

Hadoop: Taller Mapreduce/wordcount

Autor: Ing.Luis Felipe Narvaez Gomez. E-mail:luis.narvaez@usantoto.edu.co. Cod: 2312660. Facultad de Ingeniería de Sistemas.

Este taller esta realizado en Fedora 36 Linux, virtualizado en Windows 10 Home single Language , mediante el uso del software de Virtual Box. Para mas detalles de como instalar Fedora 36 y de como instalar en el Hadoop le recomendamos ver las guias anteriores.

INICIAR HADOOP

Teniendo previamente instalado y funcionando correctamente hadoop en nuestro SO de Fedora 36. Volvemos nuevamente a iniciar nuestro Hadoop.

1. Abrir el nuevo usuario de hadoop como root. Es posible que nos pida la contraseña de hadoop o la de nuestro usuario principal.

```
sudo su - hadoop
```

2. Situarnos en la raiz de este usuario.

```
cd
```

3. Ir a la carpeta de hadoop y abrir el folder de sbin.

```
cd /home/hadoop/hadoop/sbin/
```

4. Iniciar el namenodes.

```
./start-dfs.sh
```

5. Iniciar el yarn.

```
./start-yarn.sh
```

6. Abrir otra terminal nueva y acceder nuevamente como super usuario de hadoop. Luego preguntar la direccion IPv4

```
[Shift] + [CTRL] + [T] => para abrir una nueva pestaña
[Shift] + [CTRL] + [N] => para abrir una nueva ventana terminal

sudo su - hadoop
cd
ifconfig
```

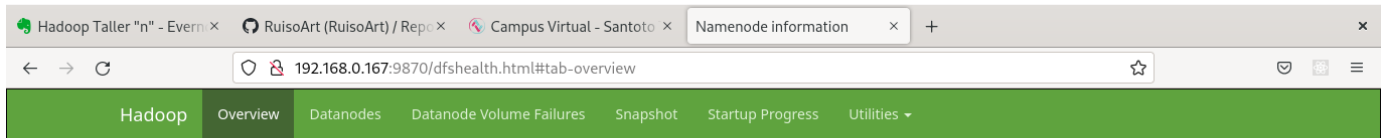
7. Rescatar la lpv4 que tengamos y guardarla para después.

```
in: ifconfig

out:
...
broadcast 192.168.0.255
...
```

8. Verificar el correcto funcionamiento en su navegador de internet predilecto el OVERVIEW

http://192.168.0.167:9870



Overview 'localhost:9000' (✓active)

| | |
|----------------|--|
| Started: | Tue Sep 20 15:39:29 -0500 2022 |
| Version: | 3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb |
| Compiled: | Fri Jul 29 07:32:00 -0500 2022 by stevel from branch-3.3.4 |
| Cluster ID: | CID-59aa29a5-3511-485e-8605-6d9c489ecc11 |
| Block Pool ID: | BP-1750418209-192.168.0.167-1663563249094 |

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

9. Verificar el correcto funcionamiento en su navegador de internet predilecto el Gestor de Recursos.

http://192.168.0.167:8042

| | |
|--|--|
| Total Vmem allocated for Containers | 16.80 GB |
| Vmem enforcement enabled | true |
| Total Pmem allocated for Container | 8 GB |
| Pmem enforcement enabled | true |
| Total VCores allocated for Containers | 8 |
| Resource types | memory-mb (unit=Mb), vcores |
| NodeHealthyStatus | true |
| LastNodeHealthTime | Tue Sep 20 16:03:23 COT 2022 |
| NodeHealthReport | |
| NodeManager started on | Tue Sep 20 15:43:13 COT 2022 |
| NodeManager Version: | 3.3.4 from a585a73c3e02ac62350c136643a5e7f6095a3dbb by stevel source checksum 17e8efaf27d922f2de51e5be9e69e9 on 2022-07-29T12:51Z |
| Hadoop Version: | 3.3.4 from a585a73c3e02ac62350c136643a5e7f6095a3dbb by stevel source checksum fb9dd8918a7b8a5b430d61af858f6ec on 2022-07-29T12:32Z |

10. Verificar el correcto funcionamiento en su navegador de internet predilecto las Piscinas de Bloqueo.

http://192.168.0.167:9864

Hadoop Taller "n" - Evern x RuisoArt (RuisoArt) / Rep x Campus Virtual - Santoto x DataNode Information x +

192.168.0.167:9864/datanode.html

Hadoop Overview Utilities

DataNode on fedora:9866

| | |
|-------------|--|
| Cluster ID: | CID-59aa29a5-3511-485e-8605-6d9c489ecc11 |
| Started: | Tue Sep 20 15:39:46 -0500 2022 |
| Version: | 3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb |

