
UNITAU
Universidade de Taubaté

Desnormalização

OLTP x OLAP

O crescimento da capacidade de processamento e do volume de informações disponíveis permitiram o surgimento de novas aplicações para o processamento de dados. Atualmente podemos dividir as aplicações de processamento de dados em duas finalidades principais:

- **OLTP** (Online Transaction Processing) – processamento das transações operacionais das empresas, essas operações são a base das aplicações que gerenciam as rotinas das empresas.
- **OLAP** (Online Analytical Processing) – processamento de grande volume de dados para gerar análises para apoio a gestão e tomada de decisões.

OLTP x OLAP

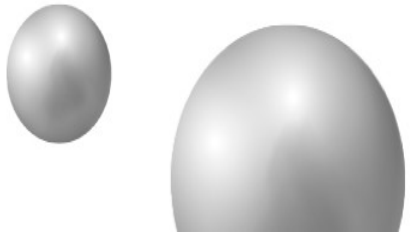
“OLTP (Online Transaction Processing ou Processamento de Transações em Tempo Real) são sistemas que se encarregam de registrar todas as transações contidas em uma determinada operação organizacional. Por exemplo: sistema de transações bancárias que registra todas as operações efetuadas [...] Em grandes aplicações, a eficiência do OLTP vai depender de um sofisticado software de gerenciamento de transações (como o CICS) e/ou otimizações táticas de base de dados de um grande número concorrente de updates em uma base de dados orientado a OLTP. “

https://pt.wikipedia.org/wiki/OLTP

OLTP x OLAP

“OLAP (Online Analytical Processing) é a capacidade para manipular e analisar um grande volume de dados sob múltiplas perspectivas. As aplicações OLAP são usadas pelos gestores em qualquer nível da organização para lhes permitir análises comparativas que facilitem a sua tomada de decisões diárias. [...] O OLAP fornece para organizações um método de acessar, visualizar, e analisar os dados corporativos com alta flexibilidade e performance.”

https://pt.wikipedia.org/wiki/OLAP

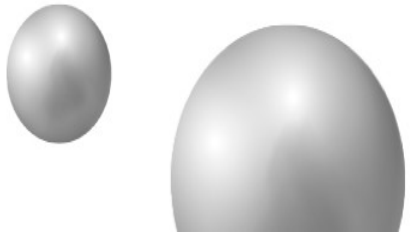


OLTP x OLAP

Aplicações OLTP tem por finalidade a entrada de dados, processamento e geração de informações das atividades operacionais das empresas. São voltadas para a rotina do negócio da empresa e tem como principal objetivo a integridade e consistência dos dados dessas atividades. Embora envolvam grande volume de consulta aos dados, a maioria dessas consultas envolvem um pequeno volume de informações. Além das consultas, essas aplicações também executam um grande volume de atualização de dados. Para atender essas características, são tradicionalmente baseadas em bancos de dados relacionais, tabelas normalizadas e uso de transações com consistência forte.

OLTP x OLAP

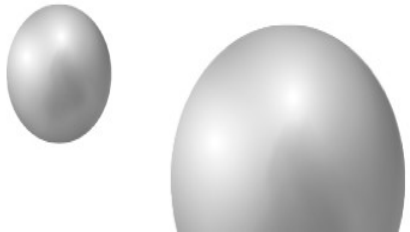
Aplicações OLAP e outras voltadas para ciência de dados são voltadas para geração de conhecimento baseados no processamento de um grande volume de dados; Não se destinam as atividades operacionais da empresa mas sim ao apoio a tomada de decisões e gerência estratégica. Normalmente envolvem consultas a enorme volume de dados e um volume muito menor de atualizações de dados. Para atender essas características, são baseadas em bancos de dados desnormalizados e que privilegiam o processamento distribuído.



Normalização x Desnormalização

“Normalização de banco de dados é um conjunto de regras que visa, principalmente, a organização de um projeto de banco de dados para reduzir a redundância de dados, aumentar a integridade de dados e o desempenho. Para normalizar o banco de dados, deve-se examinar as colunas (atributos) de uma entidade e as relações entre entidades (tabelas), com o objetivo de se evitar anomalias observadas na inclusão, exclusão e alteração de registros.”

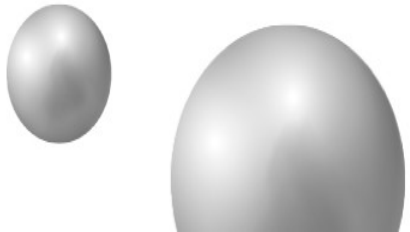
https://pt.wikipedia.org/wiki/Normalização_de_dados



Normalização x Desnormalização

“A desnormalização é uma estratégia usada em um banco de dados previamente normalizado para aumentar o desempenho. Na computação, desnormalização é o processo de tentar melhorar o desempenho de leitura de um banco de dados, às custas de perder algum desempenho de gravação, adicionando cópias redundantes de dados ou agrupando dados. Muitas vezes, é motivado pelo desempenho ou escalabilidade em um software de banco de dados relacional que precisa realizar um grande número de operações de leitura.”

https://en.wikipedia.org/wiki/Denormalization

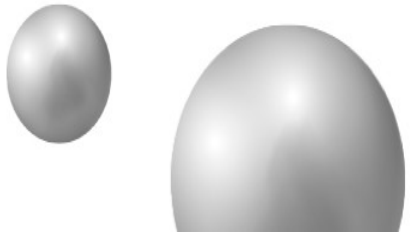


Dados

Dados Estruturados são dados que tem uma estrutura definida e rígida, como tabelas com campos definidos, todos registros das tabelas seguem as mesma estrutura

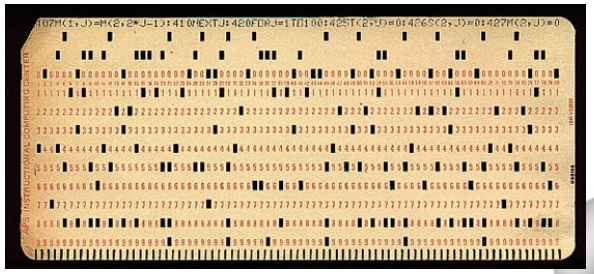
Dados não Estruturados não possuem uma estrutura previamente definidas, são baseadas em fontes que não possuem formato e conteúdo definido como documentos e vídeos

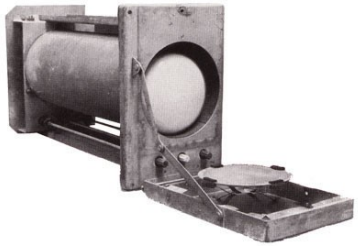
Dados Semi Estruturados são dados que não possuem uma estrutura rígida, porém possuem alguma organização com tags e metadados

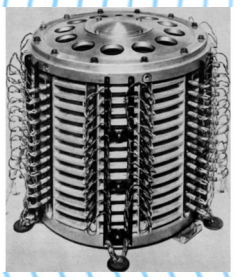


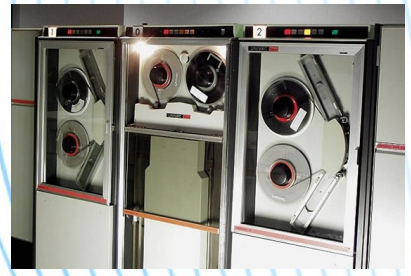
Armazenamento de Dados (OLTP)

As aplicações OLTP do início da era da informática armazenavam os dados em arquivos individuais. As primeiras formas de armazenamento de dados eram cartões perfurados e dispositivos magnéticos como tubos de Willians, tambores magnéticos e fitas magnéticas. Tecnologias com baixa capacidade de armazenamento e velocidade de leitura. Não era possível ler diversos arquivos ao mesmo tempo nesses dispositivos. Não existia o conceito de banco de dados.





