

Análise de Dados

7.^a Aula Prática Laboratorial

Mestrado Integrado em Engenharia Informática

Ano Letivo 2018/2019

Marisa Esteves

16 de Novembro de 2018



Universidade do Minho

Plano de Aula

1. Início da resolução da 5.^a ficha prática laboratorial pelos alunos em grupo.

Processo ETL

Definição

O processo ETL (*Extract, Transform, Load*) é um conjunto de processos que inclui a extração de dados de fontes de informação internas e externas, podendo estar em diferentes formatos, a transformação dos dados de acordo com as necessidades da organização e, finalmente, o carregamento dos mesmos numa estrutura de dados, como por exemplo um data mart ou um data warehouse.

Processo ETL

Definição

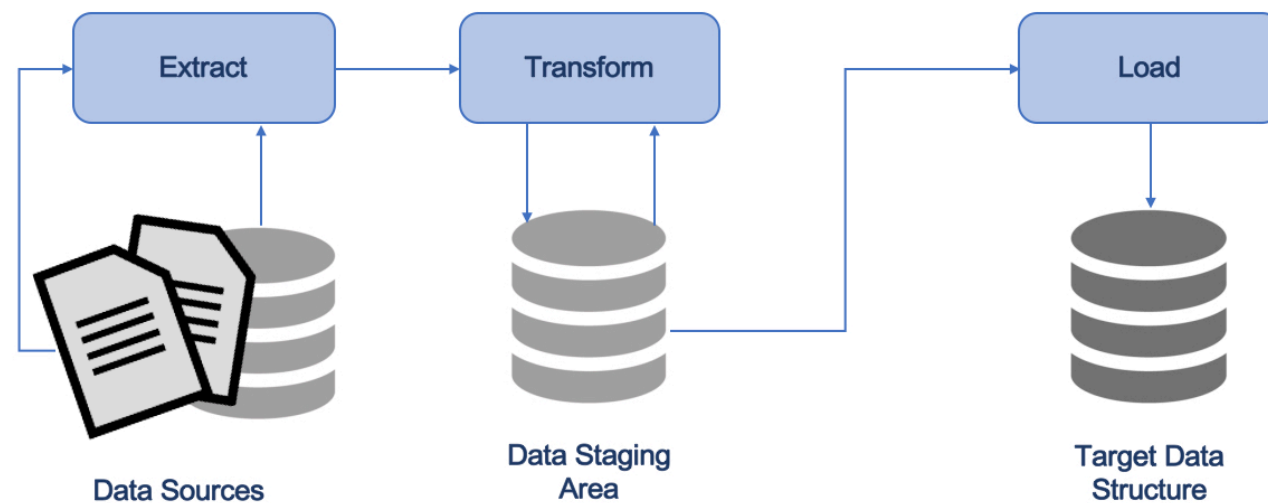


Figura 1 – Esquema do processo ETL.

Processo ETL

Porquê?

Os dados estão espalhados
por diferentes localizações

Os dados estão
armazenados em diferentes
tipos de formato

O volume de dados continua
a aumentar

Os dados podem estar
estruturados, semi-
estruturados ou não
estruturados

Data Warehousing

Definição

O processo de data warehousing enfatiza à recolha de dados de diversas fontes através do processo ETL (*Extract, Transform, Load*), correspondendo à construção de data warehouses e/ou data marts, para aceder e analisar a informação de forma útil. Os dados extraídos são processados, formatados e consolidados numa estrutura de dados única para facilitar essencialmente a análise de dados.

