

**Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ciencias y Sistemas
Seminario de Sistemas 2**



Practica 2

**Christopher Alexander Acajabon Gudiel
201404278**

PASOS PARA EL PROCESAMIENTO DE DATOS EN HADOOP

1. Comando para bajar la imagen de Hadoop
 - `docker pull sequenceiq/hadoop-docker`
2. Abrir una consola, la cual llamaremos **CMD1**.
3. Comando para correr imagen (**CMD1**)
 - `docker run --rm -it -v mihadoop:/source -p 50070-50080:50070-50080 sequenceiq/hadoop-docker /etc/bootstrap.sh -bash`

```
C:\Users\Christopher>docker run --rm -it -v mihadoop:/source -p 50070-50080:50070-50080 sequenceiq/hadoop-docker /etc/bootstrap.sh -bash
/
Starting sshd: [ OK ]
Starting namenodes on [30b4fae64085]
30b4fae64085: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-30b4fae64085.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-30b4fae64085.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-30b4fae64085.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-30b4fae64085.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-30b4fae64085.out
bash-4.1#
```

4. Si no crea el contenedor o aparece un error 139:
 - Crear un archivo `.wslconfig` en su usuario:
 - `C://users/[your user name]/.wslconfig`
 - Agregar el siguiente contenido al archivo:

```
[wsl2]
kernelCommandLine = vsyscall=emulate
```

- Reiniciar la PC.
 - Intentar correr la imagen nuevamente.
5. Verificar en el navegador: <http://localhost:50070/>
 6. Creamos carpeta (**CMD1**)
 - `mkdir practica`
 7. Abrimos otra consola, la cual llamaremos **CMD2**.
 8. Copiamos los archivos al contenedor (**CMD2**).
 - `docker cp "C:\Users\Christopher\Desktop\p2_ss2\WordCount.java" nifty_austin:/practica`
 - `docker cp "C:\Users\Christopher\Desktop\p2_ss2\Correos.txt" mystifying_jackson:/practica`
 - `docker cp "C:\Users\Christopher\Desktop\p2_ss2\Puntuacion.txt" nifty_austin:/practica`

```
bash-4.1# cd practica
bash-4.1# ls
Correos.txt  Puntuacion.txt  WordCount.java
bash-4.1#
```

- mystifying_jackson y nifty_austin es el nombre del contenedor, verificar como se llama el suyo.
- Practica es el nombre de la carpeta creada en el contenedor.
- Ejemplo de cómo ver el nombre del contenedor:

```

Seleccionar Símbolo del sistema
Microsoft Windows [Versión 10.0.19044.2728]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Users\Christopher>docker ps

CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS        PORTS
8102926df483   sequenceiq/hadoop-docker           "/etc/bootstrap.sh -..." 15 minutes ago Up 15 minutes 2122/tcp, 8030-8033/
tcp, 8040/tcp, 8042/tcp, 8088/tcp, 19888/tcp, 49707/tcp, 50010/tcp, 50020/tcp, 50090/tcp, 0.0.0.0:50070-50080->50070-500
80/tcp        beautiful_carver

```

9. Inicializamos variable HADOOP_HOME (**CMD1**)
 - `export HADOOP_HOME=/usr/local/hadoop`
10. Para verificar que se inicializo bien.
 - `ls ${HADOOP_HOME}`
11. Inicializamos variable CLASSPATH (**CMD1**)
 - Practica es el nombre de la carpeta creada en el contenedor.
 - `export CLASSPATH="$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:/practica/*:$HADOOP_HOME/lib/*"`
 - `export CLASSPATH="$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:/[Nombre Carpeta]/*:$HADOOP_HOME/lib/*"`
12. Compilamos el .java (**CMD1**)
 - `cd practica`
 - `javac -d . WordCount.java`
 - Tiene que salir el siguiente mensaje.

```

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
1 warning
bash-4.1#

```

- Nos creara otras clases .java

```
bash-4.1# ls
Correos.txt      WordCount$IntSumReducer.class  WordCount.class
Puntuacion.txt  WordCount$TokenizerMapper.class WordCount.java
bash-4.1#
```

13. Creamos archivo manifest (**CMD1**)

- `cd practica`
- `cat > Manifest.txt`
- Guardamos lo siguiente dando click derecho y pegar: [Main-class: WordCount](#)
- `Ctrl + D` para guardar (dos veces si es necesario).
- `cat Manifest.txt` para ver el contenido y verificar que se guardó.