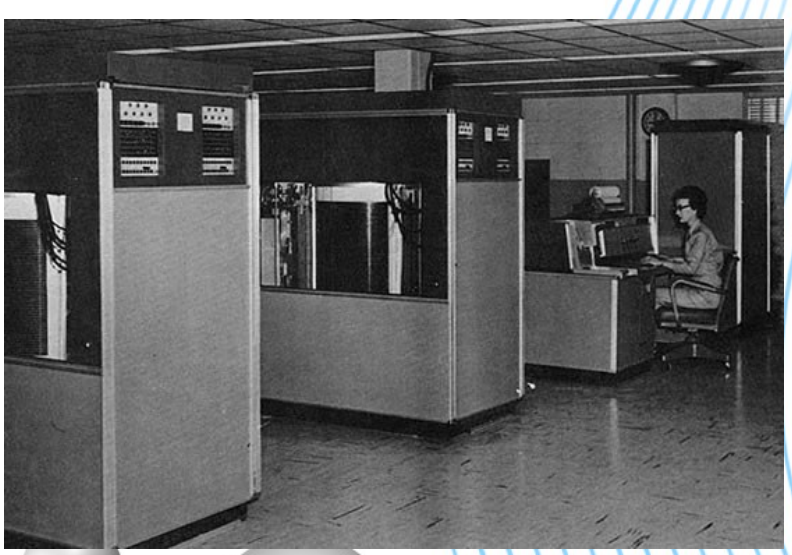




https://pplware.sapo.pt/internet/a-historia-do-armazenamento-digital/

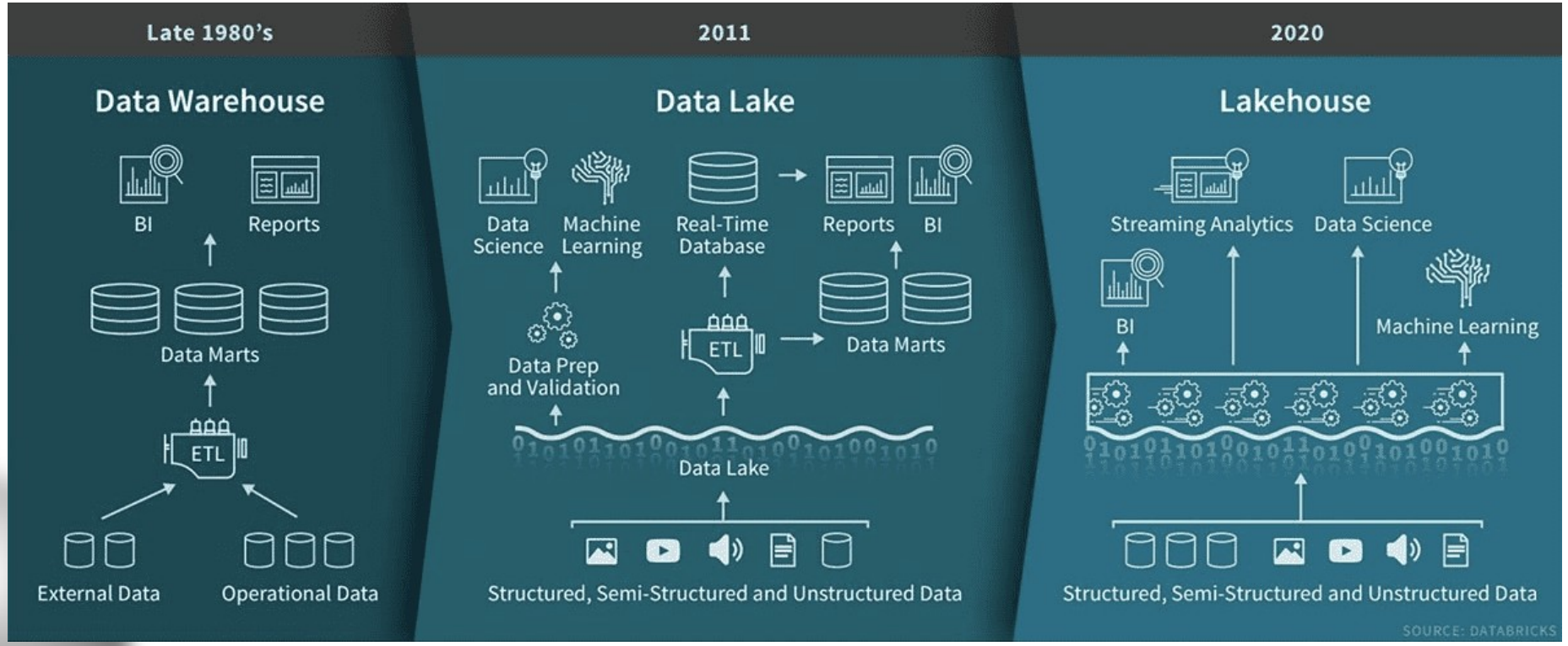
Armazenamento de Dados (OLTP)

Após o surgimento dos discos magnéticos, o aumento da capacidade de armazenamento, velocidade de leitura e a possibilidade de ler diversos arquivos ao mesmo tempo permitiu o surgimento dos Sistemas Gerenciadores de Bancos de Dados, que são a base das aplicações OLTP.



https://pplware.sapo.pt/internet/a-historia-do-armazenamento-digital/

Armazenamento de Dados (Data Science)



https://medium.com/slalom-data-analytics/the-evolution-of-the-databricks-lakehouse-paradigm-71f613c6533a

ETL

Bancos de dados normalizados são voltados para aplicações OLTP para impedir as anomalias de atualização e facilitar a consistência dos dados.

Aplicações de ciência de dados usualmente não fazem um grande volume de atualização de dados e necessitam de melhor performance na leitura dos dados para permitir o processamento de um volume maciço de informações, além de utilizarem dados não estruturados de outras fontes.

Para criar bancos de dados mais adequados para essas aplicações, são utilizados processos de ETL (Extract Transform Load) para obter dados dessas diversas fontes.

ETL

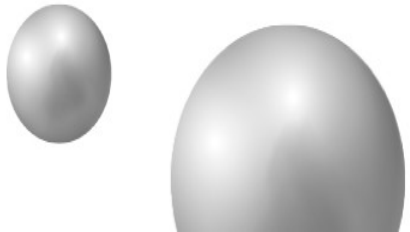
Os dados extraídos dos bancos das aplicações transacionais podem ser desnormalizados durante esse processo para otimizar as consultas.

O processo de ETL envolve as fases de:

- **Extração** – os dados são extraídos dos SGBDs operacionais e das demais fontes externas.
- **Limpeza** – devem ser feitas as correções nos dados obtidos para eliminar erros e omissões nos dados extraídos.

