

En este documento se describen los pasos que se siguieron para realizar el programa en hadoop, se tomó como base las siguientes paginas para poder realizar este taller:

1) codigo WordCount

<https://dzone.com/articles/word-count-hello-word-program-in-mapreduce>

2) instalación de hadoop

<https://tecadmin.net/install-hadoop-on-ubuntu-20-04/>

3) se realizó una modificación al archivo mapred-site.xml ya que salía el error que se muestra a continuación

```
hadoop@lgueli:~$ hadoop jar MRProgramDemo.jar hadoop.WordCount /cuenta /cuenta2
2021-05-30 22:57:35,788 INFO client.DefaultHARMFaloverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-05-30 22:57:35,995 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1622431042153_0005
2021-05-30 22:57:36,161 INFO Input.FileInputFormat: Total input files to process : 1
2021-05-30 22:57:36,261 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-30 22:57:36,361 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622431042153_0005
2021-05-30 22:57:36,361 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-30 22:57:36,457 INFO conf.Configuration: resource-types.xml not found
2021-05-30 22:57:36,457 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-30 22:57:36,783 INFO impl.YarnClientImpl: Submitted application application_1622431042153_0005
2021-05-30 22:57:36,802 INFO mapreduce.Job: The url to track the job: http://Miguel:8088/proxy/application_1622431042153_0005/
2021-05-30 22:57:36,802 INFO mapreduce.Job: Running job: job_1622431042153_0005
2021-05-30 22:57:39,616 INFO mapreduce.Job: Job job_1622431042153_0005 running in uber mode : false
2021-05-30 22:57:39,817 INFO mapreduce.Job: map 0% reduce 0%
2021-05-30 22:57:39,825 INFO mapreduce.Job: Job job_1622431042153_0005 failed with state FAILED due to: Application application_1622431042153_0005 failed 2 times due to AM Container for appattempt 1622431042153_0005_000002 exited with exitCode: 1
Failing this attempt.Diagnostics: [2021-05-30 22:57:39,323]Exception from container-launch.
Container id: container_1622431042153_0005_02_000001
Exit code: 1

[2021-05-30 22:57:39,326]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

[2021-05-30 22:57:39,326]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://Miguel:8088/cluster/app/application_1622431042153_0005 Then click on links to logs of each attempt.
Failing the application.
2021-05-30 22:57:39,835 INFO mapreduce.Job: Counters: 0
```

por lo que se tomó la siguiente solución

<https://stackoverflow.com/questions/49675782/hadoop-could-not-find-or-load-main-class-org-apache-hadoop-mapreduce-v2-app-mra>

Consideraciones:

Para realizar correctamente este ejercicio es necesario ejecutar hadoop con JDK 1.8 java ya que fue compilado con esta versión, si tanto la versión con la que se compila el el jar o la versión de hadoop no es igual, saldrá el siguiente error.

```
hadoop@lgueli:~$ hadoop jar MRProgramDemo.jar hadoop.WordCount wordcountfile cuenta1
Exception in thread "main" java.lang.UnsupportedClassVersionError: hadoop/WordCount has been compiled by a more recent version of the Java Runtime (class file version 52.0), this version of the Java Runtime only recognizes class file versions up to 52.0
    at java.lang.ClassLoader.defineClass1(Native Method)
    at java.lang.ClassLoader.defineClass(ClassLoader.java:756)
    at java.security.SecureClassLoader.defineClass(SecureClassLoader.java:142)
    at java.net.URLClassLoader.defineClass(URLClassLoader.java:468)
    at java.net.URLClassLoader.access$100(URLClassLoader.java:74)
    at java.net.URLClassLoader$1.run(URLClassLoader.java:369)
    at java.net.URLClassLoader$1.run(URLClassLoader.java:363)
    at java.security.AccessController.doPrivileged(Native Method)
    at java.net.URLClassLoader.findClass(URLClassLoader.java:362)
    at java.lang.ClassLoader.loadClass(ClassLoader.java:418)
    at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
    at java.lang.Class.forName0(Native Method)
    at java.lang.Class.forName(Class.java:348)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:316)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
```

