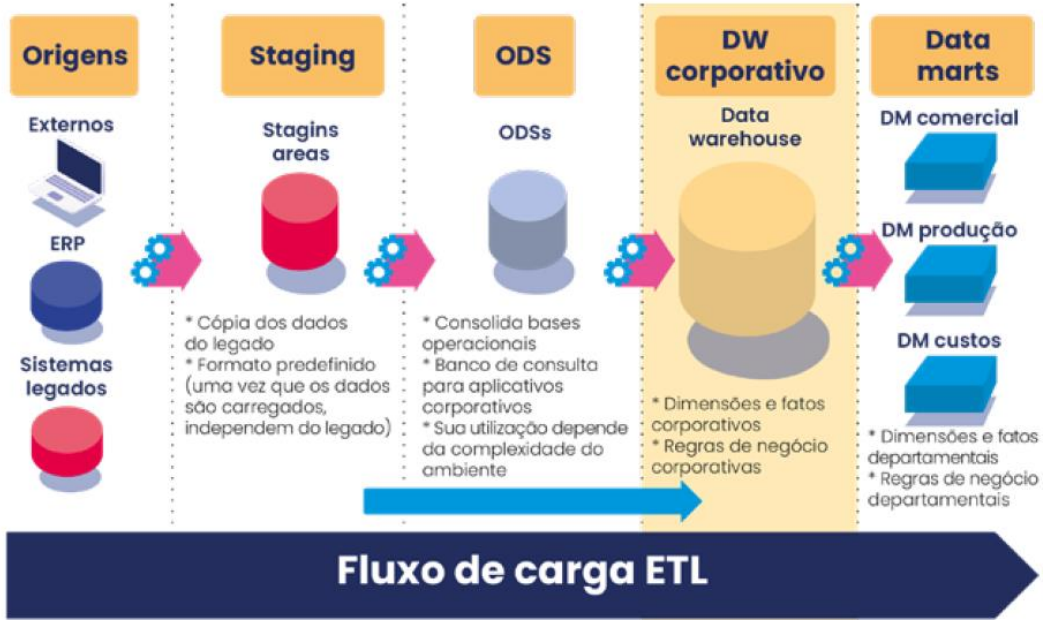
- Em geral, um ODS é **implementado para fornecer relatórios operacionais.**
- Esses relatórios são caracterizados por um conjunto limitado de consultas fixas que podem ser conectadas por ferramentas de exportação.
- Os relatórios abordam os **requisitos mais táticos de tomada de decisão da organização.**
- **Em outros casos, os ODSs são criados para suportar interações em tempo real, especialmente em aplicativos de Gerenciamento de Relacionamento com Clientes (CRM),** como acessar o itinerário da sua viagem ou em um site (ou histórico de serviços) quando se liga para o suporte ao cliente.

Data Wharehouse

Newton
Quem se prepara, não para.

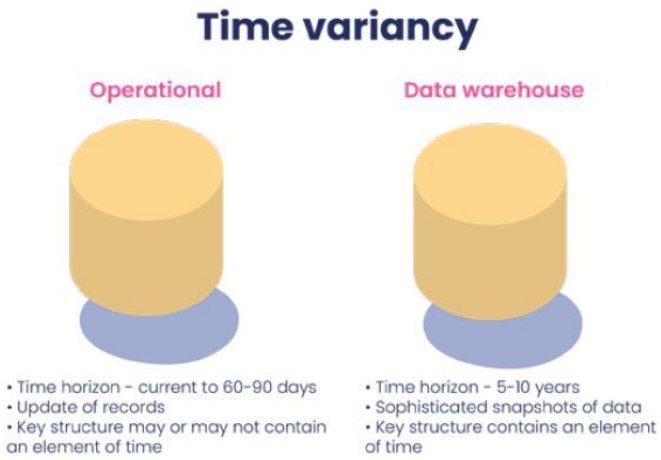
O Data Warehouse é o “coração” do ambiente de BI e é a base de todo o processamento do sistema informacional.

