

Manual de Instalación de Apache Hadoop 3.4.2 con YARN y Java 8

Entorno Pseudo-Distribuido

BSG Institute

Programa: Certified AI Data Engineer

Curso: Procesos ETL para Workloads de AI

Fecha: 04 de November de 2025

1. Introducción General a Hadoop

Apache Hadoop es un marco de software de código abierto que permite el procesamiento distribuido de grandes volúmenes de datos en clústeres de servidores mediante modelos de programación simples. Su arquitectura se basa en el paradigma MapReduce, donde los datos se dividen y procesan en paralelo. Hadoop está compuesto principalmente por tres componentes: HDFS (Hadoop Distributed File System), YARN (Yet Another Resource Negotiator) y MapReduce.

En este laboratorio instalaremos Hadoop 3.4.2 junto con YARN y Java 8, configurándolo en un solo servidor en modo pseudo-distribuido. Este entorno permite ejecutar y probar procesos de Big Data, simular clústeres reales y preparar entornos de desarrollo para integración con herramientas de AI y ETL.

2. Objetivos de Aprendizaje

- Comprender la arquitectura y componentes principales de Hadoop.
- Instalar Hadoop 3.4.2 y configurarlo con Java 8 en modo pseudo-distribuido.
- Iniciar y validar los servicios HDFS y YARN.
- Ejecutar comandos básicos de HDFS para verificar la instalación.

3. Requisitos del Entorno

- Sistema operativo: Ubuntu 22.04 o 24.04 LTS.
- Privilegios de usuario con permisos sudo.
- Conexión a Internet para descargar dependencias.
- 4 GB de RAM mínimo y 10 GB de espacio libre en disco.

- Java 8 instalado y configurado como versión predeterminada.

4. Instalación Paso a Paso

A continuación se detallan los pasos necesarios para la instalación de Hadoop 3.4.2 con YARN y Java 8.

4.1 Instalación de Java 8

```
```bash
sudo apt update
sudo apt install openjdk-8-jdk -y
java -version
````
```

4.2 Creación del usuario Hadoop (opcional)

```
```bash
sudo adduser hadoop
sudo usermod -aG sudo hadoop
su - hadoop
````
```

4.3 Configuración de SSH

```
```bash
ssh-keygen -t rsa -P ""
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ssh localhost
````
```

4.4 Descarga e instalación de Hadoop 3.4.2

```
```bash
cd /usr/local
sudo wget https://downloads.apache.org/hadoop/common/hadoop-3.4.2/hadoop-3.4.2.tar.gz
sudo tar -xzf hadoop-3.4.2.tar.gz
sudo mv hadoop-3.4.2 hadoop
sudo chown -R $USER:$USER hadoop
````
```

4.5 Variables de entorno

Editar el archivo `~/.bashrc` y agregar:

```
```bash
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
````
```

4.6 Configuración de Hadoop

Editar los archivos de configuración ubicados en `'\$HADOOP_HOME/etc/hadoop`:

core-site.xml

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

hdfs-site.xml

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/usr/local/hadoop/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop/hdfs/datanode</value>
</property>
</configuration>
```

mapred-site.xml

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

yarn-site.xml

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
```

```
<name>yarn.resourcemanager.hostname</name>
<value>localhost</value>
</property>
</configuration>
```

hadoop-env.sh

Asegurar la línea:

```
```bash
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
````
```

4.7 Formatear el NameNode

```
```bash
hdfs namenode -format
````
```

4.8 Iniciar servicios de Hadoop y YARN

```
```bash
start-dfs.sh
start-yarn.sh
````
```

5. Pruebas y Validación

Verificar los servicios en ejecución:

```
```bash
jps
````
```

Debe mostrar:

NameNode
DataNode
SecondaryNameNode
ResourceManager
NodeManager

Probar el sistema de archivos HDFS:

```
```bash
hdfs dfsadmin -report
hdfs dfs -mkdir /test
hdfs dfs -put /etc/hosts /test/
hdfs dfs -ls /test
````
```

6. Visualización y Administración

Acceder desde el navegador a las interfaces web:

- NameNode: <http://localhost:9870>
- ResourceManager: <http://localhost:8088>

7. Conclusiones y Buenas Prácticas

Con la instalación de Hadoop 3.4.2 y la configuración de YARN en modo pseudo-distribuido, los estudiantes pueden experimentar con un entorno real de Big Data en un solo equipo. Este laboratorio proporciona una base sólida para comprender cómo se gestionan los datos, cómo se distribuyen las tareas y cómo los recursos del sistema se asignan dinámicamente.

Se recomienda mantener el entorno actualizado, limpiar logs periódicamente y practicar con comandos de HDFS y YARN para entender su funcionamiento interno antes de desplegar en entornos distribuidos multi-nodo.

Anexo A. Script Automatizado de Instalación

El siguiente script realiza la instalación automática de Hadoop 3.4.2 con Java 8 en modo pseudo-distribuido.

```
```bash
bash install_hadoop_3.4.2.sh
```
```