ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

DTC-MBD-521 Data Acquisition and Transformation

SEMESTER: Spring

CREDITS: 30 hours

LANGUAGE: Spanish/English

DEGREES: Master in Big Data Technologies and Advanced Analytics

Course overview

This course is an introduction to the fundamental principles of data acquisition, transformation and load. The aim of this subject is to provide to the students the most standards and innovative techniques, methodologies and tools to successfully obtain, clean, correct, standardize and transform data from different information sources.

Prerequisites

Basic knowledge of Python and SQL languages is required.

Course contents

Theory:

1. Introduction: Types of data sources.
2. Data governance: data traceability, accessibility, integrity and management.
3. Data extraction: Relational Data Bases, No-SQL Data Bases, web extraction (Python and BeautifullSoup) and text mining.
4. Data cleansing: data correction, standardization, relation and consolidation.
5. Data transformation: different techniques to successfully transform data by using SQL, No-SQL and Python.
6. ETL tools: traditional tools as SAS, ORACLE and Microsoft as well as other newer as Pentaho Kettle and RapidMiner.

Textbook

While we will not follow a textbook, we find the following books quite remarkable in their central topics.

- **Simsion G. and Witt G., (2014),** *Data modeling essentials*. 3rd Edition. Morgan Kaufmann Publications.

- **Doan A., Halevy A., Ivez Z.,** (2012), *Principles of data integration,* 1st Edition. Morgan Kaufmann Publications.

Grading

The following conditions must be accomplished to pass the course:

- A minimum overall grade of at least 5 over 10.
- A minimum grade in the final test of 5 over 10.

The overall grade is obtained as follows:

- Final test accounts for 50% of the final grade if the grade in this exam is at least 5.
- Laboratory session work (in class and homework) accounts for 50% of the final grade.