Análise de Dados

5.ª Aula Prática Laboratorial

Mestrado Integrado em Engenharia Informática

Ano Letivo 2018/2019

Marisa Esteves

2 de Novembro de 2018

Universidade do Minho

Plano de Aula

1. Resolução da 3.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

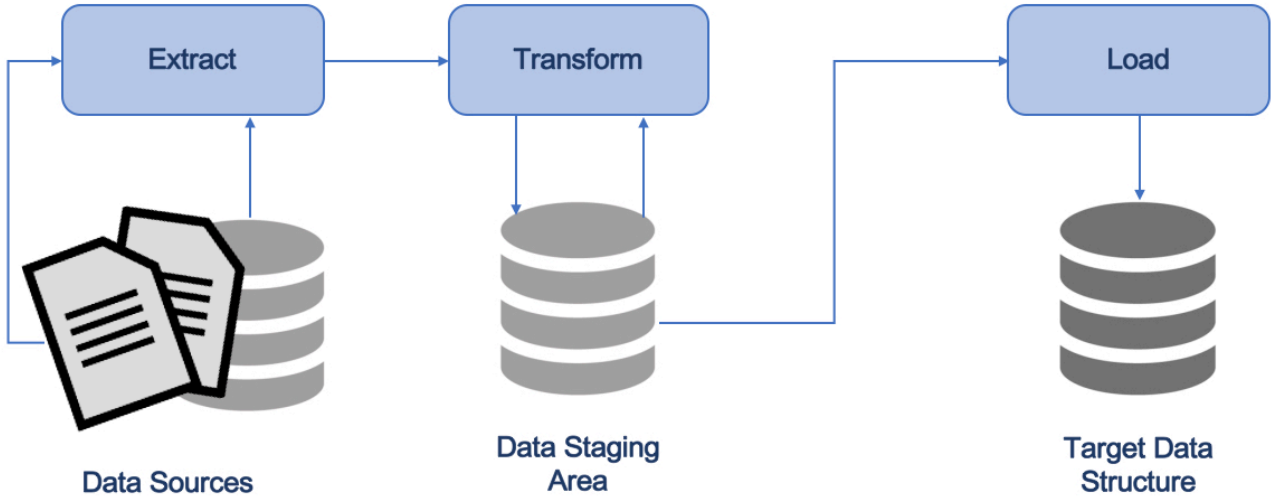

Figura 1 – Esquema do processo ETL.