14. Creamos el .jar (**CMD1**)

- `cd practica`
- `jar cfm WordCount.jar Manifest.txt *.class`

15. Creamos carpeta de entrada y salida (**CMD1**)

- `cd practica`
- `mkdir ~/input`
- `mkdir ~/output`

16. copiamos archivos de entrada a carpeta de entrada (**CMD1**)

- `cd practica`
- `cp Correos.txt ~/input`
- `cp Puntuacion.txt ~/input`

17. copiamos los archivos de entrada en el sistema de archivos de Hadoop (**CMD1**)

- `cd practica`
- `${HADOOP_HOME}/bin/hdfs dfs -copyFromLocal ~/input /`

18. Verificamos que se hallan copiado los archivos (**CMD1**)

- `${HADOOP_HOME}/bin/hdfs dfs -ls /input`

19. Realizar el conteo de palabras (**CMD1**)

- `${HADOOP_HOME}/bin/hadoop jar WordCount.jar /input /output`
- Pueda que nos salga el siguiente error

```
bash-4.1# ${HADOOP_HOME}/bin/hadoop jar WordCount.jar /input /output
Exception in thread "main" java.lang.ClassNotFoundException: /input
    at java.lang.Class.forName0(Native Method)
    at java.lang.Class.forName(Class.java:274)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:214)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
bash-4.1#
```

- Es por el Manifest, probamos con el siguiente comando indicándole el nombre del manifest.
- `${HADOOP_HOME}/bin/hadoop jar WordCount.jar WordCount /input /output`

20. Comando para ver los archivos de salida (**CMD1**)
 - `${HADOOP_HOME}/bin/hdfs dfs -ls /output`
21. Comando para ver el archivo de salida (**CMD1**)
 - `${HADOOP_HOME}/bin/hdfs dfs -cat /output/part-r-00000`
22. Comando para cambiar el nombre del archivo de conteo (**CMD1**)
 - `${HADOOP_HOME}/bin/hdfs dfs -mv /output/part-r-00000 /output/resultado.txt`
23. Comando para copiar el archivo de salida a carpeta de salida del home del usuario root (**CMD1**)
 - `${HADOOP_HOME}/bin/hdfs dfs -copyToLocal /output/resultado.txt ~/output`
24. Comando para mover el archivo de salida a la carpeta **practica** del contenedor (**CMD1**)
 - `cp ~/output/resultado.txt /practica`
25. Comando para copiar el archivo de salida del contenedor a la PC (**CMD2**)
 - `docker cp name_container:practica/resultado.txt "C:\Users\Christopher\Desktop\p2_ss2"`

CAPTURAS DEL BROWSE HDFS

1. Página principal:

The screenshot shows the Hadoop DFS Health Overview page. The browser address bar indicates the URL is `localhost:50070/dfshealth.html#tab-overview`. The page has a green header with navigation tabs: Hadoop, Overview (selected), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities.

The main content area is titled "Overview '8102926df483:9000' (active)". It contains a table with the following information:

Started:	Wed Mar 29 01:20:09 EDT 2023
Version:	2.7.0, rd4c8d4d4203c934e8074b31289a20724c0842cf
Compiled:	2015-04-10T18:40Z by jenkins from (detached from d4c8d4d)
Cluster ID:	CID-#571ad4-5145-47ad-ba01-959bc79aeebe
Block Pool ID:	BP-581371184-172.17.13.14-1437578119536

Below the table is a "Summary" section. It includes the following text:

Security is off.
 Safemode is off.
 35 files and directories, 31 blocks = 66 total filesystem object(s).
 Heap Memory used 52.62 MB of 148.5 MB Heap Memory. Max Heap Memory is 689 MB.
 Non Heap Memory used 29.63 MB of 30.94 MB Committed Non Heap Memory. Max Non Heap Memory is 130 MB.

