

PRÁCTICA Nº 2.3

Hadoop Streaming



- Nombre y apellidos: Alvaro Lucio-Villegas de Cea



Índice

Proceso Inicial.	3
Actividad:	6
Resultado final	10
rec-autos	10
sci.space	10
comp.sys.ibm.pc.hardware	10

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).

```
alvarol@hadoop-master: ~
alvarol@hadoop-master:~$ /usr/share/hadoop/sbin/start-dfs.sh
Starting namenodes on [hadoop-master]
Starting datanodes
Starting secondary namenodes [hadoop-master]
alvarol@hadoop-master:~$ /usr/share/hadoop/sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
alvarol@hadoop-master:~$
```

hadoop-master:9870/dfshealth.html#tab-overview

Overview 'hadoop-master:9000' (✓active)

Started:	Tue Mar 21 16:58:47 +0100 2023
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 14:32:00 +0200 2022 by stevel from branch-3.3.4
Cluster ID:	CID-48bbeaa9-e2b9-4994-af52-0b925ec17e0c
Block Pool ID:	BP-1937447951-192.168.13.30-1678903755661

Summary

Security is off.

Safemode is off.

21 files and directories, 8 blocks (8 replicated blocks, 0 erasure coded block groups) = 29 total filesystem object(s).

Heap Memory used 41.87 MB of 310 MB Heap Memory. Max Heap Memory is 974 MB.

Non Heap Memory used 52.86 MB of 55.69 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	57.07 GB
Configured Remote Capacity:	0 B
DFS Used:	6.27 MB (0.01%)
Non DFS Used:	43.63 GB
DFS Remaining:	10.46 GB (18.33%)
Block Pool Used:	6.27 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.01% / 0.02% / 0.00%
Live Nodes	3 (Decommissioned: 0, In Maintenance: 0)

Firefox Web Browser | mar 21 17:53

Browsing HDFS | Nodes of the cluster

localhost:8088/cluster/nodes

Nodes of the cluster

Cluster

- About
- Nodes
- Node Labels
- Applications
- NEW
- NEW SAVING
- SUBMITTED
- ACCEPTED
- RUNNING
- FINISHED
- FAILED
- KILLED
- Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	...
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:0 B, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
3	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mb), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:2>

Show 20 entries

- 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'
```

```
alvarol@hadoop-master:~$ sudo find / -name 'hadoop-streaming*.jar'
[sudo] password for alvarol:
/usr/share/hadoop/share/hadoop/tools/sources/hadoop-streaming-3.3.4-sources.jar
/usr/share/hadoop/share/hadoop/tools/sources/hadoop-streaming-3.3.4-test-sources.jar
/usr/share/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar
find: '/run/user/1000/doc': Permission denied
find: '/run/user/1000/gvfs': Permission denied
alvarol@hadoop-master:~$
```

- 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
```

```
alvarol@hadoop-master:~$ tar -xzf 20news-18828.tar.gz
alvarol@hadoop-master:~$ ls
20news-18828  data  Documents  ElQuijote.txt  mapper.py  name  Pictures  reducer.py  Templates  Videos
20news-18828.tar.gz  Desktop  Downloads  libro_salida  Music  namesecondary  Public  snap  tnp  working_dataset.xlsx
alvarol@hadoop-master:~$
```



