



Introducción a Big Data

Curso de Especialización en
Inteligencia Artificial y Big Data

Big Data Aplicado

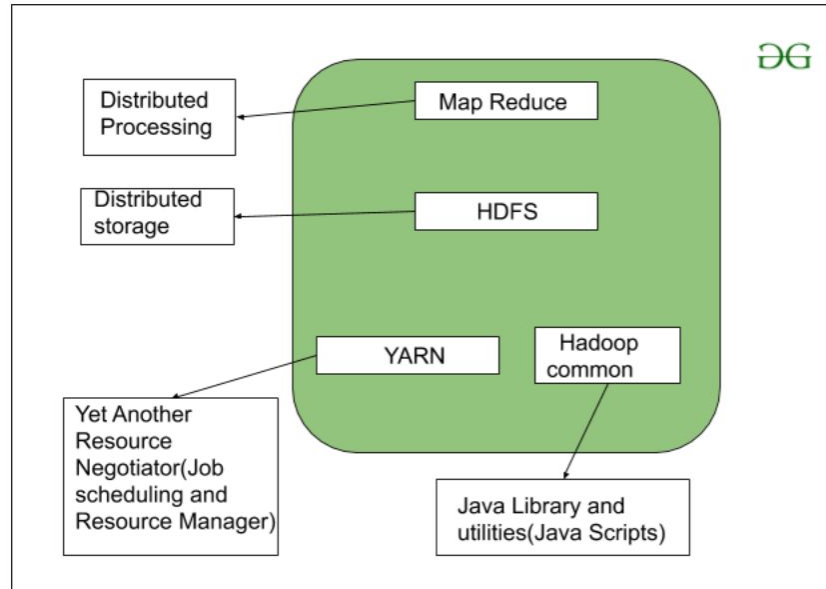
Hadoop

Proyecto de Big Data open source con los componentes principales:

- Hadoop MapReduce
- Hadoop File System (HDFS)
- Yet Another Resource Negotiator (YARN)

Apache Hadoop - <https://hadoop.apache.org/>

- HDFS
- Map Reduce
- YARN



HDFS - Hadoop Distributed File System

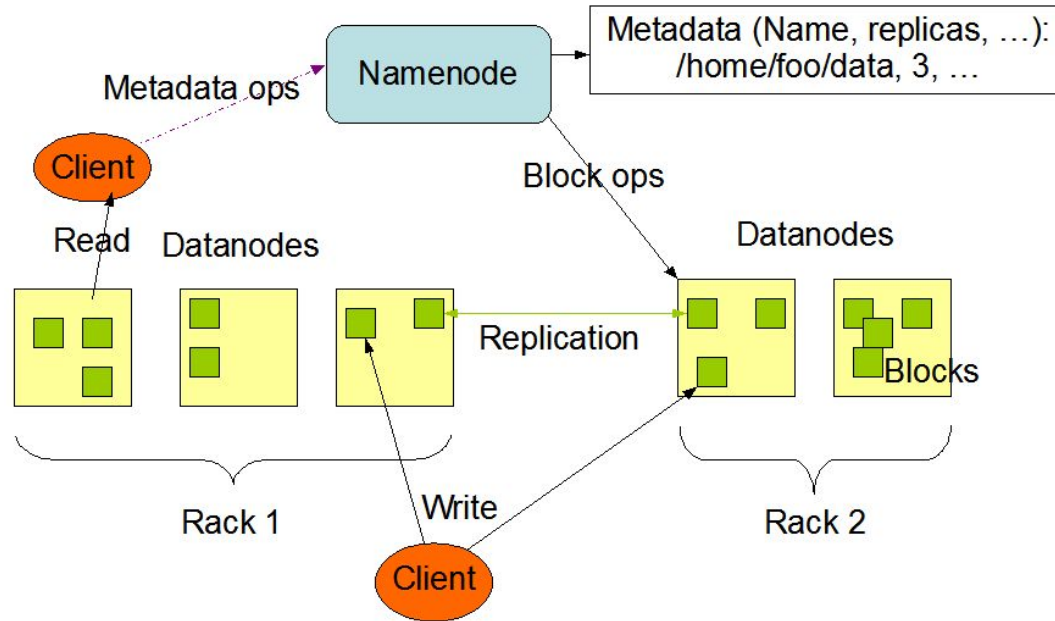
Sistema de ficheros distribuido diseñado para commodity hardware.