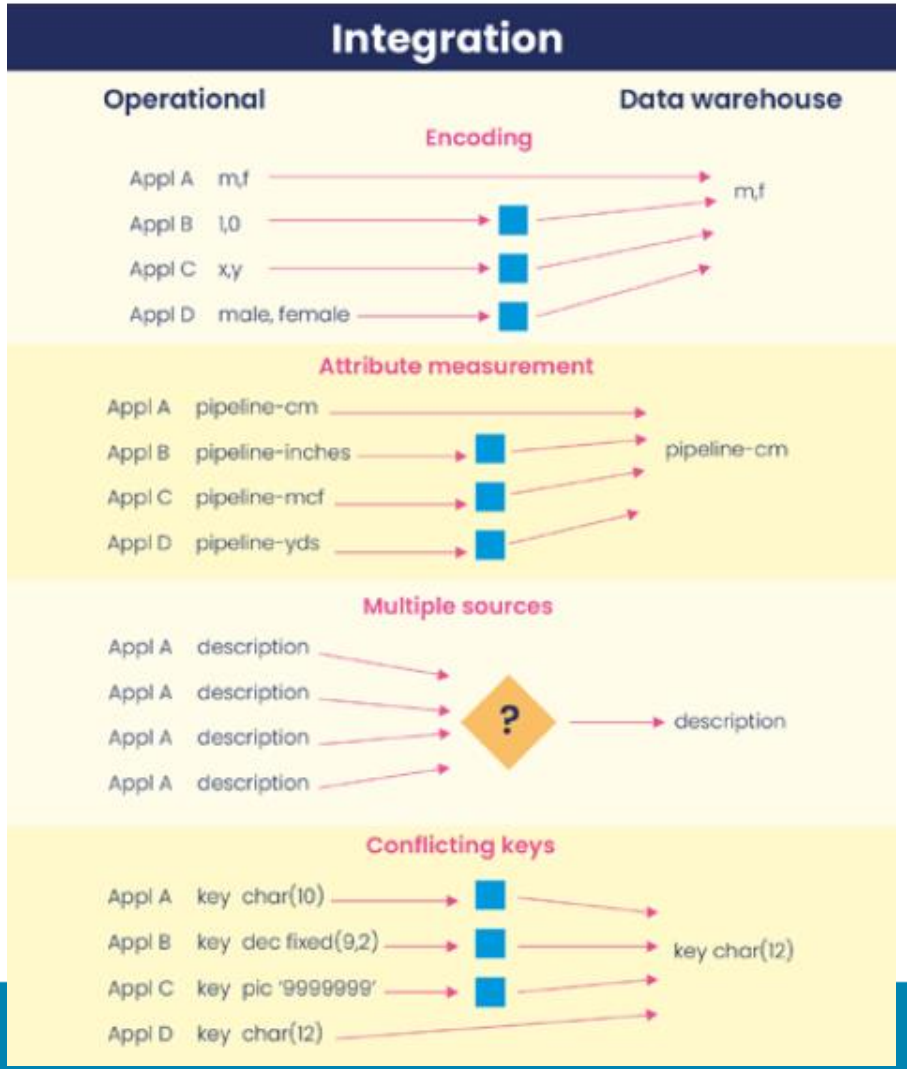
Um Data Warehouse é uma coleta de dados orientada ao assunto, integrada, não volátil e com variação de tempo para apoiar as decisões da gerência.



Data Wharehouse

A figura ilustra a integração que ocorre quando os dados passam do ambiente operacional orientado a aplicativos para o Data Warehouse.





Data Mart (Repositório de Dados)