Block Pools

| Namenode Address | Block Pool ID | Actor State | Last Heartbeat | Last Block Report | Last Block Report Size (Max Size) |
|------------------|---|-------------|----------------|-------------------|-----------------------------------|
| localhost:9000 | BP-1750418209-192.168.0.167-1663563249094 | RUNNING | 1s | 20 minutes | 0 B (128 MB) |

Volume Information

Tenga en cuenta que en caso de funcionar con la Plv4 que muestra su computadora, debe recordar que la instalación de Hadoop a veces se liga a la red con la que lo instalo inicialmente por lo que le sugerimos trabar en la misma red de instalación o dirigirse a la guía de instalación de hadoop en Fedora en la sección del paso 37 donde se obtendra el nombre del nodo al formatear el HDFS.

BUSCAR UN LIBRO

Para este taller utilizaremos un archivo de texto plano en encoding UTF-8 correspondiente a un libro que no tenga muchos problemas o ninguno en realidad con los derechos de autor, para esto visitaremos el siguiente enlace <https://gutenberg.org/> en donde buscaremos el que mas nos interese y seguiremos los siguientes pasos:

1. Entrar a <https://gutenberg.org/>

Hadoop Taller "n" - Evern x RuisoArt (RuisoArt) / Rep x Campus Virtual - Santoto x DataNode Information x Free eBooks | Project Gutenberg x +

https://gutenberg.org

Project Gutenberg About Search and Browse Help Quick search Go! Donation PayPal

Welcome to Project Gutenberg

Project Gutenberg is a library of over 60,000 free eBooks

Choose among free epub and Kindle eBooks, download them or read them online. You will find the world's great literature here, with focus on older works for which U.S. copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for you to enjoy.

| | | | | | | | | | |
|--------------------------------------|---|-----------------------|--|--------------------------------------|---|----------------------------|---------------------------------------|-------------------------------------|--------------------------|
| | | | | | | | | | |
| Közép-ázsiai utazás by Ármín Vámbéry | The prey of the strongest by Morley Roberts | De afstamming van den | Drawing in charcoal and crayon for the | Australian Fairy Tales by James Hume | Lafitte, a play in prologue and four acts | Definition by Damon Knight | The Chinese Exclusion Act by New York | Les aventures du capitaine Magon by | Blindfold by Orrick John |

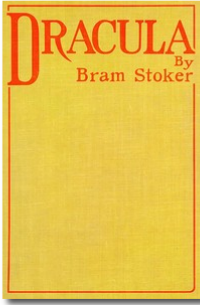
2. Buscar el libro que mas os guste, en mi caso:

Hadoop Taller "n" - Evern x RuisoArt (RuisoArt) / Rep x Campus Virtual - Santoto x DataNode Information x Dracula by Bram Stoker - x +

https://gutenberg.org/ebooks/345


Project Gutenberg About Search and Browse Help Quick search Go! Donation PayPal

Dracula by Bram Stoker



Download This eBook

| Format ? | Size | ? | ? | ? |
|---|--------|---|---|---|
| Read this book online: HTML | 906 kB | | | |
| EPUB (with images) | 591 kB | | | |
| EPUB (no images) | 427 kB | | | |
| Kindle (with images) | 758 kB | | | |
| Kindle (no images) | 592 kB | | | |
| Plain Text UTF-8 | 861 kB | | | |
| More Files... | | | | |