Data Warehousing

Definição

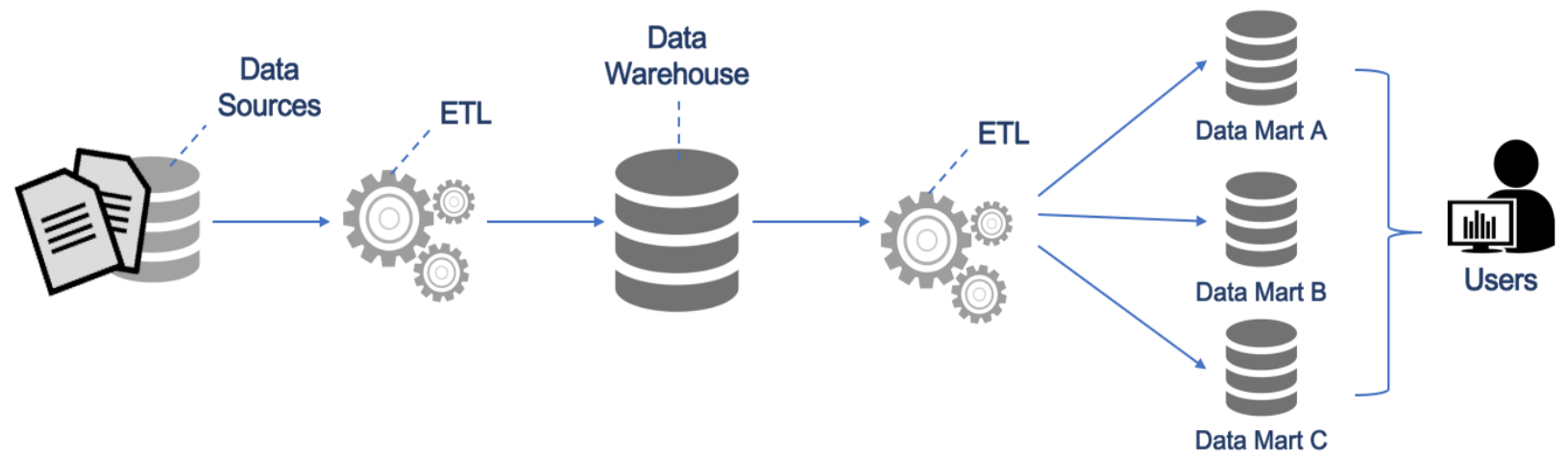


Figura 2 – Esquema do processo de data warehousing.

Data Warehousing

*Data Warehouse vs. Data
Mart*

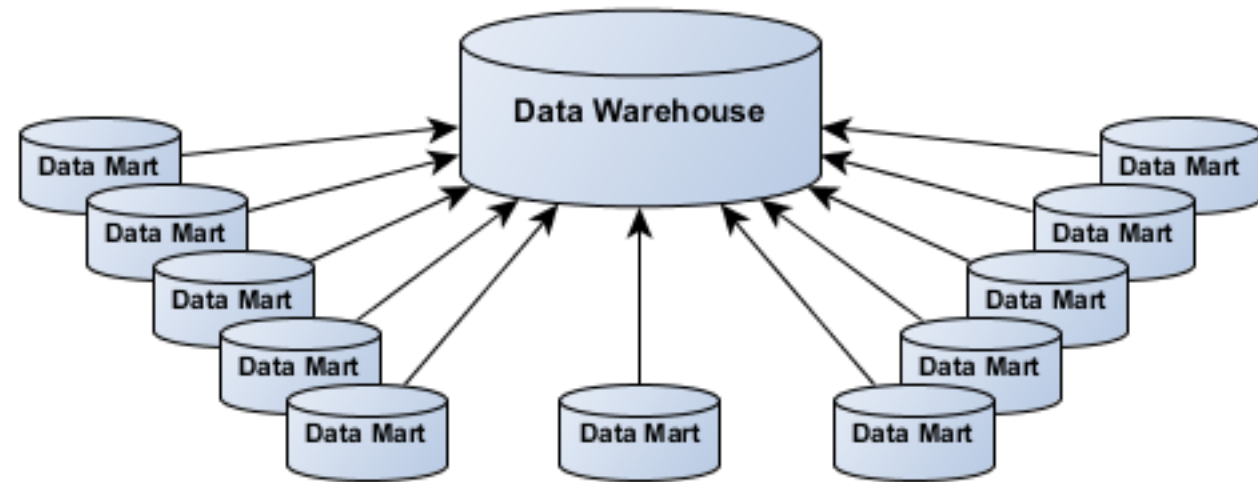


Figura 3 – Data warehouse vs. Data marts.