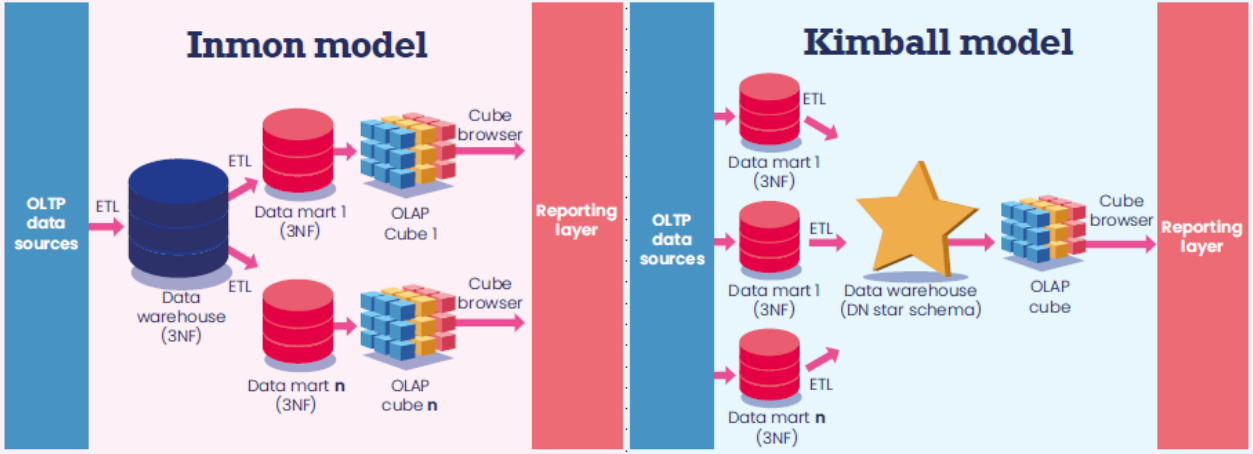
Newton
Quem se prepara, não para.

A diferença entre um **Data Mart (DM)** e um **Data Warehouse (DW)** está na relação de tamanho e escopo. Enquanto um Data Mart trata de um problema departamental ou local, um Data Warehouse envolve o esforço de toda a empresa para que o suporte às decisões atue em todos os níveis da organização.

O desenvolvimento de um Data Warehouse requer tempo, dados e investimentos gerenciais muito maiores que um Data Mart. O Data Mart é um subconjunto de dados de um DW (armazém de dados). Geralmente são dados **referentes a um assunto em especial** (exemplos: vendas, estoque, Controladoria**) ou diferentes níveis de sumarização** (exemplos: vendas por ano, vendas por mês, vendas a cada cinco anos), que focalizam uma ou mais áreas específicas.

Inmon *vs.* Kimball

Dois pioneiros em data warehouse, Bill Inmon e Ralph Kimball, diferem em suas visões sobre como os data warehouses devem ser projetados da perspectiva da organização. Veja a seguir.



Vantagens

- Herança de arquitetura: todos os DMs originados de um DW utilizam a arquitetura e os dados do DW;
- Visão de empreendimento: o DW concentra todos os negócios da empresa;
- Controle e centralização de regras: garante a existência de um único conjunto de aplicações ETL.

Desvantagens

- Maior tempo para o desenvolvimento;
- Custo elevado.

Vantagens

- Implementação rápida;
- Retorno rápido;
- Enfoque inicial nos principais negócios da empresa.

Desvantagens

- Perigo de DMs legados: um dos maiores perigos no DW é a criação de data marts independentes;
- Dificuldade em obter a visão do empreendimento;
- Coordenação de múltiplas equipes e iniciativas;
- Procedimentos de ETL mais complexos.

Newton
Quem se prepara, não para.

Bill Inmon e **Ralph Kimball** são dois precursores em data warehouse. Eles têm abordagens diferentes.

