Mestrado Integrado em Engenharia Informática, 4º ano

Análise de Dados

TP2

Daniela Oliveira

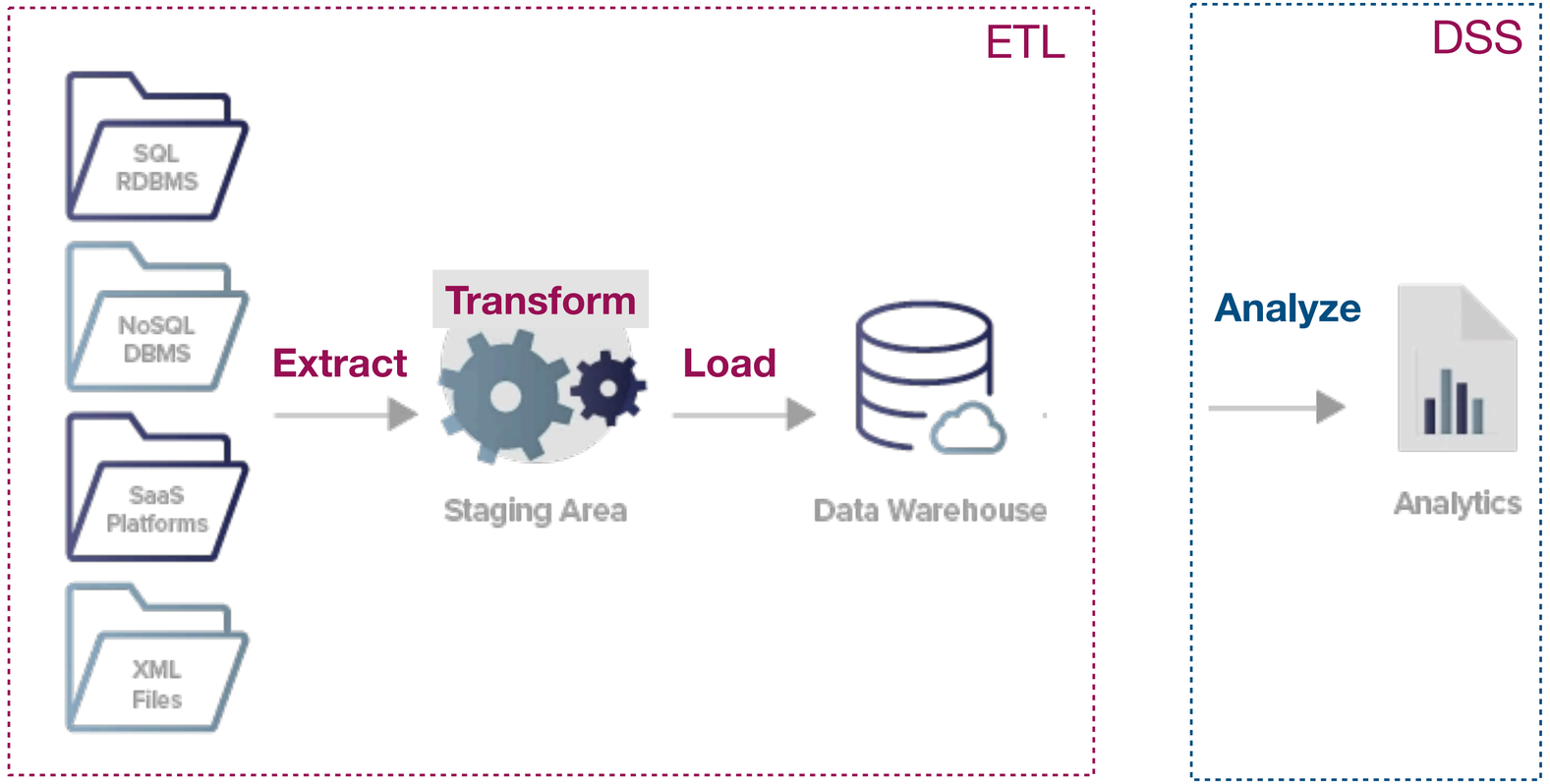
daniela.oliveira@algoritmi.uminho.pt - 2020/2021, Universidade do Minho

TP2

- Contextualização sobre os conceitos de ETL e *Data warehousing*;
- Resolução da 2ª ficha TP individual.

