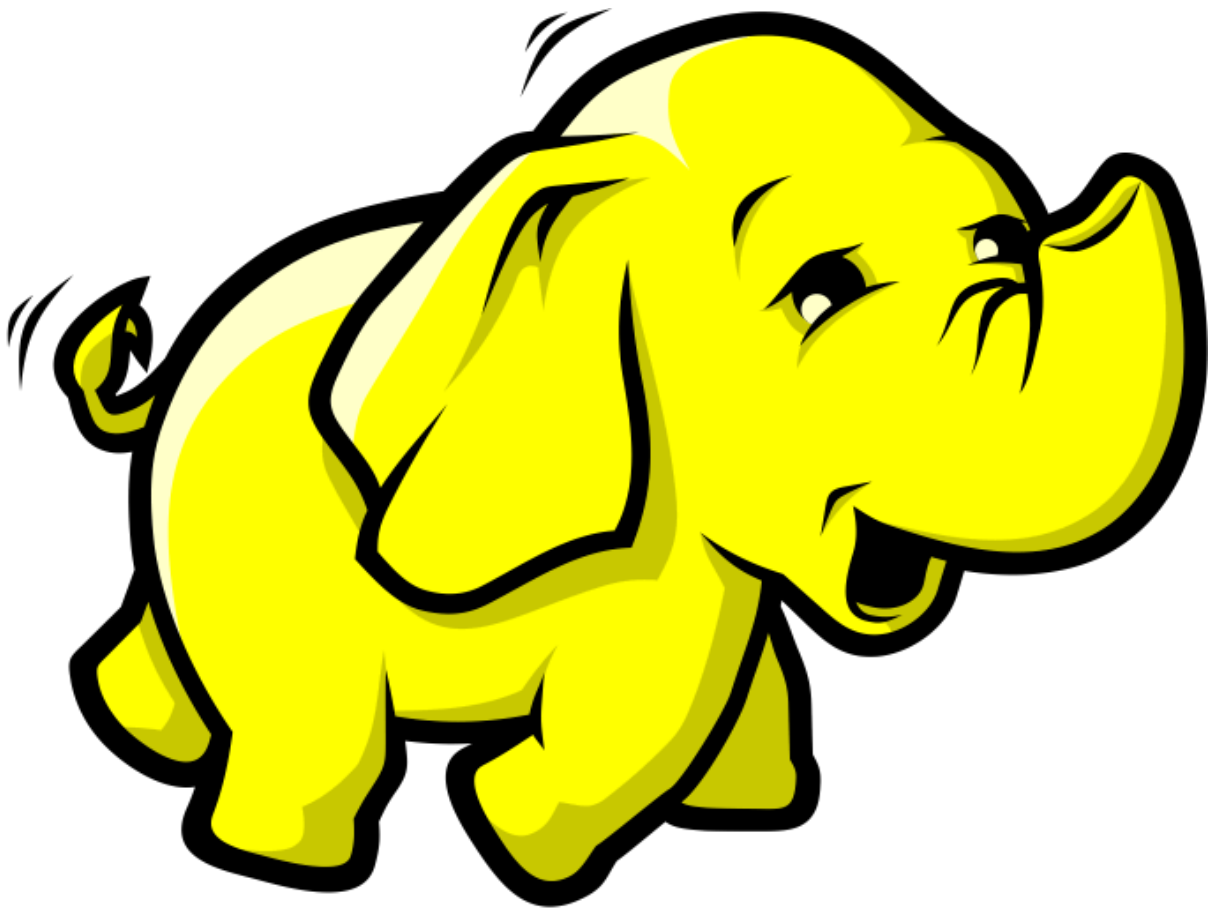


Introducción a Hadoop y HDFS



Adrián Yared Armas de la Nuez



Contenido

1. Instalación de Hadoop.....	2
1.1.1 Código.....	2
1.1.2 Prueba de ejecución.....	2
1.2. Posicionar la distribución.....	2
1.2.1 Código.....	2
1.2.2 Prueba de ejecución.....	2
1.2.3 Código.....	2
1.2.4 Prueba de ejecución.....	3
1.3 Configuración.....	3
1.3.1 Código.....	3
1.3.2 Prueba de ejecución.....	3
1.3.3 Código.....	3
1.3.4 Prueba de ejecución.....	3
1.4. Ejemplos.....	3
1.4.1 Volcado de hadoop en xml.....	4
1.4.2 Prueba de ejecución.....	4
1.4.3 Ejecución de los ejemplos.....	4
1.4.4 Prueba de ejecución.....	5
1.5 Copia.....	5
1.5.1 Código.....	5
1.5.2 Prueba de ejecución.....	5
2. HDFS.....	5
2.1 Crear el directorio prueba.....	5
2.1.1 Código.....	5
2.1.2 Prueba de ejecución.....	5
2.2 Crear un fichero local.....	5
2.2.1 Código.....	5
2.2.2 Prueba de ejecución.....	6
2.2.3 Mostrar su contenido.....	6
2.2.3.1 Código.....	6
2.2.3.2 Prueba de ejecución.....	6
3. Colab.....	7

