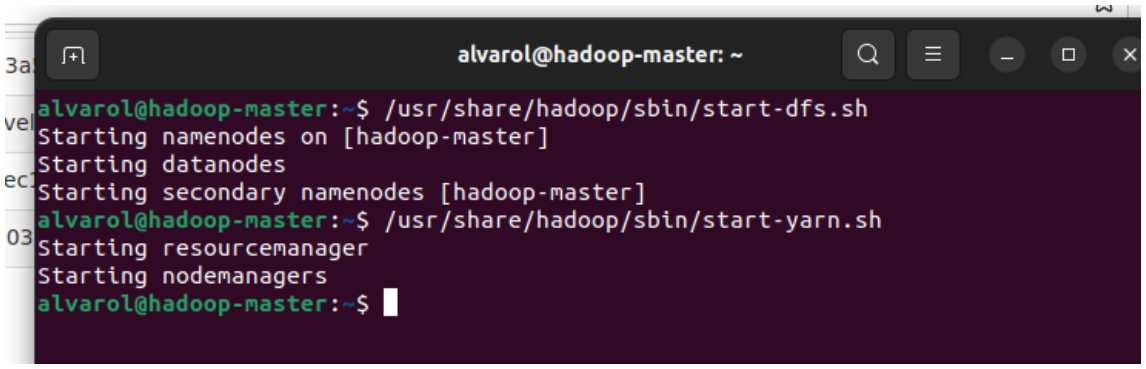
PRÁCTICA Nº 2.3
Hadoop Streaming

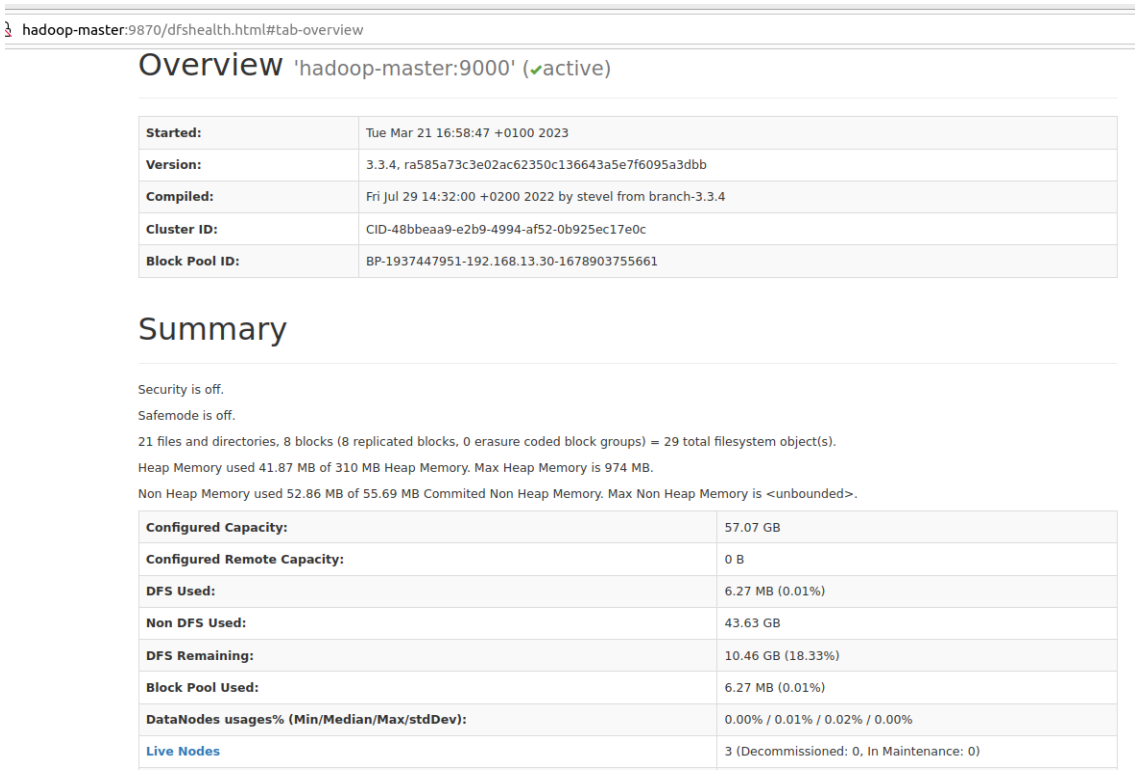
- Nombre y apellidos: Alvaro Lucio-Villegas de Cea

