INSTITUTO SUPERIOR DE TECNOLOGIAS AVANÇADAS

Ciência de Dados

Licenciatura Engenharia Informática

2º Semestre – 2021/2022

Ricardo Jesus Ferreira

ricardojesus.ferreira@my.istec.pt

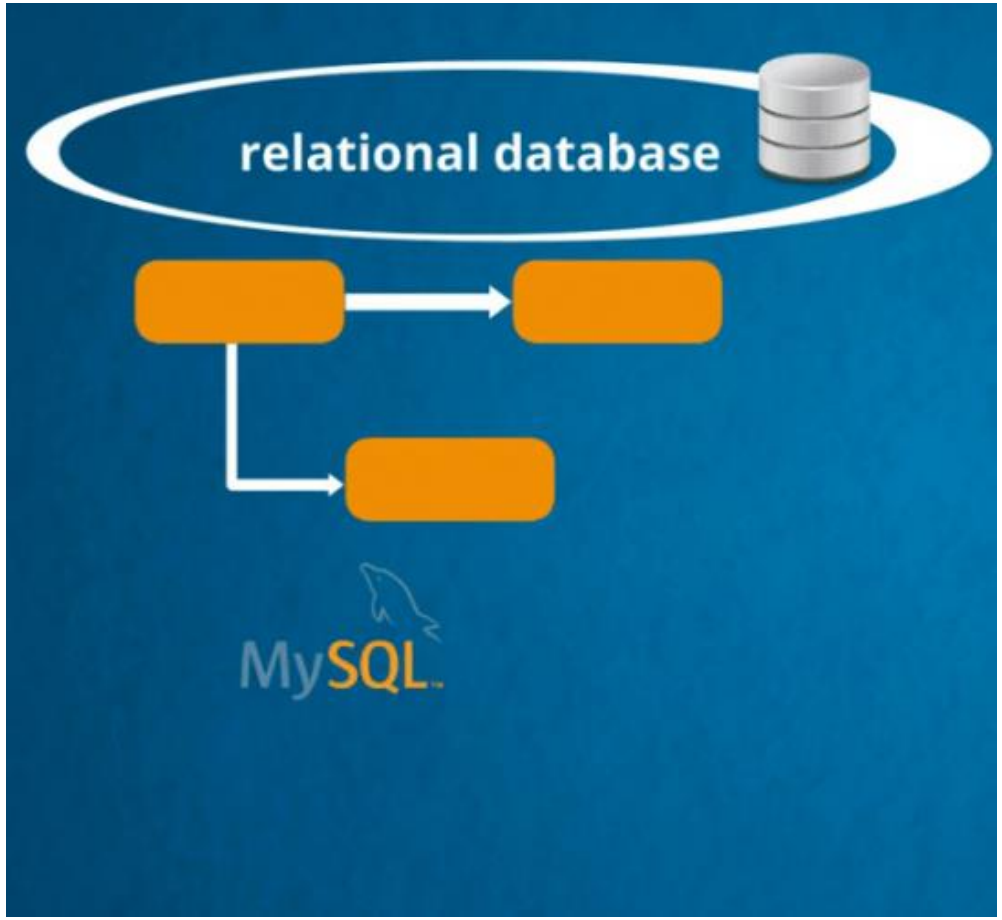
Database Schemas

- Base de dados Relacional
- Base de dados não Relacional
- Data Warehouse
- Data Lake
- Data Mart
- OLTP – Online Transactional Processing
- OLAP – Online Analytical Processing

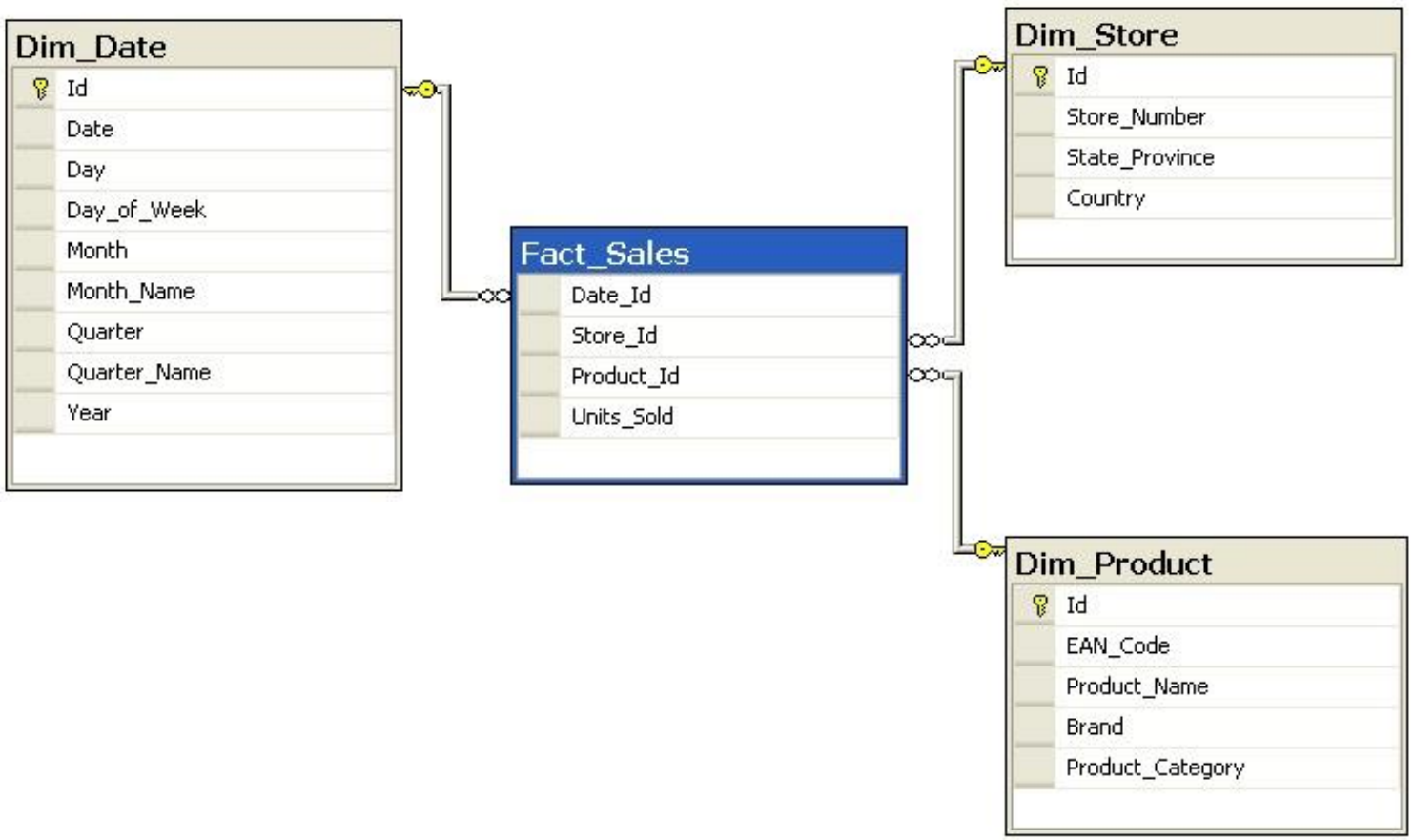
RDBMS

RDBMS

- O modelo relacional organiza a informação em tabelas
- Essas tabelas possuem relações entre si
- Isto permite que um utilizador crie uma "nova" tabela com base numa *query* realizada em uma ou múltiplas tabelas