Data Mart

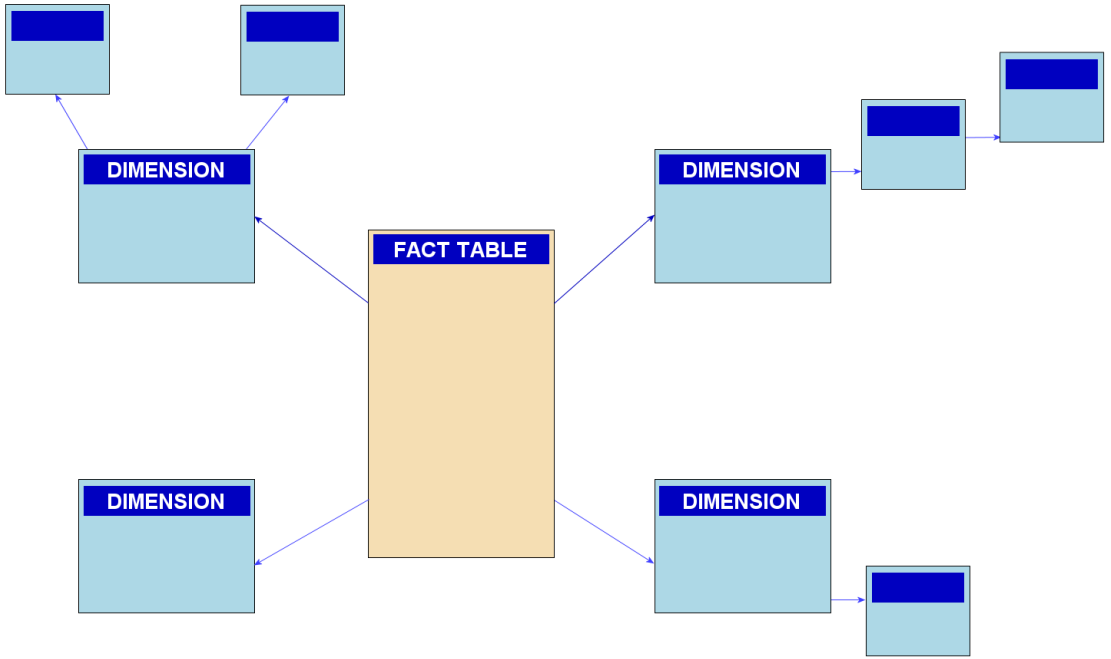
Newton
Quem se prepara, não para.

Bill Inmon favorece um **design de cima para baixo**, no qual o Data Warehouse é o repositório centralizado de dados e o componente mais importante dos sistemas de dados de uma organização. Primeiro cria o modelo de dados corporativo centralizado e o Data Warehouse é visto como a representação física desse modelo. Data Marts dimensionais relacionados a linhas de negócios específicas podem ser criados a partir do Data Warehouse quando necessário.

A abordagem de **Ralph Kimball** começa com os processos de negócios mais importantes. Nessa abordagem, uma organização **cria Data Marts** que agregam dados relevantes em torno de áreas específicas do assunto. O **Data Warehouse** é a **combinação dos data marts individuais da organização**. Com a abordagem Kimball, o Data Warehouse é o conglomerado de vários Data Marts.

Modelo SnowFlake

É um arranjo lógico de tabelas em um banco de dados multidimensional (*OLAP – Processamento Analítico Online*) de modo que o diagrama de relacionamento de entidade se assemelhe a uma forma de floco de neve . O esquema do floco de neve é representado por **tabelas de fatos centralizadas que são conectadas a várias dimensões**



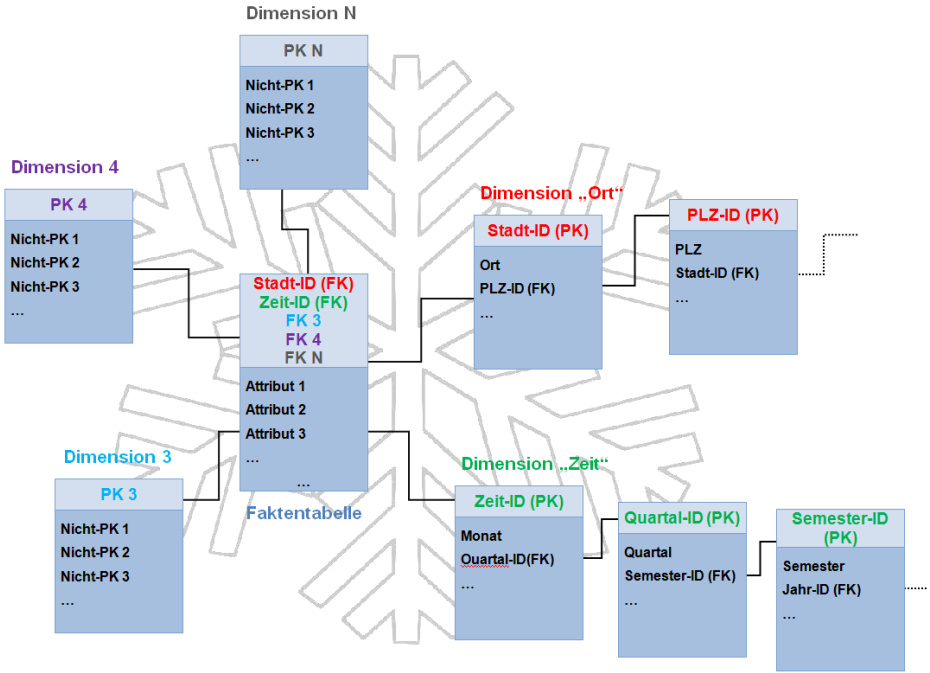
Fonte: https://en.wikipedia.org/wiki/Snowflake_schema