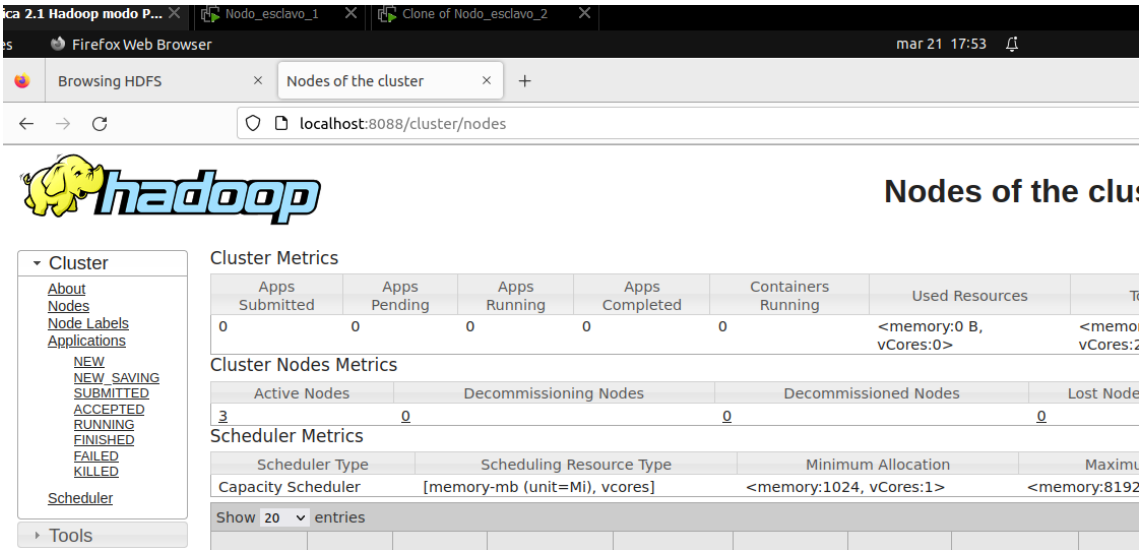
Índice

Proceso Inicial.

● Los archivos que actuarán como mapeadores/reductores deben estar en la máquina (preferiblemente en hadoop-master de tu cluster montado en la práctica 2.2).







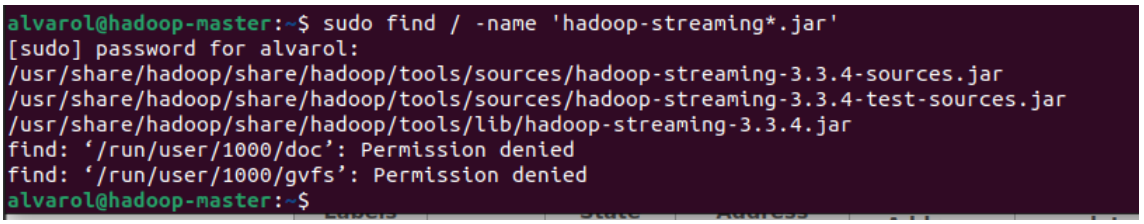
- Ajusta los permisos de los archivos:

```
chmod u+rwx mapper.py reducer.py
```

```
alvarol@hadoop-master:~$ chmod u+rwx mapper.py reducer.py
alvarol@hadoop-master:~$ ls -l
total 109548
-rw-rw-r-- 1 alvarol alvarol 14666916 mar 21 17:29 20news-18828.tar.gz
drwx------ 3 alvarol alvarol     4096 mar 15 18:15 data
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Desktop
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Documents
drwxr-xr-x 2 alvarol alvarol     4096 mar 21 17:30 Downloads
-rw------- 1 alvarol alvarol  2198936 mar  7 18:05 ElQuijote.txt
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 18:21 libro_salida
-rwx------ 1 alvarol alvarol      767 mar 21 16:53 mapper.py
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Music
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 18:35 name
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 18:35 namesecondary
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Pictures
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Public
-rwx------ 1 alvarol alvarol     1169 mar 21 16:53 reducer.py
drwx------ 4 alvarol alvarol     4096 mar  7 17:51 snap
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Templates
drwxr-xr-x 5 alvarol alvarol     4096 mar  8 20:17 tmp
drwxr-xr-x 2 alvarol alvarol     4096 mar  7 17:39 Videos
-rw-rw-r-- 1 alvarol alvarol 95238493 mar  8 19:57 working_dataset.xlsx
alvarol@hadoop-master:~$
```

