Análise de Dados

7.ª Aula Prática Laboratorial

Mestrado Integrado em Engenharia Informática

Ano Letivo 2019/2020

Marisa Esteves

15 de Novembro de 2019

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

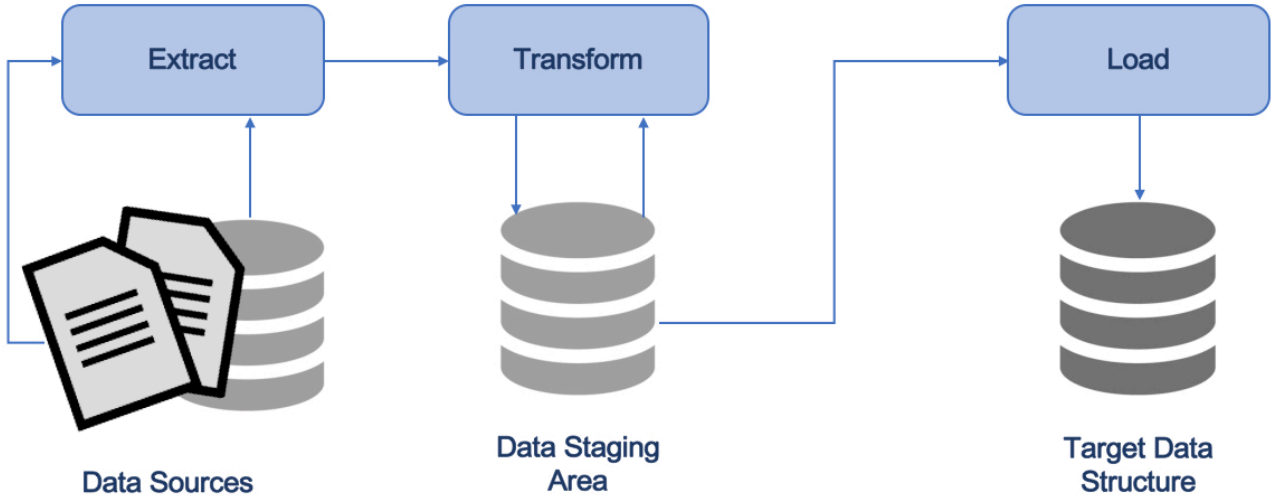

Figura 1 – Esquema do processo ETL.