Modelo SnowFlake

Newton
Quem se prepara, não para.

"**Snowflaking**" é um método de **normalizar as tabelas de dimensão** em um **esquema** em **estrela** . Quando é completamente **normalizado ao longo de todas as tabelas de dimensão, a estrutura resultante se assemelha a um floco de neve** com a tabela de fatos no meio.

O princípio por trás do floco de neve é a normalização das tabelas de dimensão removendo atributos de baixa cardinalidade e formando tabelas separadas.



Fonte: https://en.wikipedia.org/wiki/Snowflake_schema

Vantagem

Newton
Quem se prepara, não para.

A **normalização divide os dados para evitar redundância (duplicação**), movendo grupos de dados comumente repetidos para novas tabelas. A normalização, portanto, tende a **aumentar o número de tabelas** que precisam ser unidas para realizar uma determinada consulta, **mas reduz o espaço necessário para armazenar os dados e o número de locais onde precisa ser atualizado se os dados mudarem**.

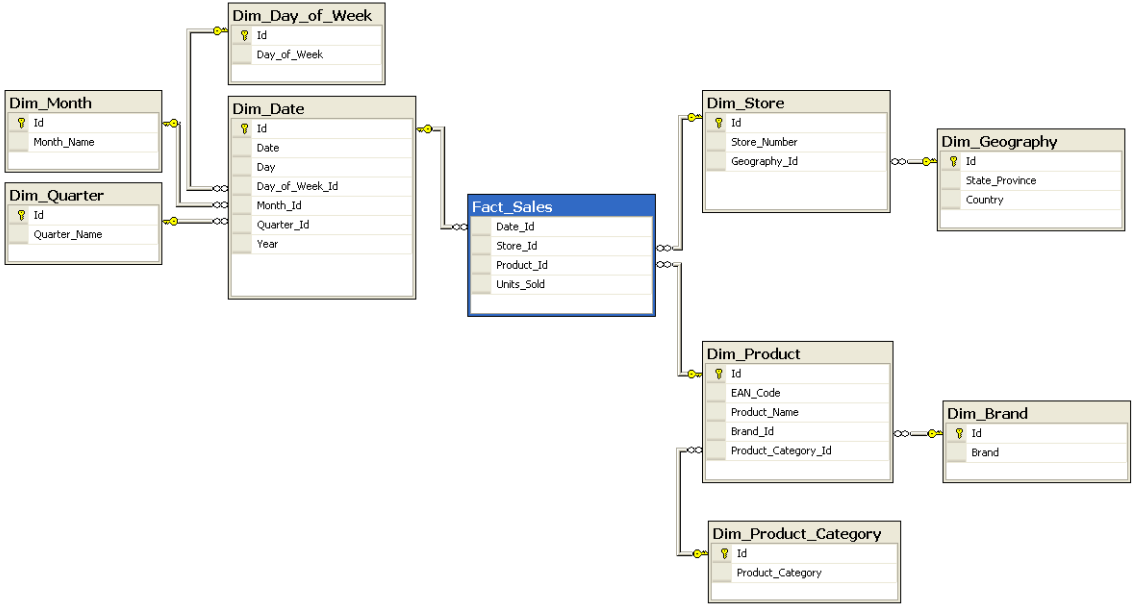
- Algumas ferramentas de modelagem de banco de dados multidimensionais OLAP são otimizadas para esquemas de floco de neve.
- A normalização de atributos resulta em economia de armazenamento, sendo a compensação uma complexidade adicional nas junções de consulta de origem.