Data Warehousing

*Modelo Dimensional –
Esquema em Estrela vs.
Esquema em Floco de Neve*

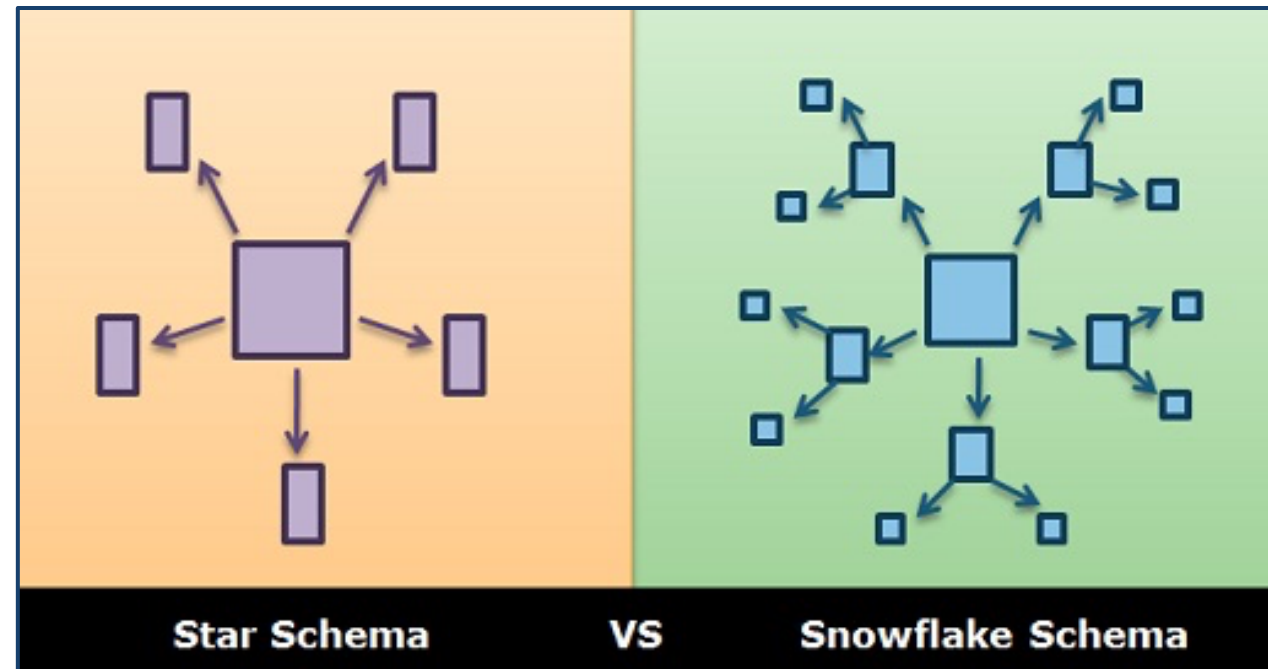


Figura 4 – Esquema em Estrela vs. Esquema em Floco de Neve.

