

Universidad de San Carlos de Guatemala  
Facultad de Ingeniería  
Departamento de Ciencias  
Segundo Semestre del 2022  
Seminario de Sistemas 2  
Profesor: Ing. Lui Alberto Vettorazzi España  
Auxiliares: Sergio Lennin Gonzalez Solis  
Edi Yovani Tomas Reynoso 201503783



**USAC**  
**TRICENTENARIA**  
Universidad de San Carlos de Guatemala

Practica 2

# Pasos para hadoop

```
//---- we are going to run the container
sudo docker run --rm -it -v Practica2:/source -p 50070-50080:50070-50080
sequenceiq/hadoop-docker /etc/bootstrap.sh -bash
//---- checkout the file
ls
// we are going to create a folder
mkdir Practica2
// checkout the file
ls
// let's copy the command in other console.
sudo docker cp "/home/tomas/Documentos/Curso_Semi2/Practica2/Correos.txt"
epic_nobel:/Practica2
sudo docker cp "/home/tomas/Documentos/Curso_Semi2/Practica2/Puntuacion.txt"
epic_nobel:/Practica2
sudo docker cp "/home/tomas/Documentos/Curso_Semi2/Practica2/WordCount.java"
epic_nobel:/Practica2
// we go back to previous console.
// let's on folder Practica 2
cd Practica2
// we check the folder
ls
// we go back to previous folder
cd ../

// Command to initialize the HADOOP_HOME variable
export HADOOP_HOME=/usr/local/hadoop
// we check the folder
ls ${HADOOP_HOME}
// Command to initialize the CLASSPATH variable
export
CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core
-2.7.0.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-2
.7.0.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-2.7.0.jar:/Practica2/*:$H
ADOOP_HOME/lib*"

// let's on folder Practica 2
cd Practica2
// we check the folder
ls
// Command to compile
javac -d . WordCount.java

// we are going to create a file manifest.
cat > manifest.txt
```

Main-class: WordCount

**//Saved with ctrl+d**

**//verify the content of file**

cat manifest.txt

**//we create a file jar**

jar cfm WordCount.jar manifest.txt \*.class

**// Verify**

ls

**// we create a folder with name input**

mkdir ~/input

mkdir ~/output

**// let's copy the file at the folder.**

cp Correos.txt ~/input

cp Puntuacion.txt ~/input

ls ~/input

**// command to copy the files of input in the system of files of hadoop**

\${HADOOP\_HOME}/bin/hdfs dfs -copyFromLocal ~/input /

**// comand to verify the files let us copy at the sistem of hadoop**

\${HADOOP\_HOME}/bin/hdfs dfs -ls /input

**// command to make the count of word**

\${HADOOP\_HOME}/bin/hadoop jar WordCount.jar /input /output

**// command of output**

\${HADOOP\_HOME}/bin/hdfs dfs -ls /output

**// comand to see the file of output**

\${HADOOP\_HOME}/bin/hdfs dfs -cat /output/part-r-00000

**// command to rename the file**

\${HADOOP\_HOME}/bin/hdfs dfs -mv /output/part-r-00000 /output/Resultado.txt

**// Command to see the file of output rename**

\${HADOOP\_HOME}/bin/hdfs dfs -cat /output/Resultado.txt

**//Command to copy the file of output to folder of output from root user home**

\${HADOOP\_HOME}/bin/hdfs dfs -copyToLocal /output/Resultado.txt ~/output

**//Command to move the output file to Practica2 folder from container**

cp ~/output/Resultado.txt /Practica2

**// Command to copy the output file from container to PC (use new console)**

sudo docker cp epic\_nobel:/Practica2/Resultado.txt  
/home/tomas/Documentos/Curso\_Semi2/Practica2