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

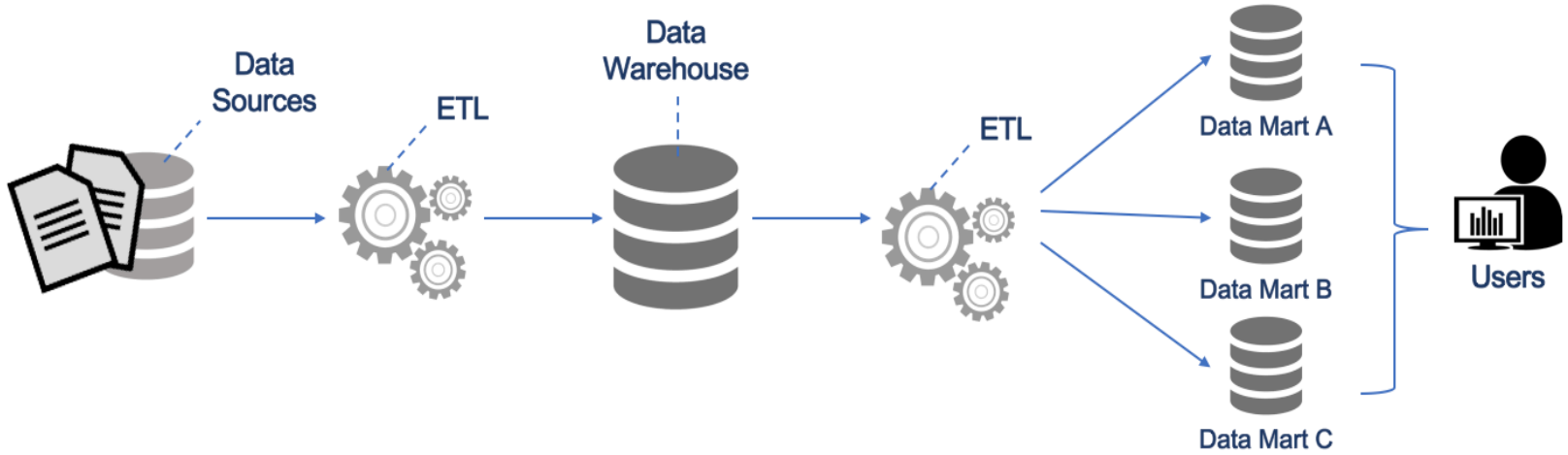

Figura 2 – Esquema do processo de data warehousing.

Data Warehousing

Data Warehouse vs. Data Mart

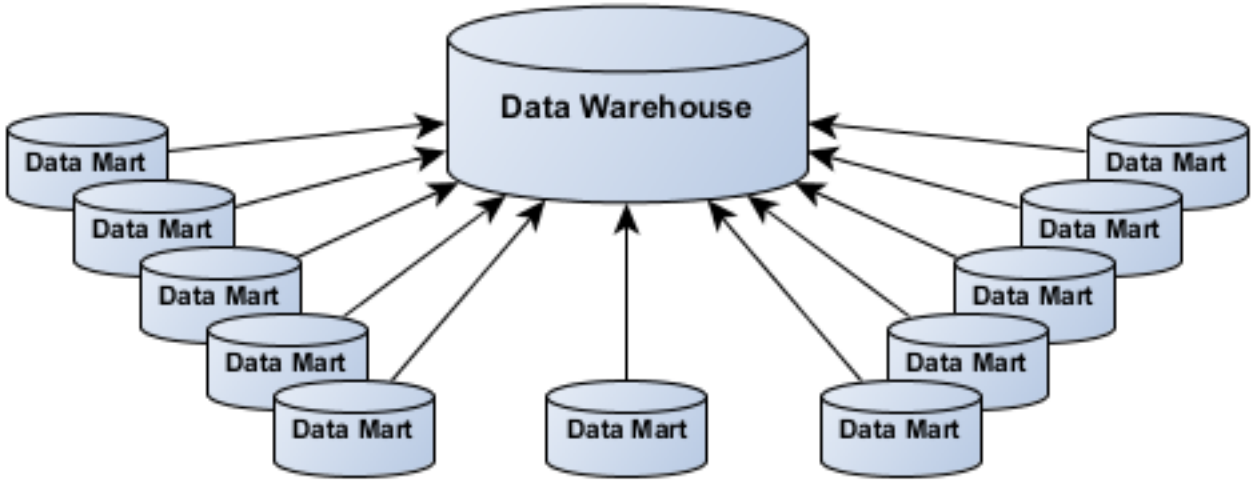

Figura 3 – Data warehouse vs. Data marts.