ETL

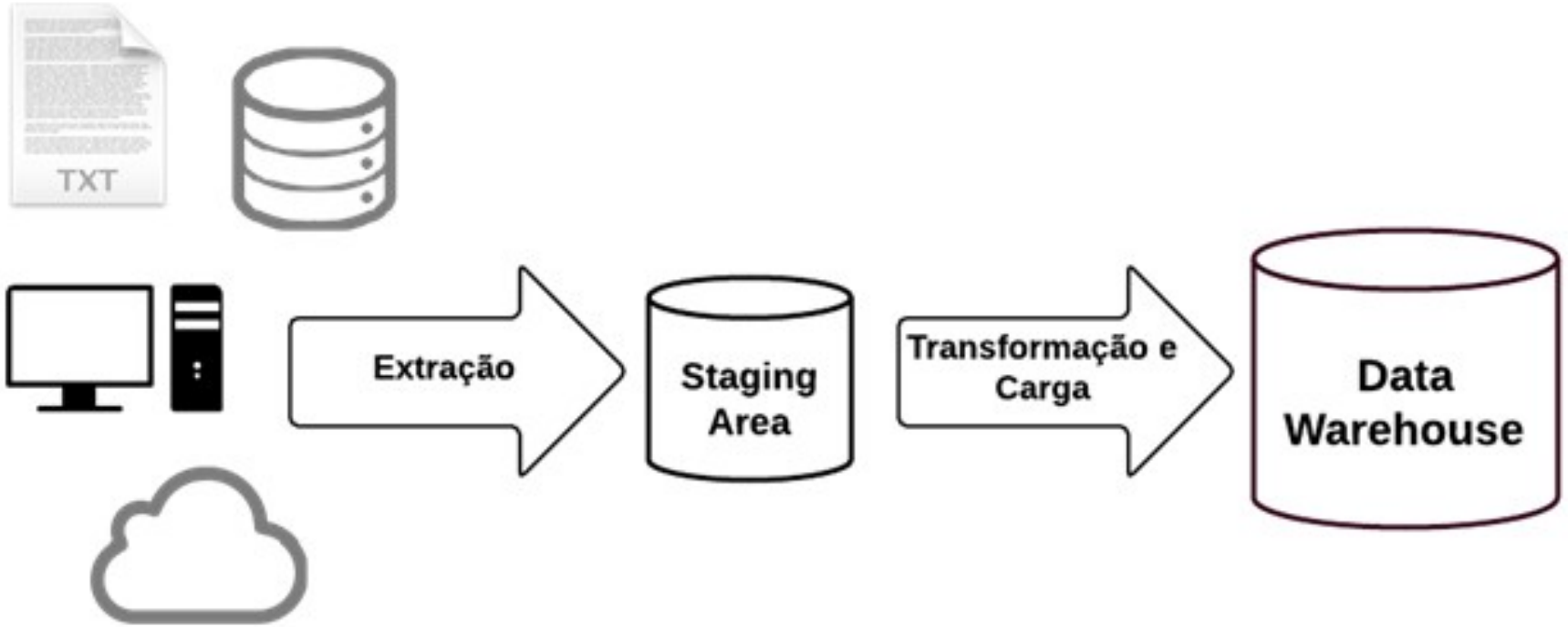
- **Transformação** – os dados devem ser transformados para o formato e semântica adotada no DataWarehouse, informações vindas de diferentes fontes podem ser tratadas de forma diferentes com a utilização de diferentes nomes, tipos de dados ou diferentes valores para os mesmos significados, essas diferenças devem ser eliminadas.
- **Carga** – os dados já limpos e transformados devem ser carregados no banco de dados, a geração de informações adicionais como resumos é feita na carga dos dados



ETL

- **Atualização – é necessário atualizar periodicamente as informações carregadas no DataWarehouse, a possibilidade de utilizar informações mais ou menos atualizadas e o impacto do processo de atualização do DataWarehouse nas demais operações da organização devem determinar a estratégia de atualização a ser utilizada.**

ETL



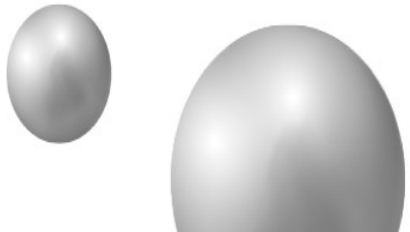
https://canaltech.com.br/business-intelligence/entendendo-o-processo-de-etl-22850/

Data Warehouse

O Data Warehouse é um sistema utilizado para armazenar dados, de uma maneira organizada. Considera-se o Data Warehouse (DW) a base para o Business Intelligence (BI).

O DW pode guardar informações relativas às atividades de uma organização em bancos de dados, de forma consolidada. O desenho da base de dados favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão

https://www.cetax.com.br/blog/o-que-e-data-warehouse/



Data Lake

Os Data Lakes são repositórios para dados brutos em uma variedade de formatos, como dados de aplicativos de linha de negócios, aplicativos móveis, mídias sociais, dispositivos IoT, etc. Armazenam uma grande quantidade de dados diferentes, não filtrados, para serem usados posteriormente para uma finalidade específica. [...]

Embora adequados para o armazenamento de dados, os tais "lagos de dados" carecem de alguns recursos essenciais: não suportam transações, não impõem a qualidade dos dados e sua falta de consistência / isolamento torna quase impossível misturar acréscimos e leituras e trabalhos em lote e streaming.

https://ilumeo.com.br/todos-posts/2021/11/01/data-lakehouse-o-avanco-do-gerenciamento-de-dados

Data Warehouse x Data Lake

Diferenças entre Data Warehouse e Data Lake

- os Data Lakes não têm objetivo definido e são conjuntos de dados brutos;
- os Data Warehouses armazenam apenas dados estruturados que já foram processados para uma finalidade específica;
- Data Lakes têm estrutura variável;
- Data Warehouses têm estruturas estáticas;
- o custo de manter um Data Lake é menor;
- Data Warehouses são menos flexíveis.

https://blog.academiain1.com.br/data-warehouse-x-data-lake-entenda-conceitos-e-diferencas/

Data Lakehouse

Um Data Lakehouse é um conceito de solução de dados que combina elementos do data warehouse com os do data lake. Data lakehouses implementam estruturas de dados de data warehouses e recursos de gerenciamento para data lakes, que normalmente são mais econômicos para armazenamento de dados.

https://www.snowflake.com/guides/what-data-lakehouse

Data Lakehouse

Características de um Data Lakehouse

- Leitura e gravação simultânea de dados
- Suporte de esquema com mecanismos para governança de dados
- Acesso direto aos dados de origem
- Separação de armazenamento e recursos de computação
- Formatos de armazenamento padronizados
- Suporte para tipos de dados estruturados e semiestruturados, incluindo dados IoT
- Streaming de ponta a ponta

https://www.snowflake.com/guides/what-data-lakehouse

Delta Lake

Para que o conceito do LakeHouse se tornasse uma realidade e pudesse funcionar sobre a camada do Data Lake, a Databricks desenvolveu um projeto open-source de uma nova tecnologia à qual deu o nome de Delta Lake.

https://everisbrasil.medium.com/as-origens-do-delta-lake-2f561894f5ad

Delta Lake

Basicamente o Delta Lake oferece uma série de recursos como:

- **Controle de Transações**
- **Processamento escalável do metadados**
- **Versionamento de Dados**
- **Formato Aberto de arquivos**
- **Aplicação de schema (metadados)**
- **Evolução de schema (metadados)**
- **Auditoria**
- **Atualização ou remoção de registros**
- **Processo único para execuções em Lote ou Stream**
- **100% compatível com o Apache Spark**
- **Open-Source**

https://everisbrasil.medium.com/as-origens-do-delta-lake-2f561894f5ad

Delta Lake

A Databricks recomenda a criação de três regiões no Lake, para organizar a evolução dos dados dentro dele:

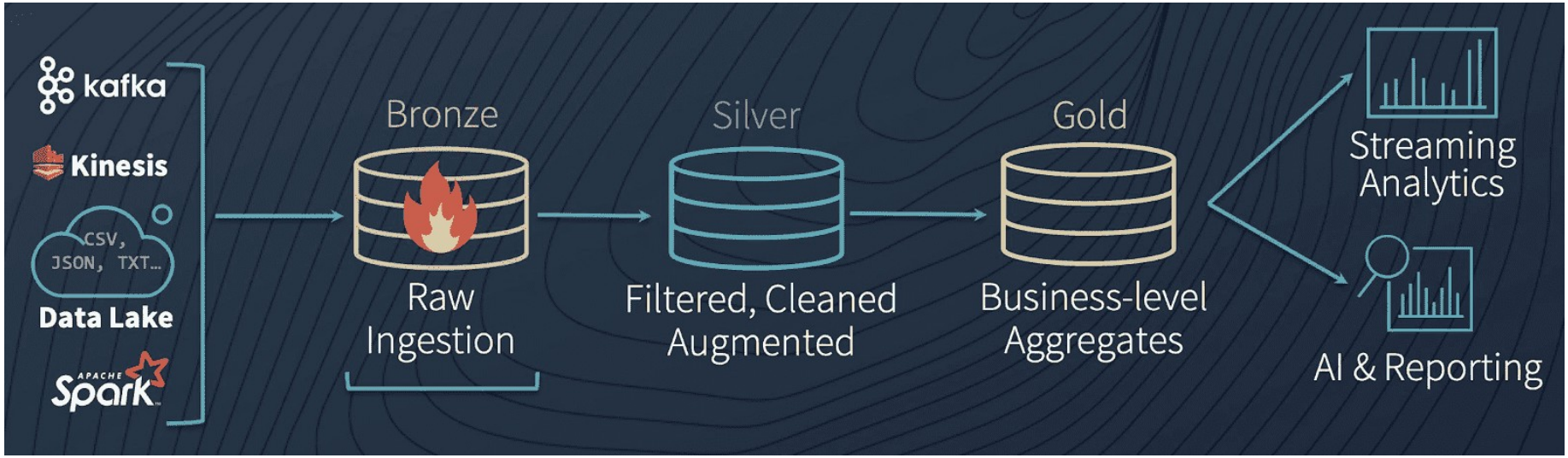
Bronze: dados brutos, exatamente como eles foram gerados em suas origens. Aqui os tipos dos campos são definidos todos como string, para evitar filtragem dos dados;

Silver: dados da Bronze que são lidos, filtrados, tratados e armazenados conforme seus data types corretos, estando prontos para uso;

Gold: consolidações e agregações originadas de tabelas da camada Silver. Destinados para consumos pelas áreas (dashboards) ou para estudos de inteligência artificial;

https://everisbrasil.medium.com/as-origens-do-delta-lake-2f561894f5ad

Delta Lake



https://databricks.com/blog/2019/08/14/productionizing-machine-learning-with-delta-lake.html

Modelos de Desnormalização

Para desnormalizar os dados dos bancos normalizados, podem ser utilizados alguns modelos:

- **Modelo One Big Table (OBT)**
- **Modelo Dimensional (ou Multidimensional)**

One Big Table (OBT)

No modelo OBT, as informações de diversas tabelas normalizadas são reunidas em uma única grande tabela desnormalizada utilizando as seguintes operações:

- **adicionar os valores dependentes das chaves estrangeiras, criando atributos redundantes**
- **adicionar atributos multivalorados com o conteúdo de relacionamentos 1:N**
- **adicionar os valores dos atributos derivados**
- **criação de visões materializadas**

One Big Table (OBT)

id	data	cliente	nome	endereco	produtos	total
1	01/08/2021	001	André	Av 9 de Julho	caneta, caderno, giz de cera	20,15
2	01/08/2021	002	Paula	XV de Novembro	borracha, lapis, compasso	12,10
3	02/08/2021	001	André	Av 9 de Julho	fichario	15,00
4	03/08/2021	003	Luis	Rua Joao Luis	caderno, regua, compasso	23,00
5	03/08/2021	002	Paula	XV de Novembro	caderno, fichario	30,00
6	04/08/2021	004	Ana	Av dos Bandeirantes	caderno, lapis, borracha	18,00

atributos redundantes
atributo multivalorado
atributo derivado

Modelo Dimensional

O modelo dimensional (ou multidimensional) foi proposto para bancos de Data Warehouse para melhorar a performance das consultas. É baseado em remodelar as tabelas baseados em fatos e dimensões.

https://dbccompany.com.br/dbc/modelagem-dimensional-star-schema-e-snowflake-schema/

Modelo Dimensional

“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. Uma tabela Fato armazena as métricas relacionadas a determinado evento, por exemplo, uma fato de Vendas pode armazenar quantidade de itens vendidos, valor dos itens vendidos, entre outras métricas. 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.”

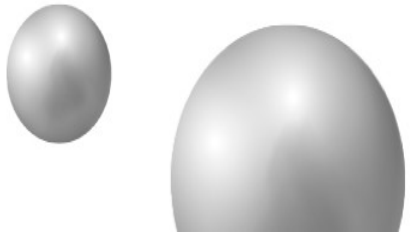
https://dbccompany.com.br/dbc/modelagem-dimensional-star-schema-e-snowflake-schema/

Modelo Dimensional

“As tabelas dimensão contêm as características de um evento. Por exemplo, quando eu faço uma venda, quero saber por onde a venda foi feita, que produto foi vendido, ou para quem.

Já a tabela fato armazena o que ocorreu, é o fato propriamente dito, por isso ela tem esse nome, porque é o fato ocorrido. A tabela fato está sempre ligada a duas ou mais dimensões, não existe tabela fato com menos de duas dimensões.”

https://rafaelpiton.com.br/blog/data-warehouse-modelagem-dimensional

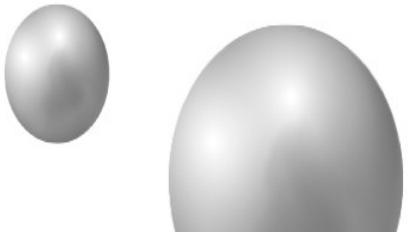


Modelo Dimensional

Tabelas de dimensões descrevem as entidades de negócios – os itens que você modela. As entidades podem incluir produtos, pessoas, locais e conceitos, incluindo o próprio tempo.

