

Taller Big Data: Introducción a la plataforma Hadoop 3 del CESGA

Descripción

En este taller se hará una introducción a la nueva plataforma Big Data del CESGA. Esta plataforma ha sido actualizada a Hadoop 3 e incluye también la nueva versión de Spark 2.4.

Este taller servirá de introducción a las herramientas disponibles dentro de la plataforma Hadoop 3.

El taller también servirá de base para posteriores talleres sobre herramientas específicas incluidas en la plataforma como por ejemplo Spark.

Duración

3 horas

Fecha

11 de junio 10:00 a 13:00

Lugar

CESGA
Avda. de Vigo s/n
Campus Vida
Santiago de compostela

Destinatarios

El taller está destinado tanto a usuarios actuales de la plataforma Big Data como a nuevos usuarios que necesiten acceso a herramientas Big Data.

¿Qué aprenderé durante el taller?

Al final del taller sabrás:

- Cómo conectarte a la plataforma Hadoop 3
- Cómo transferir datos de forma eficiente
- Las herramientas que están disponibles
- Cómo lanzar estas herramientas

¿Qué NO se enseñará durante el taller?

Dado el carácter introductorio del taller y la gran variedad de herramientas disponibles, en este taller no se enseñará a utilizar cada una de las herramientas, sino que simplemente se mostrará como se puede acceder a ellas y cómo se lanzan.

Posteriormente se realizarán talleres específicos que se centrarán en herramientas concretas como Spark donde sí se enseñará su uso en detalle.

Contenido

1. Introducción al servicio Big Data
 - 1.1. Conceptos básicos
 - 1.2. Descripción del hardware
 - 1.3. Descripción del software
 - 1.4. Portal BD|CESGA
2. Conexión al servicio Hadoop 3
 - 2.1. VPN
 - 2.2. Acceso por línea de comandos: SSH
 - 2.3. Acceso a través del interfaz web: WebUI > HUE
 - 2.4. Acceso por escritorio remoto
3. Transferencia de datos
 - 3.1. Cómo transferir datos de forma eficiente usando el servicio DTN
 - 3.2. Cómo transferir datos usando SCP
 - 3.3. Cuotas en los sistemas de ficheros
 - 3.4. Migración de datos de la antigua plataforma
4. Elementos básicos
 - 4.1. HDFS: Almacenamiento distribuido
 - 4.2. YARN: Ejecución y monitorización de trabajos
5. Herramientas disponibles
 - 5.1. Spark
 - 5.2. Jupyter
 - 5.3. Hive
 - 5.4. Impala
 - 5.5. Sqoop
 - 5.6. Modules
6. Dónde obtener información adicional
 - 6.1. Tutoriales
 - 6.2. Guía de usuario
 - 6.3. Documentación oficial