ETL
(Extract,Transform, Load)

Definição

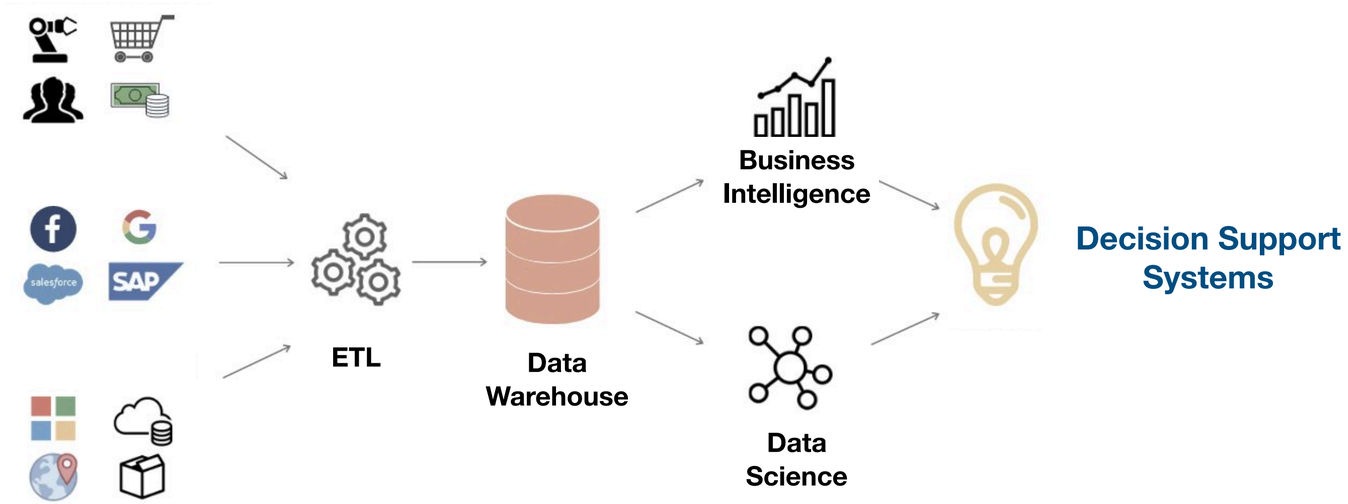


- Diferentes fontes de informação
- Constante aumento do volume de dados
- Diferentes níveis de estruturação de dados

Data Warehouse

Definição

- Define-se como um sistema de agregação de dados derivados de diferentes fontes de informação, para suporte à decisão num determinado contexto;
- Estabelece correlação entre os dados das diferentes fontes;
- Revela-se o núcleo dos processos de Business Intelligence e Data Science;
- Permite a implementação de sistemas DBMS (Database Management System);