Tabelas de fatos armazenam observações ou eventos e podem ser ordens de vendas, saldos de ações, taxas de câmbio, temperaturas, etc. Uma tabela de fatos contém colunas chave de dimensão relacionadas a tabelas de dimensões e colunas de medidas numéricas.

https://docs.microsoft.com/pt-br/power-bi/guidance/star-schema



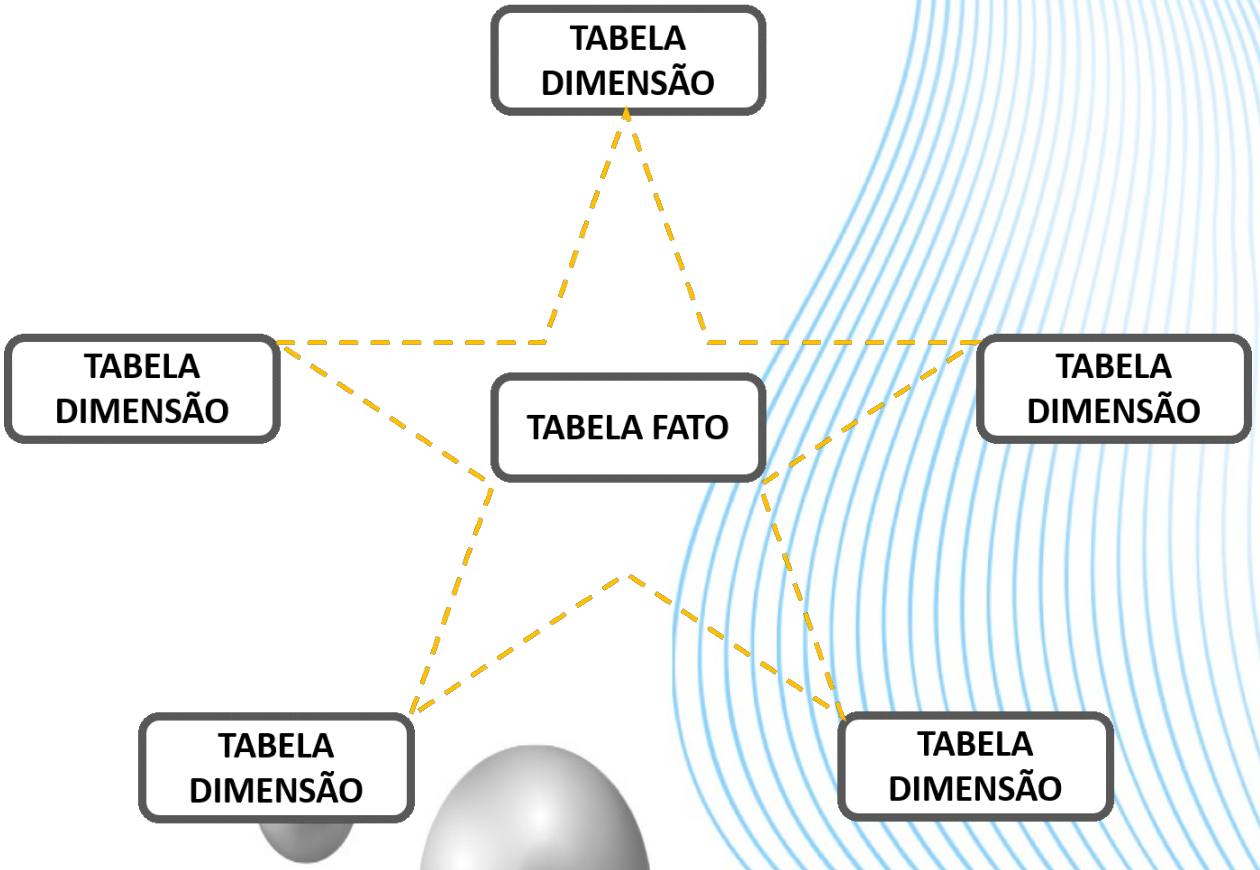
Modelo Dimensional

As tabelas de fatos e dimensões podem ser organizadas em diferentes esquemas:

- Star Schema
- Snowflake Schema
- Fat Constellation Schema (Galaxy Schema)

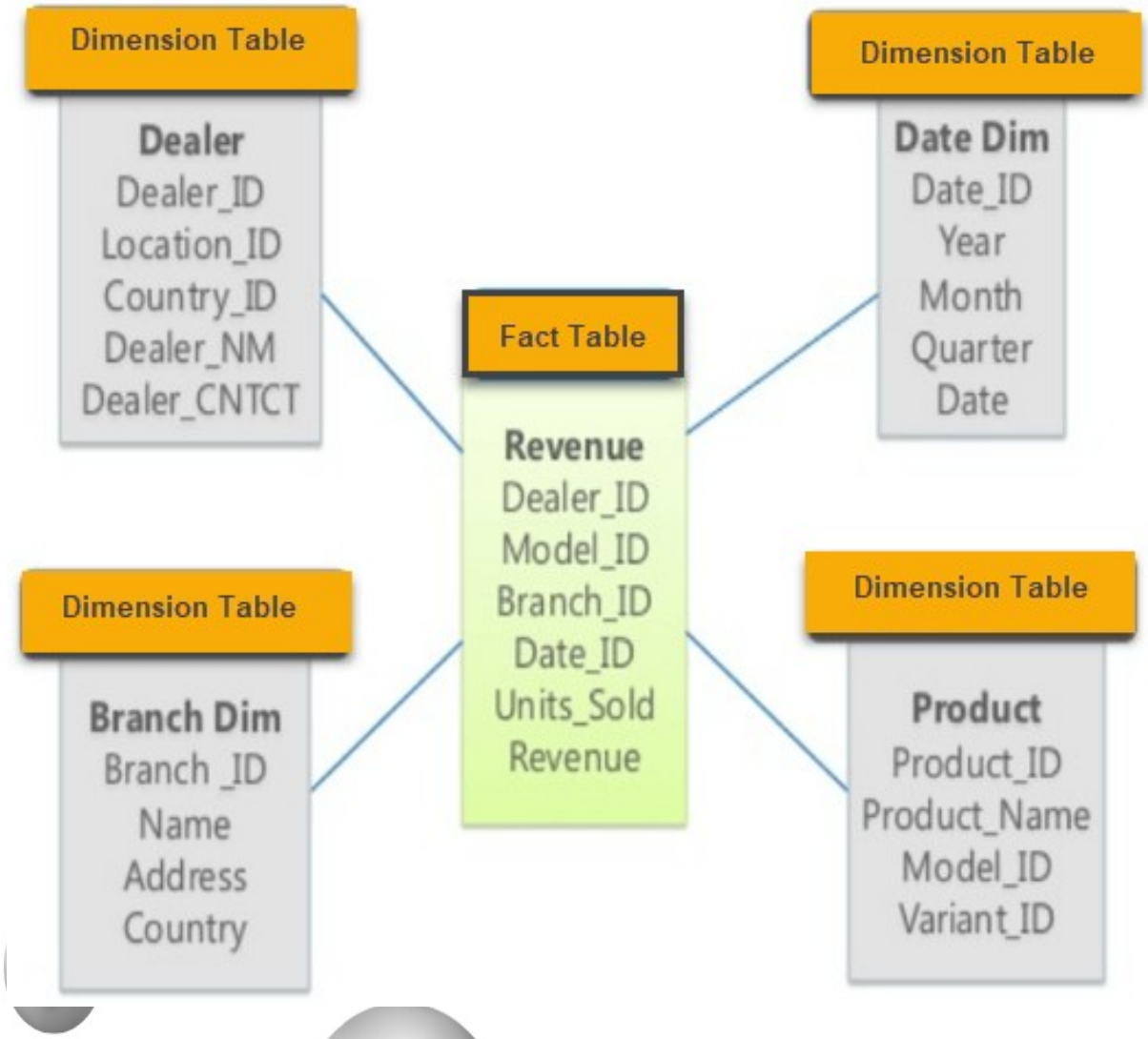
Star Schema

No esquema estrela (Star Schema) a tabela fato é ligada diretamente a cada uma das tabelas de dimensões.