● Localiza el jar de hadoop-streaming para poder ejecutarlo. Por ejemplo usando:

```
find / -name 'hadoop-streaming*.jar'
```



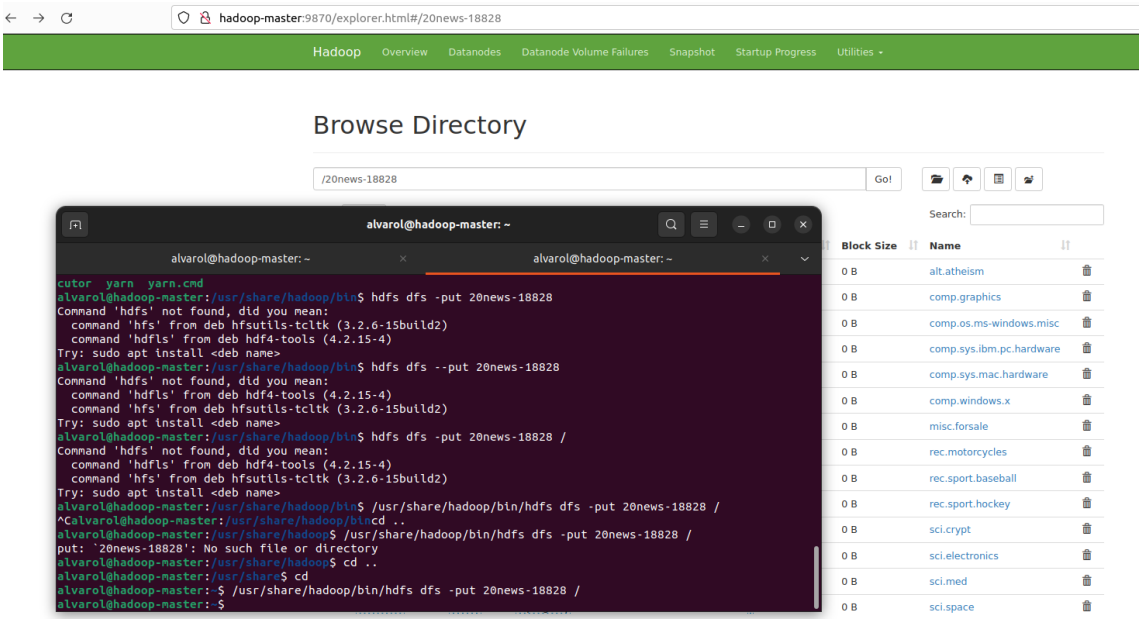
● Elige un dataset de entrada para la prueba. Por ejemplo

http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz. Descomprímelo:

```
tar -xzf 20news-18828.tar.gz
```

Actividad:

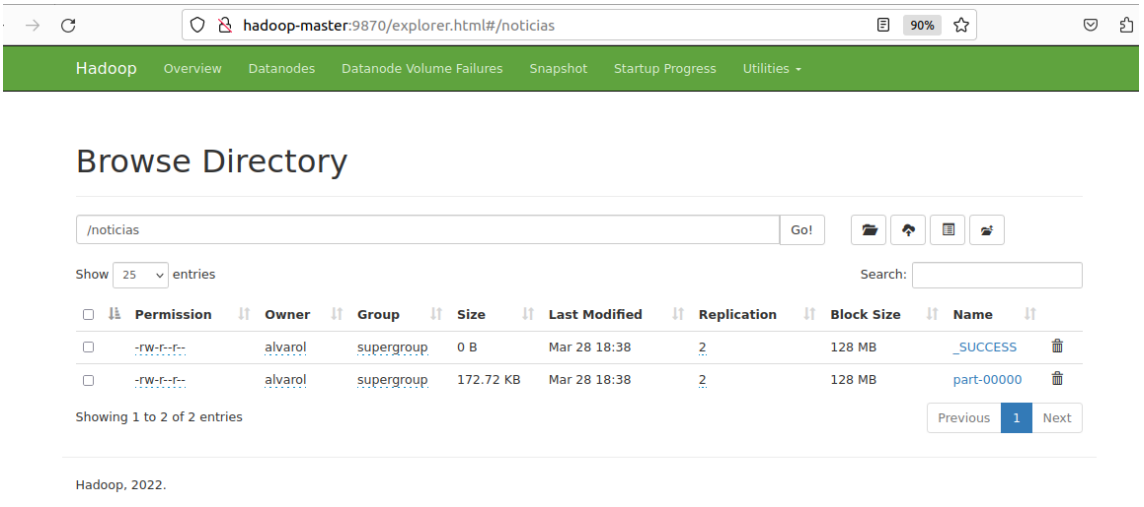
1º Cargar archivos en hdfs



2ºEjecutar los ficheros python de cada grupo

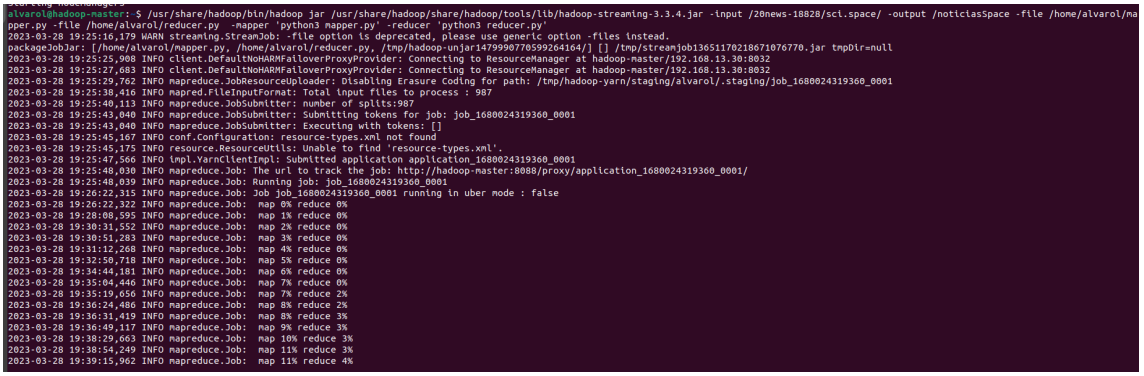
Autos

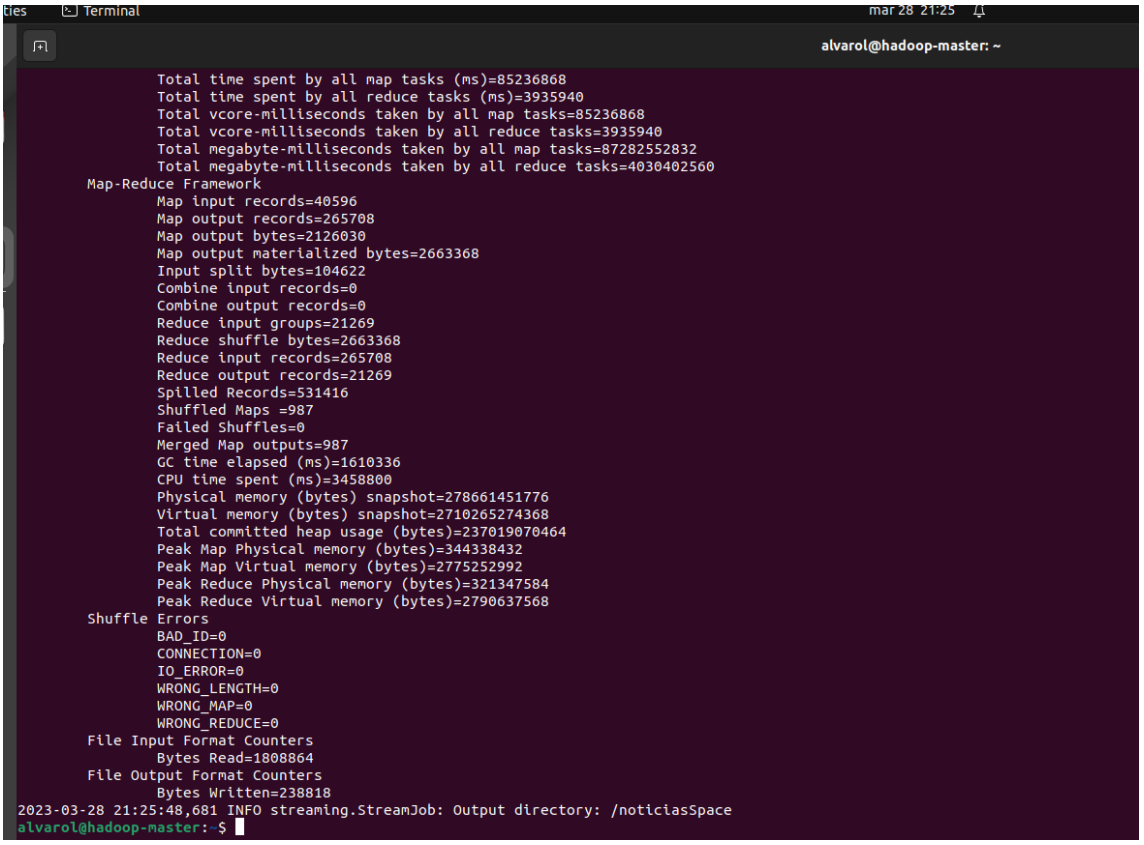
```
/usr/share/hadoop/bin/hadoop jar /usr/share/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -input /20news-18828/rec.autos/ -output /noticias -file /home/alvarol/mapper.py -file /home/alvarol/reducer.py  -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'
```