Como funciona



RDBMS

- Integridade Referencial
  - Uma chave estrangeira tem que ter uma chave primaria correspondente
  - Quando uma informação na tabela principal for eliminada todos a informação relacionada terá que ser removida
  - Caso uma chave primaria seja alterada, todas as suas referencias terão que ser alteradas

SQL

SQL

- SQL – Structured Query Language

- Linguagem declarativa
  - Serve para comunicar com a base de dados
  - Permite criar, atualizar, obter ou apagar informação da BD
  - Utiliza as chaves para realizar as relações entre as tabelas

SQL Statement

- Mostrar todas as transações do cliente com o CustomerID=1

```
Select * FROM Costumer
JOIN Sales ON
Costumer.CostumerID=Trans
action.CostumerID WHERE
Costumer = 001
```

SQL - Index

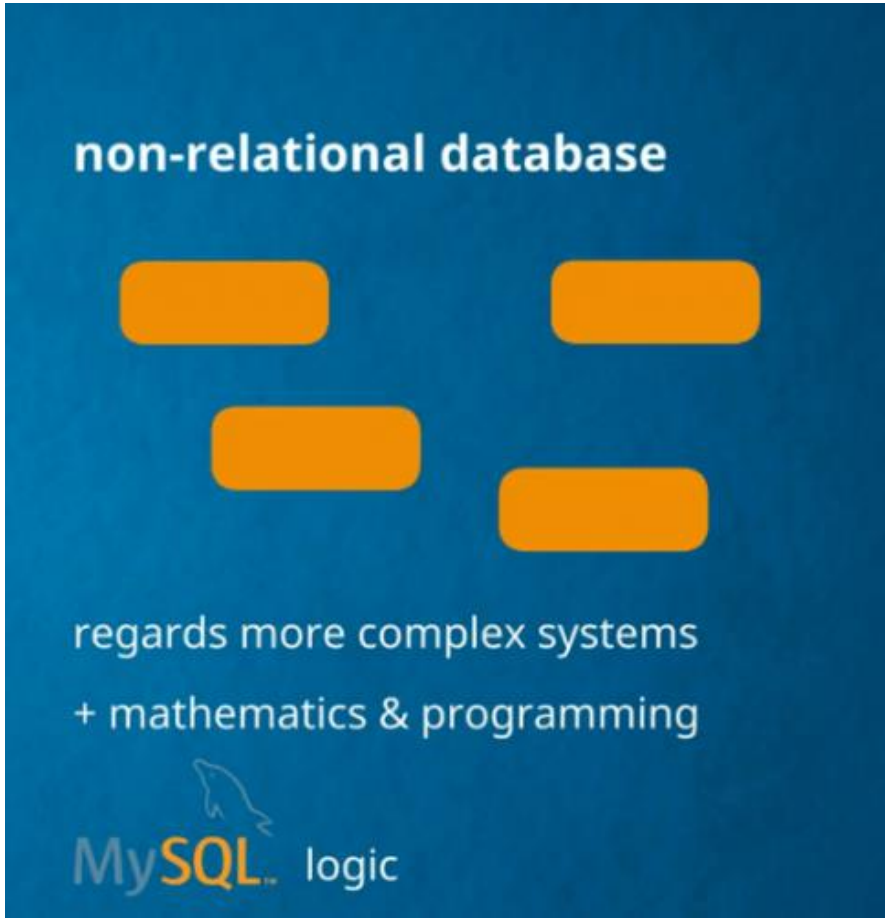
- RDBMS utilizam indexes
- Os indexes otimizam a performance das *queries* à DB
- Maior parte das *queries* são feitas com base nos indexes
- Estes indexes são tipicamente associados a *queires* frequentes e à união entre tabelas (JOIN)

Non-RDBMS

Non-RDBMS

- Não existe o conceito Tabela
- Não existe o conceito Chave Primária
- Não existe o conceito Chave Estrangeira

- Este modelo utiliza uma mecanismo de armazenamento especifico para o tipo de dados a serem guardados



Document

- Uma base de dados de documentos armazena a recolha de documentos, onde cada documento é composto por campos e dados.
- Os dados podem ser valores simples ou elementos complexos, tais como listas e coleções.
- Os documentos são identificados por chaves únicas

Key	Document
1001	{ "CustomerID": 99, "OrderItems": [{ "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" }
1002	{ "CustomerID": 220, "OrderItems": [{ "ProductID": 1285, "Quantity": 1, "Cost": 120 }], "OrderDate": "05/08/2017" }

Column

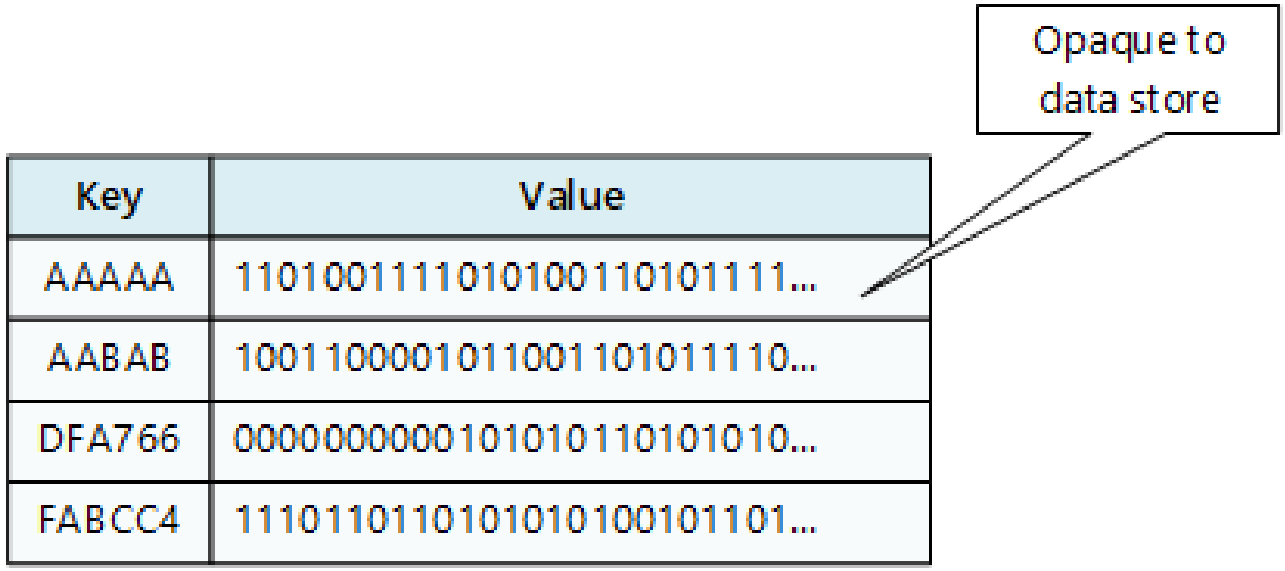
- As bases de dados de família de colunas organizam os dados em linhas e colunas

