Práctica 6: Apache Hadoop/Apache Spark

Iago Domínguez Cameán <iago.dominguez.camean@udc.es>

Iker García Calviño <iker.gcalvino@udc.es>

Ejercicio 1

```
vagrant@idc-aisi2223-master:~$ cat /etc/ansible/hosts
[masters]
idc-aisi2223-master
[workers]
idc-aisi2223-worker-1
# Group 'cluster' with all nodes
[cluster:children]
masters
workers
vagrant@idc-aisi2223-master:~$ ansible cluster -m ping
idc-aisi2223-master | SUCCESS => {
    "ansible_facts": {
        "discovered_interpreter_python": "/usr/bin/python3"
    },
    "changed": false,
    "ping": "pong"
}
idc-aisi2223-worker-1 | SUCCESS => {
    "ansible_facts": {
        "discovered_interpreter_python": "/usr/bin/python3"
    },
    "changed": false,
    "ping": "pong"
}
vagrant@idc-aisi2223-master:~$ ansible cluster -m shell -a "df -h | grep /data"
idc-aisi2223-master | CHANGED | rc=0 >>
/dev/sdb        9.8G   24K  9.3G   1% /data/disk0
/dev/sdc        9.8G   24K  9.3G   1% /data/disk1
idc-aisi2223-worker-1 | CHANGED | rc=0 >>
/dev/sdb        9.8G   24K  9.3G   1% /data/disk0
/dev/sdc        9.8G   24K  9.3G   1% /data/disk1
vagrant@idc-aisi2223-master:~$
```

Figure 1: Despliegue del clúster virtual

Ejercicio 2

```
vagrant@idc-aisi2223-master:~$ sudo exportfs
/share/nfs      10.10.1.10/24
vagrant@idc-aisi2223-master:~$ ansible workers -a "cat /proc/mounts" | grep "share/nfs"
idc-aisi2223-master:/share/nfs /share/nfs nfs4 rw,sync,relatime,vers=4.2,rsize=262144,wsize=262144,namlen=255,hard,proto=tcp,timeo=600,retrans=2,sec=sys,clientaddr=10.10.1.11,local_lock=none,addr=10.10.1.10 0 0
vagrant@idc-aisi2223-master:~$
```

