

Tecnología de Computación de Datos Masivos. Presentación

Diego Sevilla

Dpto. Ingeniería y Tecnología de Computadores
Facultad de Informática
Universidad de Murcia

fgarcia@um.es, dsevilla@um.es

2024

CURSO ACADÉMICO	2024/2025
TITULACIÓN	MÁSTER BIG DATA
CUATRIMESTRE	PRIMERO
CURSO	PRIMERO
CARÁCTER	OBLIGATORIA
CRÉDITOS ECTS	4,5
DEPARTAMENTO	INGENIERÍA Y TECNOLOGÍA DE COMPUTADORES

Profesores y horario

	Profesor	Horario	Aula
Teoría	D. Sevilla	Viernes 17:30–19:00	Lab. 2.4
Prácticas	D. Sevilla	Viernes 19:00–20:30	Lab. 2.4

Profesor	Tutorías	Contacto
D. Sevilla	Miércoles 11:00–14:00 (y tutorías electrónicas)	Despacho 3.31 dsevilla@um.es 868 88 7571

1 Big Data y MapReduce

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

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

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

4 Hadoop en el Cloud

5 MapReduce en Hadoop

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

6 Spark

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

7 Introducción al procesamiento en streaming con Apache Flink

Planificación del curso

Fecha	Sesión Teoría	Sesión Prácticas
27/09	Introducción a la asignatura	–
04/10	Tema 1 Big Data y MapReduce	Práctica 1 Hadoop y HDFS (i)
11/10	Tema 2 Hadoop y HDFS (i)	Práctica 1 Hadoop y HDFS (ii)
18/10	Tema 2 Hadoop y HDFS (ii)	Práctica 2 HDFS
25/10	Finalización Prácticas 1 y 2	Finalización Prácticas 1 y 2
01/11	–	–
08/11	Práctica 3 MapReduce (i)	Práctica 3 MapReduce (ii)
15/11	–	–
22/11	Tema 3 Apache Spark (i)	Práctica 4 Apache Spark (i)
29/11	Tema 3 Apache Spark (ii) / Flink	Práctica 4 Apache Spark (ii)
06/12	–	–

Evaluación de teoría

- Examen teórico: tipo test
- Ponderación: 30 %

Evaluación de prácticas

- Documentación y entrevista final
- Ponderación: 70 %

Hay que superar cada parte por separado

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

- Guiones de prácticas e información adicional
- El repositorio está alojado en GitHub y se llama 'tcdm-public', dirección `https://github.com/dsevilla/tcdm-public`
- Para obtenerlo (rama **24-25**):

```
$ git clone https://github.com/dsevilla/tcdm-public.git  
$ cd tcdm-public
```
- Para algunas cuestiones no hace falta bajarlo (usaremos **Google Colab**)
- (Esto requiere una cuenta Google)
- Los *Notebooks* se podrán guardar en **Drive** o en un repositorio **GitHub** y luego enviar al profesor