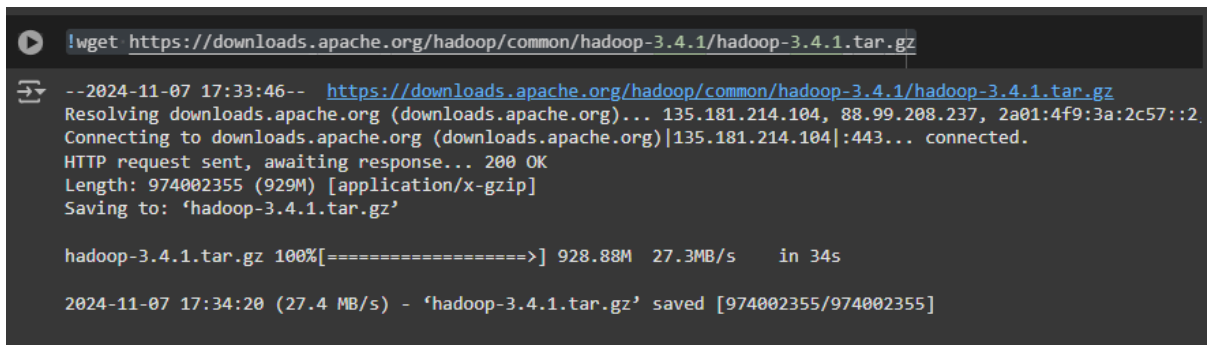
1. Instalación de Hadoop

1.1.1 Código

!wget

<https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz>

1.1.2 Prueba de ejecución



```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
--2024-11-07 17:33:46-- https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.208.237, 2a01:4f9:3a:2c57::2
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 974002355 (929M) [application/x-gzip]
Saving to: 'hadoop-3.4.1.tar.gz'

hadoop-3.4.1.tar.gz 100%[=====>] 928.88M 27.3MB/s in 34s

2024-11-07 17:34:20 (27.4 MB/s) - 'hadoop-3.4.1.tar.gz' saved [974002355/974002355]
```

1.2. Posicionar la distribución

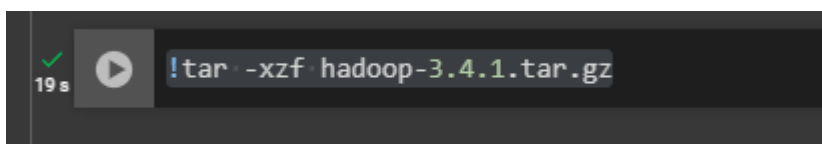
Extraer la distribución descargada en el sistema de archivos de colab y moverla a /usr/local

1.2.1 Código

//Extraer

!tar -xzf hadoop-3.4.1.tar.gz

1.2.2 Prueba de ejecución



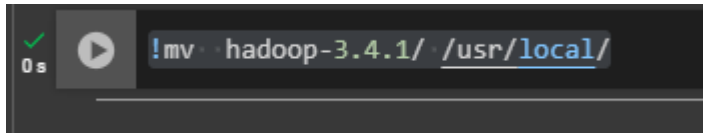
```
!tar -xzf hadoop-3.4.1.tar.gz
```

1.2.3 Código

//Mover a /usr/local/

!mv hadoop-3.4.1/ /usr/local/

1.2.4 Prueba de ejecución



```
0 s !mv hadoop-3.4.1/ /usr/local/
```

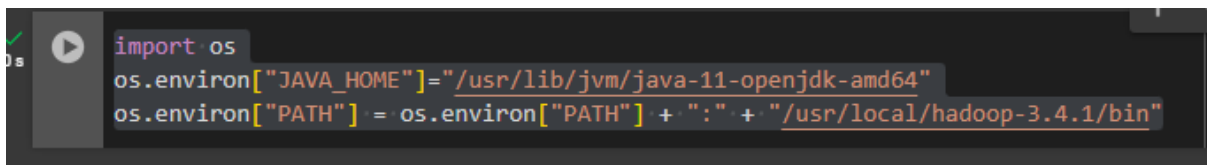
1.3 Configuración

Actualización variables de entorno (JAVA_HOME, PATH)