Similar Books

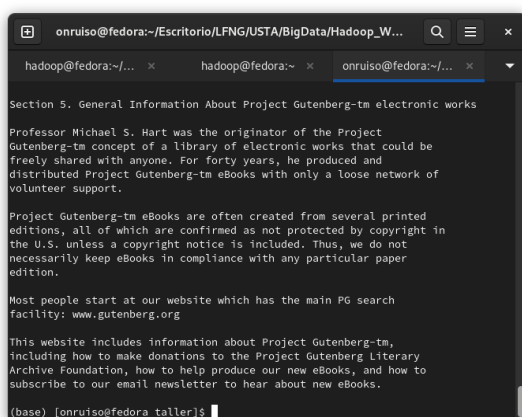
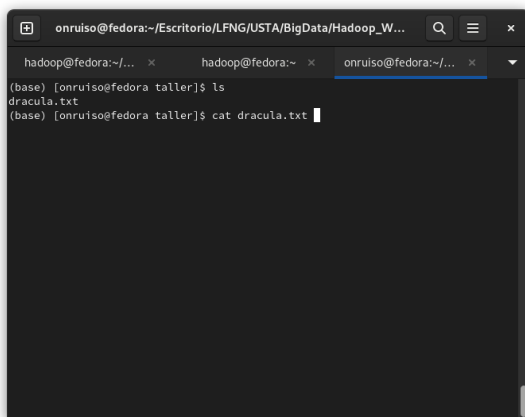
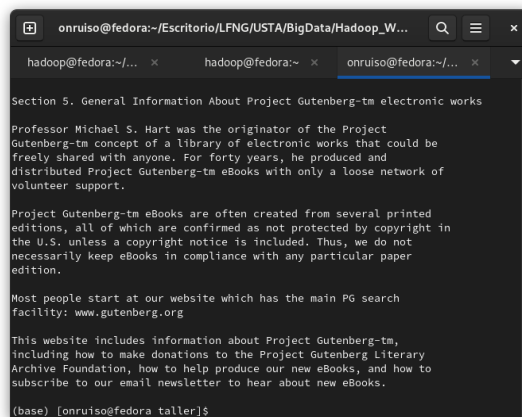
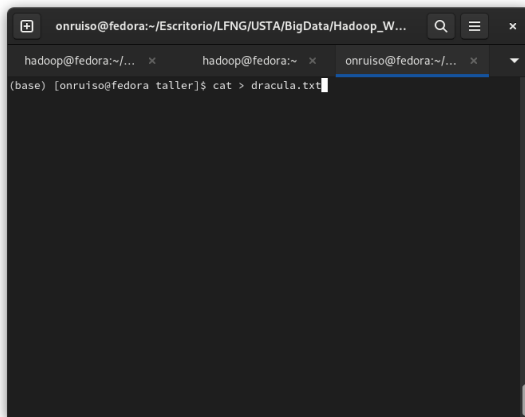
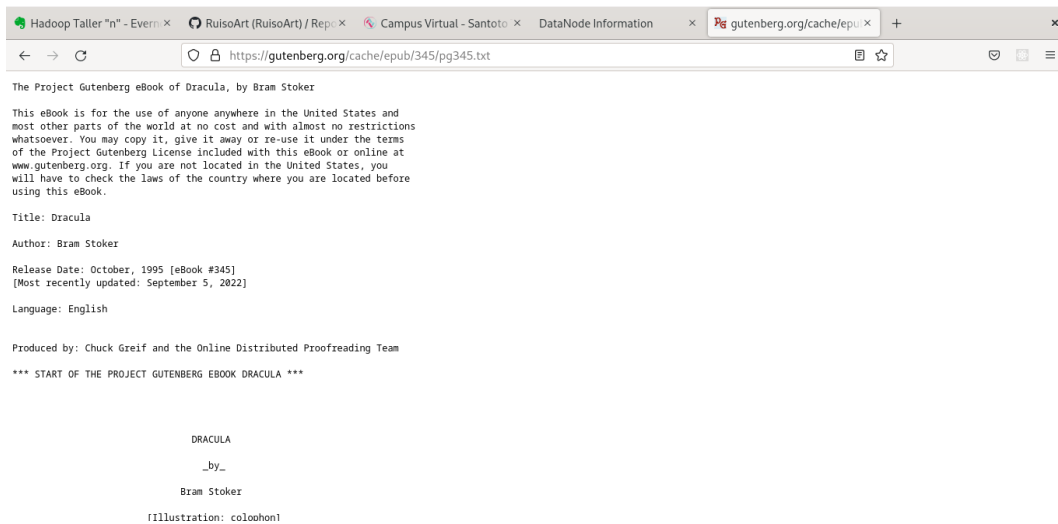
3. Bajamos la versión del archivo en UTF-8, al seleccionarlo se nos abrirá una nueva ventana con el texto, allí seleccionaremos todo y rescataremos en un archivo de texto en nuestra maquina. El procedimiento que vera en las imágenes es el siguiente:
 1. Dirigirse a la pagina con el texto plano abierto.
 2. seleccionar todo con [CTRL]+[A] copiar todo el texto [CTRL]+[C].
 3. Dirigirnos al sitio donde queremos que se guarde el libro en txt por medio de terminal o administrador de archivos.
 4. Crear un nuevo documento con el comando

```
cat > documento_nombre.txt
```

Al dar clic en la tecla enter el cursor se ira inmediatamente abajo, allí podremos escribir el contenido del documento.