CustomerID	Column Family: Identity
001	**First name:** Mu Bae **Last name:** Min
002	**First name:** Francisco **Last name:** Vila Nova **Suffix:** Jr.
003	**First name:** Lena **Last name:** Adamcyz **Title:** Dr.

CustomerID	Column Family: Contact Info
001	**Phone number:** 555-0100 **Email:** someone@example.com
002	**Email:** vilanova@contoso.com
003	**Phone number:** 555-0120

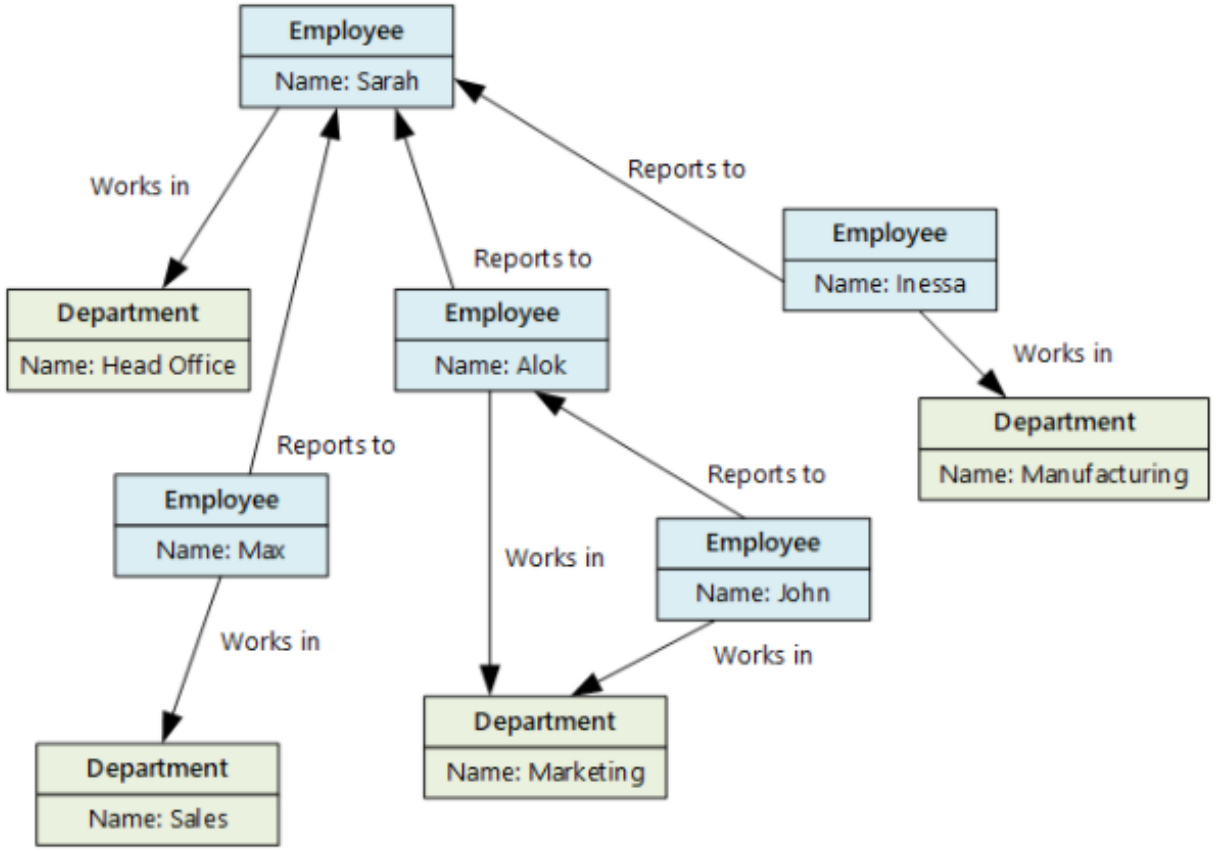
Key-value

- Uma BD chave/valor associa cada valor de dados a uma chave única.



Graph

- As bases de dados de grafos armazenam dois tipos de informações, *nodes* e *edges*
- Os *nodes* especificam relações entre nós
- Os *edges* indicam uma relação podendo ter uma direção que indica a sua natureza



RDBMS vs Non-RDBMS

- Microsoft SQL Server
- Oracle Database
- MySQL
- IBM DB2
- SQL Server Express
- PostgreSQL
- SQLite
- …

- MongoDB
- Apache Cassandra
- Redis
- Couchbase
- Neo4j
- GraphQL
- …

Resumo

RDBMS

- Funciona com dados estruturados
- Relações baseadas em indexes promovendo a integridade dos dados
- Possibilidade de criar indexes promovendo a rapidez na consulta dos dados
- Possibilidade de escrever instruções complexas para análise, manipulação e *reporting*
- A integridade estrutural dos dados promove a integridade aplicacional