Fonte: https://en.wikipedia.org/wiki/Snowflake_schema

Desvantagem

A principal desvantagem do esquema floco de neve é que os níveis adicionais de normalização de atributos adicionam complexidade às junções de consulta de origem, quando comparados ao esquema em estrela .

```
SELECT
    B.Brand,
    G.Country,
    SUM(F.Units_Sold)
FROM Fact_Sales F
INNER JOIN Dim_Date D             ON F.Date_Id = D.Id
INNER JOIN Dim_Store S            ON F.Store_Id = S.Id
INNER JOIN Dim_Geography G        ON S.Geography_Id = G.Id
INNER JOIN Dim_Product P          ON F.Product_Id = P.Id
INNER JOIN Dim_Brand B            ON P.Brand_Id = B.Id
INNER JOIN Dim_Product_Category C ON P.Product_Category_Id = C.Id
WHERE
    D.Year = 1997 AND
    C.Product_Category = 'tv'
GROUP BY
    B.Brand,
    G.Country
```

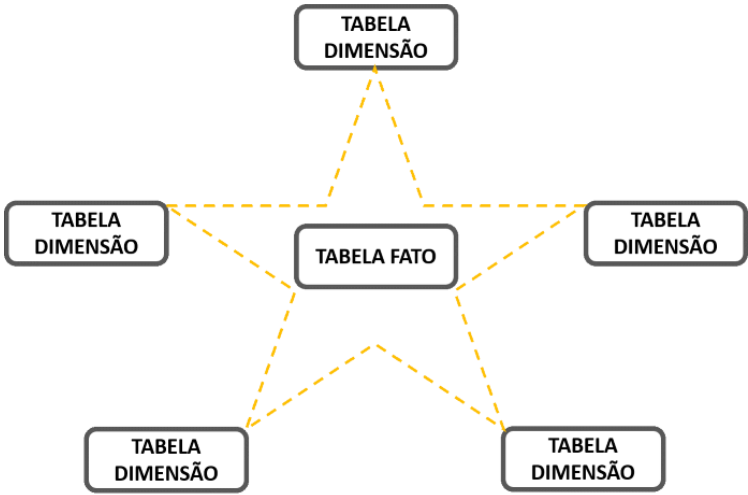


Fonte: https://en.wikipedia.org/wiki/Snowflake_schema

Star Schema vs SnowFlake

Newton
Quem se prepara, não para.

O **Star Schema** é, sem dúvidas, o modelo mais difundido e utilizado na criação de um Data Warehouse (DW). Este foi um modelo proposto por **Ralph Kimball** com o objetivo de simplificar a visualização dimensional, facilitando a distinção entre as dimensões e aos fatos.



Fatos são **métricas** (algo que pode ser medido ou quantificado) resultantes de um evento do processo de negócio. Ou seja, um acontecimento do negócio, que traz uma métrica (ou medida) associada a ele.

Já as **dimensões** representam os **contextos para análise de um fato,** proporcionando diferentes perspectivas de análise para o usuário e normalmente interpretadas como os “filtros possíveis” para determinada tabela fato.

Fonte: https://www.linkedin.com/pulse/snowflake-x-star-damian-carvalho/?originalSubdomain=pt

Star Schema vs SnowFlake

o **Snowflake Schema** adiciona complexidade ao modelo, com o objetivo de reduzir a redundância no armazenamento.

Essa **complexidade não é apenas em nível de armazenamento, mas também na consulta e extração das informações,** pois este modelo tende a aproximar novamente a **modelagem dimensional** da modelagem utilizada nos sistemas transacionais e isto dificulta o entendimento por parte dos usuários de negócio.

Granularidade

A granularidade está diretamente ligada na criação das fatos, impactando e definindo o volume de dados a ser armazenado e processado em cada fato.

- **Exemplo**: cenário de vendas de uma loja varejista, onde em uma **fato com baixa granularidade teremos o armazenamento de dados de vendas em nível de cupom fiscal,** resultando em um grande número de linhas armazenadas, porém possibilitando a **visualização individual de cada venda**. Já em uma fato determinada com **alta granularidade, poderíamos armazenar os dados de vendas consolidados por dia**, assim reduziríamos a quantidade de linhas armazenadas na tabela, mas perderíamos a capacidade de ver detalhadamente cada venda. É possível ainda ter os dois cenários dentro do mesmo modelo, onde a fato seria selecionada de acordo com a necessidade da consulta, permitindo assim tornar o modelo mais eficiente.