1. Cerrar la edicion del documento con [CTRL]+[D]
2. verificar el documento con el comando

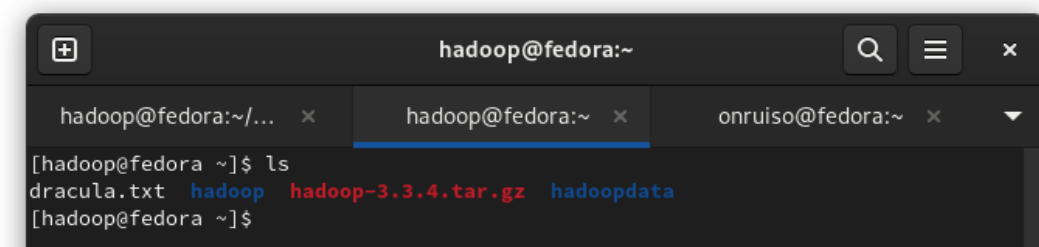
```
cat documento_nombre.txt
```



Debemos tener en cuenta que antes de crear todo este archivo y demás, todo se debe estar realizando dentro del super usuario de hadoop en la terminal.

SUBIR EL LIBRO A HADOOP

Aquí vamos a subir nuestro libro al cluster de hadoop. Primero confirmemos la ruta donde guardamos nuestro libro, en mi caso lo tengo directamente en mi "home" del sistema.

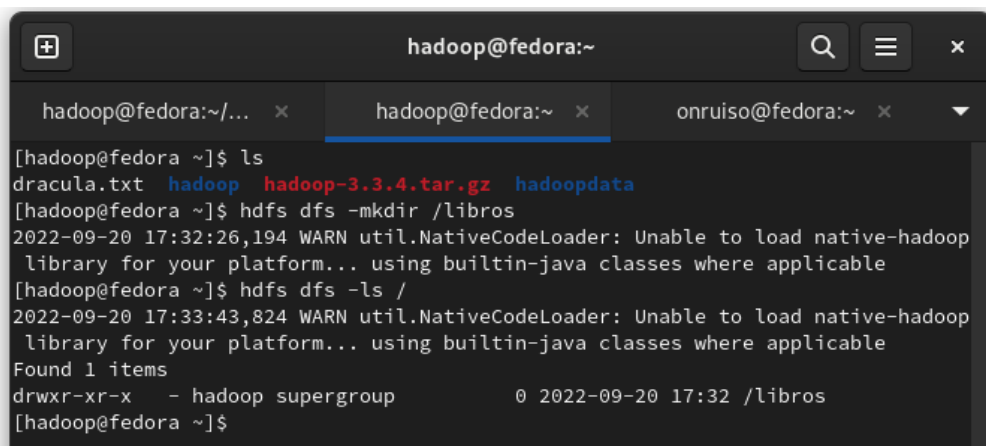


Ahora lo que haremos es crear una nueva carpeta con ayuda de HDFS. Para esto utilizamos la funcion dfs y lo haremos a nivel de raíz.

```
hdfs dfs -mkdir /libros
```

Comprobamos la creación de la carpeta con:

```
hdfs dfs -ls /
```



```
hadoop@fedora:~$ ls
dracula.txt  hadoop  hadoop-3.3.4.tar.gz  hadoopdata
[hadoop@fedora ~]$ hdfs dfs -mkdir /libros
2022-09-20 17:32:26,194 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
[hadoop@fedora ~]$ hdfs dfs -ls /
2022-09-20 17:33:43,824 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x  - hadoop supergroup          0 2022-09-20 17:32 /libros
[hadoop@fedora ~]$
```

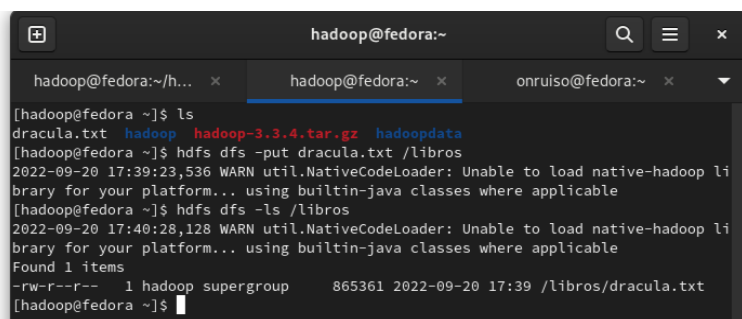
Ahora si vamos a subir nuestro archivo a la carpeta de hadoop libros, para esto utilizamos el siguiente comando:

```
hdfs dfs -put el_libro.txt /libros
```

El comando nos dice "con el componente de hadoop, la función dfs, súbeme un archivo llamado x justo donde estoy en este momento, a la ruta que te digo". Hay que tener cuidado de en que parte de la ruta del sistema se ejecuta este comando, pues buscara ahí el archivo que estamos queriendo subir.