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

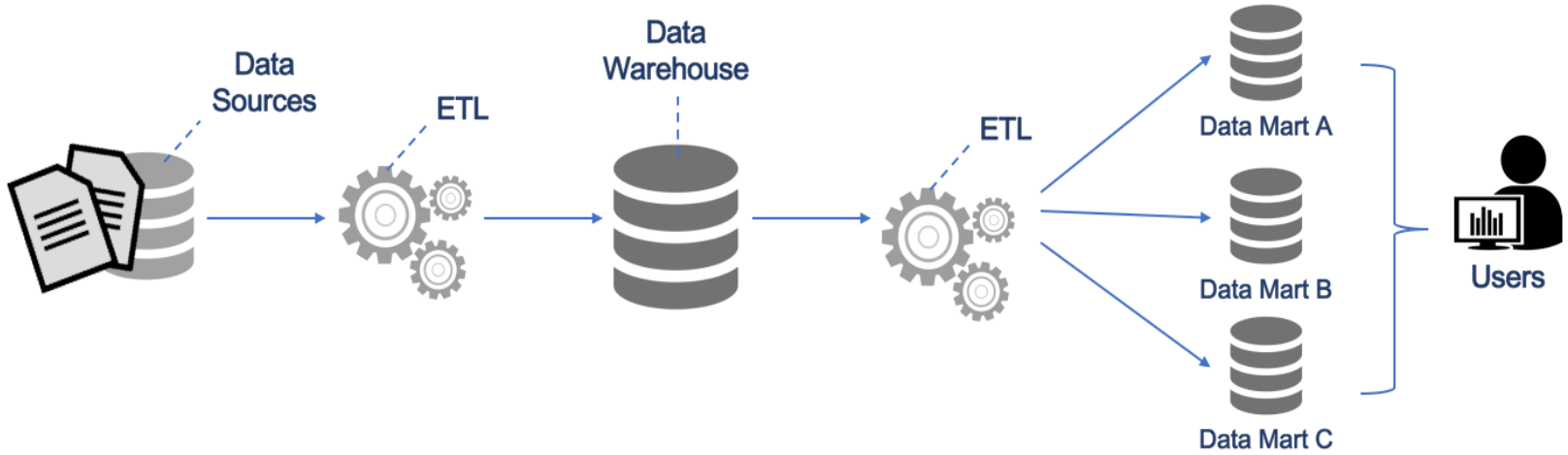

Figura 2 – Esquema do processo de data warehousing.

Data Warehousing

Data Warehouse vs. Data Mart

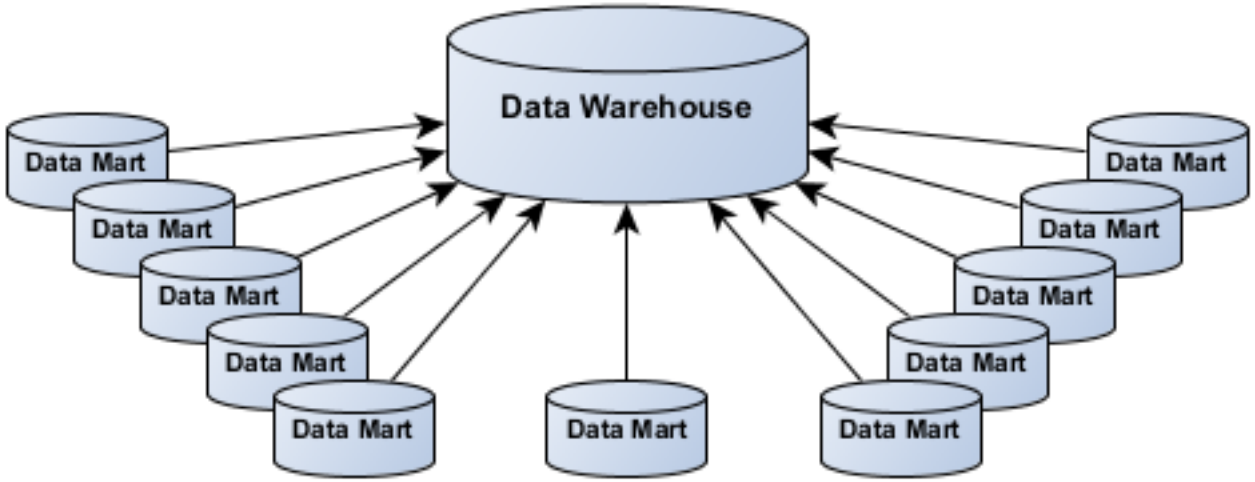

Figura 3 – Data warehouse vs. Data marts.

