ESCUELA TÉCNICA SUPERIOR DE
INGENIERÍA

DOI-MBD-XXX 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.

This document is a brief outline of the course and does not replace the official program of study

Course contents

Theory

1. Data scientist Toolbox
   - 1.1. Introduction to ecosystem
   - 1.2. Python first steps
   - 1.3. Scientific Python
   - 1.4. SQL
   - 1.5. git and GitHub
   - 1.6. Markdown
2. Hadoop Ecosystem
   - 2.1. HDFS and Hadoop client
   - 2.2. Sqoop
   - 2.3. Flume
   - 2.4. Hive
   - 2.5. Pig
   - 2.6. Kafka
3. Apache Spark
   - 3.1. Introduction to RDD
   - 3.2. Spark DataFrame
   - 3.3. Spark ML (Machine Learning)
   - 3.4. Spark packages
4. Introduction to search engines: Lucene
5. Schedule Hadoop Jobs with Apache Oozie

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

Final Grade = 0.6 * A + 0.4 * B

Where:

- A: Mean of student’s practices (0-10 points)
- B: Mark of the final exam (0-10 points)