UNIVERSIDAD COMPLUTENSE DE MADRID

ntic master School

Hadoop / Spark

Presentación del curso

Dr. Pablo J. Villacorta

Marzo de 2021

Agenda

- **Presentación**
- **Temario del curso**
- **Método de evaluación**
- **Infraestructura**

Temario del curso

Vamos a estudiar dos tecnologías fundamentales:

- **HDFS (Hadoop Distributed File System)** para almacenar ficheros muy grandes en un cluster de ordenadores
- Veremos cómo está hecho por dentro y comandos fundamentales

- **Apache Spark** para procesar de forma distribuida esos datos, aprovechando los nodos del cluster
- Lo estudiaremos en profundidad, todos sus módulos y lo que se puede hacer con él (transformar datos, ETL, agregaciones, Machine Learning, procesamiento de datos en streaming…)

- **Además** veremos una introducción para motivar la necesidad de las tecnologías Big Data en la actualidad

Método de evaluación

- Ejercicio práctico (10 puntos): resolveréis un notebook en pyspark sobre un dataset personalizado para cada uno de vosotros

- El ejercicio consistirá en transformaciones sencillas similares a las que estudiaremos en clase

Infraestructura

- Utilizaremos Google Cloud para desplegar un cluster al vuelo, que se desmontará también automáticamente pasado cierto tiempo (el que vayamos a dedicarle al estudio en esa sesión)

- Está todo en la guía de instalación que tenéis en moodle.

- Tenéis almacenamiento persistente en un **bucket** de Google Cloud Storage, donde subiremos los datasets desde nuestro PC personal