Confirmamos con el comando:

```
hdfs dfs -ls /libros
```



```
hadoop@fedora:~$ ls
dracula.txt  hadoop  hadoop-3.3.4.tar.gz  hadoopdata
[hadoop@fedora ~]$ hdfs dfs -put dracula.txt /libros
2022-09-20 17:39:23,536 WARN util.NativeCodeLoader: Unable to load native-hadoop li
brary for your platform... using builtin-java classes where applicable
[hadoop@fedora ~]$ hdfs dfs -ls /libros
2022-09-20 17:40:28,128 WARN util.NativeCodeLoader: Unable to load native-hadoop li
brary for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--  1 hadoop supergroup    865361 2022-09-20 17:39 /libros/dracula.txt
[hadoop@fedora ~]$
```

MAPREDUCE - WORDCOUNT

Lo primero es ir a donde tenemos instalado hadoop en nuestra maquina, en mi caso es lo siguiente:

```
hadoop@fedora:~/hadoop
hadoop@fedora:~/h... x hadoop@fedora:~/h... x onruiso@fedora:~ x
[hadoop@fedora ~]$ pwd
/home/hadoop
[hadoop@fedora ~]$ ls
dracula.txt hadoop-3.3.4.tar.gz hadoopdata
[hadoop@fedora ~]$ cd hadoop/
[hadoop@fedora hadoop]$ ls -al
total 84
drwxr-xr-x. 1 hadoop hadoop 218 sep 18 23:54 .
drwx----- 1 hadoop hadoop 218 sep 20 17:28 ..
drwxr-xr-x. 1 hadoop hadoop 218 jul 29 08:44 bin
drwxr-xr-x. 1 hadoop hadoop 12 jul 29 07:35 etc
drwxr-xr-x. 1 hadoop hadoop 120 jul 29 08:44 include
drwxr-xr-x. 1 hadoop hadoop 12 jul 29 08:44 lib
drwxr-xr-x. 1 hadoop hadoop 372 jul 29 08:44 libexec
-rw-r--r-- 1 hadoop hadoop 24707 jul 28 15:30 LICENSE-binary
drwxr-xr-x. 1 hadoop hadoop 1702 jul 29 08:44 licenses-binary
-rw-r--r-- 1 hadoop hadoop 15217 jul 16 13:20 LICENSE.txt
drwxr-xr-x. 1 hadoop hadoop 1190 sep 20 15:43 logs
-rw-r--r-- 1 hadoop hadoop 29473 jul 16 13:20 NOTICE-binary
-rw-r--r-- 1 hadoop hadoop 1541 abr 22 09:58 NOTICE.txt
-rw-r--r-- 1 hadoop hadoop 175 abr 22 09:58 README.txt
drwxr-xr-x. 1 hadoop hadoop 778 jul 29 07:35 sbin
drwxr-xr-x. 1 hadoop hadoop 18 jul 29 09:21 share
[hadoop@fedora hadoop]$
```

En esta ruta encontraremos una carpeta que se llama SHARE aquí encontraremos varios ejemplos y utilidades de hadoop que podemos utilizar. La ruta a lo que nos dirigiremos es:

```
cd /home/hadoop/hadoop/share/hadoop/mapreduce
```

Tenga en cuenta que la dirección es tal en mi caso, cambiara dependiendo si usted tiene la instalación de hadoop en otra parte de su sistema operativo. Las utilidades que tenemos de mapreduce son las siguientes:

```
hadoop@fedora:~/hadoop/share/hadoop/mapreduce
hadoop@fedora:~/hadoop/... x hadoop@fedora:~/hadoop/... x onruiso@fedora:~ x
[hadoop@fedora mapreduce]$ pwd
/home/hadoop/hadoop/share/hadoop/mapreduce
[hadoop@fedora mapreduce]$ ls -al
total 5292
drwxr-xr-x. 1 hadoop hadoop 946 jul 29 08:44 .
drwxr-xr-x. 1 hadoop hadoop 68 jul 29 08:44 ..
-rw-r--r-- 1 hadoop hadoop 590752 jul 29 08:22 hadoop-mapreduce-client-app-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 805750 jul 29 08:22 hadoop-mapreduce-client-common-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 1636329 jul 29 08:22 hadoop-mapreduce-client-core-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 181707 jul 29 08:22 hadoop-mapreduce-client-hs-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 9966 jul 29 08:22 hadoop-mapreduce-client-hs-plugins-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 49783 jul 29 08:22 hadoop-mapreduce-client-jobclient-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 1658927 jul 29 08:22 hadoop-mapreduce-client-jobclient-3.3.4-tests.jar
-rw-r--r-- 1 hadoop hadoop 90704 jul 29 08:22 hadoop-mapreduce-client-nativetask-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 62093 jul 29 08:22 hadoop-mapreduce-client-shuffle-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 22263 jul 29 08:22 hadoop-mapreduce-client-uploader-3.3.4.jar
-rw-r--r-- 1 hadoop hadoop 280990 jul 29 08:22 hadoop-mapreduce-examples-3.3.4.jar
drwxr-xr-x. 1 hadoop hadoop 2364 jul 29 08:44 idiff
drwxr-xr-x. 1 hadoop hadoop 32 jul 29 08:44 lib-examples
drwxr-xr-x. 1 hadoop hadoop 1810 jul 29 08:44 sources
[hadoop@fedora mapreduce]$
```