Data Warehousing

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

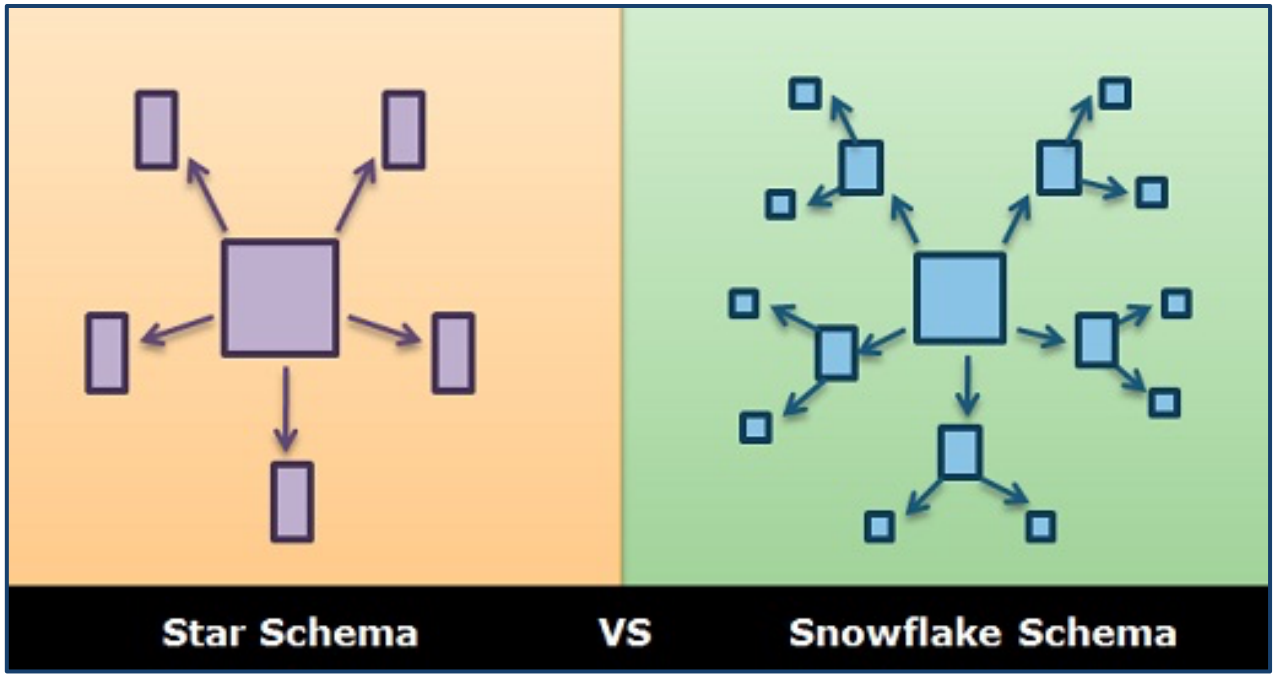

Figura 4 – Esquema em Estrela vs. Esquema em Floco de Neve.

