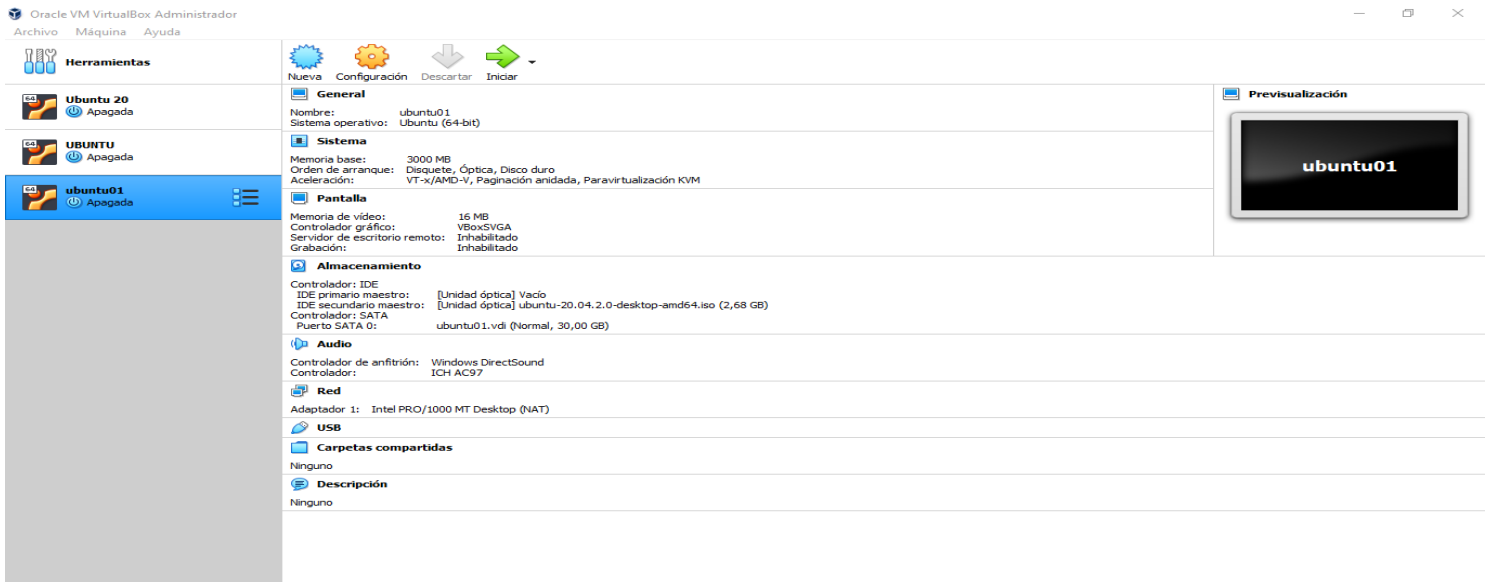




Informe Taller 1 – Ecosistema de Hadoop

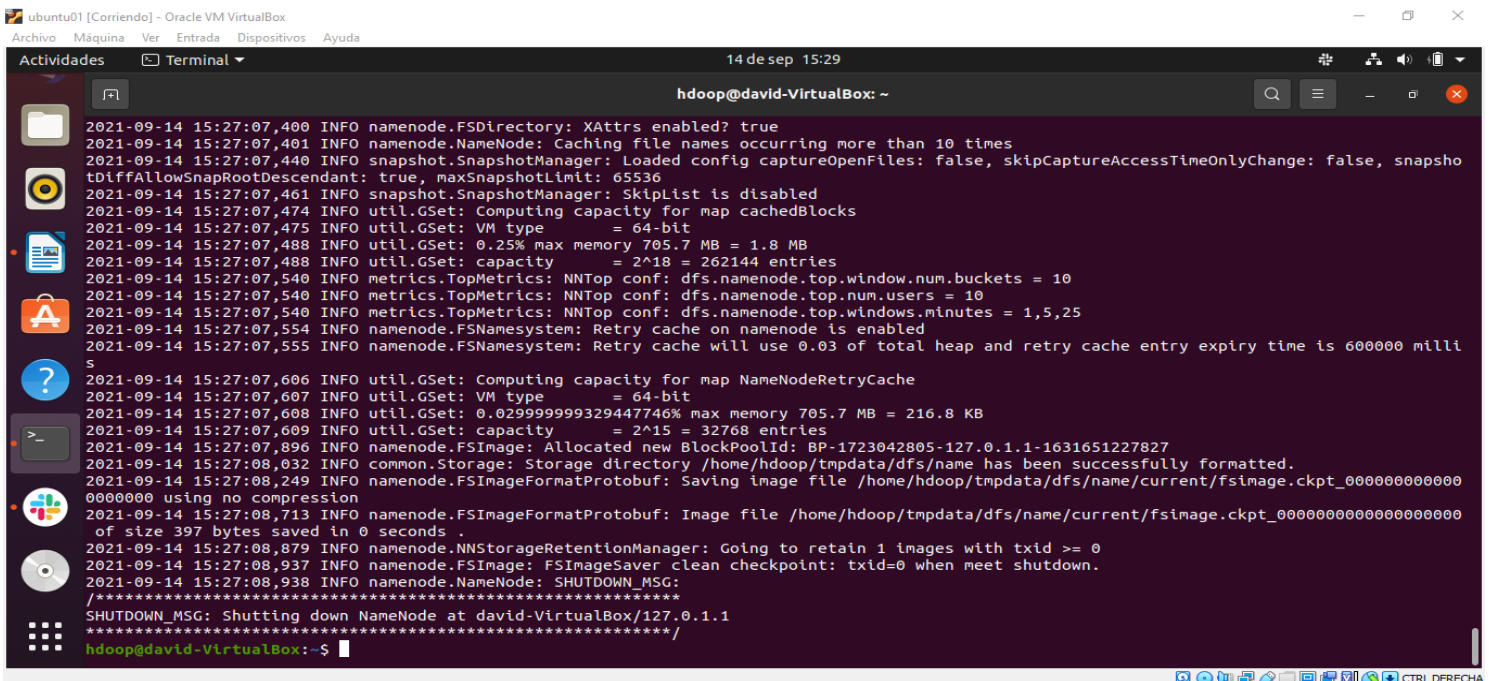
- **Parte 1 – Configuración Ecosistema Hadoop**