Data Warehouse

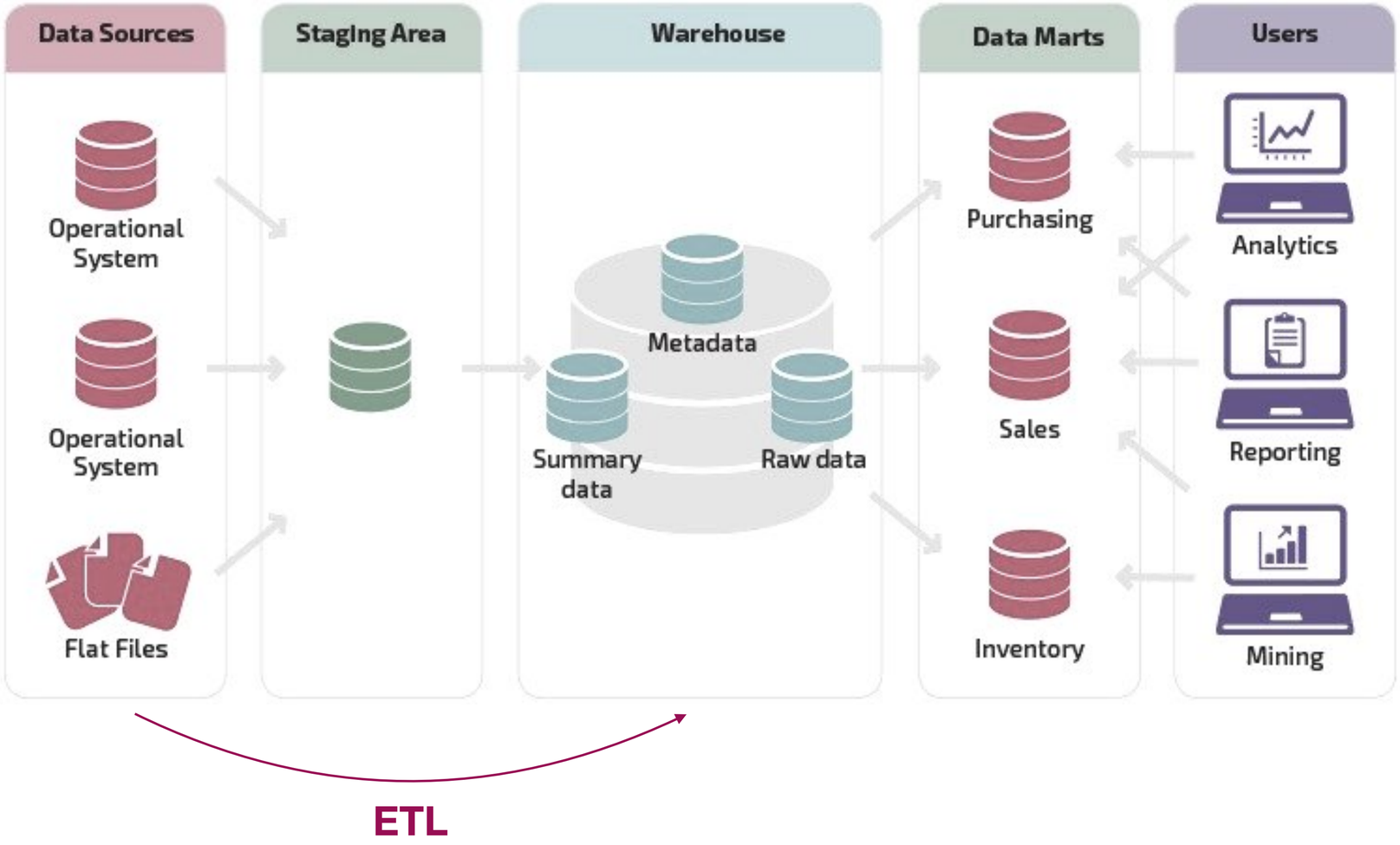
Características

- **Subject-Oriented**
  - Orientado à modelação e análise de dados para a tomada de decisões;
  - Visão simples e concisa sobre um assunto específico para apoiar o processo de decisão.
- **Integrated**
  - Unidade de medida comum para dados provenientes de diferentes bases de dados;
  - Consistência de nomenclaturas, formatos e codificação dos dados;
- **Time-variant**
  - Dados estão relacionados com um determinado período de tempo, revelando informações do ponto de vista histórico;
  - Uma vez inseridos os dados no *data warehouse*, estes não sofrem mutações.
- **Non-volatile**
  - Os dados não são apagados aquando inserção de novos;
  - Apenas operações de *loading* e de *access* são permitidas sobre um data warehouse.

Data Warehouse

Arquitetura

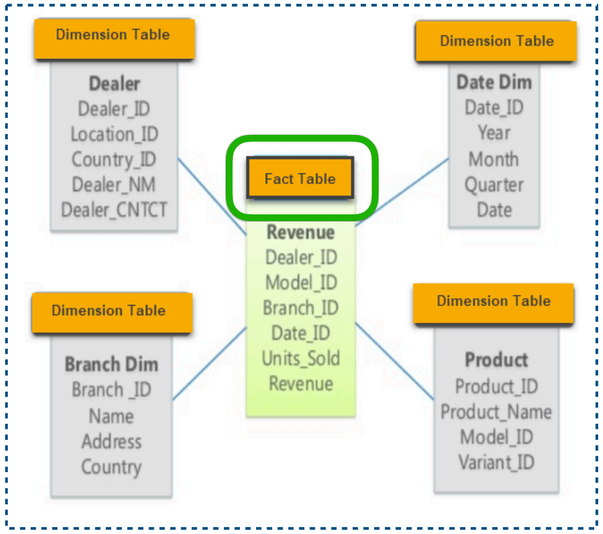
Arquitetura complexa constituída por dados históricos de várias fontes de informação.