Data Warehousing

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

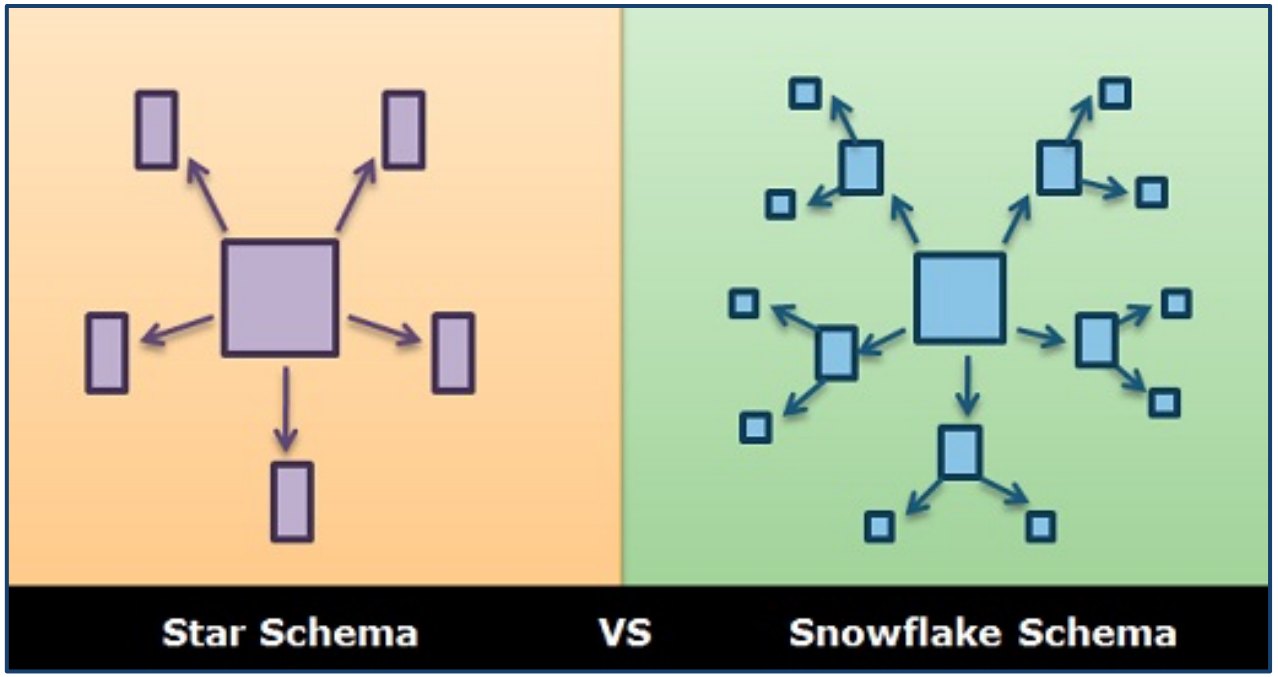

Figura 4 – Esquema em Estrela vs. Esquema em Floco de Neve.