Below this text is a table with the following data:

Configured Capacity:	256.98 GB
DFS Used:	324 KB (0%)
Non DFS Used:	16.31 GB
DFS Remaining:	234.67 GB (93.5%)
Block Pool Used:	324 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)

2. Archivo correos:

The screenshot shows a web browser window with the address bar at `localhost:50070/explorer.html#/`. The Hadoop web interface has a green header with navigation links: Hadoop, Overview, Datanodes, Snapshot, Startup Progress, and Utilities. The main content area is titled "Browse Directory" and shows a table of files in the root directory. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The files listed are `input`, `output`, `tmp`, and `user`, all owned by `root` and `supergroup` with a size of 0 B. A search bar at the top contains the character `/`. Below the table, it says "Hadoop, 2014." A Windows taskbar is visible at the bottom with a terminal window open showing the command prompt.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	29/3/2023, 16:46:12	0	0 B	input
drwxr-xr-x	root	supergroup	0 B	29/3/2023, 16:49:30	0	0 B	output
drwx-----	root	supergroup	0 B	29/3/2023, 16:47:34	0	0 B	tmp
drwxr-xr-x	root	supergroup	0 B	22/7/2015, 9:17:26	0	0 B	user

The screenshot shows the same Hadoop web interface, but the address bar now points to `localhost:50070/explorer.html#/input`. The "Browse Directory" section shows a table with one file: `Correos.txt`, which is 30.62 KB in size, owned by `root` and `supergroup`, and has a replication factor of 1. The search bar at the top contains `/input`. The Windows taskbar at the bottom shows the terminal window with the same command prompt.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	30.62 KB	29/3/2023, 16:46:12	1	128 MB	Correos.txt

Browsing HDFS

localhost:50070/explorer.html#/output

Aplicaciones INVESTIGACIÓN DE... EJERCICIOS DE MO... Learn SQL | SoloLea... Log in to Overleaf ~... Quiz 2 Chapters 5 a... SQL - mostrar porc... MySQL, Hosting Gr...

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

/output

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	29/3/2023, 16:48:11	1	128 MB	_SUCCESS
-rw-r--r--	root	supergroup	18.19 KB	29/3/2023, 16:48:10	1	128 MB	resultado.txt

Hadoop, 2014

Símbolo del sistema

Microsoft Windows [Versión 10.0.19044.2728]
(c) Microsoft Corporation. Todos los derechos reservados.
C:\Users\Christopher>Christopher Acajabon - 201404278

16:55
29/03/2023

3. Archivo de puntuaciones:

Browsing HDFS

localhost:50070/explorer.html#/

Aplicaciones INVESTIGACIÓN DE... EJERCICIOS DE MO... Learn SQL | SoloLea... Log in to Overleaf ~... Quiz 2 Chapters 5 a... SQL - mostrar porc... MySQL, Hosting Gr...

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

/

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	29/3/2023, 17:05:06	0	0 B	input
drwxr-xr-x	root	supergroup	0 B	29/3/2023, 17:07:32	0	0 B	output
drwx-----	root	supergroup	0 B	29/3/2023, 17:06:01	0	0 B	tmp
drwxr-xr-x	root	supergroup	0 B	22/7/2015, 9:17:26	0	0 B	user

Hadoop, 2014

Símbolo del sistema

Microsoft Windows [Versión 10.0.19044.2728]
(c) Microsoft Corporation. Todos los derechos reservados.
C:\Users\Christopher>Christopher Acajabon - 201404278

17:09
29/03/2023

Browsing HDFS

localhost:50070/explorer.html#/input

Aplicaciones INVESTIGACIÓN DE... EJERCICIOS DE MO... Learn SQL | SoloLea... Log in to Overleaf ~... Quiz 2 Chapters 5 a... SQL - mostrar porc... MySQL, Hosting Gr...

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

/input

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	18 KB	29/3/2023, 17:05:06	1	128 MB	Puntuacion.txt

Hadoop, 2014.

17:10 29/03/2023

Browsing HDFS

localhost:50070/explorer.html#/output

Aplicaciones INVESTIGACIÓN DE... EJERCICIOS DE MO... Learn SQL | SoloLea... Log in to Overleaf ~... Quiz 2 Chapters 5 a... SQL - mostrar porc... MySQL, Hosting Gr...