# Capturas del Procedimiento

## Descripción

En las imágenes se muestra todo el proceso que se realizó con los comandos, además se muestra una página web con los detalles de los archivos de salida y los archivos de entrada.

```
practica2@practica2:~/Practica2$ sudo docker run --rm -it -v class:/source -p 50070-50080:50070-50080 sequencetq/hadoop-docker /etc/passwd
bootstrap.sh -bash
[sudo] contraseña para tomas:
/
starting sshd: [ OK ]
starting namenodes on [05d61382700a]
05d61382700a: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-05d61382700a.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-05d61382700a.out
starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-05d61382700a.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-05d61382700a.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-05d61382700a.out
bash-4.1# ls
bin dev home lib64 mnt proc sbin source sys usr
boot etc lib media opt root selinux srv tmp var
bash-4.1# mkdir Practica2
bash-4.1# ls
Practica2 boot etc lib media opt root selinux srv tmp var
bin dev home lib64 mnt proc sbin source sys usr
bash-4.1# cd Practica2
bash-4.1# ls
Practica2 boot etc lib media opt root selinux srv tmp var
bash-4.1# cd ../
bash-4.1# ls
Practica2 bin boot dev etc home lib lib64 media mnt opt proc root sbin selinux source srv sys tmp usr var
bash-4.1# cd Practica2
bash-4.1# ls
Correos.txt Puntuacion.txt WordCount.java
bash-4.1# cd ../
bash-4.1# export HADOOP_HOME=/usr/local/hadoop
bash-4.1# ls
Practica2 bin boot dev etc home lib lib64 media mnt opt proc root sbin selinux source srv sys tmp usr var
bash-4.1# ls ${HADOOP_HOME}
LICENSE.txt NOTICE.txt README.txt bin etc include input lib libexec logs sbin share
bash-4.1# export CLASSPATH="${HADOOP_HOME}/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.7.0.jar:${HADOOP_HOME}/share/hadoop/mapreduce/had
hadoop-mapreduce-client-common-2.7.0.jar:${HADOOP_HOME}/share/hadoop/common/hadoop-common-2.7.0.jar:/Practica2/*:${HADOOP_HOME}/lib*"
bash-4.1# cd Practica2
bash-4.1# ls
Correos.txt Puntuacion.txt WordCount.java
bash-4.1# javac -d . WordCount.java
/usr/local/hadoop/share/hadoop/common/hadoop-common-2.7.0.jar(org/apache/hadoop/fs/Path.class): warning: Cannot find annotation method 'value()' in type 'LimitedPrivate': class file for org.apache.hadoop.classification.InterfaceAudience not found
warning
bash-4.1# cat > manifest.txt
Main-class: WordCount
bash-4.1# cat manifest.txt
Main-class: WordCount
bash-4.1# jar cfm WordCount.jar manifest.txt *.class
bash-4.1# ls
Correos.txt WordCount$IntSumReducer.class WordCount.class WordCount.java
Puntuacion.txt WordCount$TokenizerMapper.class WordCount.jar manifest.txt
bash-4.1# mkdir ~/input
bash-4.1# cp Correo.txt ~/input
cp: cannot stat 'Correo.txt': No such file or directory
bash-4.1# cp Correos.txt ~/input
bash-4.1# cp Puntuacion.txt ~/input
bash-4.1# ls ~/input
Correos.txt Puntuacion.txt
bash-4.1# ${HADOOP_HOME}/bin/hdfs dfs -copyFromLocal ~/input /
bash-4.1# ${HADOOP_HOME}/bin/hdfs dfs -ls /input
found 2 items
-rw-r--r-- 1 root supergroup 31354 2022-10-04 20:46 /input/Correos.txt
-rw-r--r-- 1 root supergroup 18429 2022-10-04 20:46 /input/Puntuacion.txt
bash-4.1# ${HADOOP_HOME}/bin/hadoop jar WordCount.jar /input /output
22/10/04 21:56:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/10/04 21:56:05 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/10/04 21:56:05 INFO input.FileInputFormat: Total input paths to process : 2
22/10/04 21:56:05 INFO mapreduce.JobSubmitter: number of splits:2
22/10/04 21:56:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664889465966_0001
22/10/04 21:56:07 INFO impl.YarnClientImpl: Submitted application application_1664889465966_0001
```

```

bash-4.1# $(HADOOP_HOME)/bin/hadoop jar WordCount.jar /input /output
22/10/04 21:56:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/10/04 21:56:05 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/10/04 21:56:05 INFO input.FileInputFormat: Total input paths to process : 2
22/10/04 21:56:05 INFO mapreduce.JobSubmitter: number of splits:2
22/10/04 21:56:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664889465966_0001
22/10/04 21:56:07 INFO impl.YarnClientImpl: Submitted application application_1664889465966_0001
22/10/04 21:56:07 INFO mapreduce.Job: The url to track the job: http://05d61382700a:8088/proxy/application_1664889465966_0001/
22/10/04 21:56:07 INFO mapreduce.Job: Running job: job_1664889465966_0001
22/10/04 21:56:19 INFO mapreduce.Job: Job job_1664889465966_0001 running in uber mode : false
22/10/04 21:56:19 INFO mapreduce.Job:  map 0% reduce 0%
22/10/04 21:56:32 INFO mapreduce.Job:  map 100% reduce 0%
22/10/04 21:56:40 INFO mapreduce.Job:  map 100% reduce 100%
22/10/04 21:56:41 INFO mapreduce.Job: Job job_1664889465966_0001 completed successfully
22/10/04 21:56:41 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=26299
        FILE: Number of bytes written=397267
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=50000
        HDFS: Number of bytes written=18637
        HDFS: Number of read operations=9
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=19445
        Total time spent by all reduces in occupied slots (ms)=5130
        Total time spent by all map tasks (ms)=19445
        Total time spent by all reduce tasks (ms)=5130
        Total vcore-seconds taken by all map tasks=19445