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'
```

```
py -file /home/alvarol/reducer.py -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'
```

Browse Directory

/noticias Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	alvarol	supergroup	0 B	Mar 28 18:38	2	128 MB	_SUCCESS
-rw-r--r--	alvarol	supergroup	172.72 KB	Mar 28 18:38	2	128 MB	part-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2022.

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'
```

```
alvarol@hadoop-master:~$ /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'
2023-03-28 19:25:16,179 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/alvarol/mapper.py, /home/alvarol/reducer.py, /tmp/hadoop-unjar199907059264164/] [] /tmp/streamjob13651170218671076770.jar tmpDir=null
2023-03-28 19:25:25,908 INFO client.DefaultHMAHAFAILOVERProxyProvider: Connecting to ResourceManager at hadoop-master/192.168.13.30:8032
2023-03-28 19:25:29,702 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/alvarol/.staging/job_1680024319360_0001
2023-03-28 19:25:38,416 INFO mapred.FileInputFormat: Total input files to process : 987
2023-03-28 19:25:40,113 INFO mapreduce.JobSubmitter: number of splits:987
2023-03-28 19:25:43,040 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680024319360_0001
2023-03-28 19:25:43,040 INFO mapreduce.JobSubmitter: Executing with tokens: {}
2023-03-28 19:25:45,167 INFO conf.Configuration: resource-types.xml not found
2023-03-28 19:25:45,175 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-03-28 19:25:47,560 INFO Impl.VarsClientImpl: Submitted application application_1680024319360_0001
2023-03-28 19:25:48,030 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1680024319360_0001/
2023-03-28 19:25:48,039 INFO mapreduce.Job: Running job: job_1680024319360_0001
2023-03-28 19:26:22,315 INFO mapreduce.Job: Job job_1680024319360_0001 running in uber mode : false
2023-03-28 19:26:22,322 INFO mapreduce.Job: map 0% reduce 0%
2023-03-28 19:28:08,595 INFO mapreduce.Job: map 1% reduce 0%
2023-03-28 19:30:31,552 INFO mapreduce.Job: map 2% reduce 0%
2023-03-28 19:30:51,283 INFO mapreduce.Job: map 3% reduce 0%
2023-03-28 19:31:12,268 INFO mapreduce.Job: map 4% reduce 0%
2023-03-28 19:32:50,710 INFO mapreduce.Job: map 5% reduce 0%
2023-03-28 19:34:44,181 INFO mapreduce.Job: map 6% reduce 0%
2023-03-28 19:35:04,446 INFO mapreduce.Job: map 7% reduce 0%
2023-03-28 19:35:19,050 INFO mapreduce.Job: map 7% reduce 2%
2023-03-28 19:36:24,486 INFO mapreduce.Job: map 8% reduce 2%
2023-03-28 19:36:31,419 INFO mapreduce.Job: map 8% reduce 3%
2023-03-28 19:36:49,117 INFO mapreduce.Job: map 9% reduce 3%
2023-03-28 19:38:29,663 INFO mapreduce.Job: map 10% reduce 3%
2023-03-28 19:38:54,249 INFO mapreduce.Job: map 11% reduce 3%
2023-03-28 19:39:15,962 INFO mapreduce.Job: map 11% reduce 4%
```

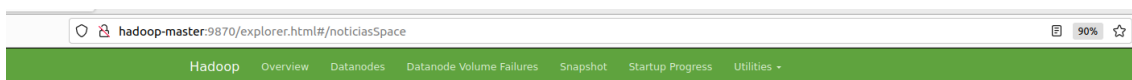
```

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:~$
    
```



Browse Directory

/noticiasSpace

Show entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	alvarol	supergroup	0 B	Mar 28 21:25	2	128 MB	_SUCCESS
<input type="checkbox"/>	-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'
```

```
alvarol@hadoop-master: $ /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'
```

```
Total vcore-milliseconds taken by all map tasks=45733645
Total vcore-milliseconds taken by all reduce tasks=1878657
Total megabyte-milliseconds taken by all map tasks=46831252480
Total megabyte-milliseconds taken by all reduce tasks=1923744768

Map-Reduce Framework
  Map input records=28853
  Map output records=177918
  Map output bytes=1363503
  Map output materialized bytes=1725231
  Input split bytes=118822
  Combine input records=0
  Combine output records=0
  Reduce input groups=15108
  Reduce shuffle bytes=1725231
  Reduce input records=177918
  Reduce output records=15108
  Spilled Records=355836
  Shuffled Maps =982
  Failed Shuffles=0
  Merged Map outputs=982
  GC time elapsed (ms)=545716
  CPU time spent (ms)=1608790
  Physical memory (bytes) snapshot=280661307392
  Virtual memory (bytes) snapshot=2695715786752
  Total committed heap usage (bytes)=211358978560
  Peak Map Physical memory (bytes)=342523904
  Peak Map Virtual memory (bytes)=2772779008
  Peak Reduce Physical memory (bytes)=275386368
  Peak Reduce Virtual memory (bytes)=2780667904

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=1181264
File Output Format Counters
  Bytes Written=170657

2023-03-29 18:56:56,474 INFO streaming.StreamJob: Output directory: /noticiasHardware
alvarol@hadoop-master: $
```

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$'\t' -k2 -r -n | head -n 10
the      7921
a        4091
to       3876
i        3799
and      3150
is       2604
of       2338
it       2077
in       1943
for     1838
alvarol@hadoop-master:~$
```