Como podemos observar en la imagen, encontramos muchos proceso de extencion jar , esto quiere decir que los mismos utilizan JAVA para poder funcionar y en caso de tener un proceso al cual ejecutar con hadoop el mismo obligatoriamente debe tener esta extencion. Todo proceso desarrollado en Java por nuestra cuenta o descargado por otro lado, debe tener esta extencion. En nuestro caso utilizaremos uno de los ejemplos ya desarrollados llamado EXAMPLES.

```
hadoop jar hadoop-mapreduce-examples-3.3.4.jar
```

```
hadoop@fedora:~/hadoop/share/hadoop/mapreduce

hadoop@fedora:~/hadoop/... x hadoop@fedora:~/hadoop/... x onruiso@fedora:~ x
aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files
.
aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words
in the input files.
bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
dbcount: An example job that count the pageview counts from a database.
distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
grep: A map/reduce program that counts the matches of a regex in the input.
join: A job that effects a join over sorted, equally partitioned datasets
multifilewc: A job that counts words from several files.
pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
randomwriter: A map/reduce program that writes 10GB of random data per node.
secondarysort: An example defining a secondary sort to the reduce.
sort: A map/reduce program that sorts the data written by the random writer.
sudoku: A sudoku solver.
terasgen: Generate data for the terasort
terasort: Run the terasort
teravalidate: Checking results of terasort
wordcount: A map/reduce program that counts the words in the input files.
wordmean: A map/reduce program that counts the average length of the words in the input files.
wordmedian: A map/reduce program that counts the median length of the words in the input files.
wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of th
e words in the input files.
[hadoop@fedora mapreduce]$
```

Creamos una nueva carpeta en la raíz de hadoop que contendrá los resultados de nuestro conteo de palabras:

```
hdfs dfs -mkdir /result_libros
```

Confirmamos con:

```
hdfs dfs -ls /
```

Podemos observar que dentro de la clase comprimida JAR tenemos varios sub-programas que podemos utilizar, entre ellos el que nos interesa llamado WORDCOUNT.

Vamos a utilizar este WORDCOUNT con nuestro libro.

```
hadoop jar hadoop-mapreduce-examples-3.3.4.jar wordcount /libros /result_libros
```

El comando seria algo como "utilizando hadoop, con la funcion JAR para leer clases comprimidas en JAVA, ejecutar el EXAMPLE, el metodo de WORDCOUNT, para todo lo que esta en la dirección A, y luego exporta me el resultado en la carpeta B". Tenga en cuenta que al ejecutarlo se demorara un tiempo variable entre la complejidad del proceso y los recursos de su maquina.

```
hadoop@fedora:~/hadoop/share/hadoop/mapreduce

hadoop@fedora:~/hadoop/sbin x hadoop@fedora:~/hadoop/share/had... x onruiso@fedora:~ x
[hadoop@fedora mapreduce]$ hadoop jar hadoop-mapreduce-examples-3.3.4.jar wordcount /libros /result_libros
```

