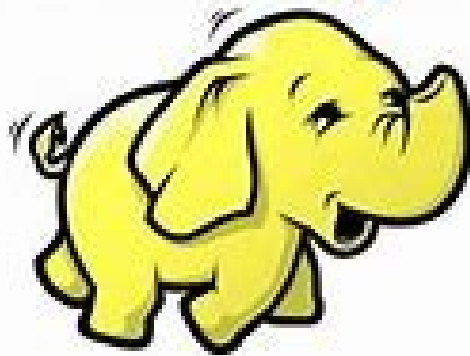


hadoop



Introducción a Hadoop y HDFS

Javier Díaz Machado

ÍNDICE

- Instalación y comandos HDFS

Instalación y comandos HDFS

Google Colab es una herramienta online gratuita basada en la nube que permite desplegar modelos de aprendizaje automático de forma remota en CPUs y GPUs

Se basa en la tecnología de código abierto Jupyter, con la que se puede crear un cuaderno Ipython. El cuaderno Ipython no sólo te ofrece la posibilidad de escribir código, sino también de contar una historia a través de él.

Puedes ejecutar código, crear visualizaciones de datos, y escribir sobre cada paso que se haga en texto plano o markdown. Esto hace que el código incluya explicaciones y visualizaciones de lo que hace.

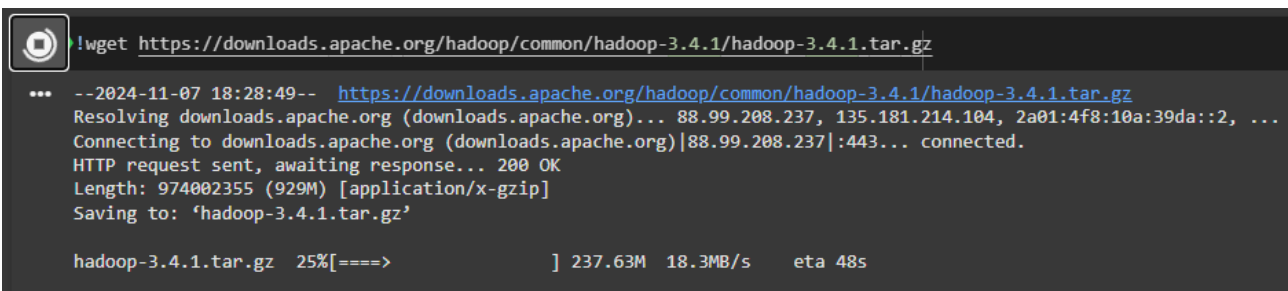
Hadoop

Hadoop es un marco de programación basado en Java que permite procesar y almacenar conjuntos de datos extremadamente grandes en un clúster de máquinas de bajo coste. Fue el primer gran proyecto de código abierto en el ámbito del Big Data y está patrocinado por la Apache Software Foundation.

Instalación de Hadoop:

Descargamos la distribución:

```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
```

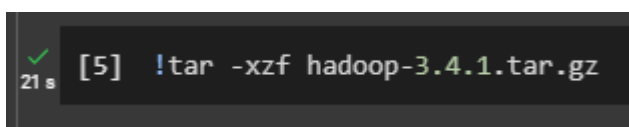


```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
... --2024-11-07 18:28:49-- https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.208.237, 135.181.214.104, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.208.237|: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  25%[====>                ] 237.63M  18.3MB/s   eta 48s
```

Extraerla:

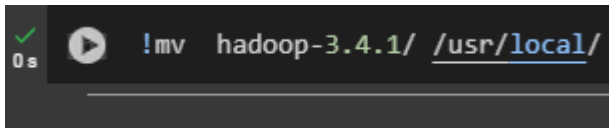
```
!tar -xzf hadoop-3.4.1.tar.gz
```



```
✓ 21 s [5] !tar -xzf hadoop-3.4.1.tar.gz
```

Mover la distribución a [/usr/local](#):

```
!mv hadoop-3.3.2/ /usr/local/
```



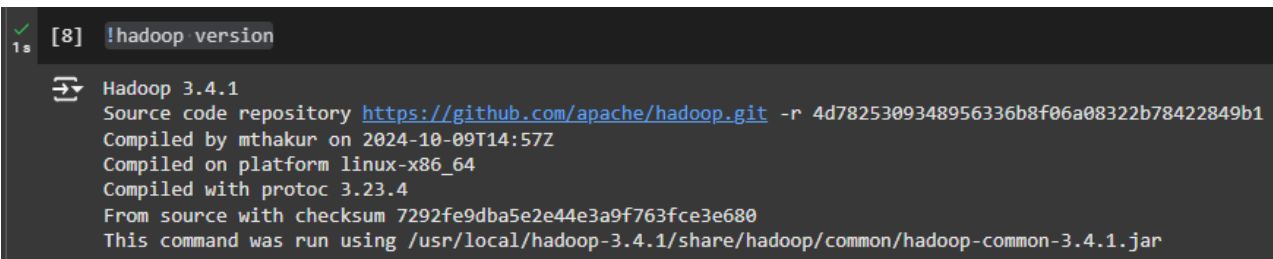
Configuración

Actualizamos variables de entorno (JAVA_HOME, PATH)

```
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"
```