Data Warehousing

*Modelo Dimensional –
Esquema em Constelação de Factos*

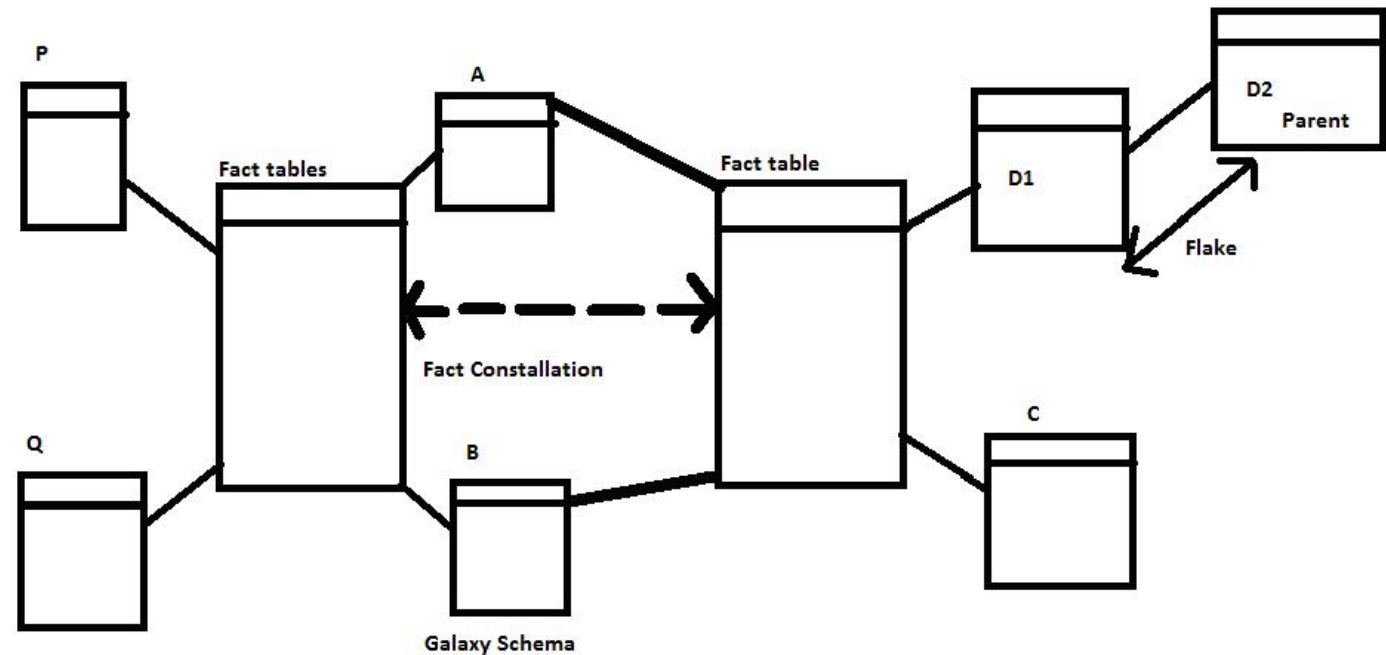


Figura 5 – Esquema em Constelação de Factos.