https://dbccompany.com.br/dbc/modelagem-dimensional-star-schema-e-snowflake-schema/

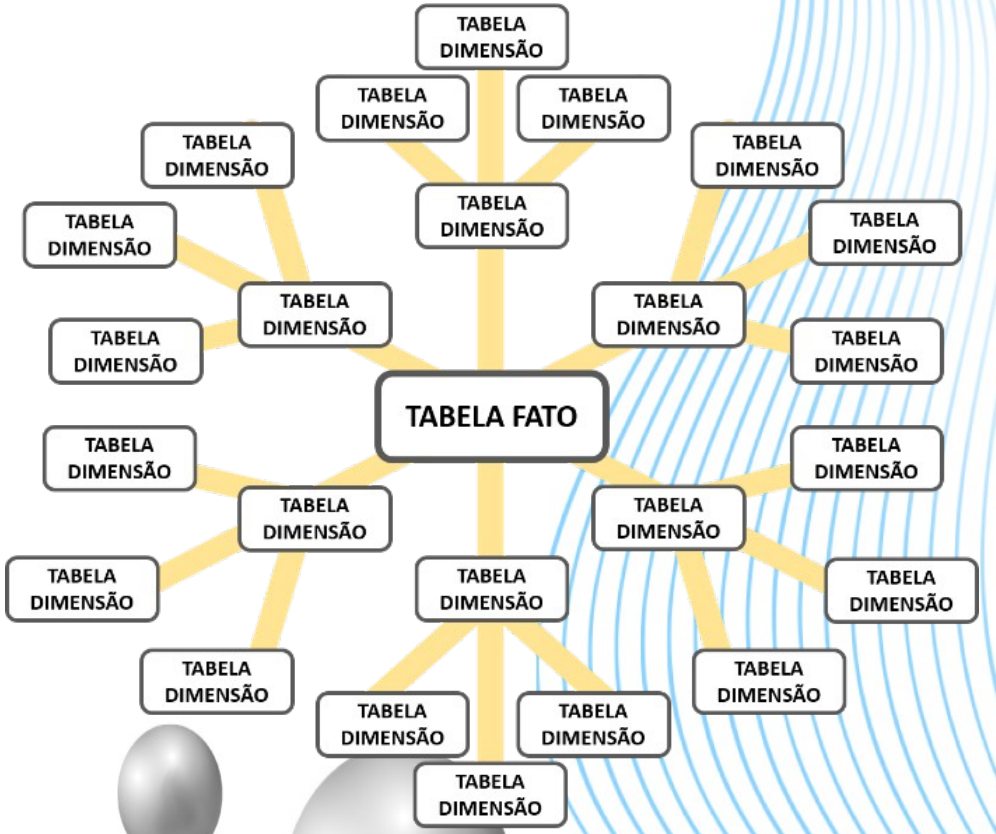
Star Schema



https://www.guru99.com/star-snowflake-data-warehousing.html

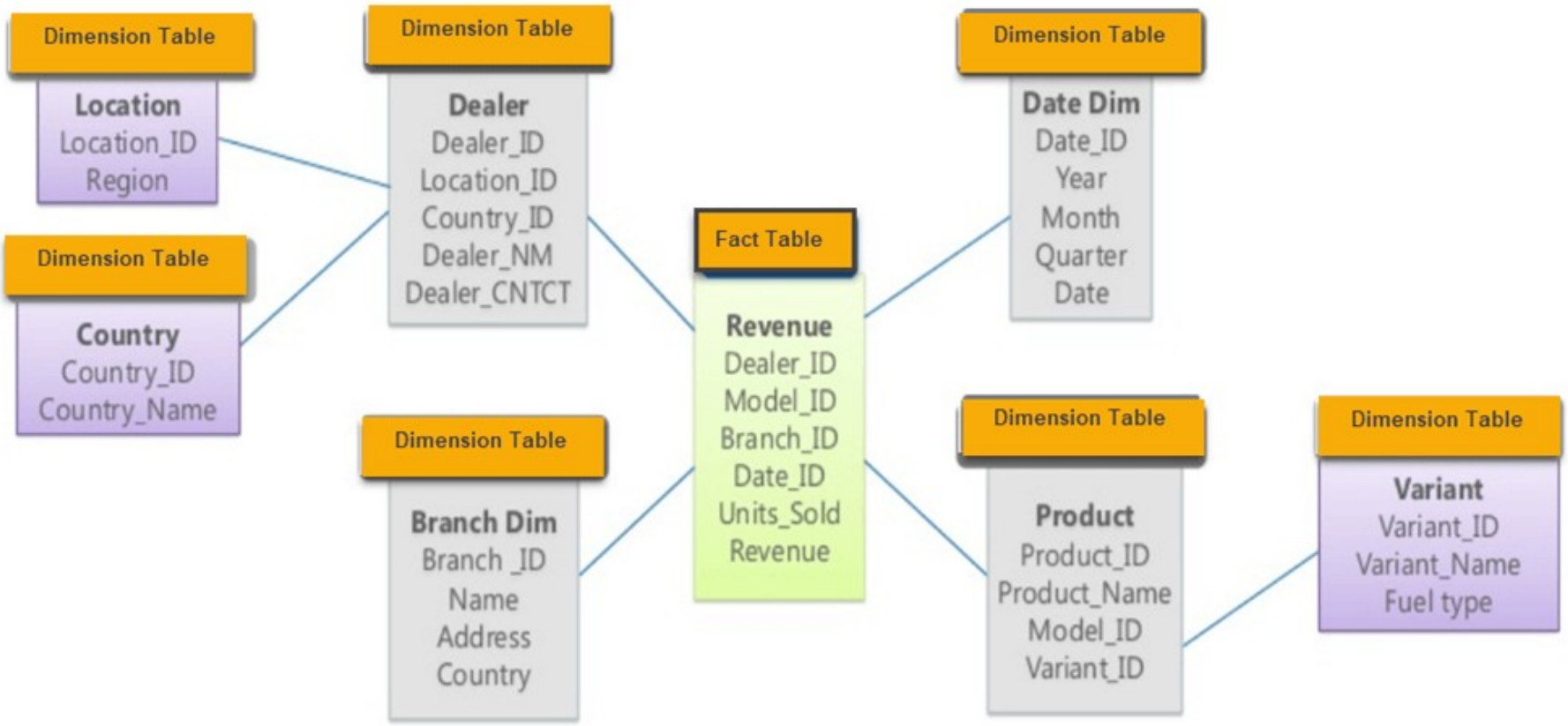
Snowflake Schema

No esquema floco de neve (Snowflake Schema) as tabelas de dimensões são normalizadas, podem estar ligadas a outras tabelas de dimensões



https://dbccompany.com.br/dbc/modelagem-dimensional-star-schema-e-snowflake-schema/