1.3.1 Código

```
import os
os.environ["JAVA_HOME"]="/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["PATH"] = os.environ["PATH"] + ":" + "/usr/local/hadoop-3.4.1/bin"
```

1.3.2 Prueba de ejecución

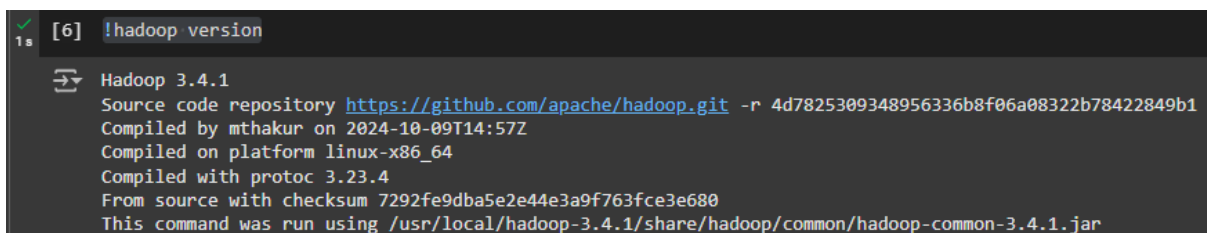


```
import os
os.environ["JAVA_HOME"]="/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["PATH"] = os.environ["PATH"] + ":" + "/usr/local/hadoop-3.4.1/bin"
```

1.3.3 Código

```
//Comprobación de la instalación
!hadoop version
```

1.3.4 Prueba de ejecución



```
[6] !hadoop version

Hadoop 3.4.1
Source code repository https://github.com/apache/hadoop.git -r 4d7825309348956336b8f06a08322b78422849b1
Compiled by mthakur on 2024-10-09T14:57Z
Compiled on platform linux-x86_64
Compiled with protoc 3.23.4
From source with checksum 7292fe9dba5e2e44e3a9f763f3e3e680
This command was run using /usr/local/hadoop-3.4.1/share/hadoop/common/hadoop-common-3.4.1.jar
```

1.4. Ejemplos

Una de las formas tradicionales de asegurarnos que un ambiente de Hadoop recién instalado funciona correctamente, es ejecutando el jar de ejemplos map-reduce incluido con toda instalación de hadoop (hadoop-mapreduce-examples.jar).

[Hadoop Map Reduce Examples](#)

1.4.1 Volcado de hadoop en xml

//Código:

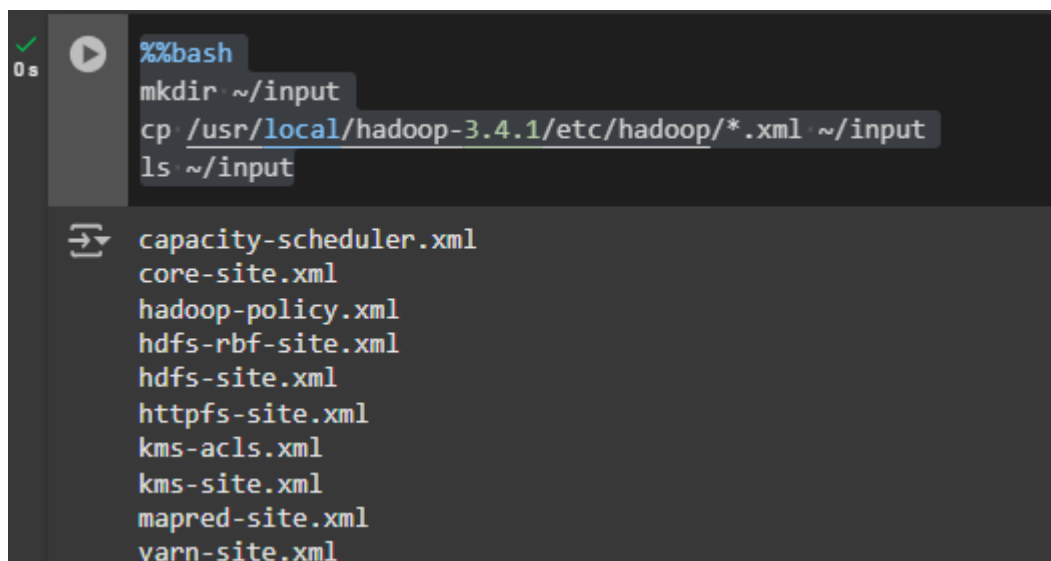
```
%%bash
```

```
mkdir ~/input
```

```
cp /usr/local/hadoop-3.4.1/etc/hadoop/*.xml ~/input
```

```
ls ~/input
```