```

```

    File System Counters
        FILE: Number of bytes read=26299
        FILE: Number of bytes written=397267
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=50000
        HDFS: Number of bytes written=18637
        HDFS: Number of read operations=9
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=19445
        Total time spent by all reduces in occupied slots (ms)=5130
        Total time spent by all map tasks (ms)=19445
        Total time spent by all reduce tasks (ms)=5130
        Total vcore-seconds taken by all map tasks=19445
        Total vcore-seconds taken by all reduce tasks=5130
        Total megabyte-seconds taken by all map tasks=19911680
        Total megabyte-seconds taken by all reduce tasks=5253120
    Map-Reduce Framework
        Map input records=51
        Map output records=13709
        Map output bytes=104523
        Map output materialized bytes=26305
        Input split bytes=217
        Combine input records=13709
        Combine output records=1928
        Reduce input groups=1923
        Reduce shuffle bytes=26305
        Reduce input records=1928
        Reduce output records=1923
        Spilled Records=3856

```

```
Total vcore-seconds taken by all reduce tasks=5130
Total megabyte-seconds taken by all map tasks=19911680
Total megabyte-seconds taken by all reduce tasks=5253120
Map-Reduce Framework
  Map input records=51
  Map output records=13709
  Map output bytes=104523
  Map output materialized bytes=26305
  Input split bytes=217
  Combine input records=13709
  Combine output records=1928
  Reduce input groups=1923
  Reduce shuffle bytes=26305
  Reduce input records=1928
  Reduce output records=1923
  Spilled Records=3856
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=237
  CPU time spent (ms)=4490
  Physical memory (bytes) snapshot=663838720
  Virtual memory (bytes) snapshot=2216423424
  Total committed heap usage (bytes)=559939584
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=49783
File Output Format Counters
  Bytes Written=18637
bash-4.1#
```

Actividades Terminal 4 de oct 20:06

tomas@tomas-HP-ENVY-15-Notebook-PC: ~

tomas@tomas-HP-ENVY-15-Notebook-PC: ~

tomas@tomas-HP-ENVY-15-Notebook-PC: ~

```
File System Counters
  FILE: Number of bytes read=26299
  FILE: Number of bytes written=397267
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=50000
  HDFS: Number of bytes written=18637
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=19445
  Total time spent by all reduces in occupied slots (ms)=5130
  Total time spent by all map tasks (ms)=19445
  Total time spent by all reduce tasks (ms)=5130
  Total vcore-seconds taken by all map tasks=19445
  Total vcore-seconds taken by all reduce tasks=5130
  Total megabyte-seconds taken by all map tasks=19911680
  Total megabyte-seconds taken by all reduce tasks=5253120
Map-Reduce Framework
  Map input records=51
  Map output records=13709
  Map output bytes=104523
  Map output materialized bytes=26305
  Input split bytes=217
  Combine input records=13709
  Combine output records=1928
  Reduce input groups=1923
  Reduce shuffle bytes=26305
  Reduce input records=1928
  Reduce output records=1923
  Spilled Records=3856
```

```

Total vcore-seconds taken by all reduce tasks=5130
Total megabyte-seconds taken by all map tasks=19911680
Total megabyte-seconds taken by all reduce tasks=5253120
Map-Reduce Framework
  Map input records=51
  Map output records=13709
  Map output bytes=104523
  Map output materialized bytes=26305
  Input split bytes=217
  Combine input records=13709
  Combine output records=1928
  Reduce input groups=1923
  Reduce shuffle bytes=26305
  Reduce input records=1928
  Reduce output records=1923
  Spilled Records=3856
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=237
  CPU time spent (ms)=4490
  Physical memory (bytes) snapshot=663838720
  Virtual memory (bytes) snapshot=2216423424
  Total committed heap usage (bytes)=559939584
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=49783
File Output Format Counters
  Bytes Written=18637

```

bash-4.1#

```

-rw-r--r-- 1 root supergroup 18637 2022-10-04 21:56 /output/part-r-00000