Data Warehousing

Modelo Dimensional – Esquema em Constelação de Factos

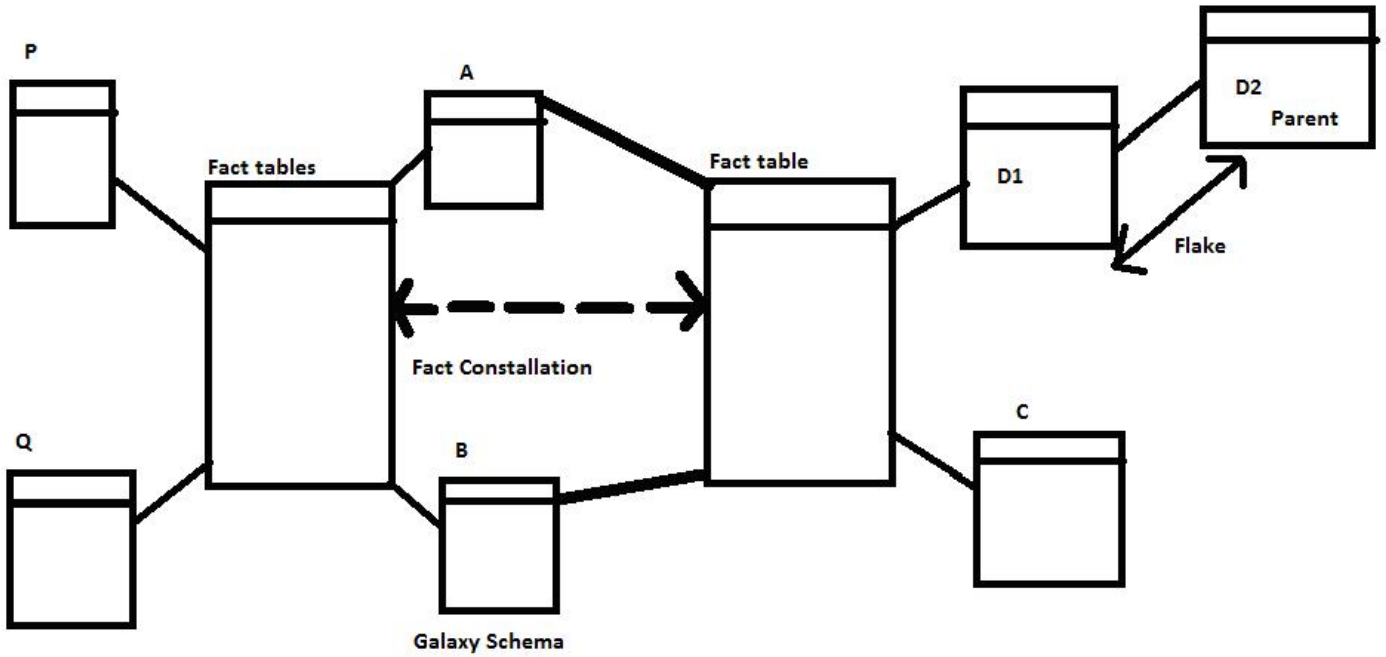

Figura 5 – Esquema em Constelação de Factos.

