

DTC-MBD-515 Big Data Processing Technologies

SEMESTER: Fall
CREDITS: 7.5 ECTS
LANGUAGE: Spanish
DEGREES: MBD

Course overview

The purpose of the course is to give an overview of the ingestion and processing tools of the big data environment, especially focused on Spark and Hive.

By the end of the course, students will:

- Be able to choose which is the most appropriate tool to extract data from different sources and take it to a Hadoop cluster.
- Have experience with some processing tools and languages (python,hql,etc).
- Have deep knowledge of spark with python and how to optimize jobs.

Prerequisites

Students willing to take this course should be familiar with any programming language, preferably python or SQL and with Linux commands and utilities.

Course contents

Theory

1. Introduction to SQL
2. Hadoop Ecosystem
 - HDFS and Hadoop client
 - Sqoop
 - Flume
 - Hive
 - Kafka

- Spark

- 3. Introduction to Data scientist
 - Python first steps
 - Scientific Python

- 4. Apache Spark for Data scientist
 - Spark DataFrame
 - Spark ML (Machine Learning)
 - Spark packages

Practice

All sessions will have a hands-on approach. In the development of the course will be proposed to students practices that will be 60% of the final grade.

Textbook

- Notes and notebooks prepared by the lecturer (available in Moodle).
- White, T. (2015). *Hadoop: The definitive guide 4th edition*. " O'Reilly Media, Inc."
- Shreedharan, Hari (2014). *Using Flume* " O'Reilly Media, Inc."
- Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). *Learning spark: lightning-fast big data analysis*. " O'Reilly Media, Inc."
- VanderPlas, J. (2016). *Python Data Science Handbook*.

Grading

$$\text{Final Grade} = 0.6 * A + 0.4 * B$$

Where:

- € A: Mean of student's practices (0-10 points), it must be passed.
- € B: Mark of the final exam (0-10 points), it must be passed.