Data Warehousing

Modelo Dimensional – Esquema em Constelação de Factos

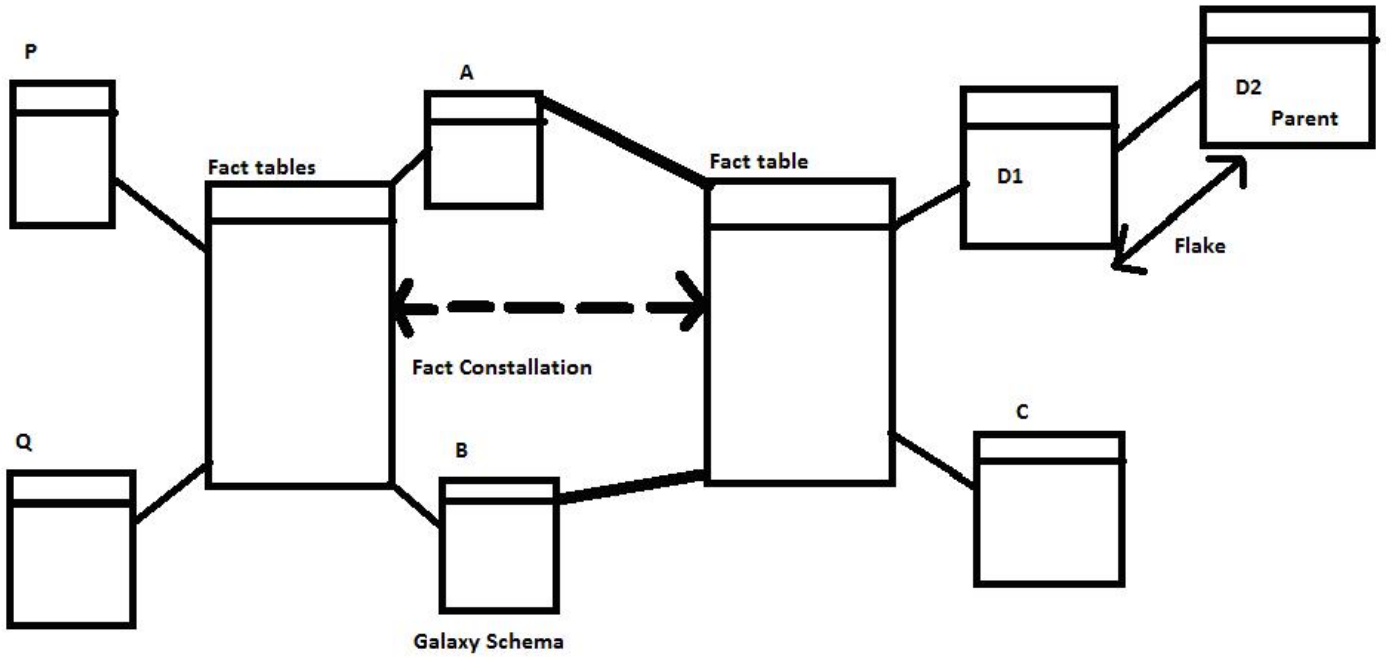

Figura 5 – Esquema em Constelação de Factos.