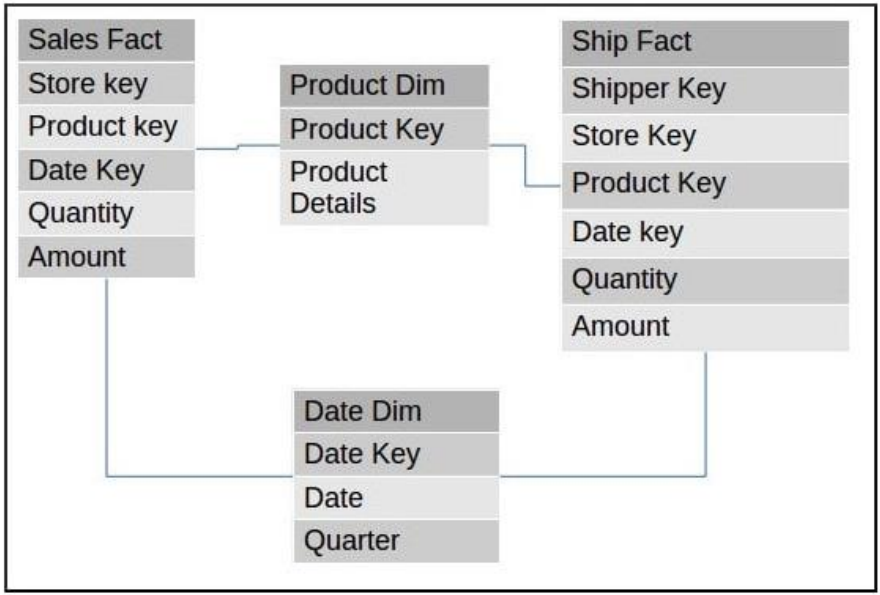
Fonte: https://www.linkedin.com/pulse/snowflake-x-star-damian-carvalho/?originalSubdomain=pt

Fact Constellation

Newton
Quem se prepara, não para.

A Constelação de fatos pode ser referida como **uma coleção de várias tabelas de fatos que compartilham tabelas de dimensões**. Portanto, pode até ser referido como uma **coleção de estrelas** que também é chamada de galáxia. Esse tipo específico de esquema geralmente é usado para aplicativos sofisticados.

Um exemplo que se refere a este esquema seria normalmente um cenário de vendas, onde há duas tabelas de fatos e ambas compartilham as tabelas de dimensão Product e Data. Portanto, o modelo de data warehouse é uma combinação de dois esquemas em estrela.



Fonte: https://amrithaf.wordpress.com/2017/07/11/star-schema-vs-fact-constellation/

Fact Constellation

Newton
Quem se prepara, não para.

Vantagem dos armazéns de dados do esquema de constelação de fatos

- Fornece um esquema flexível
- Diferentes tabelas de fatos são explicitamente atribuídas às dimensões

Desvantagem dos armazéns de dados do esquema de constelação de fatos

- A solução Constellation é difícil de manter
- Complexidade do esquema envolvido devido ao número de agregações

Fonte: https://amrithaf.wordpress.com/2017/07/11/star-schema-vs-fact-constellation/