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

/output

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	29/3/2023, 17:06:35	1	128 MB	_SUCCESS
-rw-r--r--	root	supergroup	35 B	29/3/2023, 17:06:33	1	128 MB	resultado.txt

Hadoop, 2014.

17:10 29/03/2023

Tabla 1. Puntuaciones

Puntuación	Cantidad	Promedio	
1	1381	0,150	14,99%
2	1070	0,116	11,61%
3	1245	0,135	13,51%
4	2550	0,277	27,67%
5	2969	0,322	32,22%

Figura 1. Correos



ANÁLISIS

En la tabla de puntuaciones se observa que aproximadamente el 25% de las personas que van al hotel no quedan satisfechos con el mismo, ya que dan una puntuación de 1 y 2, lo cual es una puntuación mala para el hotel. Mientras que un 41% de los huéspedes dan una puntuación aceptable y buena, que seria 3 y 4 respectivamente, un 32% aproximadamente de los huéspedes queda satisfecha con el hotel puntuando excelentemente con un 5, lo cual parece ser un porcentaje bajo, ya que probablemente la mitad de las personas que llegan al hotel ya no vuelvan más.

En la figura de correos, se observa la relación que tiene el hotel con dichas palabras que aparecen frecuentemente en los correos, lo cual sirve para poder interpretar que es lo que les gusta y que no a los huéspedes del hotel.

Las palabras positivas como bonito, buena, excelente, deben permanecer a las personas que quedan muy satisfechas, las que dieron puntuación muy alto.

Sin embargo, hay otras palabras como cama, habitaciones, personal, servicio, que es muy poco probable que se hablen cosas positivas de estas, ya que la palabra NO de igual forma se repite mucho, pueda que estas estén relacionadas entre sí y ese sea el problema del porque se le da una calificación baja al hotel.

Por lo tanto, el problema del hotel es que las habitaciones no sean las esperadas por los huéspedes, así como las camas, o el servicio del personal no sea tan bueno, la palabra estacionamiento y ubicación es una de las más mencionadas también, puede ser que sea muy difícil llegar al hotel o muy confuso, así como el estacionamiento debe estar muy limitado o poco seguro al perderse algún tipo de objeto de los mismos o bien algún golpe en los carros.

Las habitaciones son un elemento fundamental para un hotel, por lo que es probable que los huéspedes presten mucha atención a la comodidad, la limpieza, el tamaño y la calidad de las habitaciones, la cama es esencial para una buena noche de sueño, por lo que es probable que los huéspedes valoren camas cómodas, si el hotel cuenta con un estacionamiento privado y seguro, puede ser un gran beneficio para los huéspedes que llegan en automóvil, el personal puede ser clave para garantizar una estancia agradable, si el personal es amable, servicial y eficiente, es probable que los huéspedes tengan una experiencia positiva, si el hotel ofrece servicios adicionales como desayuno, limpieza de habitaciones, servicio de lavandería, entre otros, es probable que los huéspedes valoren estos servicios y la calidad del servicio prestado.

CONCLUSIONES

- Los factores que parecen ser más importantes para los huéspedes en un hotel son la calidad de las habitaciones, la comodidad de la cama, la seguridad del estacionamiento, la amabilidad y eficiencia del personal, la calidad del servicio, la ubicación y la impresión general del hotel.
- los factores de la figura 1 de correos, puede ayudar al hotel a mejorar la experiencia del huésped y a garantizar su satisfacción.
- La mayoría de los huéspedes dieron al hotel una puntuación alta, por lo que tuvieron una experiencia positiva y están satisfechos con su estancia.
- Se interpreto las posibles razones del porque hubo puntuaciones bajas, por lo cual es necesario mejorar esos factores.
- Las puntuaciones proporcionan una idea general de la satisfacción de los huéspedes en el hotel.
- Hadoop permite procesar grandes conjuntos de datos no estructurados en paralelo, lo que puede mejorar la eficiencia y velocidad de procesamiento.
- La función de conteo de palabras es una tarea común en el procesamiento de texto no estructurado y Hadoop puede realizar esta tarea con facilidad.
- El conteo de palabras puede ser una tarea útil para identificar patrones o tendencias en grandes conjuntos de datos de texto no estructurado, como la frecuencia de palabras clave, en este caso un conjunto de comentarios de huéspedes de un hotel y puntuaciones del mismo.