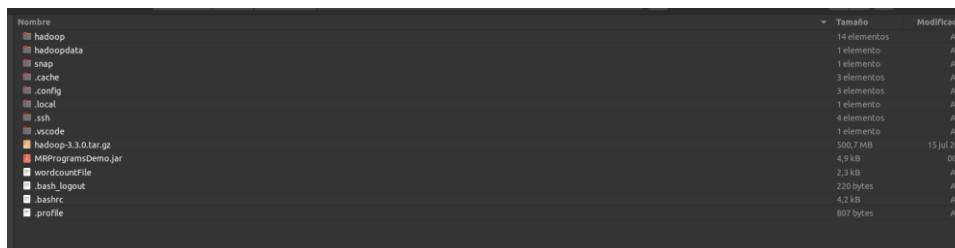
**Universidad el Bosque**  
**Ingeniería de Sistemas**  
**Curso: Big Data**  
**Profesor: Fabián Camilo Peña**  
**Tarea: Hadoop mapping**  
**Estudiante: Miguel Ángel Caro Boyacá**



1) Se realizó la modificación al código del link 1 (esta modificación va desde la línea 36 hasta la línea 47 (clase map), se verifica que los caracteres que se van a leer son tanto alfabéticos o son espacios, con el fin de remover caracteres especiales como el punto, la coma o “»”.

```
33 public static class MapForWordCount extends Mapper<LongWritable, Text, Text, IntWritable> {
34     public void map(LongWritable key, Text value, Context con) throws IOException, InterruptedException {
35         String line = value.toString();
36         String sinCaracteres="";
37         char[] ch = new char[line.length()];
38         for (int i = 0; i < line.length(); i++) {
39             ch[i] = line.charAt(i);
40         }
41         for(char c : ch){
42             if(Character.isAlphabetic(c) || c==' '){
43                 sinCaracteres+=c;
44             }
45         }
46         String[] words = sinCaracteres.split(" ");
47         for (String word : words) {
48             Text outputKey = new Text(word.toUpperCase().trim());
49             IntWritable outputValue = new IntWritable(1);
50             con.write(outputKey, outputValue);
51         }
52         System.out.println();
53     }
54 }
55 }
```

2) se realiza el jar correspondiente al proyecto. en este caso este jar lo dejo en la carpeta home del usuario hadoop

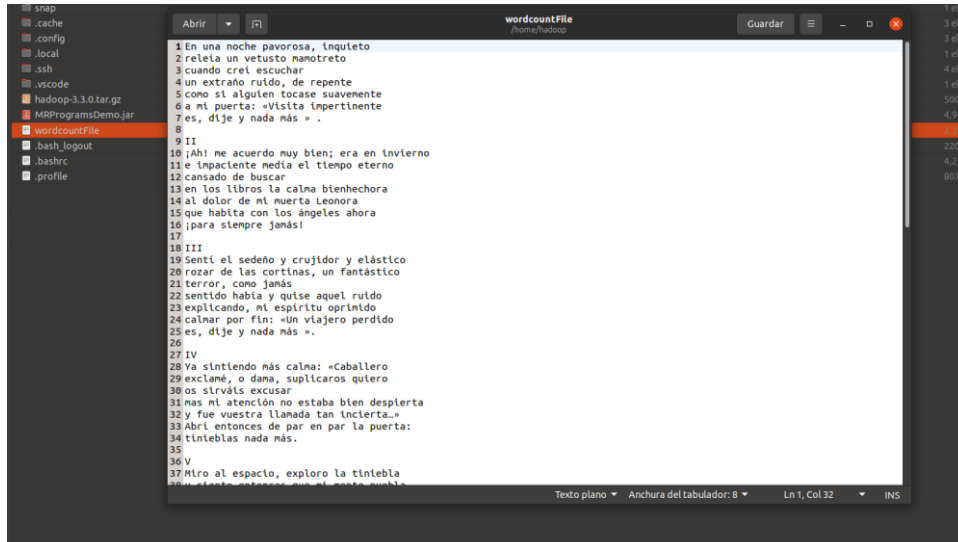


Nombre	Tamaño	Modificación
hadoop	14 elementos	Ayer
hadoopdata	1 elemento	Ayer
snap	1 elemento	Ayer
.cache	3 elementos	Ayer
.config	3 elementos	Ayer
.local	1 elemento	Ayer
.ssh	4 elementos	Ayer
.vscode	1 elemento	Ayer
hadoop-3.3.0.tar.gz	500.7 MB	15 Jul 2020
MRProgramsDemo.jar	4.9 kB	00:31
wordcountFile	2.3 kB	Ayer
.bash_logout	220 bytes	Ayer
.bashrc	4.2 kB	Ayer
.profile	807 bytes	Ayer