Figure 2: Configuración de NFS

Ejercicio 3

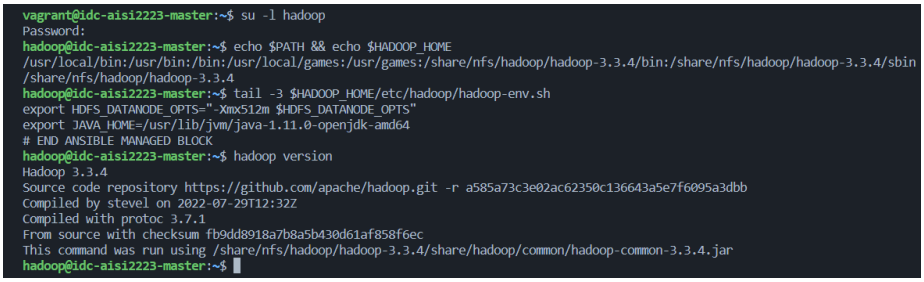

Figure 3: Instalación de Hadoop

Ejercicio 4

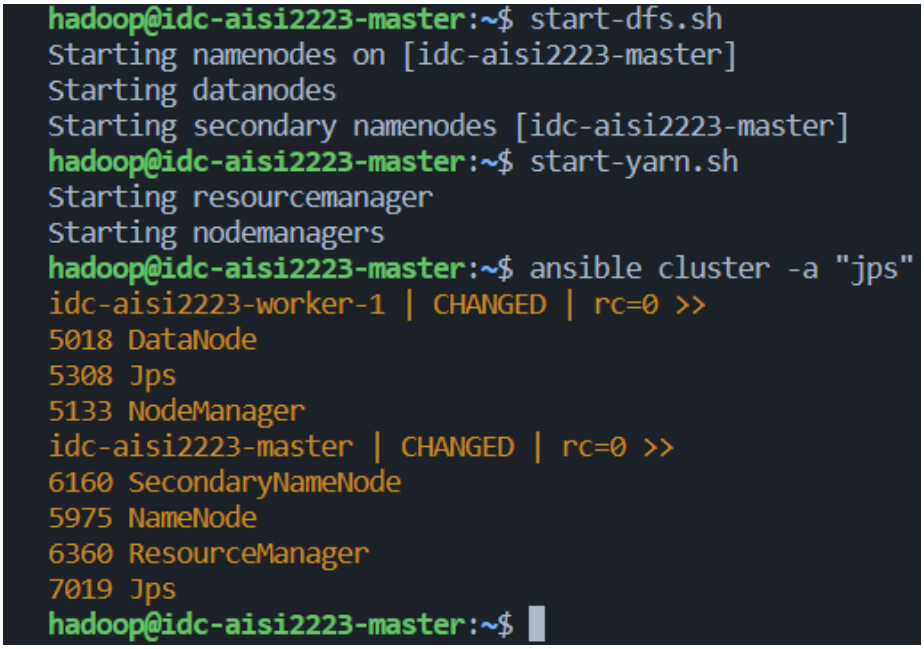

Figure 4: Despliegue de Hadoop

Ejercicio 5

```
hadoop@idc-aisi2223-master:~$ ls -lh $HOME
total 453M
-rw-r--r-- 1 hadoop hadoop 453M Jan 13  2022 file.xml
hadoop@idc-aisi2223-master:~$ hdfs dfs -ls -R -h /
drwxr-xr-x   - hadoop supergroup          0 2023-05-02 23:11 /hadoop-input
-rw-r--r--   2 hadoop supergroup    452.5 M 2023-05-02 23:11 /hadoop-input/file.xml
hadoop@idc-aisi2223-master:~$ ansible workers -m shell -a "df -h | grep /data"
idc-aisi2223-worker-1 | CHANGED | rc=0 >>
/dev/sdb                      9.8G  231M  9.1G   3% /data/disk0
/dev/sdc                      9.8G  226M  9.1G   3% /data/disk1
hadoop@idc-aisi2223-master:~$ hdfs dfs -df -h
Filesystem                          Size     Used  Available  Use%
hdfs://idc-aisi2223-master:9000  19.5 G  456.2 M     18.0 G    2%
hadoop@idc-aisi2223-master:~$
```

Figure 5: Información del HDFS

```
hadoop@idc-aisi2223-master:~$ hdfs dfs -ls -R -h /hadoop-output1
-rw-r--r--   2 hadoop supergroup          0 2023-05-03 00:14 /hadoop-output1/_SUCCESS
-rw-r--r--   2 hadoop supergroup    144.9 M 2023-05-03 00:14 /hadoop-output1/part-r-00000
hadoop@idc-aisi2223-master:~$ hdfs dfs -ls -R -h /hadoop-output4
-rw-r--r--   2 hadoop supergroup          0 2023-05-03 00:19 /hadoop-output4/_SUCCESS
-rw-r--r--   2 hadoop supergroup     36.0 M 2023-05-03 00:18 /hadoop-output4/part-r-00000
-rw-r--r--   2 hadoop supergroup     36.0 M 2023-05-03 00:18 /hadoop-output4/part-r-00001
-rw-r--r--   2 hadoop supergroup     36.3 M 2023-05-03 00:19 /hadoop-output4/part-r-00002
-rw-r--r--   2 hadoop supergroup     36.6 M 2023-05-03 00:19 /hadoop-output4/part-r-00003
hadoop@idc-aisi2223-master:~$
```

Figure 6: Directorios de salida

```
hadoop@idc-aisi2223-master:~$ hdfs dfs -getmerge /hadoop-output4 /tmp/hadoopOutputR4
hadoop@idc-aisi2223-master:~$
hadoop@idc-aisi2223-master:~$ ls -lh /tmp/hadoopOutputR*
-rw-r--r-- 1 hadoop hadoop 145M May  3 00:24 /tmp/hadoopOutputR1
-rw-r--r-- 1 hadoop hadoop 145M May  3 00:26 /tmp/hadoopOutputR4
hadoop@idc-aisi2223-master:~$
hadoop@idc-aisi2223-master:~$ cat /tmp/hadoopOutputR1 | grep "<page>"
<page>  134342
hadoop@idc-aisi2223-master:~$
hadoop@idc-aisi2223-master:~$ cat /tmp/hadoopOutputR4 | grep "<page>"
<page>  134342
hadoop@idc-aisi2223-master:~$
```

Figure 7: Número de ocurrencias que tiene en el fichero de entrada la etiqueta XML “<page>”