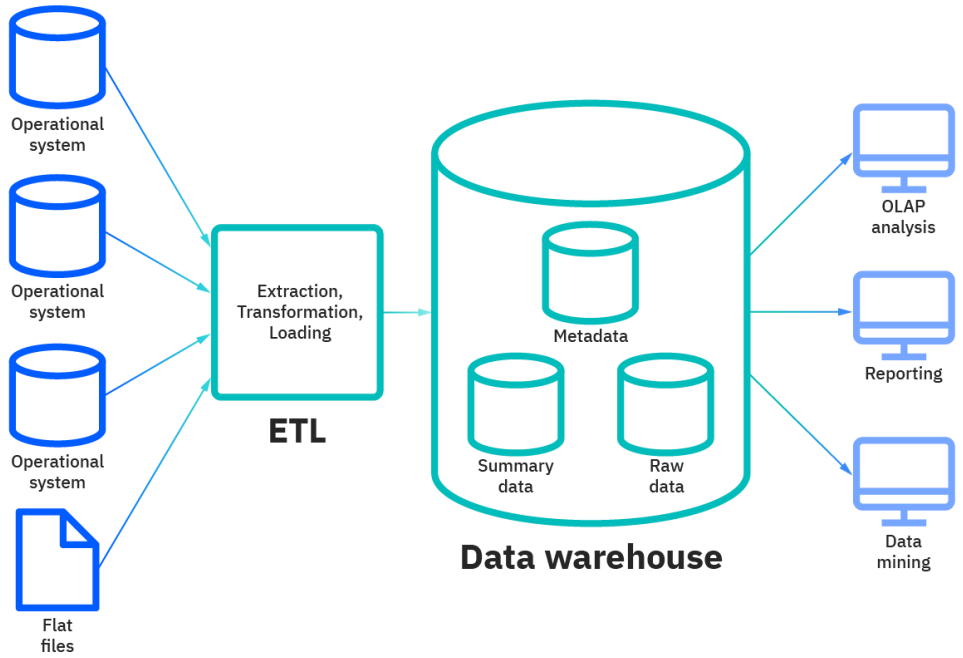
Non-RDBMS

- Consegue-se armazenar bastante informação com uma estrutura menos coesa
- Oferecem maior flexibilidade e escalabilidade
- Oferecem *schema-free*
- Conseguem armazenar todo o tipo de dados incluindo os dados não estruturados

Data Warehouse

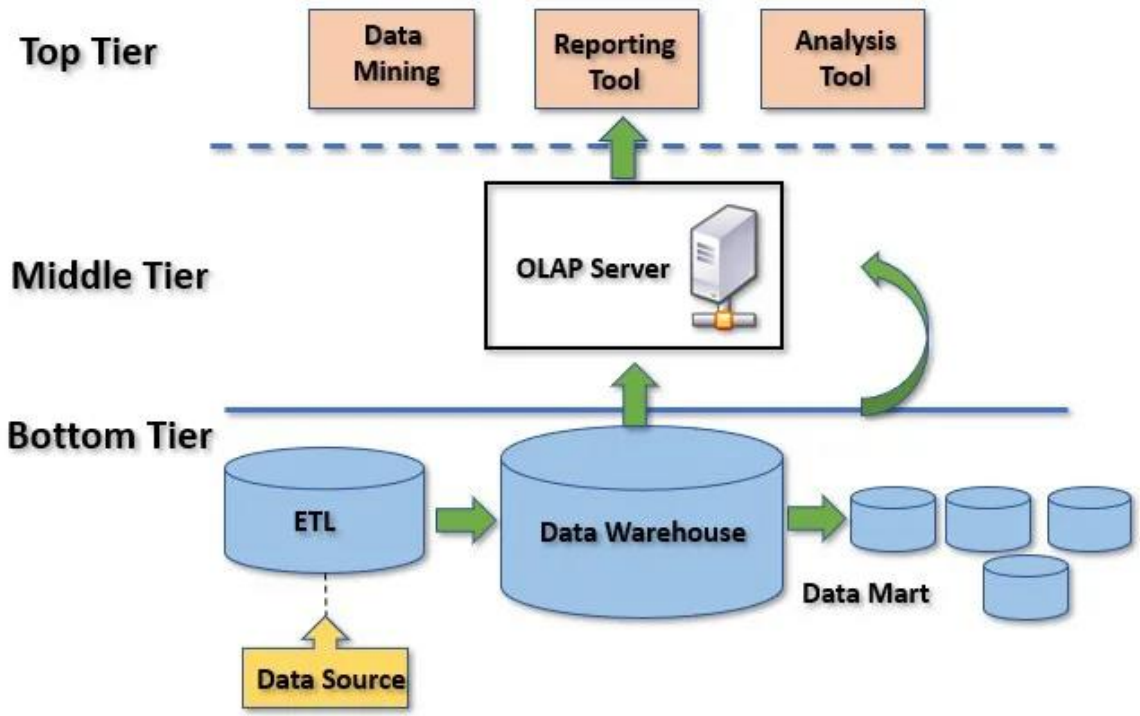
Data Warehouse

- Repositório central onde são guardados todos os dados
- Permite a uma organização centralizar todos os dados e com base nisso elaborar análises avançadas
- Os provedores de *cloud* já possuem este serviços com recursos extremamente avançados



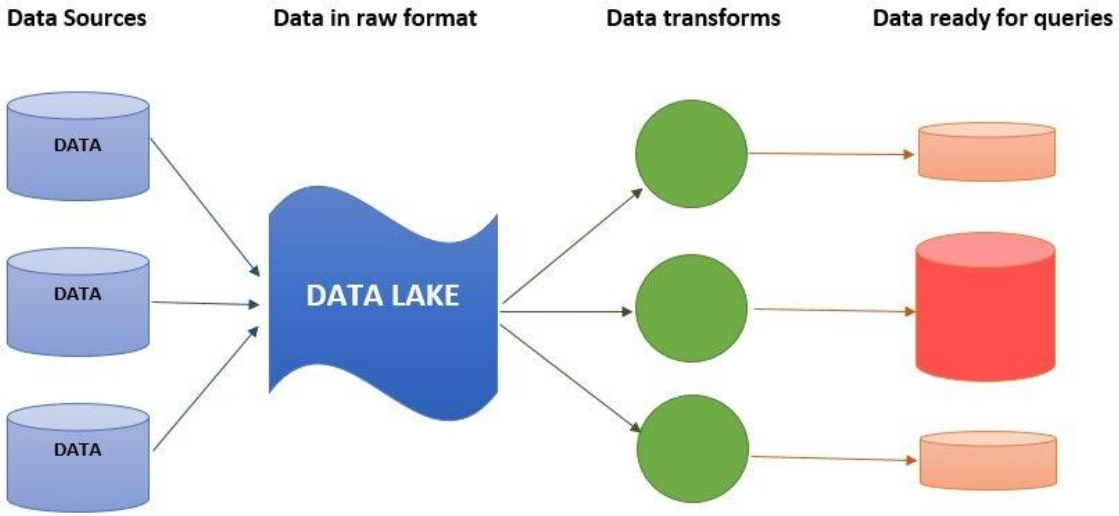
Data Warehouse - Tiers

- Superior
  - Interface de *Frontend*
- Média
  - Servidor OLAP – favorece a performance
- Inferior
  - Recolhe, trata e transforma os dados provenientes de diversas fontes