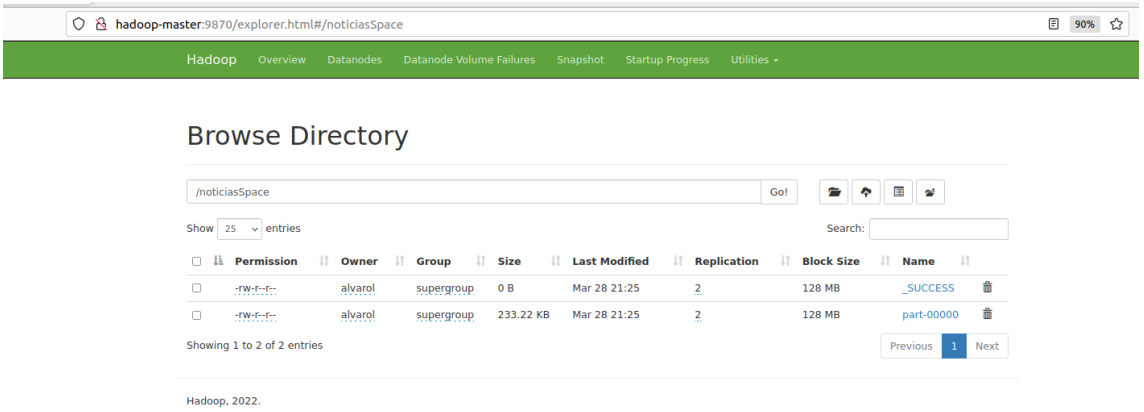


Space

```
/usr/share/hadoop/bin/hadoop jar /usr/share/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -input /20news-18828/sci.space/ -output /noticiasSpace -file /home/alvarol/mapper.py -file /home/alvarol/reducer.py  -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'
```