3) se crea el directorio cuenta, en donde se va a guardar el archivo txt

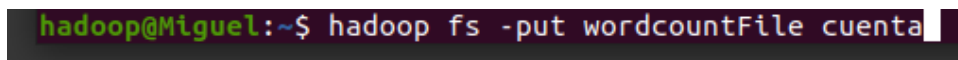
```
hadoop@Miguel:~$ hadoop fs -mkdir /cuenta
```

4) se crea el archivo



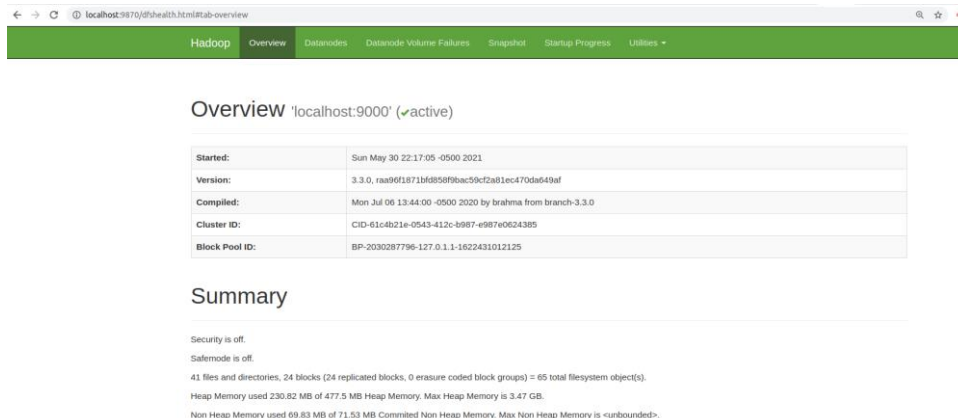
```
1 En una noche pavorosa, inquieto
2 relea un vetusto namotreto
3 cuando creí escuchar
4 un extraño ruido, de repente
5 como si alguien tocara suavemente
6 a mi puerta: «Visita impertinente
7 es, dije y nada más».
8
9 II
10 ¡Ah! me acuerdo muy bien; era en invierno
11 e impaciente media el tiempo eterno
12 cansado de buscar
13 en los libros la calma bienhechora
14 al dolor de mi muerta Leonora
15 que habita con los ángeles ahora
16 ¡para siempre jamás!
17
18 III
19 Sentí el sedoso y crujidor y elástico
20 rozar de las cortinas, un fantástico
21 terror, como jamás
22 sentido había y quise aquel ruido
23 explicando, mi espíritu oprimido
24 calmar por fin: «Un viajero perdido
25 es, dije y nada más».
26
27 IV
28 Ya sintiendo más calma: «Caballero
29 exclame, o dama, suplicaros quiero
30 os stróváis excusar
31 mas mi atención no estaba bien despierta
32 y fue vuestra llanada tan incierta.»
33 Abrí entonces de par en par la puerta:
34 tinieblas nada más.
35
36 V
37 Miro al espacio, exploro la tiniebla
38 y siento, entonces, en el pecho, súbita
```

5) se copia el archivo a la carpeta cuenta



```
hadoop@Miguel:~$ hadoop fs -put wordcountFile cuenta
```

6) se puede validar el documento en la interfaz que provee haddop en este caso, localhost:9870



localhost:9870/shell.html#tab-overview

Hadoop Overview Datanodes Datacode Volume Failures Snapshot Startup Progress Utilities

### Overview 'localhost:9000' (✓active)

Started:	Sun May 30 22:17:05 -0500 2021
Version:	3.3.0, raa90f1871bf6858f9bac59c72a01ec470da649af
Compiled:	Mon Jul 06 13:44:00 -0500 2020 by brahma from branch-3.3.0
Cluster ID:	CID-61c4b21e-0543-412c-8987-e987e0624385
Block Pool ID:	BP-2030287796-127.0.1.1-1622431012125

### Summary

Security is off.  
Safe mode is off.  
41 files and directories, 24 blocks (24 replicated blocks, 0 erasure coded block groups) = 65 total filesystem object(s).  
Heap Memory used 230.82 MB of 477.5 MB Heap Memory. Max Heap Memory is 3.47 GB.  
Non Heap Memory used 69.83 MB of 71.53 MB Committed Non Heap Memory. Max Non Heap Memory is «unbounded».