Ejercicio 6

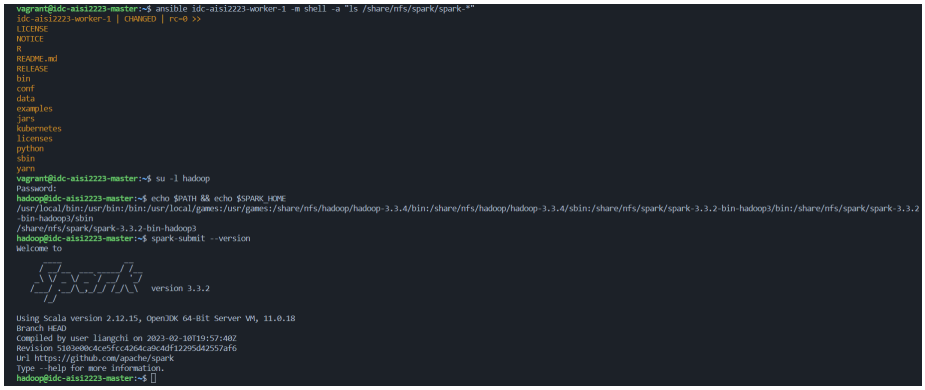

Figure 8: Instalación de Spark

Ejercicio 7

```
hadoop@idc-aisi2223-master:~$ hdfs dfs -ls -R -h /spark-output1
-rw-r--r--   2 hadoop supergroup          0 2023-05-04 09:40 /spark-output1/_SUCCESS
-rw-r--r--   2 hadoop supergroup    144.9 M 2023-05-04 09:40 /spark-output1/part-r-00000
hadoop@idc-aisi2223-master:~$ hdfs dfs -ls -R -h /spark-output4
-rw-r--r--   2 hadoop supergroup          0 2023-05-04 09:43 /spark-output4/_SUCCESS
-rw-r--r--   2 hadoop supergroup     36.4 M 2023-05-04 09:43 /spark-output4/part-r-00000
-rw-r--r--   2 hadoop supergroup     36.0 M 2023-05-04 09:43 /spark-output4/part-r-00001
-rw-r--r--   2 hadoop supergroup     36.2 M 2023-05-04 09:43 /spark-output4/part-r-00002
-rw-r--r--   2 hadoop supergroup     36.4 M 2023-05-04 09:43 /spark-output4/part-r-00003
hadoop@idc-aisi2223-master:~$
```

Figure 9: Directorios de salida

```
hadoop@idc-aisi2223-master:~$ hdfs dfs -getmerge /spark-output4 /tmp/sparkOutputR4
hadoop@idc-aisi2223-master:~$ ls -lh /tmp/sparkOutputR4
-rw-r--r-- 1 hadoop hadoop 145M May  4 09:53 /tmp/sparkOutputR4
hadoop@idc-aisi2223-master:~$ cat /tmp/sparkOutputR4 | grep "<page>"
<page>  134342
hadoop@idc-aisi2223-master:~$
```

Figure 10: Número de ocurrencias que tiene en el fichero de entrada la etiqueta XML “<page>”

Ejercicio 8

```
hadoop@idc-aisi2223-master:~$ hdfs dfs -rm -R /tmp
Deleted /tmp
hadoop@idc-aisi2223-master:~$ hdfs dfs -ls -R -h /
drwxr-xr-x   - hadoop supergroup          0 2023-05-02 23:11 /hadoop-input
-rw-r--r--   2 hadoop supergroup    452.5 M 2023-05-02 23:11 /hadoop-input/file.xml
drwxr-xr-x   - hadoop supergroup          0 2023-05-03 00:14 /hadoop-output1
-rw-r--r--   2 hadoop supergroup          0 2023-05-03 00:14 /hadoop-output1/_SUCCESS
-rw-r--r--   2 hadoop supergroup    144.9 M 2023-05-03 00:14 /hadoop-output1/part-r-00000
drwxr-xr-x   - hadoop supergroup          0 2023-05-03 00:19 /hadoop-output4
-rw-r--r--   2 hadoop supergroup          0 2023-05-03 00:19 /hadoop-output4/_SUCCESS
-rw-r--r--   2 hadoop supergroup     36.0 M 2023-05-03 00:18 /hadoop-output4/part-r-00000
-rw-r--r--   2 hadoop supergroup     36.0 M 2023-05-03 00:18 /hadoop-output4/part-r-00001
-rw-r--r--   2 hadoop supergroup     36.3 M 2023-05-03 00:19 /hadoop-output4/part-r-00002
-rw-r--r--   2 hadoop supergroup     36.6 M 2023-05-03 00:19 /hadoop-output4/part-r-00003
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 09:40 /spark-output1
-rw-r--r--   2 hadoop supergroup          0 2023-05-04 09:40 /spark-output1/_SUCCESS
-rw-r--r--   2 hadoop supergroup    144.9 M 2023-05-04 09:40 /spark-output1/part-r-00000
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 09:43 /spark-output4
-rw-r--r--   2 hadoop supergroup          0 2023-05-04 09:43 /spark-output4/_SUCCESS
-rw-r--r--   2 hadoop supergroup     36.4 M 2023-05-04 09:43 /spark-output4/part-r-00000
-rw-r--r--   2 hadoop supergroup     36.0 M 2023-05-04 09:43 /spark-output4/part-r-00001
-rw-r--r--   2 hadoop supergroup     36.2 M 2023-05-04 09:43 /spark-output4/part-r-00002
-rw-r--r--   2 hadoop supergroup     36.4 M 2023-05-04 09:43 /spark-output4/part-r-00003
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 09:37 /user
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 09:37 /user/hadoop
drwxr-xr-x   - hadoop supergroup          0 2023-05-04 09:43 /user/hadoop/.sparkStaging
hadoop@idc-aisi2223-master:~$ hdfs dfs -df -h
Filesystem                      Size   Used  Available  Use%
hdfs://idc-aisi2223-master:9000  19.5 G  1.0 G     16.9 G    5%
hadoop@idc-aisi2223-master:~$
```

Figure 11: Limpieza