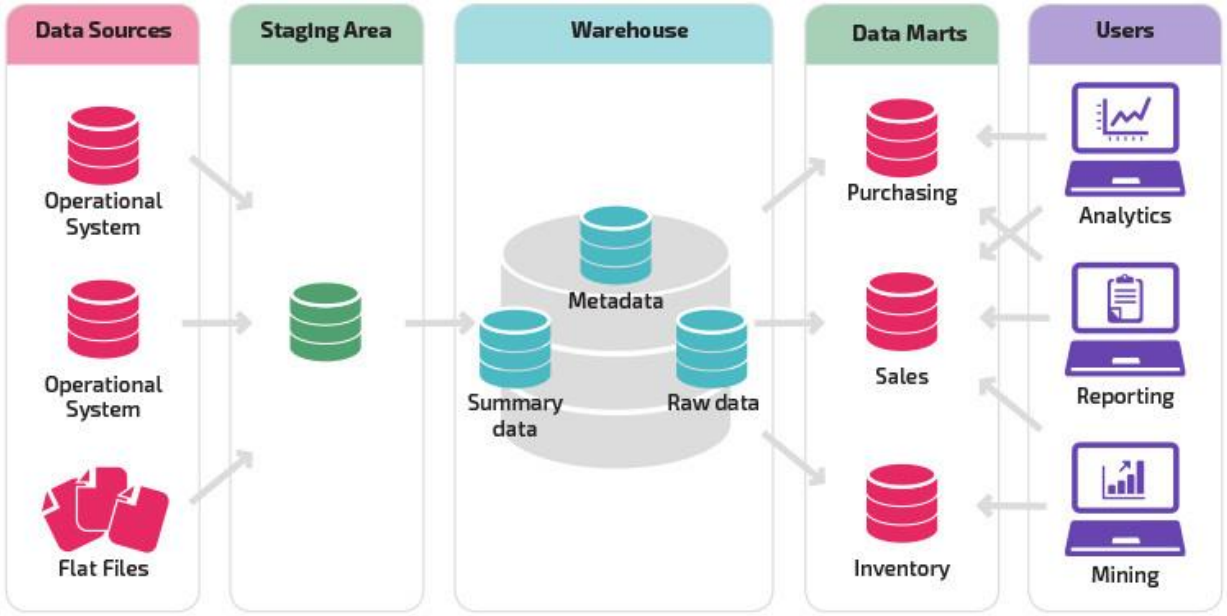
Data Lake

- Repositório com diferentes fontes e de diferentes formatos
- Igual ao Data Warehouse mas sem as BD relacionais
- Mais versátil - Permite outro tipo de análise de dados
- Construído em Apache Hadoop



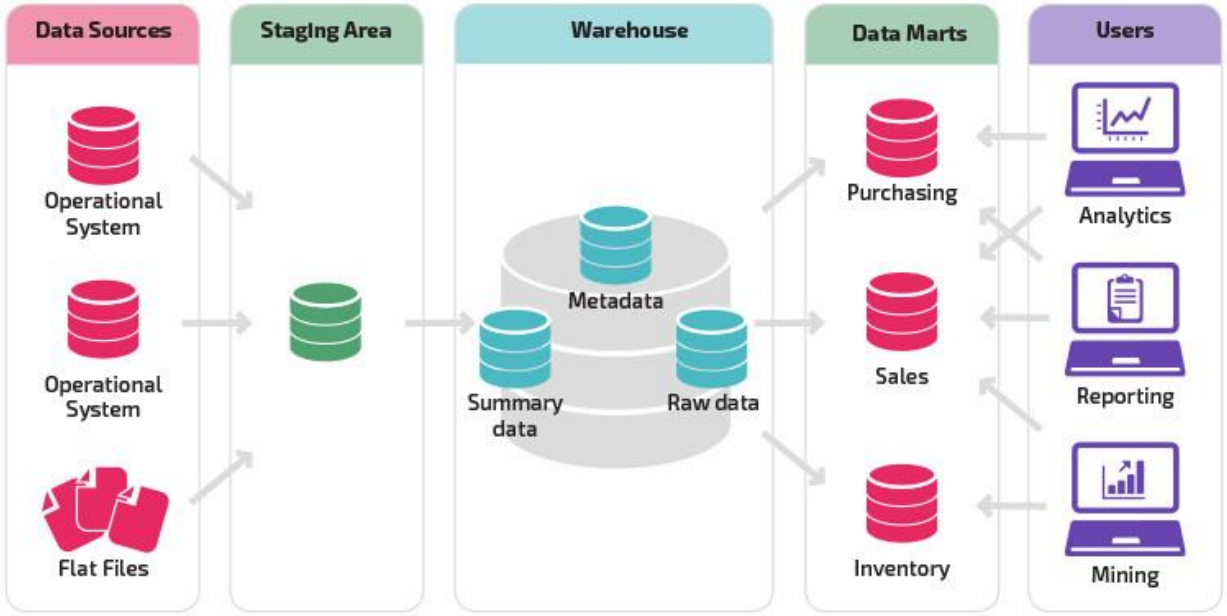
Data Mart

- Subset/tipo de *Data Warehouse*
- Criado com o propósito de responder às necessidades de determinados utilizadores
- Maior agilidade e rapidez na pesquisa de informação
- Por exemplo: Marketing

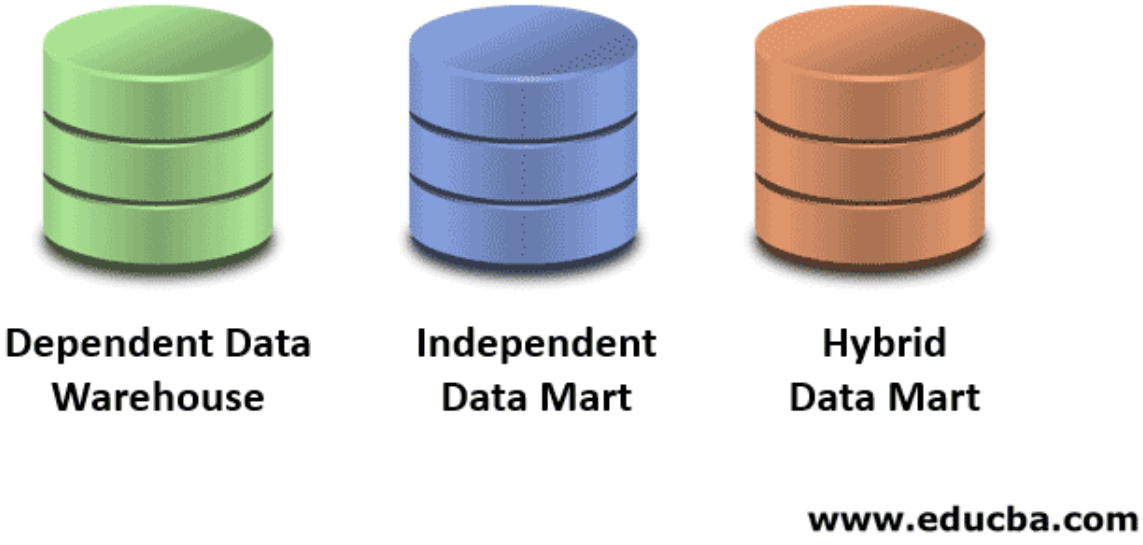


Data Mart – Benefícios

- Sistema menos dispendioso
- Acesso a dados mais simples
- Acesso a informações com mais performance
- Manutenção mais simples
- Simples e rápida implementação



