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

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
CAMPUS VIDA CAMPUS DE EXCELENCIA INTERNACIONAL
CiTiUS Centro Singular de Investigación en Tecnoloxías da Información

Temario de la materia

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 de la materia

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 de Docker

- Usaremos Docker (`https://www.docker.com/`) para desplegar un cluster Hadoop
- Podéis instalar Docker en vuestro PC
- Requisitos para la práctica:
  - +8 GB RAM recomendable
  - Preferible disponer de Linux
  - Si usáis una MV con Linux, la MV debería tener 2 cores y 8+ GB de RAM y disponer de aceleración por hardware
  - Si tenéis Windows, se necesita versión 10 o superior con WSL2 o Hyper-V activado
  - Con macOS, version 10.15 o superior
- Alternativa: usar una máquina virtual en la nube de AWS

Uso del CESGA

- Cuentas del CESGA disponibles en breve
- 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:
  - Instrucciones: `https://cesga-docs.gitlab.io/ft3-user-guide/how_to_connect.html#configure-vpn`

Más información

Localización de las transparencias y las prácticas:

- En el Campus Virtual de la USC

Información oficial de la materia: https://www.usc.gal/gl/estudos/masteres/enxenaria-arquitectura/master-universitario-tecnoloxias-analise-datos-masiv20232024/tecnoloxias-computacion-datos-masivos-16519-15865-2-