El curso será eminentemente práctico y los asistentes tendrán que resolver problemas diversos utilizando Spark.

Al final del curso los asistentes tendrán los conocimientos necesarios para comenzar a utilizar Spark en sus tareas cotidianas de análisis de datos.

Contenidos

1. Herramientas necesarias

- Jupyter
- HDFS
- YARN

2. Conceptos básicos de Spark
3. Programando con RDDs
4. Programando con PairRDDs
4. Programando con Spark SQL
5. Programando con DataFrames
6. Lanzando aplicaciones
7. Monitorizando y depurando la ejecución de aplicaciones

Prerrequisitos

Los asistentes deben poseer una cuenta de usuario en el CESGA y deberán traer su propio portátil. Esta cuenta será la que utilizarán para resolver los distintos problemas que se planteen durante el curso. En caso de no disponer de cuenta esta puede solicitarse a través del siguiente enlace: https://www.altausuarios.cesga.es/

Es importante que el portátil tenga configurada la VPN del CESGA ya que será necesaria para lanzar los notebooks de Jupyter.

El curso require conocimientos básicos de programación con Python ya que será el lenguaje que se usará durante el curso.

Python es un lenguaje muy popular y que se puede aprender rápidamente, por lo que a los alumnos que no estén familarizados con este lenguaje, les recomendamos que realicen antes del curso alguno de los numerosos tutoriales de Python 2 existentes, por ejemplo:

- http://www.learnpython.org/
- https://docs.python.org/2/tutorial/

Es aconsejable para todos los participantes realizar el siguiente test de python antes del curso: http://www.mypythonquiz.com

Resultarán de utilidad, aunque no imprescindibles, conocimientos de GNU/Linux y familiaridad con SQL.