Data Mart – Tipos



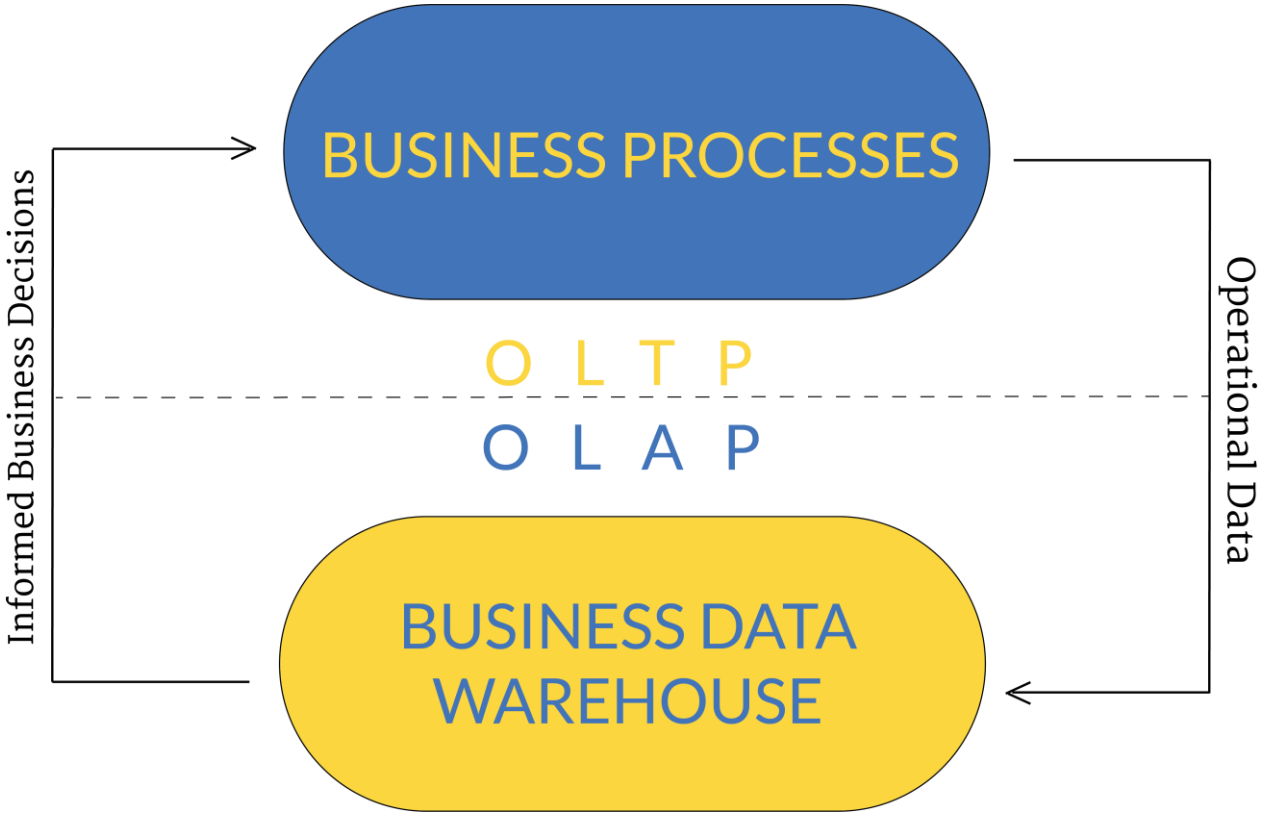
- Dependente
  - Parte de uma *data warehouse* já existente. Construída com o objetivo de dividir o grande problema
- Independente
  - Contruída para acesso rápido a informação. Não existe uma *data warehouse* criada
- Hibrida
  - Combinação de diversas *data warehouses*

