

Tecnologías de Computación para Datos Masivos

Tomás Fernández Pena

tf.pena@usc.es twitter:@tfpena

Máster en Tecnologías de Análisis de Datos Masivos: Big Data

Universidade de Santiago de Compostela

Material bajo licencia Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

citius.usc.es

Temario

Tema 1 Big Data y MapReduce

- ▷ Introducción al BigData
- ▷ Modelo de programación MapReduce: ejemplos de uso, ejecución, optimizaciones, implementaciones

Tema 2 Introducción a Hadoop

- ▷ Introducción e instalación de Hadoop
- ▷ Introducción a HDFS
- ▷ Gestor de recursos y planificador de tareas: YARN
- ▷ Introducción a MapReduce en Hadoop

Tema 3 HDFS

- ▷ Filesystems en Hadoop
- ▷ Interfaces principales: línea de comandos y Java
- ▷ Herramientas para la gestión del HDFS
- ▷ Namenode principal y de checkpoint
- ▷ Otras interfaces a HDFS

Temario (cont.)

Tema 4 MapReduce en Hadoop

- ▷ Java MapReduce en Hadoop
- ▷ Serialización y entrada/salida
- ▷ Tareas MapReduce
- ▷ Otros aspectos
- ▷ Alternativas a Java

Tema 5 Spark

- ▷ Introducción a Apache Spark
- ▷ API estructurada: DataFrames y DataSets
- ▷ API de bajo nivel: RDDs
- ▷ Despliegue y optimización de aplicaciones
- ▷ Extensiones: Streaming, MLLib, GraphX

Tema 6 Introducción al procesamiento en streaming con Apache Flink

Bibliografía recomendada

- Tom White, *Hadoop: The Definitive Guide*, 4th Edition, O'Reilly, 2015
- Bill Chambers, Matei Zaharia, *Spark: The Definitive Guide*, O'Reilly, 2018
- Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, *Learning Spark. Lightning-Fast Big Data Analysis*, O'Reilly, 2015
- Hueske F., Kalavri V, *Stream Processing with Apache Flink*, O'Reilly, 2019

Otros libros

- P. Zečević, M. Bonaći, *Spark in action*, Manning Pubs, 2017
- H. Karau, R. Warren, *High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark*, O'Reilly, 2017
- S. Ryza, U. Laserson, S. Owen, J. Wills, *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*, O'Reilly, 2017

Uso del CESGA

- Cuentas del CESGA ya disponibles
- Usaremos la plataforma Big Data del CESGA:
 - ▷ Acceso por ssh: `hadoop3.cesga.es`
 - ▷ Interfaz web: `https://bigdata.cesga.es/`
 - ▷ IMPORTANTE: cambiad la contraseña lo antes posible
- Necesitáis tener instalada la VPN del CESGA:
 - ▷ Id a `https://portalusuarios.cesga.es/` y autenticaros
 - ▷ En **Información** → **Conexión Remota** → **Conexión desde un centro no autorizado** os indica como instalarla

Localización de las transparencias y las prácticas:

- En el Campus Virtual de la USC

Información oficial de la materia:

<https://www.usc.gal/gl/estudios/masteres/enxenaria-arquitectura/master-universitario-tecnoloxias-analise-datos-masivos-big-data/20212022/tecnoloxias-computacion-datos-masivos-16519-15865-2-93301>