Puede que nos salga el siguiente error:

```
hadoop@fedora:~/hadoop/share/hadoop/mapreduce

hadoop@fedora:~/hadoop/sbin x hadoop@fedora:~/hadoop/share/had... x hadoop@fedora:~ x
2022-09-20 18:46:46,279 INFO mapreduce.Job: Running job: job_1663706599371_0004
2022-09-20 18:46:50,328 INFO mapreduce.Job: Job job_1663706599371_0004 running in uber mode : false
2022-09-20 18:46:50,329 INFO mapreduce.Job: map 0% reduce 0%
2022-09-20 18:46:50,372 INFO mapreduce.Job: Job job_1663706599371_0004 failed with state FAILED due to: Application applicat
ion_1663706599371_0004 failed 2 times due to AM Container for appattempt_1663706599371_0004_000002 exited with exitCode: 1
Failing this attempt.Diagnostics: [2022-09-20 18:46:49.775]Exception from container-launch.
Container id: container_1663706599371_0004_02_000001
Exit code: 1

[2022-09-20 18:46:49.782]Container exited with a non-zero exit code 1. Error file: prelaunch.err.
Last 4096 bytes of prelaunch.err :
Last 4096 bytes of stderr :
Error: no se ha encontrado o cargado la clase principal org.apache.hadoop.mapreduce.v2.app.MRAppMaster

[2022-09-20 18:46:49.782]Container exited with a non-zero exit code 1. Error file: prelaunch.err.
Last 4096 bytes of prelaunch.err :
Last 4096 bytes of stderr :
Error: no se ha encontrado o cargado la clase principal org.apache.hadoop.mapreduce.v2.app.MRAppMaster

For more detailed output, check the application tracking page: http://fedora:8088/cluster/app/application_1663706599371_0004
Then click on links to logs of each attempt.
. Failing the application.
2022-09-20 18:46:50,457 INFO mapreduce.Job: Counters: 0
[hadoop@fedora mapreduce]$
```

Nueva Terminal e ingresamos al sitio donde se instalo hadoop.


```
cd
cd hadoop/
hadoop classpath
```

Obtendremos como salida algo similar a esto:

```
/home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop/common/lib/*:/home/hadoop/hadoop/share/hadoop/common/*:/home/hadoop/hadoop/share/hadoop/hdfs:/home/hadoop/hadoop/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop/share/hadoop/hdfs/*:/home/hadoop/hadoop/share/hadoop/mapreduce/*:/home/hadoop/hadoop/share/hadoop/yarn:/home/hadoop/hadoop/share/hadoop/yarn/lib/*:/home/hadoop/hadoop/share/hadoop/yarn/*
```

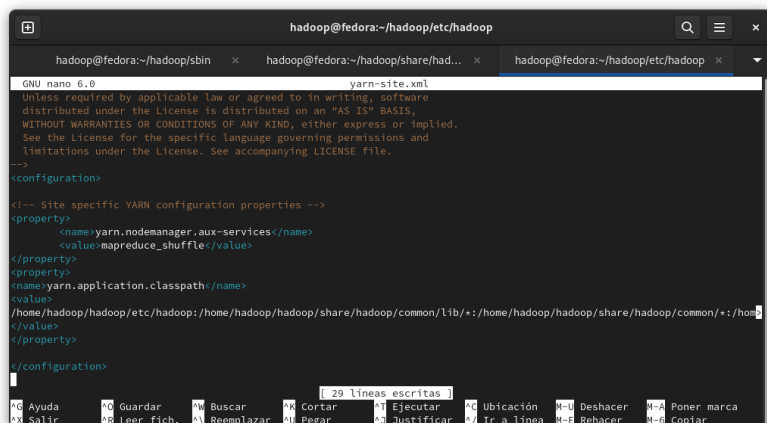
Agregue el valor de la salida anterior al atributo yarn.application.classpath correspondiente al archivo yarn-site.xml.

```
cd /home/hadoop/hadoop/etc/hadoop/
nano
nano yarn-site.xml
```

Se abrirá nano y tendremos la siguiente propiedad que pondremos dentro.

```
<property>
<name>yarn.application.classpath</name>
<value>
/home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop/common/lib/*:/home/hadoop/hadoop/share/hadoop/common/*:/home/hadoop/hadoop/share/hadoop/hdfs:/home/hadoop/hadoop/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop/share/hadoop/hdfs/*:/home/hadoop/hadoop/share/hadoop/mapreduce/*:/home/hadoop/hadoop/share/hadoop/yarn:/home/hadoop/hadoop/share/hadoop/yarn/lib/*:/home/hadoop/hadoop/share/hadoop/yarn/*
</value>
</property>
```

Guardamos con [CTRL]+[O] y salimos con [CTRL]+ 



Ahora queda reiniciar yarn.

```
cd /home/hadoop/hadoop/sbin
./start-dfs.sh
./start-yarn.sh
./start-all.sh
```

Ahora volvemos a probar.

```
cd /home/hadoop/hadoop/share/hadoop/mapreduce
hadoop jar hadoop-mapreduce-examples-3.3.4.jar wordcount /libros/dracula.txt /result_libros2
```