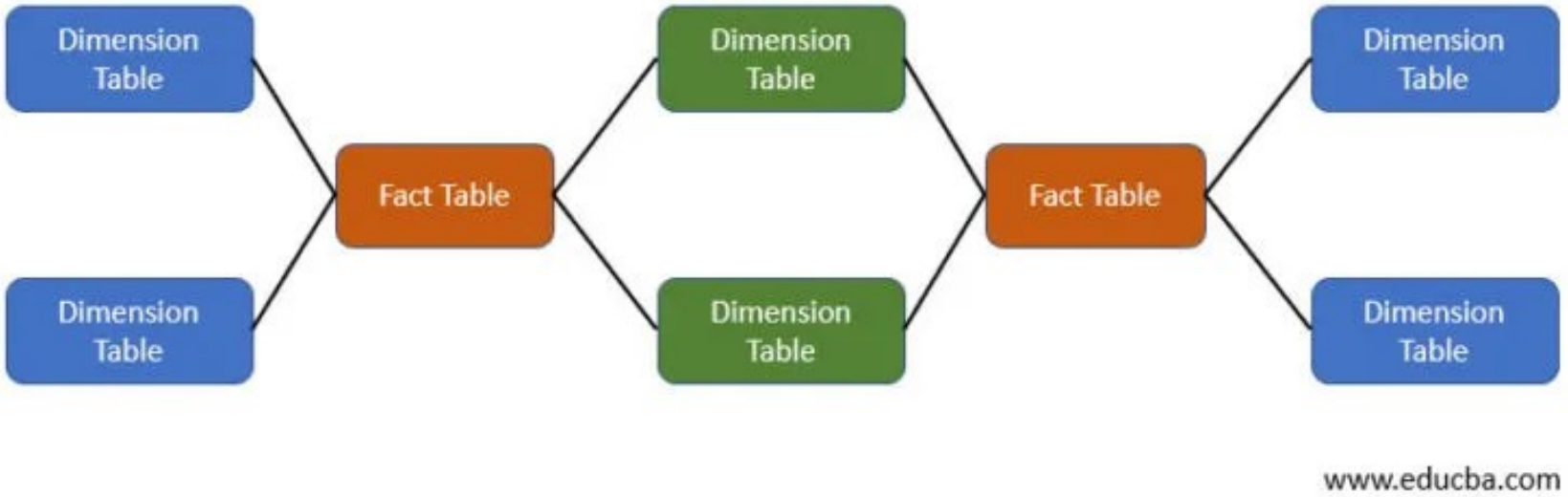
Snowflake Schema



https://www.guru99.com/star-snowflake-data-warehousing.html

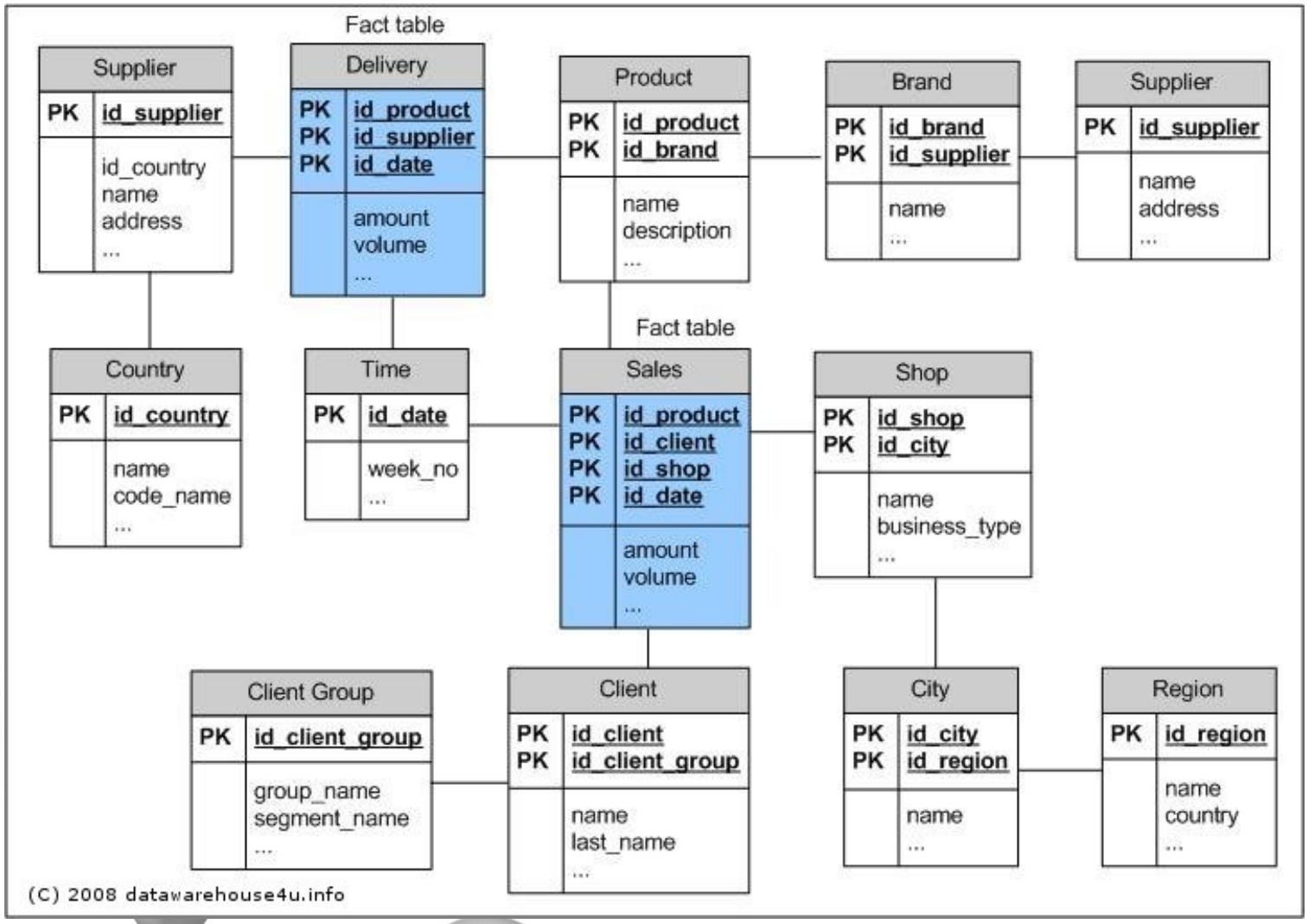
Fact Constellation (Galaxy) Schema

No Fact Constellation Schema ou Galaxy Schema, tabelas de fatos compartilham tabelas de dimensões, criando várias estrelas (ou flocos de neve) interligadas pelas dimensões.



https://www.educba.com/fact-constellation-schema/

Fact Constellation (Galaxy) Schema



https://www.datawarehouse4u.info/Data-warehouse-schema-architecture-fact-constellation-schema.html

OBT x Modelo Dimensional

Apesar da proposta do modelo dimensional ter por objetivo melhorar a performance das consultas, porém não há consenso quanto ao modelo mais adequado a ser usado em um Data Warehouse.