OLTP e OLAP

OLTP – Online Transaction Processing

OLAP – Online Analytical Processing

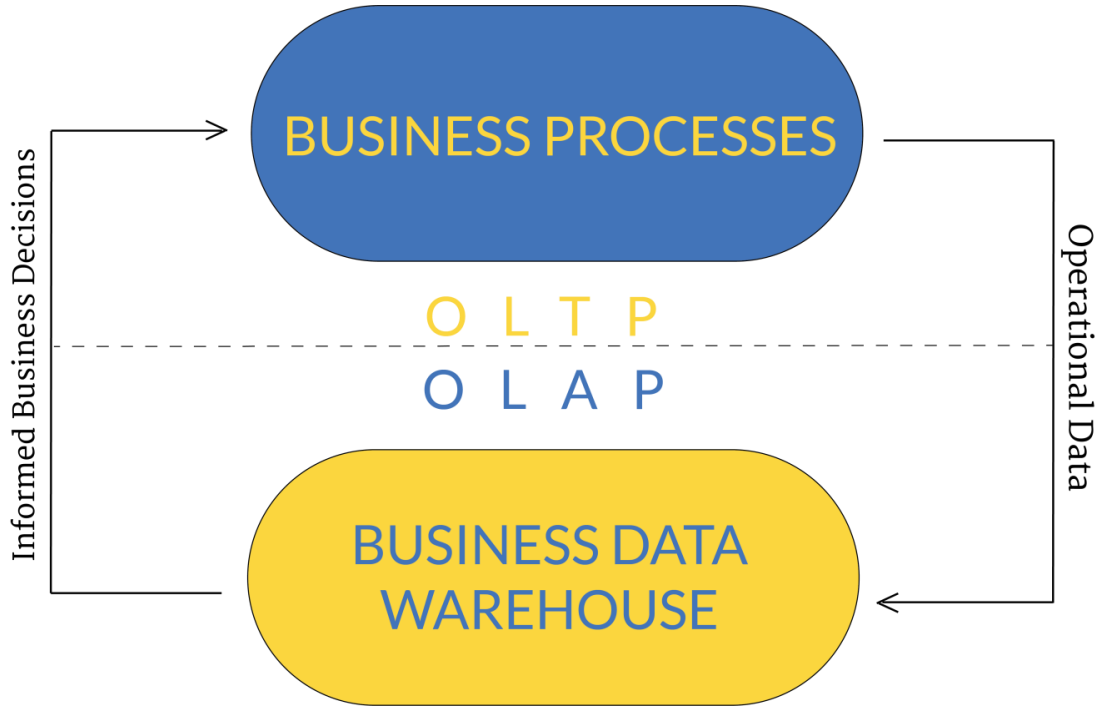
OLTP – Online Transaction Processing



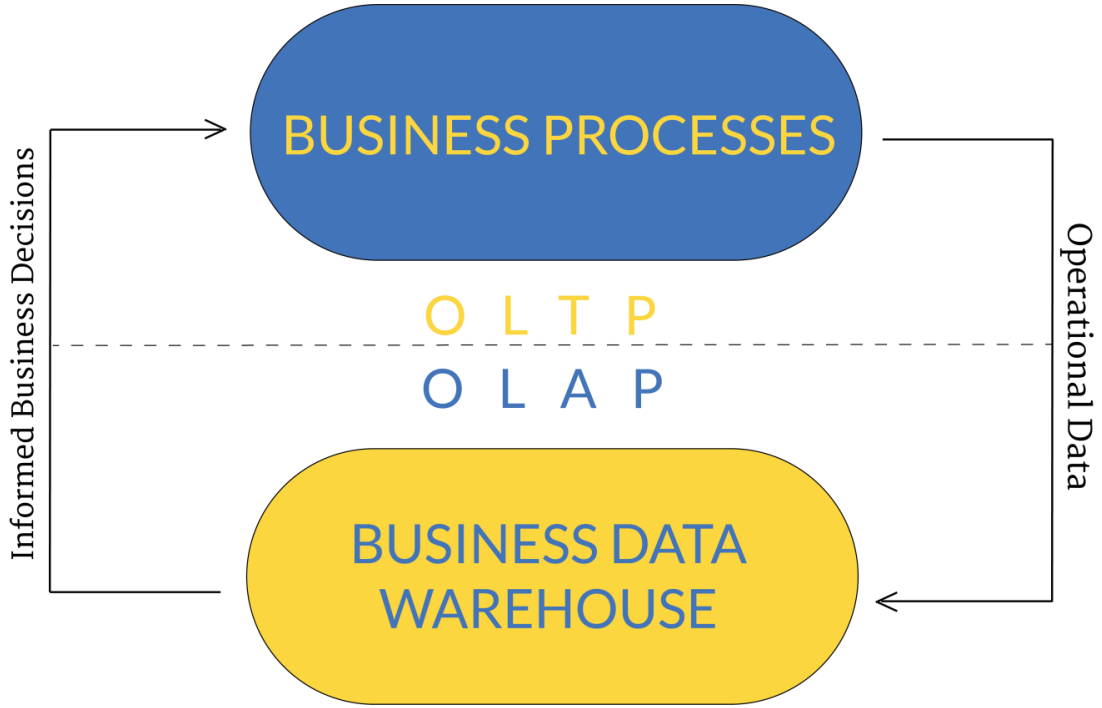
- Update, Insert, Delete são os comandos mais executados
- OLTP permite transações em tempo real e em larga escala
- As *queries* OLTP são curtas, simples e requerem menos tempo de processamento
- Exemplo: Compra de voos, compras online, levantamentos multibando

OLAP – Online Analytical Processing

- Usado para obtenção de dados, com alta performance, que foram guardados centralmente
- Armazena os dados que foram inseridos pelo OLTP
- Permite que os utilizadores realizem uma análise multidimensional de dados
- Exemplo: *data mining*, cenários preditivos, analise matemática, etc…



OLTP vs OLAP



Characteristics	OLTP (Transactional system)	OLAP (decision support system)
Application	Ordinary management, production	Analysis / Decision-support
Users	Information system experts	Decision-makers
Data schema	Entity / Relationship	Star / Snowflake / Constellation
Normalization	Frequent	Scare
Data	Up to date / Raw	Archived / Aggregated
Up dating	Immediate / Real time	Delay or postpone
Queries	Simple / Regular / Predefine / Predictable	Complex / Irregular / Non-Predictable / Ad-hoc
Query language	SQL, QBE, QUEL	MDX, XQuery, XMLA
Analysis axis	Uni- or bi-directional	Multidimensional or multi-axes
Operations	Modification / Up to date / Cancelling / Insertion	Lecture / Cross analysis / Refreshment
Data size	Mega or Gigabytes	Tera, Peta or Zetabytes

OLTP vs OLAP

OLTP characteristics:

- Many transactions
- Latency sensitive
- Small payloads
- Balanced read/write or Heavy write workloads

OLAP characteristics:

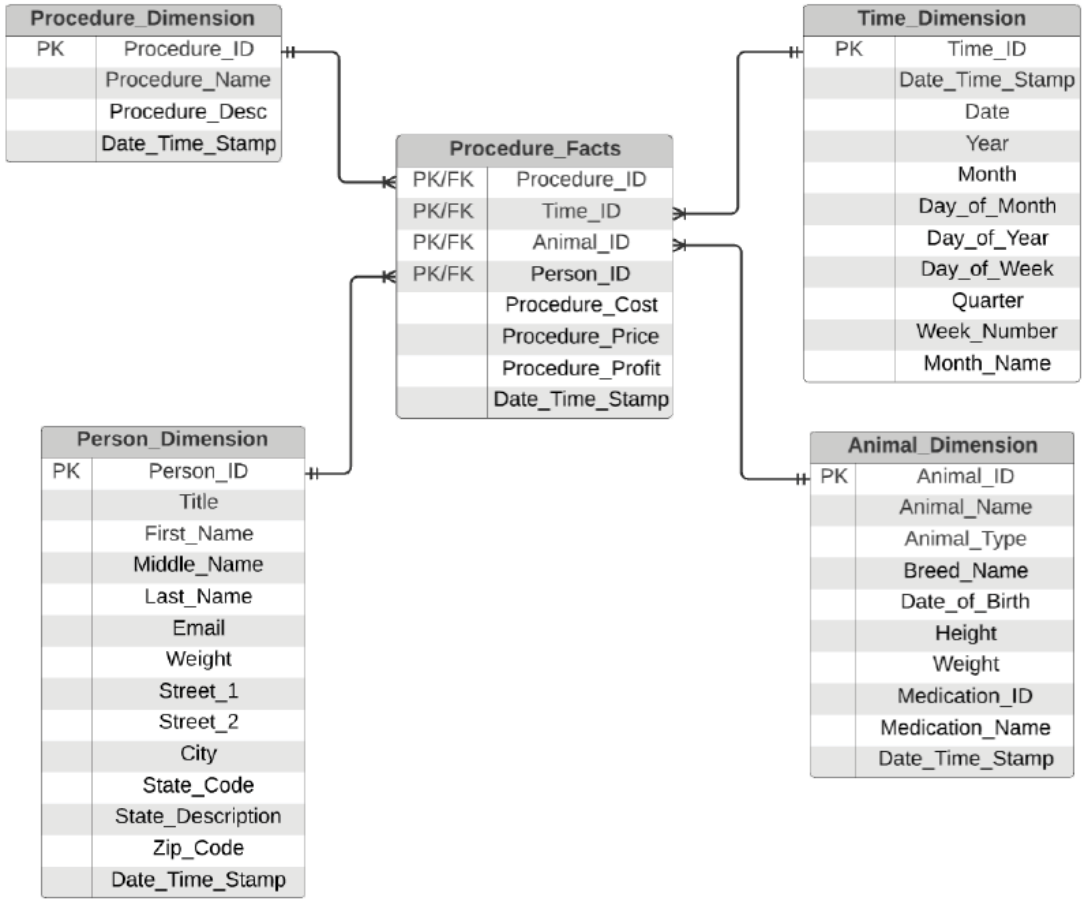
- Few transactions
- Throughput sensitive
- Large (return) payloads
- Heavy read workloads (including full table scans)