```

```

bash-4.1# ${HADOOP_HOME}/bin/hdfs dfs -cat /output/part-r-00000

```

```

1 1386

```

```

1,000 1

```

```

1/2 1

```

```

10 3

```

```

100- 1

```

```

10am 2

```

```

10am, 1

```

```

10th 1

```

```

11 1

```

```

12 2

```

```

12/1a 1

```

```

125 1

```

```

12th 1

```

```

150 1

```

```

175. 1

```

```

17th 2

```

```

18 1

```

```

18-19 1

```

```

1970 1

```

```

1:30 1

```

```

1st 2

```

```

1st, 1

```

```

2 1084

```

```

2.5 1

```

```

20 3

```

```

200/night 1

```

```

2005. 1

```

```

2007. 1

```

```

2007my 1

```

```

2008. 1

```

```

20th 2

```

```

21 1

```

```

21, 1

```

```

21/day 1

```

2	1084	
2.5	1	
20	3	
200/night		1
2005.	1	
2007.	1	
2007my	1	
2008.	1	
20th	2	
21	1	
21,	1	
21/day	1	
24	1	
25	1	
25.	1	
250+/night,		1
28	1	
29/night,		1
2nd	2	
2x,	1	
3	1255	
30	3	
30.the	1	
300+	1	
38	1	
38.	1	
3pm	1	
3rd	1	
4	2558	
4*	2	
4/23-5/1,		1
40	1	
400	1	
45	1	
4pm	1	
4th	4	
5	2075	



tomas@tomas-HP-ENVY-15-Notebook-PC: ~

tomas@tomas-HP-ENVY-15-Noteboo...

tomas@tomas-HP-ENVY-15-Noteboo...

wont 1  
wood 1  
word 1  
work 1  
work, 3  
worked 6  
worked, 1  
working 1  
worse, 1  
worst 3  
worth 2  
write 1  
wrong 1  
x 1  
year 2  
years 4  
yep 1  
yes 2  
yikes, 1  
yoga 1  
young 3  
zone 1  
bash-4.1#

# Browse Directory

/

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	4/10/2022, 18:46:32	0	0 B	<a href="#">input</a>
drwxr-xr-x	root	supergroup	0 B	4/10/2022, 20:29:51	0	0 B	<a href="#">output</a>
drwx-----	root	supergroup	0 B	4/10/2022, 19:56:04	0	0 B	<a href="#">tmp</a>
drwxr-xr-x	root	supergroup	0 B	22/7/2015, 09:17:26	0	0 B	<a href="#">user</a>

# Browse Directory

/output								Go!
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
-rw-r--r--	root	supergroup	18.2 KB	4/10/2022, 19:56:38	1	128 MB	<a href="#">Resultado.txt</a>	
-rw-r--r--	root	supergroup	0 B	4/10/2022, 19:56:39	1	128 MB	<a href="#">_SUCCESS</a>	

Hadoop, 2014.

Actividades

Navegador web Firefox ▾

4 de oct. 20:55

WhatsApp hadoo Facult New C (1612) REPRCE 0798A [SS2]Prác Comand Browse X seque (1612) Docum Tradu + - ⌵ ⌵ ⌵

← → ↻ localhost:50070/explorer.html#/ ☆ ⌵ ⌵ ⌵

Hadoop Overview Datanodes Snapshot Startup Progress Utilities ▾

Browse Directory

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	4/10/2022, 18:46:32	0	0 B	<a href="#">input</a>
drwxr-xr-x	root	supergroup	0 B	4/10/2022, 20:29:51	0	0 B	<a href="#">output</a>
drwx-----	root	supergroup	0 B	4/10/2022, 19:56:04	0	0 B	<a href="#">tmp</a>
drwxr-xr-x	root	supergroup	0 B	22/7/2015, 09:17:26	0	0 B	<a href="#">user</a>

Hadoop, 2014.

## File information - Resultado.txt



[Download](#)

Block information -- Block 0 ▾

Block ID: 1073741864

Block Pool ID: BP-581371184-172.17.13.14-1437578119536

Generation Stamp: 1040

Size: 18637

Availability:

- eac36ed5051d

Close

WhatsAppHorarios d[SS2]PrácticaComandos.pd[SS2]Pract(1615) SIA REPRODUCCI...Browsing H XTraductorFacultad dRecibidos+--x

←→↻

localhost:50070/explorer.html#/input

☆📄⬇️☰

HadoopOverviewDatanodesSnapshotStartup ProgressUtilities ▾

# Browse Directory

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	30.62 KB	5/10/2022, 13:12:18	1	128 MB	<a href="#">Correos.txt</a>
-rw-r--r--	root	supergroup	18 KB	5/10/2022, 13:12:19	1	128 MB	<a href="#">Puntuacion.txt</a>

Hadoop, 2014.

## **Conclusiones acerca de los resultados de cada archivo**

Muestra el resultado del conteo de cada palabra donde indica cual es son las palabras que se repiten más y muestra el total de palabra en cada archivo que hay.

## **Conclusiones acerca del uso de Hadoop en BigData.**

para simples solicitudes de información y problemas que se pueden dividir en unidades independientes, pero no es eficiente para realizar tareas analíticas iterativas e interactivas. MapReduce trabaja con muchos archivos. Como los nodos no se intercomunican salvo a través de procesos de clasificación y mezcla, los algoritmos iterativos requieren múltiples fases de mapeo-mezcla/clasificación-reducción para completarse. Esto da origen a múltiples archivos entre fases de MapReduce y no es eficiente para el cómputo analítico avanzado.