OLTP vs. OLAP

Definição

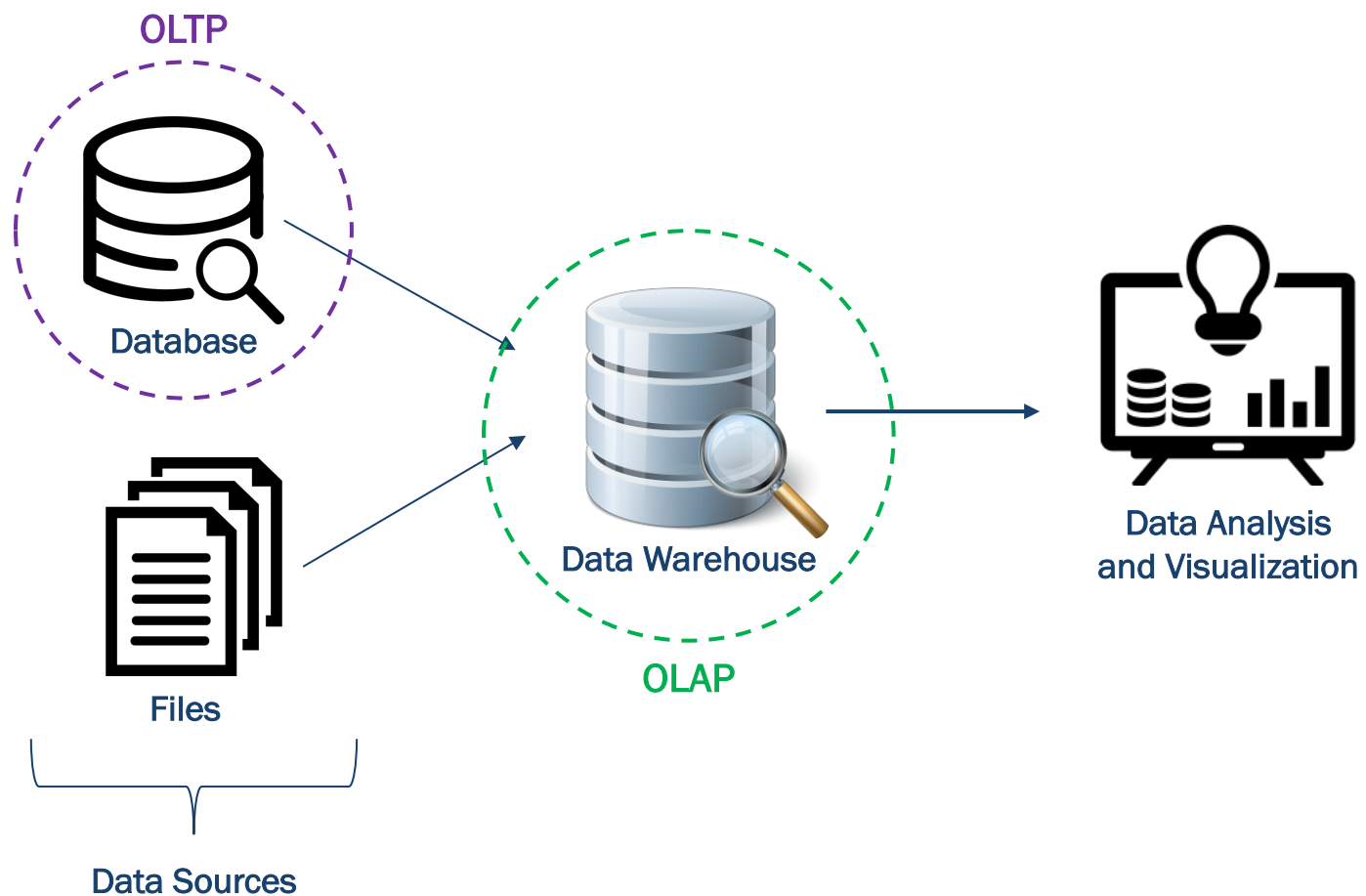


Figura 6 – OLTP (*Online Transaction Processing*) vs. OLAP (*Online Analytical Processing*).

OLTP vs. OLAP

Definição

Relational Database (OLTP)	Analytical Data Warehouse (OLAP)
Contains current data	Contains historical data
Useful in running the business	Useful in analysing the business
Based on Entity Relationship Model	Based on Star, Snowflake or Galaxy Schema
Provides primitive and highly detailed data	Provides summarized and consolidated data
Used for writing into the database	Used for reading data from the data warehouse
Database size ranges from 100 MB to 1 GB	Data warehouse ranges from 100 GB to 1 TB
Fast and it provides high performance	Highly flexible but it is not fast
Number of records accessed is in tens	Number of records accessed is in millions
Example: all bank transactions made by a customer	Example: bank transactions made by a customer at a particular time

Figura 7 – Diferenças entre OLTP e OLAP.

Resolução da 5.ª Ficha Prática Laboratorial

1 Modelação Dimensional em Constelação de Factos

O principal objetivo da resolução da segunda parte deste exercício é proceder ao povoamento do *data warehouse* definido e implementado na 4.ª ficha prática laboratorial (carregamento inicial), bem como, seguidamente, à gestão dos seus processos.

O *data warehouse* deverá ser povoado recorrendo à base de dados sakila, bem como ao ficheiro *calendario.xlsx*, disponibilizados durante as aulas teóricas e práticas laboratoriais desta unidade curricular.

É de notar que pode consultar mais informação de apoio sobre a base de dados sakila disponibilizada na seguinte referência: <https://dev.mysql.com/doc/sakila/en/>.

Com base no caso apresentado, pretende-se que:

1. Defina o mapa lógico de dados para o povoamento do *data warehouse* definido e implementado.
2. Crie uma nova base de dados denominada “sakila_dsa” no MySQL Workbench.
3. Faça o *import* dos dados no ficheiro *calendario.xlsx* para uma nova tabela denominada “calendario” na base de dados criada no passo anterior.
4. Povoie o *data warehouse* em SQL recorrendo à base de dados sakila, bem como à tabela *calendario* da base de dados sakila_dsa. No entanto, deverá ter em consideração os seguintes dois pontos:
 - (a) Adicione uma nova coluna denominada “etiqueta_DTA” em cada tabela do modelo dimensional de modo a guardar o *datetime* de povoamento de cada linha.
 - (b) Deverá antes povoar todas as suas tabelas de dimensão e, seguidamente, as suas duas tabelas de factos.
5. Defina e implemente os processos que acha necessários para a gestão do *data warehouse*.