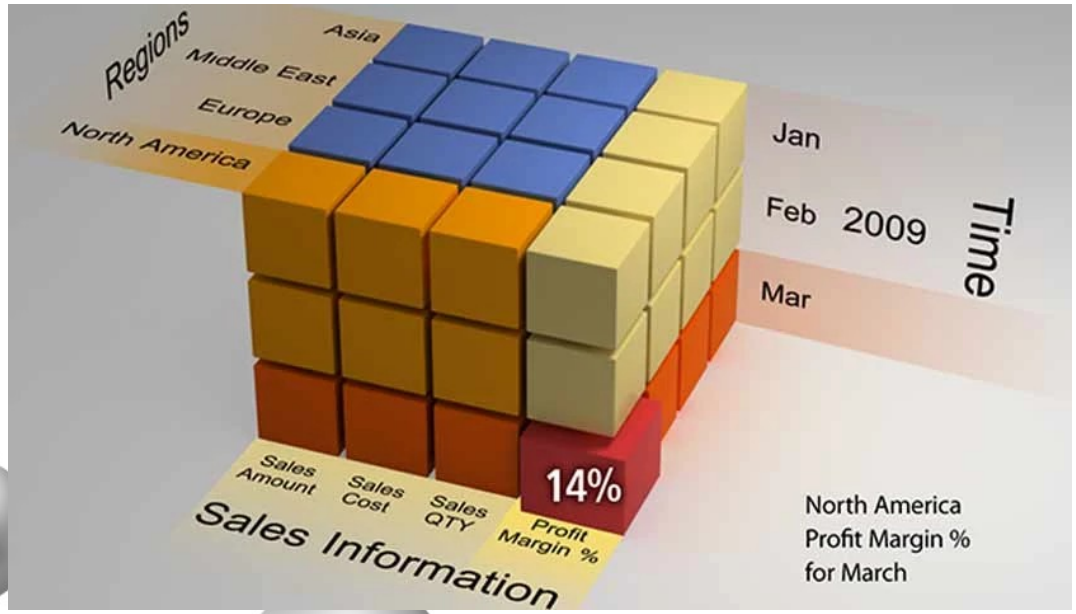
“The speed improvement of using a single denormalized table represents an improvement of 25%-50% depending on which warehouse you're using. This amounts to a difference of about 10 seconds on a single-node cluster in Redshift. Excluding redshift query compilation time, the improvements are:

- **Redshift: 25%-30% (depending on warehouse size and number of clusters)**
- **Snowflake: ~25%**
- **Bigquery: ~50%”**

https://fivetran.com/blog/star-schema-vs-obt

OLAP

As aplicações de OLAP se caracterizam por apresentar visões multidimensionais das informações. O foco e a apresentação das informações em formatos de cubos, onde as dimensões dos cubos representam os componentes do negócio da organização como Tempo, Produtos, Locais, etc.



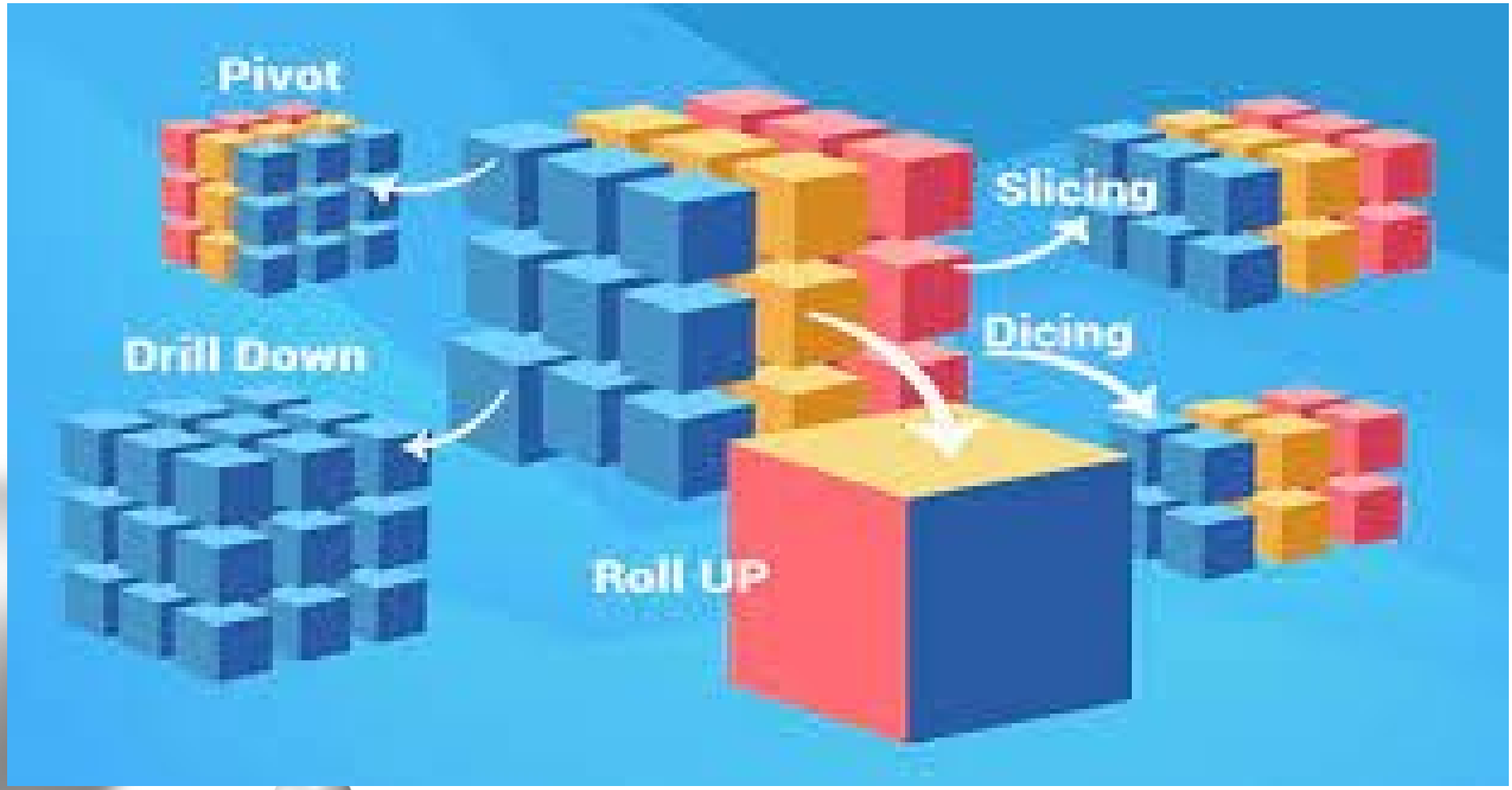
https://ayltoninacio.com.br/blog/criando-e-alimentando-um-cubo-olap-fisico-no-mysql

OLAP

As ferramentas OLAP apresentam as informações no formato de cubos com agregações e sumarização das informações e permitem algumas operações sobre as análises apresentadas:

- **Roll-Up** – apresenta os dados agrupando os valores de uma dimensão.
- **Drill-Down** – apresenta os dados com maior detalhando de uma dimensão.
- **Slice** (fatiar) - apresenta os dados separados por grupos em uma dimensão
- **Dice** (cortar) - apresentar os valores selecionando algum intervalo de valores
- **Pivoting** (rotação) - permite inverter as dimensões da apresentação

OLAP



https://www.devmedia.com.br/um-estudo-sobre-as-ferramentas-olap/6691