Si tenemos la siguiente salida por pantalla es que ya tenemos todo bien.

```
hadoop@fedora:~/hadoop/share/hadoop/mapreduce

Reduce output records=19806
Spilled Records=38012
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=131
CPU time spent (ms)=2460
Physical memory (bytes) snapshot=434618368
Virtual memory (bytes) snapshot=5876568064
Total committed heap usage (bytes)=360452096
Peak Map Physical memory (bytes)=277368832
Peak Map Virtual memory (bytes)=2934898088
Peak Reduce Physical memory (bytes)=157240536
Peak Reduce Virtual memory (bytes)=2941669376

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=865361
File Output Format Counters
  Bytes Written=198010
[hadoop@fedora mapreduce]$
```

Comprobamos la salida en nuestro cluster de hadoop

```
hdfs dfs -ls /result_libros2
```

Teniendo una salida como esta:

```
hadoop@fedora:~/hadoop/share/hadoop/mapreduce

IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=865361
File Output Format Counters
  Bytes Written=198010
[hadoop@fedora mapreduce]$
[hadoop@fedora mapreduce]$ hdfs dfs -ls /
2022-09-20 19:15:26,497 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
java classes where applicable
Found 4 items
drwxr-xr-x - hadoop supergroup      0 2022-09-20 17:39 /libros
drwxr-xr-x - hadoop supergroup      0 2022-09-20 18:09 /result_libros
drwxr-xr-x - hadoop supergroup      0 2022-09-20 19:13 /result_libros2
drwx----- - hadoop supergroup      0 2022-09-20 18:04 /tmp
[hadoop@fedora mapreduce]$ hdfs dfs -ls /result_libros2
2022-09-20 19:15:50,329 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-
java classes where applicable
Found 2 items
-rw-r--r-- 1 hadoop supergroup      0 2022-09-20 19:13 /result_libros2/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 198010 2022-09-20 19:13 /result_libros2/part-r-00000
[hadoop@fedora mapreduce]$
```

Podemos ver el contenido de dos maneras, descargarlo o verlo directamente. Primero descargarlo:

```
hdfs dfs -get /result_libros2/part-r-00000 /home/hadoop
```

La ruta seria algo como "con HDFS, y la funcion DFS, traeme con la funcion GET, el archivo de la direccion que te digo,y descargamela en la otra ruta que te digo".

Tenemos nuestro archivo descargado:

```
[hadoop@fedora ~]$ ls
dracula.txt  hadoop  hadoop-3.3.4.tar.gz  hadoopdata  part-r-00000
[hadoop@fedora ~]$
```

Ahora podemos verlo así:

```
more part-r-00000
```

Observaremos que tenemos un conteo de cada una de las diferentes palabras, caracteres individuales, espacios y demás strings individuales de nuestro libro. En el caso de querer generar etiquetas a partir de este archivo, hay varias palabras que no nos interesan, por lo que es recomendable hacer un proceso de limpieza al texto puro de forma externa antes de utilizar esta herramienta de hadoop.

```
hadoop@fedora:~  
hadoop@fedora:~/hadoop/sbin x hadoop@fedora:~/hadoop/share/had... x hadoop@fedora:~  
''Are 1  
''E's 1  
''I 1  
''Ittin' 1  
''Little 1  
''Lucy, 1  
''Maybe 1  
''Miss 1  
''My 2  
''Never 1  
''No' 1  
''Ow 1  
''Silence! 1  
''That's 1  
''Tyke 1  
''Wilhelmina'--I 1  
''Yes, 1  
''A 8  
''ABRAHAM 1  
''ART. 1  
''ARTHUR. 1  
''About 1  
''Afraid 1  
''Again 1  
''Agreed! 1  
--Hds-- (0%)
```

También podemos visualizar lo mismo en hadoop de la siguiente manera.

```
hdfs dfs -cat /result_libros2/part-r-00000
```

Obteniendo una salida como la siguiente.

```
hadoop@fedora:~  
hadoop@fedora:~/hadoop/sbin x hadoop@fedora:~/hadoop/share/had... x hadoop@fedora:~  
yours, 5  
yours. 3  
yours." 1  
yours; 1  
yourself 8  
yourself, 4  
yourself," 1  
yourself. 6  
yourself?" 2  
yourself?" 1  
yourselves 2  
youth 5  
youthful 2  
zeal; 1  
zealous 1  
zoophagous 3  
zoophagous, 1  
zoophagy!" 1  
{pg 8  
{pg}184 1  
El 1  
E10 1  
at. 1  
atat 1  
"The 1  
[hadoop@fedora ~]$
```