

Instalación de Hadoop y configuración de HDFS

Introducción al Big Data Analytics(BDA)

Máster Universitario en Inteligencia Computacional e Internet de las Cosas

Universidad de Córdoba, EPSC

2022/2023



Autor:

Antonio Gómez Giménez (i72gogia@uco.es)

Índice:

1. Ejercicio de comprobación de la instalación:	2
1.1. Comprobar el estado del sistema HDFS	2
1.2. Crear el directorio de usuario en HDFS	2
1.3. Crear un directorio de prueba para el ejemplo	3
1.4. Copiar un fichero en el directorio de usuario	3
1.5. Ejecución de un ejemplo MapReduce	4
1.6. Visualización del resultado	5
1.7. Terminar la ejecución	5

1. Ejercicio de comprobación de la instalación:

1.1. Comprobar el estado del sistema HDFS

El comando usado ha sido el siguiente:

-> bin/hdfs fsck /

La salida obtenida es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hdfs fsck /
Connecting to namenode via http://localhost:9870/fsck?ugi=root&path=%2F
FSCK started by root (auth:SIMPLE) from /127.0.0.1 for path / at Wed Dec 21 17:5
3:10 CET 2022

Status: HEALTHY
Number of data-nodes: 1
Number of racks: 1
Total dirs: 1
Total symlinks: 0

Replicated Blocks:
Total size: 0 B
Total files: 0
Total blocks (validated): 0
Minimally replicated blocks: 0
Over-replicated blocks: 0
Under-replicated blocks: 0
Mis-replicated blocks: 0
Default replication factor: 1
Average block replication: 0.0
Missing blocks: 0
Corrupt blocks: 0
```

1.2. Crear el directorio de usuario en HDFS

Los comandos usados han sido los siguientes:

-> bin/hdfs dfs -mkdir /user

-> bin/hdfs dfs -mkdir /user/root

La salida obtenida es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hdfs dfs -mkdir /user
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hdfs dfs -mkdir /user/root
```

1.3. Crear un directorio de prueba para el ejemplo

El comando usado ha sido el siguiente:

-> bin/hdfs dfs -mkdir /prueba

La salida obtenida es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hdfs dfs -mkdir /prueba
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1#
```

1.4. Copiar un fichero en el directorio de usuario

El comando usado para realizar la copia de un fichero ha sido el siguiente:

-> bin/hdfs dfs -put /home/antoniogg/hadoop-3.3.1/LICENSE.txt /prueba

El comando usado para mostrar que se ha copiado correctamente es el siguiente:

-> bin/hdfs dfs -ls /prueba

La salida obtenida es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hdfs dfs -ls /prueba
Found 1 items
-rw-r--r--  1 root supergroup      15217 2022-12-23 17:02 /prueba/LICENSE.txt
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1#
```

1.5. Ejecución de un ejemplo MapReduce

El comando usado ha sido el siguiente:

```
-> bin/hdfs dfs -put /home/antoniogg/hadoop-3.3.1/LICENSE.txt /prueba
```

Para la compilación y código de java se recurrió al tutorial dado por hadoop, en el siguiente enlace se puede encontrar:

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example: WordCount_v1.0

Como el directorio prueba existía y contenía el fichero LICENSE.txt copiado anteriormente, la salida se volcó en un nuevo directorio llamado output.

El comando final usado es el siguiente:

```
-> bin/hadoop jar wc.jar WordCount /prueba/LICENSE.txt /prueba/output
```

La salida obtenida es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hadoop jar w
c.jar WordCount /prueba/LICENSE.txt /prueba/output
2022-12-23 18:27:04,961 INFO impl.MetricsConfig: Loaded properties from hadoop
-metrics2.properties
2022-12-23 18:27:05,204 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot
period at 10 second(s).
2022-12-23 18:27:05,205 INFO impl.MetricsSystemImpl: JobTracker metrics system
started
2022-12-23 18:27:05,478 WARN mapreduce.JobResourceUploader: Hadoop command-lin
e option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
2022-12-23 18:27:05,673 INFO input.FileInputFormat: Total input files to proce
ss : 1
2022-12-23 18:27:05,881 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-23 18:27:06,595 INFO mapreduce.JobSubmitter: Submitting tokens for job
: job_local1413369068_0001
2022-12-23 18:27:06,595 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-23 18:27:06,890 INFO mapreduce.Job: The url to track the job: http://l
ocalhost:8080/
```

1.6. Visualización del resultado

El comando usado para mostrar los resultados del apartado anterior ha sido el siguiente:

-> /home/antoniogg/hadoop-3.3.1# bin/hadoop fs -cat /prueba/output/part-r-00000

La salida obtenida es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# bin/hadoop fs -cat /prueba/output/part-r-00000
"AS" 2
"Contribution" 1
"Contributor" 1
"Derivative" 1
"Legal" 1
"License" 1
"License"); 1
"Licensor" 1
"NOTICE" 1
"Not" 1
"Object" 1
"Source" 1
"Work" 1
"You" 1
"Your" 1
```

1.7. Terminar la ejecución

El comando usado para terminar la ejecución ha sido el siguiente:

-> sbin/stop-dfs.sh

La salida obtenida por consola es la siguiente:

```
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1# sbin/stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [antoniogg-SATELLITE-L50D-C]
root@antoniogg-SATELLITE-L50D-C:/home/antoniogg/hadoop-3.3.1#
```