- Altamente tolerante a fallos
- Datasets de gran tamaño (gigabytes - terabytes)
- Batch processing (procesamiento por lotes)
- Escalabilidad horizontal
- Modelo write-once-read-many
- Arquitectura maestro/esclavo

<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>

HDFS - Hadoop Distributed File System

HDFS Architecture



HDFS - Hadoop Distributed File System

Arquitectura maestro/esclavo:

- **NameNode:** Servidor **maestro** que gestiona el espacio de nombres del sistema de ficheros y regula el acceso a dichos ficheros por parte de clientes.
- **DataNode:** Nodos que almacenan la información del clúster en bloques de datos.

HDFS - Hadoop Distributed File System

Arquitectura maestro/esclavo

- **Escritura:**

- 1) El cliente envía una petición de escritura al NameNode.
- 2) El NameNode responde con los DataNodes a escribir.
- 3) Al escribir en un DataNode, se replica la información al resto.
- 4) Se confirma la escritura al cliente.

- **Lectura:**

- 1) El cliente solicita una ubicación de fichero al NameNode.
- 2) El NameNode le envía los DataNodes donde se encuentran los bloques del fichero.

Instalación Hadoop en GNU/Linux

Instalación de Java:

```
sudo apt update
```

```
sudo apt install openjdk-8-jdk -y
```

```
java -version; javac -version
```


Instalación Hadoop en GNU/Linux

Creación de usuario hdoop y configuración ssh:

```
sudo apt install openssh-server openssh-client -y
```

```
sudo adduser hdoop
```

```
su - hdoop
```

```
ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
```

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

```
chmod 0600 ~/.ssh/authorized_keys
```

```
ssh localhost
```

Instalación de Hadoop en GNU/Linux

Descarga de Hadoop:

wget

<https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz>


tar xzf hadoop-3.3.4.tar.gz

Instalación de Hadoop en GNU/Linux

Edición de ficheros de configuración:

sudo adduser hdoop sudo

sudo nano .bashrc



```
export HADOOP_HOME=/home/hdoop/hadoop-3.3.4
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
```

source ~/.bashrc

```
export
HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export
HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

Instalación de Hadoop en GNU/Linux

Edición de ficheros de configuración:

```
sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

Añadir al final:

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


Instalación de Hadoop en GNU/Linux

Edición de ficheros de configuración:

sudo nano \$HADOOP_HOME/etc/hadoop/core-site.xml

Entre las etiquetas

<configuration></configuration>



```
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/hadoop/tmpdata</value>
  <description>A base for other temporary
directories.</description>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
  <description>The name of the default file
system></description>
</property>
```

Instalación de Hadoop en GNU/Linux

Edición de ficheros de configuración:

sudo nano \$HADOOP_HOME/etc/hadoop/hdfs-site.xml

```
<property>
  <name>dfs.data.dir</name>
  <value>/home/hadoop/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hadoop/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
```

Instalación de Hadoop en GNU/Linux

Edición de ficheros de configuración:

sudo nano \$HADOOP_HOME/etc/hadoop/mapred-site.xml

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

Instalación de Hadoop en GNU/Linux

Edición de ficheros de configuración:

sudo nano \$HADOOP_HOME/etc/hadoop/yarn-site.xml

```
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
```

CONTINÚA

Instalación de Hadoop en GNU/Linux

```
<property>
```

```
  <name>yarn.acl.enable</name>
```

```
  <value>0</value>
```

```
</property>
```

```
<property>
```

```
  <name>yarn.nodemanager.env-whitelist</name>
```

```
  <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
```

```
</property>
```

Ejecutando Hadoop

hdfs namenode -format → Formatea el sistema de ficheros especificado en hdfs-site.xml.

./start-dfs.sh → Dentro del directorio sbin, inicia el sistema de ficheros de hadoop.

Trabajando con HDFS

1. Lista el directorio raíz.
2. Crea un directorio en hdfs llamado hdp-test1.
3. Copia un fichero csv desde tu equipo al directorio hdp-test1 dentro de HDFS.
4. Copia el mismo fichero desde HDFS hasta tu equipo local, concretamente dentro del directorio /tmp.
5. Lista el directorio hdp-test1.
6. Crea dos directorios distintos dentro de hdp-test1, llamados hdp-test2 y hdp-test3.
7. Mueve el fichero csv que contiene hdp-test1 hacia hdp-test2.
8. Cambia los permisos del fichero csv movido a 777.
9. Copia el fichero csv a hdp-test1.
10. Elimina los 3 directorios creados.