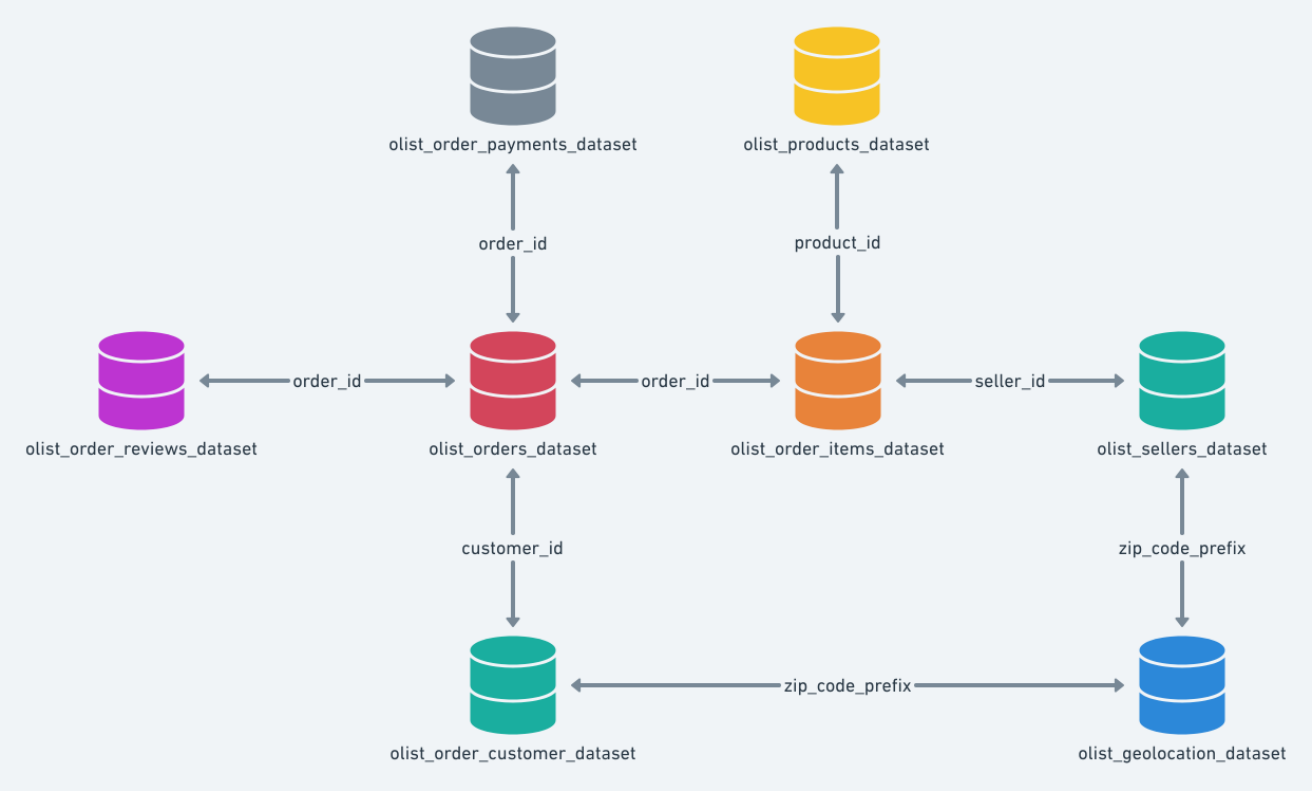
Exercício

Este é um conjunto de dados público de comércio eletrônico brasileiro de pedidos feitos na Olist Store . O conjunto de dados contém informações de **100 mil pedidos de 2016 a 2018 feitos em vários marketplaces no Brasil.** Seus recursos permitem visualizar um pedido de várias dimensões: desde o status do pedido, preço, pagamento e desempenho do frete até a localização do cliente, atributos do produto e, finalmente, avaliações escritas pelos clientes. Também lançamos um conjunto de dados de geolocalização que relaciona os códigos postais brasileiros às coordenadas lat/lng.

Disponível no Kaggle para download: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Fonte: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Exercício



Atenção

1.Um pedido pode ter vários itens.
2.Cada item pode ser atendido por um vendedor distinto.
3.Todo o texto identificando lojas e parceiros foi substituído pelos nomes das grandes casas de Game of Thrones.

Fonte: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Exercício

Newton
Quem se prepara, não para.

1) Fazer o download das Tabelas no site https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

2) Importar para o PowerBI – Escolher a opção IMPORTAR TEXTO/CSV, importe um arquivo por vez.

3) Realizar a modelagem dos dados no formato Snow Flake

Exercício

Newton
Quem se prepara, não para.

3) Realizar a modelagem dos dados no formato Snow Flake

Organizar os relacionamentos
Criar a tabela dimensão

Adicionar os gráficos:
Dot Plot da MAQ
GrowthRate