OLTP vs. OLAP

Definição

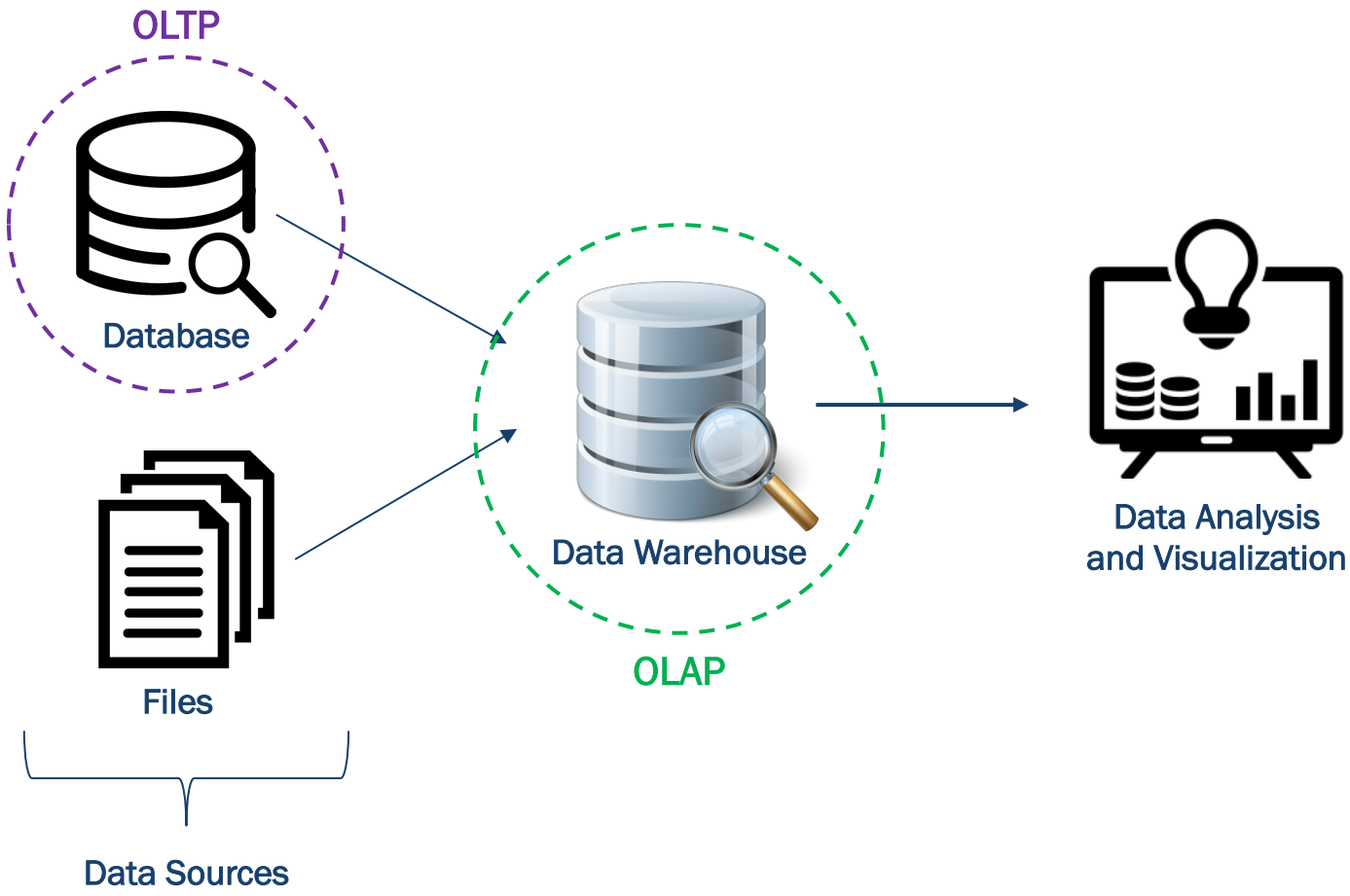

Figura 6 – OLTP (*Online Transaction Processing*) vs. OLAP (*Online Analytical Processing*).

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 Fact Constellation 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.

MySQL

INSERT INTO SELECT FROM

Permite copiar dados de uma tabela e os inserir noutra tabela.
No entanto, este comando SQL requer que os tipos de dados na tabela de origem (table1) e na tabela destino (table2) sejam iguais.

- **INSERT INTO** *table2* (*column1*, *column2*, *column3*, ...)
  SELECT *column1*, *column2*, *column3*, ...
  FROM *table1*
  WHERE *condition*

MySQL

Cursores

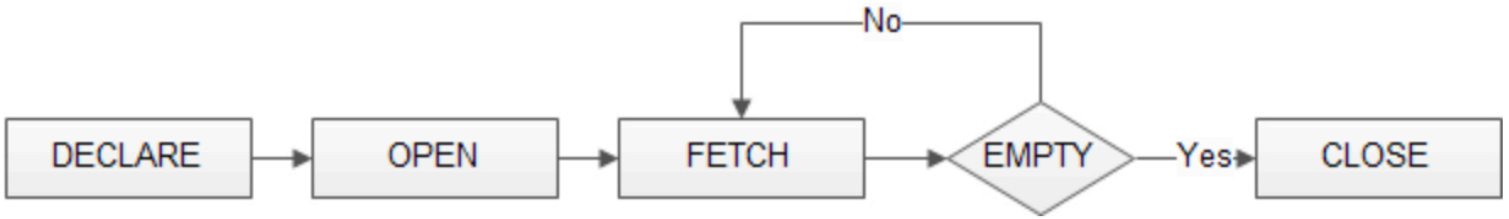

Figura 8 – Modo de funcionamento de cursores em MySQL.

MySQL

Cursores

```
1  CREATE PROCEDURE curdemo()
2  BEGIN
3    DECLARE done INT DEFAULT FALSE;
4    DECLARE a CHAR(16);
5    DECLARE b, c INT;
6    DECLARE cur1 CURSOR FOR SELECT id,data FROM test.t1;
7    DECLARE cur2 CURSOR FOR SELECT i FROM test.t2;
8    DECLARE CONTINUE HANDLER FOR NOT FOUND SET done = TRUE;
9
10   OPEN cur1;
11   OPEN cur2;
12
13   read_loop: LOOP
14     FETCH cur1 INTO a, b;
15     FETCH cur2 INTO c;
16     IF done THEN
17       LEAVE read_loop;
18     END IF;
19     IF b < c THEN
20       INSERT INTO test.t3 VALUES (a,b);
21     ELSE
22       INSERT INTO test.t3 VALUES (a,c);
23     END IF;
24   END LOOP;
25
26   CLOSE cur1;
27   CLOSE cur2;
28 END;
```

Figura 9 – Exemplo de um procedimento com cursores em MySQL.

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 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. Povoe o *data warehouse* em SQL recorrendo à base de dados *sakila*, à tabela *calendario*, bem como à base de dados *sakila_dsa*. No entanto, deverá ter em atenção que todas as suas tabelas de dimensão deverão ser povoadas antes das suas duas tabelas de factos.
5. Defina e implemente todos os processos que acha necessários para garantir a gestão contínua do *data warehouse* definido.