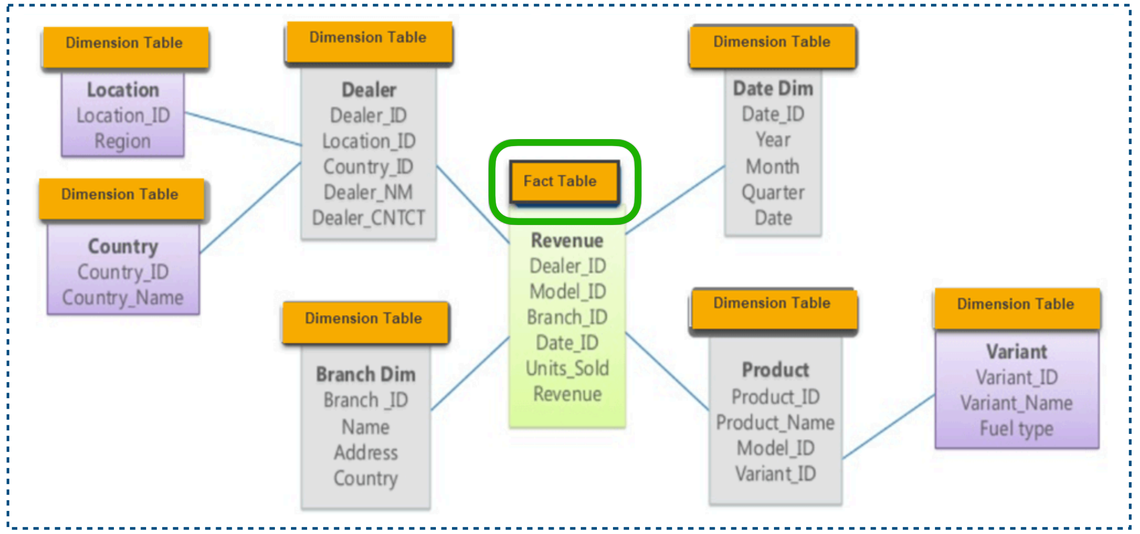
Data Warehouse

Modelação Dimensional

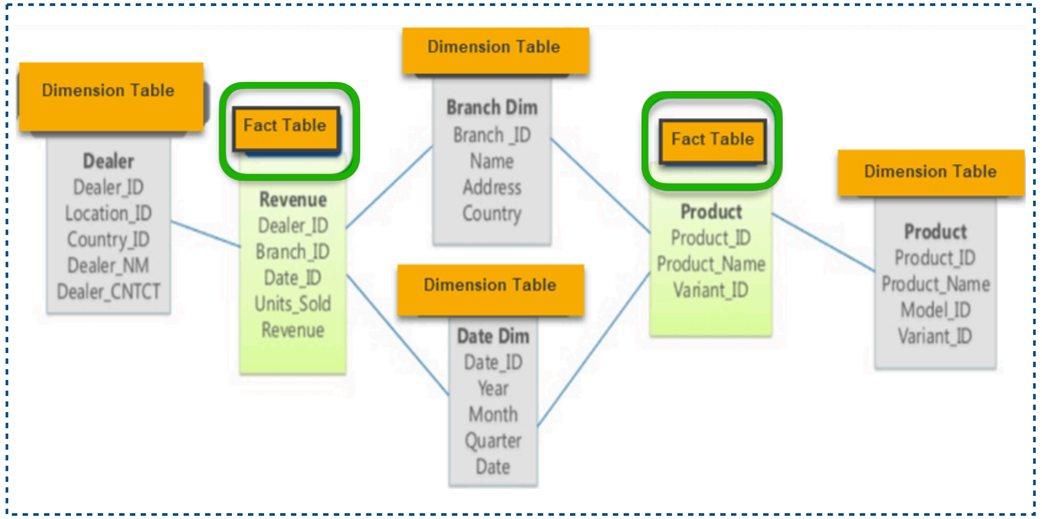
Star Schema



Snowflake Schema



Fact Constellation Schema



Send to: **daniela.oliveira@algoritmi.uminho.pt** with subject: **"AD/TP2/pgXXXX"**

Ficha N.º 2

1 Datawarehouse COVID-19 da Coreia do Sul

Pretende-se o desenvolvimento de um processo de *data warehousing*, transformando os dados presentes no ficheiro *South_Korea_Covid19.csv* para um modelo dimensional com o seu esquema em estrela.

No *dataset* em estudo, a Coreia do Sul está dividida em províncias e as suas respetivas cidades, com os diferentes tipos de foco de infeção COVID-19 bem como o seu respetivo país de origem.

Cada caso está caracterizado com o seu respetivo range de idades, o seu género e a sua proveniência. Ao nível sintomático, encontra-se registada a data do início dos mesmos, bem como a sua data de cura e/ou óbito.

Ferramentas: MySQL e MySQL Workbench

Requisitos a desenvolver

1. Criar um modelo lógico cujo o seu modelo dimensional seja em esquema de estrela, a partir da análise dos dados disponibilizados (com a sua tabela de factos e respetivas tabelas de dimensões);
2. Converter o modelo lógico para o seu modelo físico através da opção *Forward Engineer*;
3. Povoar as tabelas com os dados presentes nos ficheiro *South_Korea_Covid19.csv* - Para facilitar a transição de dados, pode criar um schema com os dataset disponibilizado;

Bibliografia

- https://www.kaggle.com/kimjihoo/coronavirusdataset?select=PatientInfo.csv
- Kimball, Ralph, and Margy Ross. ***The data warehouse toolkit: The definitive guide to dimensional modeling.*** John Wiley & Sons, 2013.
- https://www.mysqltutorial.org/