1.4.2 Prueba de ejecución



```
0s %%bash
mkdir ~/input
cp /usr/local/hadoop-3.4.1/etc/hadoop/*.xml ~/input
ls ~/input
capacity-scheduler.xml
core-site.xml
hadoop-policy.xml
hdfs-rbf-site.xml
hdfs-site.xml
httpfs-site.xml
kms-acls.xml
kms-site.xml
mapred-site.xml
yarn-site.xml
```

1.4.3 Ejecución de los ejemplos

Ejecutamos `hadoop jar` con el fin de ejecutar uno de los ejemplos por defecto, en este caso el `grep` que busca expresiones regulares dentro de los ficheros que le especifiquemos.

//Código:

```
%%bash
```

```
hadoop jar \
```

```
/usr/local/hadoop-3.4.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar \
```

```
grep ~/input ~/grep_example 'allowed[.]'
```

1.4.4 Prueba de ejecución

```
%%bash
hadoop jar \
  /usr/local/hadoop-3.4.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar \
  grep ~/input ~/grep_example 'allowed[.]*'

2024-11-12 19:22:19,186 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-11-12 19:22:19,821 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-12 19:22:19,823 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-12 19:22:20,769 INFO input.FileInputFormat: Total input files to process : 10
```

1.5 Copia

1.5.1 Código

```
!cat ~/grep_example/*
```

1.5.2 Prueba de ejecución

```
!cat ~/grep_example/*

27      allowed.
1       allowed
```

2. HDFS

2.1 Crear el directorio prueba

2.1.1 Código

```
!hdfs dfs -mkdir prueba
```

2.1.2 Prueba de ejecución

```
!hdfs dfs -mkdir prueba
```

2.2 Crear un fichero local

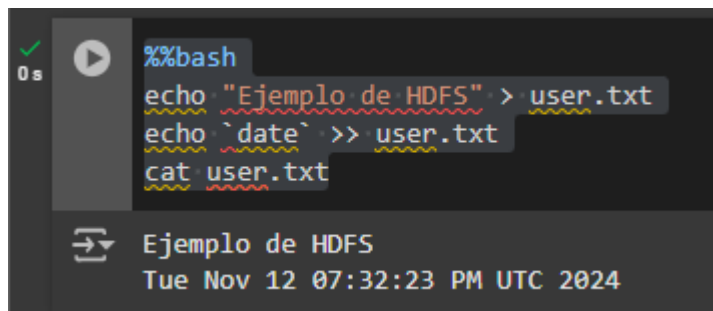
2.2.1 Código

```
%%bash
echo "Ejemplo de HDFS" > user.txt
```

```
echo `date` >> user.txt  
cat user.txt
```

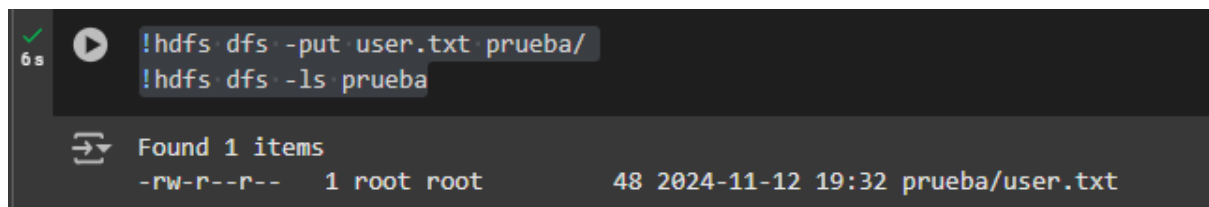
```
!hdfs dfs -put user.txt prueba/  
!hdfs dfs -ls prueba
```

2.2.2 Prueba de ejecución



```
%%bash  
echo "Ejemplo de HDFS" > user.txt  
echo `date` >> user.txt  
cat user.txt
```

Ejemplo de HDFS
Tue Nov 12 07:32:23 PM UTC 2024



```
!hdfs dfs -put user.txt prueba/  
!hdfs dfs -ls prueba
```

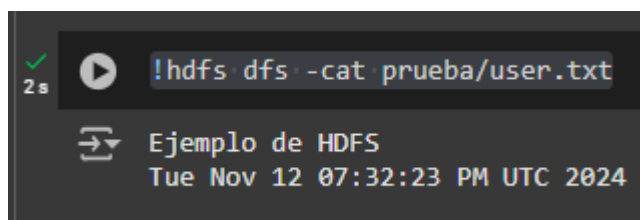
Found 1 items
-rw-r--r-- 1 root root 48 2024-11-12 19:32 prueba/user.txt

