Universitat d'Alacant
Universidad de Alicante

Guía básica de instalación y uso: Docker, Hadoop y Spark en Ubuntu

Introducción

Esta guía ha sido diseñada como un punto de partida sencillo y práctico para que puedas preparar por tu cuenta un entorno básico de trabajo con **Docker, Hadoop y Apache Spark en sistemas Ubuntu**. Su propósito es ayudarte a instalar, lanzar y comprender los elementos fundamentales de estas herramientas utilizadas en entornos distribuidos de análisis de datos.

Ten en cuenta que, aunque los pasos están explicados de forma clara, **pueden surgir errores dependiendo de la versión del sistema, las configuraciones previas o permisos del usuario**. Estos errores son comunes en contextos reales y forman parte del aprendizaje técnico. Por ello, te animamos a adoptar una actitud activa y autónoma en su resolución, usando recursos como:

- **IA generativa** (como Copilot o ChatGPT) para entender y resolver mensajes de error,
- la **documentación oficial** de cada herramienta,
- **foros técnicos y comunidades** (Stack Overflow, GitHub...),
- y las **guías de instalación específicas de tu equipo o sistema operativo**.

Aprender a enfrentarte a estos retos es una parte esencial del trabajo en entornos reales de computación distribuida y desarrollo profesional.

¿Qué es Docker y para qué se usa?

Docker es una plataforma que permite crear, ejecutar y gestionar entornos aislados llamados contenedores. Estos contenedores permiten ejecutar aplicaciones de forma consistente sin preocuparse por las dependencias del sistema. En nuestro contexto, Docker nos permite lanzar rápidamente entornos preconfigurados para usar Spark o Hadoop sin instalar cada componente por separado.

Instalación de Docker en Ubuntu

1. Abrir una terminal.

2. Ejecutar:

1. sudo apt update
2. sudo apt install docker.io

3. Habilitar Docker:

1. sudo systemctl enable docker
2. sudo systemctl start docker

4. (Opcional) Para evitar usar 'sudo' en cada comando:

1. sudo usermod -aG docker $USER
2. Luego cerrar sesión y volver a entrar.

¿Qué es Hadoop y para qué se usa?

Hadoop es un framework de código abierto diseñado para el almacenamiento y procesamiento distribuido de grandes volúmenes de datos. Se basa en dos componentes principales:

- **HDFS (Hadoop Distributed File System),** que divide los datos en bloques y los reparte entre nodos.
- **MapReduce**, un modelo de programación para procesarlos en paralelo.

En esta asignatura lo trabajamos como un ejemplo clásico de middleware orientado a datos, destacando su papel en la distribución, sincronización, balanceo de carga y tolerancia a fallos.

NOTA: Además, Hadoop se profundiza desde otras perspectivas en asignaturas como Big Data y Análisis Masivo de Datos, donde el alumnado del Grado en IA de la Universidad de Alicante lo utiliza para ejecutar procesos de análisis sobre grandes conjuntos de datos reales, integrándolo con Spark, Jupyter o Hive. Allí el foco se desplaza hacia el procesamiento eficiente y la extracción de conocimiento, completando así la visión arquitectónica que se introduce aquí.

Instalación básica de Hadoop en Ubuntu (modo pseudo-distribuido)

1. Instalar Java:

1. sudo apt install openjdk-11-jdk

2. Descargar Hadoop desde:

2. https://hadoop.apache.org/releases.html

3. Extraer el archivo y configurar las variables de entorno:

3. export HADOOP_HOME=/ruta/a/hadoop
4. export PATH=$PATH:$HADOOP_HOME/bin

4. Modificar archivos de configuración: core-site.xml, hdfs-site.xml (básico).

5. Iniciar los servicios de Hadoop:

start-dfs.sh

¿Qué es Apache Spark y para qué se usa?

Apache Spark es una plataforma para el procesamiento de datos en memoria, muy rápida y flexible. Permite transformar grandes volúmenes de datos en paralelo, y es más eficiente que Hadoop MapReduce en muchos casos. Incluye librerías para SQL, machine learning, procesamiento de grafos y más.

Ejecutar Spark fácilmente con Docker

1. Descargar imagen de Spark:

    1. docker pull bitnami/spark

2. Ejecutar Spark en modo interactivo:

    1. docker run -it --rm -p 4040:4040 -v "$PWD":/app bitnami/spark spark-shell

3. Abrir el navegador en http://localhost:4040 para ver el monitor de tareas.

Consejos finales

• Siempre que sea posible, usar Docker para evitar problemas de configuración.

• No olvidar revisar los logs y la interfaz web de Spark para entender cómo se ejecutan las tareas.

• Para pruebas pequeñas, Spark funciona perfectamente en modo local.

• Recuerda que Hadoop y Spark no son excluyentes: Spark puede leer datos desde HDFS.