7) Nos dirigimos a Utilities-> browse the file system

## Browse Directory

/






Go!

Show

25

entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	May 30 22:40	0	0 B	cuenta	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	May 30 22:47	0	0 B	cuenta1	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	May 31 00:11	0	0 B	cuenta2	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	May 31 00:32	0	0 B	cuenta3	
<input type="checkbox"/>	drwx-----	hadoop	supergroup	0 B	May 30 22:53	0	0 B	tmp	

Showing 1 to 5 of 5 entries

Previous

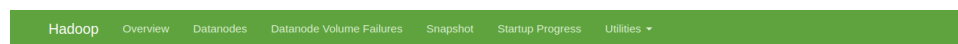
1

Next

Hadoop, 2020.

Hadoop, 2020.

En esta se visualizan todas las carpetas y archivos hdfs (hadoop distributed file system) que se encuentran.



## Browse Directory

/cuenta

Go!

Show

25

entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		-rw-r--r--		<a href="#">hadoop</a>		<a href="#">supergroup</a>		2.28 KB		May 30 22:40		<a href="#">1</a>		128 MB		<a href="#">wordcountFile</a>	

Showing 1 to 1 of 1 entries

Previous

1

Next

Hadoop\_2020

Hadoop, 2020.

8) Después de validar el archivo, se procede a ejecutar el jar

```

hadoop@Miguel:~$ hadoop jar MRProgramasDemo.jar hadoop.WordCount /cuenta /cuenta3
2021-05-31 00:32:25,582 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2021-05-31 00:32:25,600 INFO impl.MetricsSystemImpl: Scheduled metric snapshot period at 10 second(s).
2021-05-31 00:32:25,620 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2021-05-31 00:32:25,764 INFO input.FileInputFormat: Total input files to process : 1
2021-05-31 00:32:25,775 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-31 00:32:25,834 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1973347842_0001
2021-05-31 00:32:25,834 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-31 00:32:25,894 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2021-05-31 00:32:25,895 INFO mapreduce.Job: Running job: job_local1973347842_0001
2021-05-31 00:32:25,895 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2021-05-31 00:32:25,898 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2021-05-31 00:32:25,898 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false
, ignore cleanup failures: false
2021-05-31 00:32:25,898 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2021-05-31 00:32:25,935 INFO mapred.LocalJobRunner: Waiting for map tasks
2021-05-31 00:32:25,935 INFO mapred.LocalJobRunner: Starting task: attempt_local1973347842_0001_m_000000_0
2021-05-31 00:32:25,945 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2021-05-31 00:32:25,946 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false
, ignore cleanup failures: false
2021-05-31 00:32:25,952 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2021-05-31 00:32:25,954 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/cuenta/wordcountFile:0+2337
2021-05-31 00:32:25,986 INFO mapred.MapTask: (EQUATOR) 0 kvt 26214396(104857504)
2021-05-31 00:32:25,986 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2021-05-31 00:32:25,986 INFO mapred.MapTask: soft limit at 83886080
2021-05-31 00:32:25,986 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2021-05-31 00:32:25,986 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2021-05-31 00:32:25,988 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer

```

En esta se crea un nuevo archivo dentro de la carpeta cuenta3

## Browse Directory

/cuenta3

Show  entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	0 B	May 31 00:32	1	128 MB	<a href="#">_SUCCESS</a>	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	2.05 KB	May 31 00:32	1	128 MB	<a href="#">part-r-00000</a>	

Showing 1 to 2 of 2 entries

Hadoop, 2020.

9) Se descarga el archivo part-r-00000 y se puede ver la solución