2.2.3 Mostrar su contenido

2.2.3.1 Código

```
!hdfs dfs -cat prueba/user.txt  
%%bash  
hdfs dfs -tail prueba/user.txt
```

2.2.3.2 Prueba de ejecución

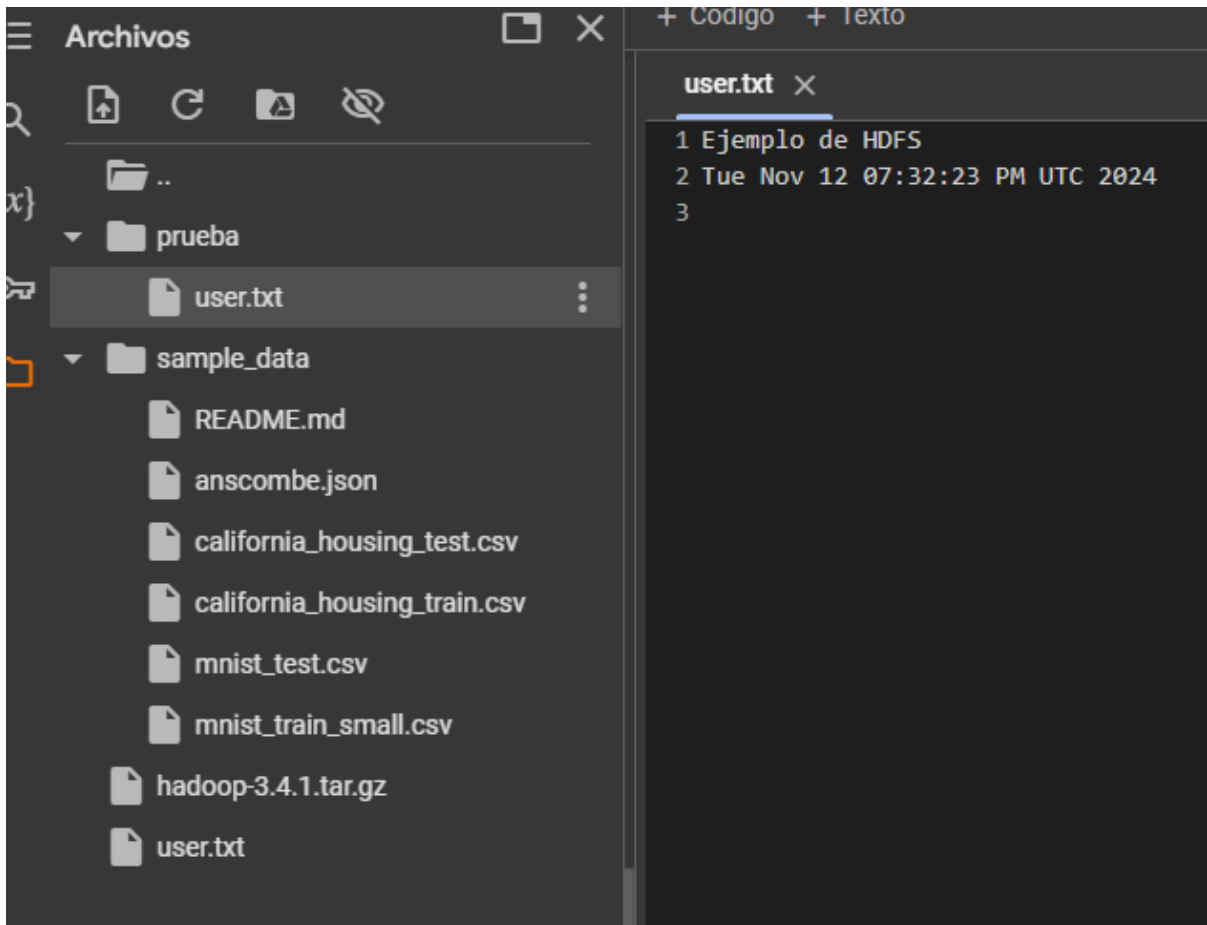


```
!hdfs dfs -cat prueba/user.txt
```

Ejemplo de HDFS
Tue Nov 12 07:32:23 PM UTC 2024

```
6 s  %%bash
      hdfs dfs -tail prueba/user.txt
```

Ejemplo de HDFS
Tue Nov 12 07:32:23 PM UTC 2024



3. Colab

