V Máster en Advanced Analytics on Big Data

Módulo: Apache Spark (2 semanas)

Docentes:

Antonio Jesús Nebro Urbaneja (antonio@lcc.uma.es). Tutorías presenciales Martes 16:30 a 18:30. Tutorías virtuales (email, foros, chat, Skype) Miércoles 16.30 a 18.30.

Distribución horaria:

Clase 1: Antonio Jesús Nebro Urbaneja

Clase 2: Antonio Jesús Nebro Urbaneja

Clase 3: Antonio Jesús Nebro Urbaneja

Clase 4: Antonio Jesús Nebro Urbaneja

Clase 5: Antonio Jesús Nebro Urbaneja

Clase 6: Cristóbal Barba González

Objetivos: El objetivo de este módulo es dar a conocer el sistema de procesamiento de datos escalable Apache Spark. Se ofrecerá una visión global de las características de Spark y se trabajará con las dos APIs que ofrece en la actualidad, una basada en RDDs (*Resilient Distributed Datasets*) y otra basada en *dataframes*. Se incluirá además una introducción a Apache Hadoop, prestando particular atención al sistema de ficheros HDFS.

Tecnologías: Spark, Java, Python, Intellij Idea, PyCharm

Pre-Requisitos: Conocimientos básicos de programación en Java y Python

V Máster en Advanced Analytics on Big Data

Planificación docente:

Clase 1: Introducción a Apache Spark. API basada en RDDs

Trabajo previo necesario:

- Página Web del proyecto Apache Spark (http://spark.apache.org/)
- https://www.infoq.com/articles/apache-spark-introduction
- Instalación de Spark

Tarea 1 (Mandatory): Análisis de un fichero de contraseñas de un sistema Linux.

Tarea 2 (Mandatory): Comparar el rendimiento de una aplicación Spark que sume números contenidos en ficheros en las versiones Java y Python.

Clase 2: Aplicaciones basadas en el uso de pares clave-valor.

Trabajo previo necesario:

- Página Web del proyecto Apache Spark (http://spark.apache.org/)

Tarea 3 (Mandatory): Análisis de datos de aeropuertos usando RDDs.

Tarea 4 (Mandatory): Análisis de tweets.

Clase 3: Aspectos avanzados: acumuladores y *broadcasters*

Tarea 5 (Advanced): Aplicación que filtra datos incorrectos.

Tarea 6 (Advanced): Análisis de datos de electricidad (fuente: Banco Mundial).

Clase 4: Aplicaciones basadas en dataframes I. Operaciones básicas. Acceso a fuentes de datos de diferentes tipos. Inferencia de esquemas.

Trabajo previo necesario

- Introducción a dataframes en Spark (https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html)
- Documentación sobre SQL, dataset y dataframes (https://spark.apache.org/docs/latest/sql-programming-guide.html)

Tarea 7 (Mandatory): análisis de datos de aeropuertos usando dataframes.

Tarea 8 (Mandatory): análisis de datos de cáncer con dataframes.

V Máster en Advanced Analytics on Big Data

Clase 5: Aplicaciones basadas en dataframes II. Interoperabilidad con RDDs. Especificación de esquemas.

Trabajo previo necesario

- Introducción a dataframes en Spark (https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html)
- Documentación sobre SQL, dataset y dataframes (https://spark.apache.org/docs/latest/sql-programming-guide.html)

Tarea 9 (Advanced): A partir de un fichero CSV de gran tamaño (como el de crímenes de Chicago), realizar dos aplicaciones que realicen la misma consulta usando RDDs y dataframes y comparar el rendimiento de ambas en un sistema multi-core.

Tarea 10 (Challenge): A partir de un fichero de datos de crímenes de la ciudad de Chicago (u otro similar) se propone realizar una aplicación que combine datos almacenados en MongoDB, procesamiento con Spark y visualización en un notebook Jupyter.

Clase 6: Hadoop

Tarea 11 (Mandatory): Operaciones con HDFS.