Schemas Concepts

Star

Snowflake

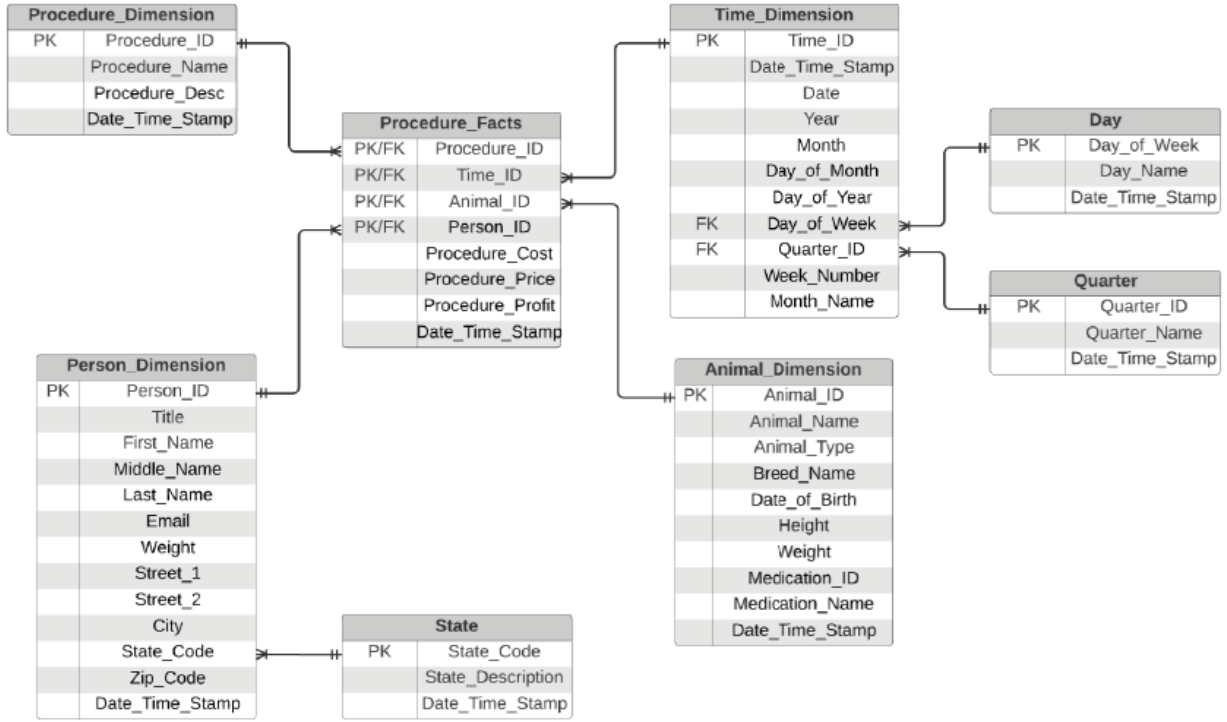
Star Schema



- Contém uma tabela de *facts* que relaciona eventos de negocio específicos.
- Esta tabela reside no centro de toda as tabelas
- De forma geral não existe dependências entre as tabelas *dimension*. Requer menos JOINS
- Esta estrutura é extremamente eficiente para análise de dados em *datasets* com grande volumes de dados

Snowflake Schema

- Extensão do modelo Star, mas com mais tabelas *dimension.*
- Existe uma maior normalização dos dados, potenciando a integridade dos dados e reduzindo a redundância.
- Requer menos espaço de armazenamento das tabelas *dimension*
- Modelo mais complexo e de difícil manutenção



Pequenas alterações

Type 1: Update Changes

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	IL

Type 2: Keep Historical

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date
123	ABC	Acme Supply Co	CA	01-Jan-2000	21-Dec-2004
124	ABC	Acme Supply Co	IL	22-Dec-2004	

Os dados mudam e a informação que queremos obter também muda!

Type 3: Preserve Limited History

Supplier_Key	Supplier_Code	Supplier_Name	Original_Supplier_State	Effective_Date	Current_Supplier_State
123	ABC	Acme Supply Co	CA	22-Dec-2004	IL

Tipo 1

Chave Dimensão	Código Item	Nome Setor	Nome Responsável
001	25	Coordenação de Produtos	Carlos José

Chave Dimensão	Código Item	Nome Setor	Nome Responsável
001	25	Coordenação de Produtos	João Pereira

- Re-escreve um novo valor na tabela
- Não mantém o histórico
- Usado quando o antigo valor não é importante para a informação

Tipo 2

- Adiciona uma nova linha de informação

Chave Dimensão	Código Item	Nome Setor	Nome Responsável
001	25	Coordenação de Produtos	Carlos José

- Mantém o histórico

- Forma mais utilizada porque permite realizar análises com base na evolução ocorrida

Chave Dimensão	Código Item	Nome Setor	Nome Responsável
001	25	Coordenação de Produtos	Carlos José
002	25	Coordenação de Produtos	João Pereira

Tipo 3

Chave Dimensão	Código Item	Nome Setor	Nome Responsável
001	25	Coordenação de Produtos	Carlos José

Chave Dimensão	Código Item	Nome Setor	Nome Responsável	Nome Responsável Atual
001	25	Coordenação de Produtos	Carlos José	João Pereira

- Cria uma nova coluna, mantendo a informação original
- Mantém o histórico
- Porém, quando os restantes atributos forem alterados irá perder informação relativa ao histórico

Bibliografia

- B. Gomez,(2020) “Resolviendo problemas de Big Data”, Alfaomega.
- D. Insua, (2019)“Big data: Conceptos, tecnologías y aplicaciones”, CSIC.
- H. Jones, (2019)“Analítica de datos”, HJ,.
- J. Somed, (2020)“Big Data Analytics”, JLC.
- D. Petković (2020)“Microsoft® SQL Server® 2019 A Beginner’s Guide - Seventh Edition”, McGraw Hill.