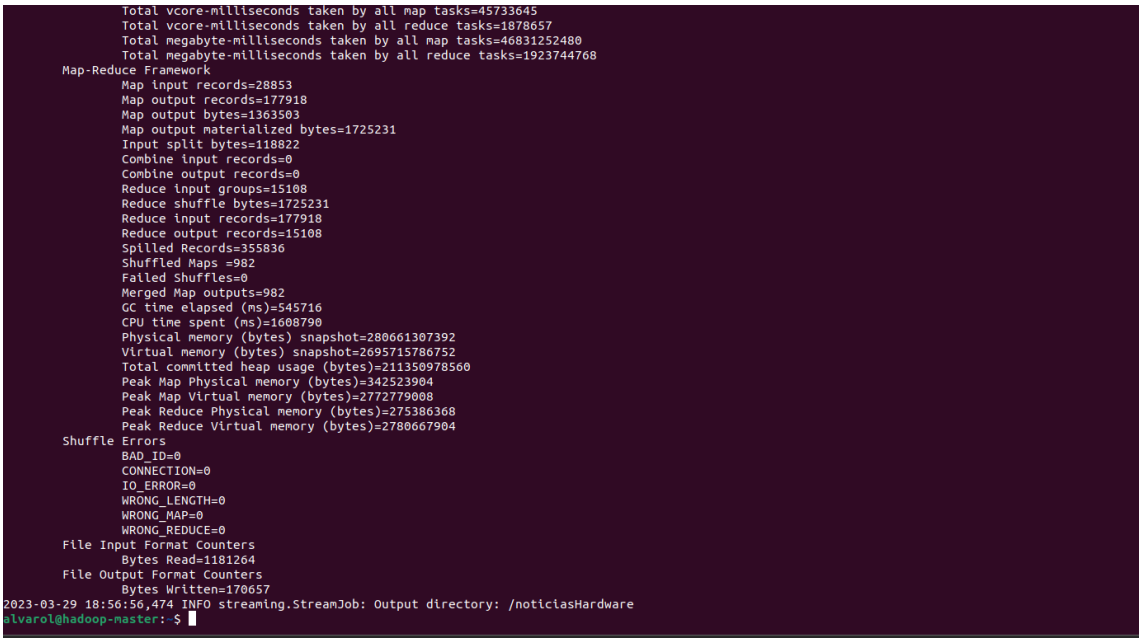



ties
Terminal
mar 28 21:25
alvarol@hadoop-master: ~
Total time spent by all map tasks (ms)=85236868
Total time spent by all reduce tasks (ms)=3935940
Total vcore-milliseconds taken by all map tasks=85236868
Total vcore-milliseconds taken by all reduce tasks=3935940
Total megabyte-milliseconds taken by all map tasks=87282552832
Total megabyte-milliseconds taken by all reduce tasks=4030402560
Map-Reduce Framework
Map input records=40596
Map output records=265708
Map output bytes=2126030
Map output materialized bytes=2663368
Input split bytes=104622
Combine input records=0
Combine output records=0
Reduce input groups=21269
Reduce shuffle bytes=2663368
Reduce input records=265708
Reduce output records=21269
Spilled Records=531416
Shuffled Maps =987
Failed Shuffles=0
Merged Map outputs=987
GC time elapsed (ms)=1610336
CPU time spent (ms)=3458800
Physical memory (bytes) snapshot=278661451776
Virtual memory (bytes) snapshot=2710265274368
Total committed heap usage (bytes)=237019070464
Peak Map Physical memory (bytes)=344338432
Peak Map Virtual memory (bytes)=2775252992
Peak Reduce Physical memory (bytes)=321347584
Peak Reduce Virtual memory (bytes)=2790637568
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=1808864
File Output Format Counters
Bytes Written=238818
2023-03-28 21:25:48,681 INFO streaming.StreamJob: Output directory: /noticiasSpace
alvarol@hadoop-master:~$


hadoop-master:9870/explorer.html#/noticiasSpace
90%
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/noticiasSpace
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
alvarol
supergroup
0 B
Mar 28 21:25
2
128 MB
_SUCCESS
-rw-r--r--
alvarol
supergroup
233.22 KB
Mar 28 21:25
2
128 MB
part-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2022.

comp.sys.ibm.pc.hardware

```
/usr/share/hadoop/bin/hadoop jar /usr/share/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -input /20news-18828/comp.sys.ibm.pc.hardware -output /noticiasHardware -file /home/alvarol/mapper.py -file /home/alvarol/reducer.py  -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'
```



Resultado final

- Muestra un ranking con las 10 palabras más empleadas en los grupos 'rec.autos','comp.sys.ibm.pc.hardware' y 'sci.space'.

```
cat rec.autos | sort -t$'\t' -k2 -r -n | head -n 10

cat comp.sys.ibm.pc.hardware | sort -t$'\t' -k2 -r -n | head -n 10

cat sci.space | sort -t$'\t' -k2 -r -n | head -n 10
```

rec-autos

```
alvarol@hadoop-master:~$ /usr/share/hadoop/bin/hdfs dfs -cat /noticias/part-00000 | sort -t$'\t' -k2 -r -n | head -n 10
the     9986
a       5023
to      4097
i       3916
and     3654
in      3409
of      3357
is      2668
it      2358
that    2236
alvarol@hadoop-master:~$
```

sci.space

```
alvarol@hadoop-master:~$ /usr/share/hadoop/bin/hdfs dfs -cat /noticiasSpace/part-00000 | sort -t$'\t' -k2 -r -n | head -n 10
the     13861
of      6427
to      6217
a       5619
and     5162
in      4396
is      3414
that    2795
for     2545
i       2362
alvarol@hadoop-master:~$
```

comp.sys.ibm.pc.hardware

```
alvarol@hadoop-master:~$ /usr/share/hadoop/bin/hdfs dfs -cat /noticiasHardware/part-00000 | sort -t$'\t' -k2 -r -n | head -n 10
the     7921
a       4091
to      3876
i       3799
and     3150
is      2604
of      2338
it      2077
in      1943
for     1838
alvarol@hadoop-master:~$
```