OLTP vs. OLAP

Definição

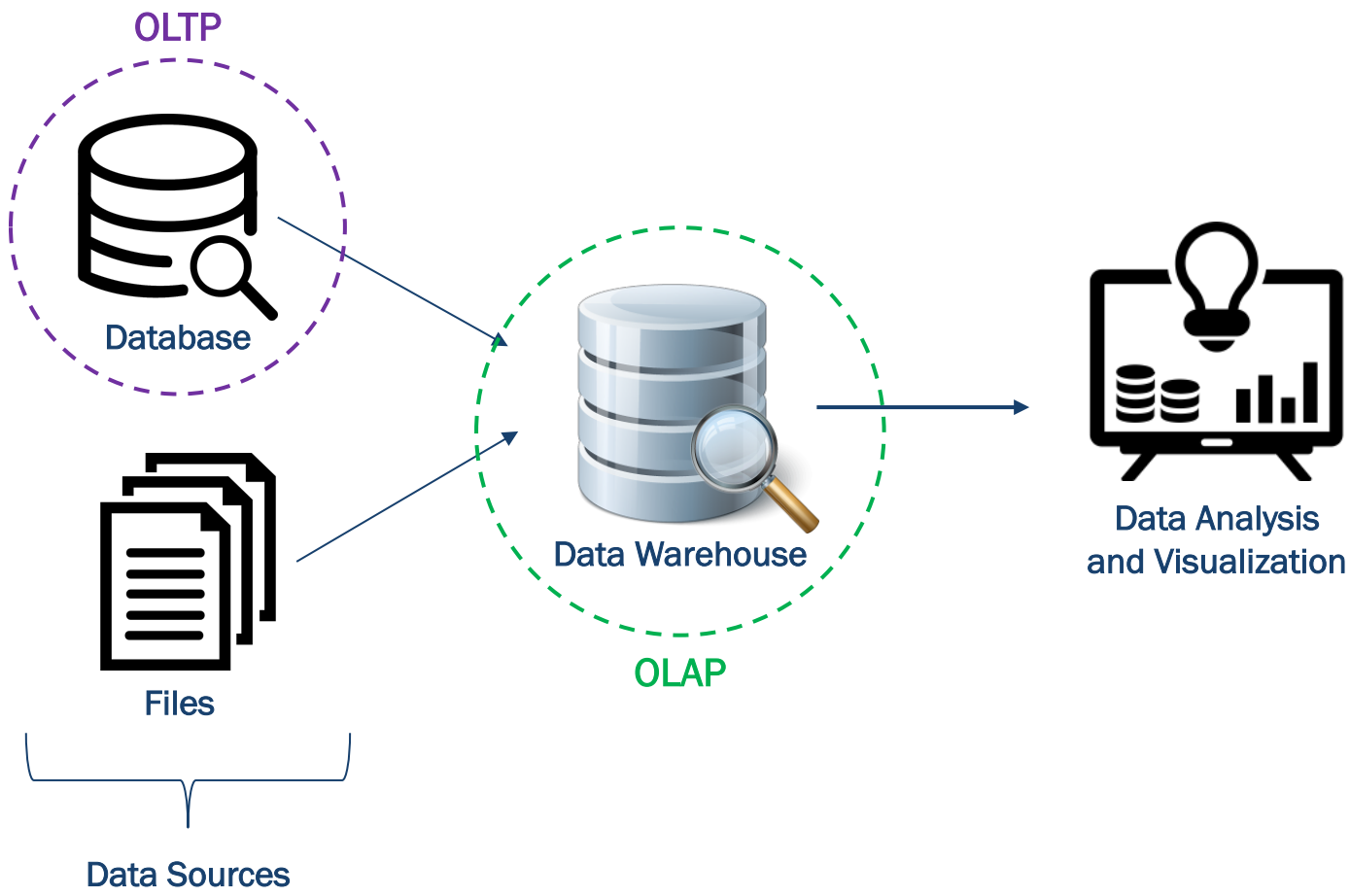

Figura 6 – OLTP (*Online Transaction Processing*) vs. OLAP (*Online Analytical Processing*).

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 3.ª Ficha Prática Laboratorial

1 Esquema em Estrela vs. Esquema em Floco de Neve

O assassinato em 2014 de Michael Brown em Ferguson, Missouri, Estados Unidos da América (EUA), iniciou um movimento de protesto que culminou com o *Black Lives Matter* e um foco maior na responsabilidade dos polícias em todo o país.

Desde o dia 1 de Janeiro de 2015, *The Washington Post* tem vindo a recolher dados numa base de dados relativos a todos os disparos fatais nos EUA por um polícia durante o seu cumprimento de dever legal.

É interessante referir que é difícil encontrar dados confiáveis antes do dia 1 de Janeiro de 2015 uma vez que este tipo de acontecimentos não era documentado de forma abrangente, e estatísticas sobre a brutalidade policial estão ainda muito menos disponíveis. Como resultado, um grande número deste tipo de casos não está relatado.

The Washington Post está a recolher mais de uma dúzia de detalhes sobre cada assassinato, incluindo a raça, a idade e o género do falecido, se a pessoa estava armada e se a vítima estava num estado de crise (saúde mental).

O ficheiro disponibilizado juntamente com esta ficha prática laboratorial, nomeadamente police_killings_-us.csv, contém os dados reais dessa recolha realizada pelo *The Washington Post*. Cada linha do ficheiro corresponde a um disparo fatal por um polícia nos EUA desde 2015. A informação representada inclui 14 colunas, nomeadamente: id (identificador único de cada disparo fatal), name (da vítima), date (da morte da vítima), manner_of_death, armed, age, gender, race, city, state, signs_of_mental_illness, threat_level, flee e body_-camera.

Assim, este conjunto de dados representa uma oportunidade única para fazer questões relevantes sobre a brutalidade policial nos últimos anos nos EUA.

Resolução da 3.ª Ficha Prática Laboratorial

Com base no caso apresentado, pretende-se que:

1. Analise a estrutura da tabela police_killings_us.csv e, consequentemente, define um modelo dimensional no formato de esquema em estrela.
2. Analise a estrutura da tabela police_killings_us.csv e, consequentemente, define um modelo dimensional no formato de esquema em floco de neve.
3. Construa cada um dos modelos dimensionais definidos nas alíneas anteriores no MySQL Workbench (*EER diagram*).
4. Descreva as vantagens e as desvantagens entre os dois diferentes tipos de modelo dimensional.
5. Defina dez questões de interesse que poderia colocar a esta base de dados. Justique cada uma das questões definidas.