Comprobamos instalación

```
!hadoop version
```

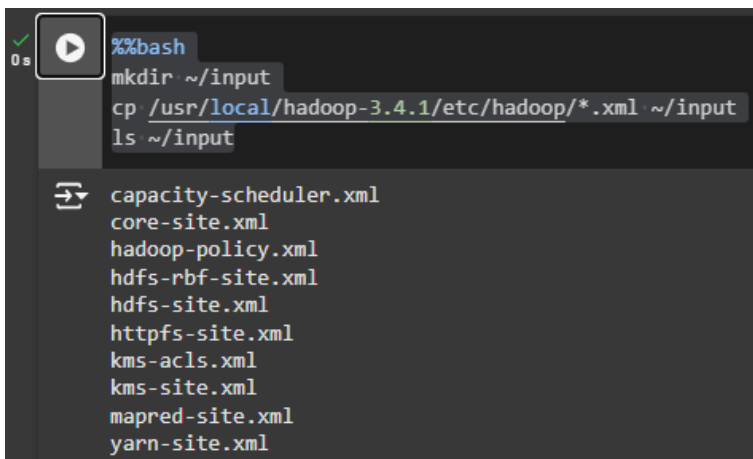


Ejecución de Ejemplos

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*).

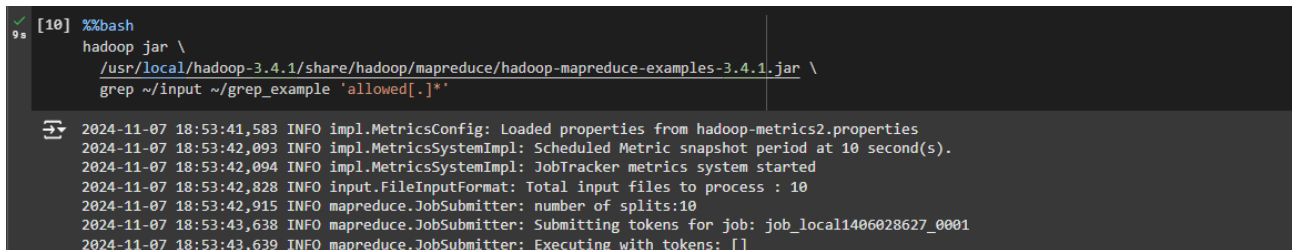
Creamos un directorio de ficheros en los que volquemos los xml de hadoop

```
%%bash
mkdir ~/input
cp /usr/local/hadoop-3.4.1/etc/hadoop/*.xml ~/input
ls ~/input
```



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.

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

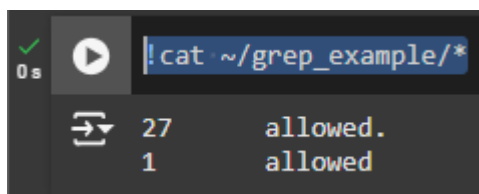


```
[10] %%bash
hadoop jar \
  /usr/local/hadoop-3.4.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.1.jar \
  grep ~/input ~/grep_example 'allowed[.]*'

2024-11-07 18:53:41,583 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-11-07 18:53:42,093 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-07 18:53:42,094 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-07 18:53:42,828 INFO input.FileInputFormat: Total input files to process : 10
2024-11-07 18:53:42,915 INFO mapreduce.JobSubmitter: number of splits:10
2024-11-07 18:53:43,638 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1406028627_0001
2024-11-07 18:53:43,639 INFO mapreduce.JobSubmitter: Executing with tokens: []
```

Mostramos el resultado

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



```
!cat ~/grep_example/*

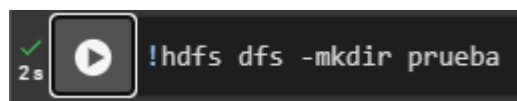
27    allowed.
1      allowed
```

HDFS

Las siguientes sentencias únicamente sirven para probar comandos básicos de HDFS no para gestionar una Infraestructura que en Google Colab no existe, en este caso el sistema de archivos HDFS es el mismo que el local.

Crear el directorio prueba

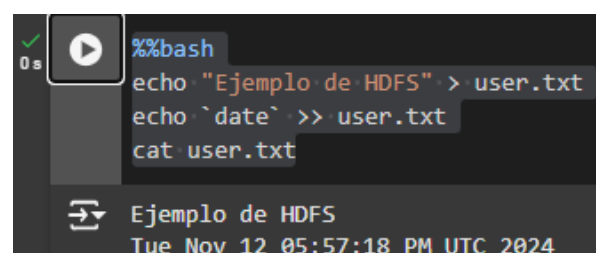
```
!hdfs dfs -mkdir prueba
```



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

Crear un fichero local :

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



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

Ejemplo de HDFS
Tue Nov 12 05:57:18 PM UTC 2024
```

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

```
!hdfs dfs -ls prueba
```

```
5s [play] !hdfs dfs -put user.txt prueba/
!hdfs dfs -ls prueba

Found 1 items
-rw-r--r-- 1 root root 48 2024-11-12 17:57 prueba/user.txt
```

Mostrar su contenido

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

```
2s [play] !hdfs dfs -cat prueba/user.txt

Ejemplo de HDFS
Tue Nov 12 05:57:18 PM UTC 2024
```

```
%%bash
```

```
hdfs dfs -tail prueba/user.txt
```

```
2s [play] %%bash
hdfs dfs -tail prueba/user.txt

Ejemplo de HDFS
Tue Nov 12 05:57:18 PM UTC 2024
```