Se inicio con la configuración de la maquina virtual mediante la utilización del VirtualBox. Para este caso se utilizo la maquina con nombre “ubuntu01”.



- **Posterior a esto se realizó la configuración del Hadoop**

Finalizando el proceso de configuración de Hadoop. Se procede a realizar el formateo del archivo HDFS y vemos que es la salida esperada es decir todo quedo bien configurado y el Hadoop quedo funcionando de manera correcta.





- Luego de formatear el archivo HDFS. Se realizo la ejecución de los archivos yarn.sh y dfs.sh del entorno Hadoop, para verificar y comprobación de que se ejecutan de manera correcta.

```
ubuntu01 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades Terminal 14 de sep 15:36
hadoop@david-VirtualBox: ~/hadoop-3.2.2/sbin

2021-09-14 15:27:07,609 INFO util.GSet: capacity = 2^15 = 32768 entries
2021-09-14 15:27:07,896 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1723042805-127.0.1.1-1631651227827
2021-09-14 15:27:08,032 INFO common.Storage: Storage directory /home/hadoop/tmpdata/dfs/name has been successfully formatted.
2021-09-14 15:27:08,249 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/tmpdata/dfs/name/current/fsimage.ckpt_00000000000000000000 using no compression
2021-09-14 15:27:08,713 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/tmpdata/dfs/name/current/fsimage.ckpt_00000000000000000000 of size 397 bytes saved in 0 seconds.
2021-09-14 15:27:08,879 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-09-14 15:27:08,937 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-09-14 15:27:08,938 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at david-VirtualBox/127.0.1.1
*****/
hadoop@david-VirtualBox:~/hadoop-3.2.2/bin$ cd ..
hadoop@david-VirtualBox:~/hadoop-3.2.2$ cd /sbin
hadoop@david-VirtualBox:/sbin$ cd ..
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ cd /
-bash: cd: /hadoop-3.2.2/sbin/: No existe el archivo o el directorio
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ exit
cerrar sesión
Hay trabajos detenidos.
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ cd
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [david-VirtualBox]
david-VirtualBox: Warning: Permanently added 'david-virtualbox' (ECDSA) to the list of known hosts.
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ ./start-yarn.sh
```

- Finalmente ejecutamos el comando jps, para verificar la existencia de los archivos que vienen con el Hadoop.

```
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$ jps
27810 DataNode
27972 SecondaryNameNode
26983 ResourceManager
27177 NodeManager
28570 Jps
27693 NameNode
hadoop@david-VirtualBox:~/hadoop-3.2.2/sbin$
```

Access Hadoop UI from Browser

Use your preferred browser and navigate to your localhost URL or IP. The



- Se procede a ejecutar el Localhost:9000 en el entorno web para la ejecución del Hadoop.

Overview 'localhost:9000' (active)

Started:	Tue Sep 14 15:34:40 -0500 2021
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled:	Sun Jan 03 04:26:00 -0500 2021 by hexiaoqiao from branch-3.2.2
Cluster ID:	CID-a1772f26-8568-4e08-b13d-c4ae5ab3f5b5
Block Pool ID:	BP-1723042805-127.0.1.1-1631651227827

Summary

Security is off.
Safemode is off.

- Se procede a ejecutar el puerto localhost en el puerto 9866 para la ejecución de los entornos de Hadoop.

DataNode on david-VirtualBox:9866

Cluster ID:	CID-a1772f26-8568-4e08-b13d-c4ae5ab3f5b5
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-1723042805-127.0.1.1-1631651227827	RUNNING	0s	9 minutes	0 B (64 MB)

Volume Information

David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II



UNIVERSIDAD
EL BOSQUE

- Características del Hadoop en la maquina virtual.

Heap Memory used 31.7 MB of 61.49 MB Heap Memory. Max Heap Memory is 705.69 MB.
Non Heap Memory used 48.53 MB of 49.8 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	28.91 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	10.29 GB
DFS Remaining:	17.13 GB (59.26%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0

- Finalmente, se ejecuto el localhost 8088 para la ejecución de los entornos de Hadoop.

hadoop

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	0	0	0	0	<memory:0, vCores:0>	<memory:8192, vCores:8>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	0

Scheduler Metrics

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

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB
No data available in table													

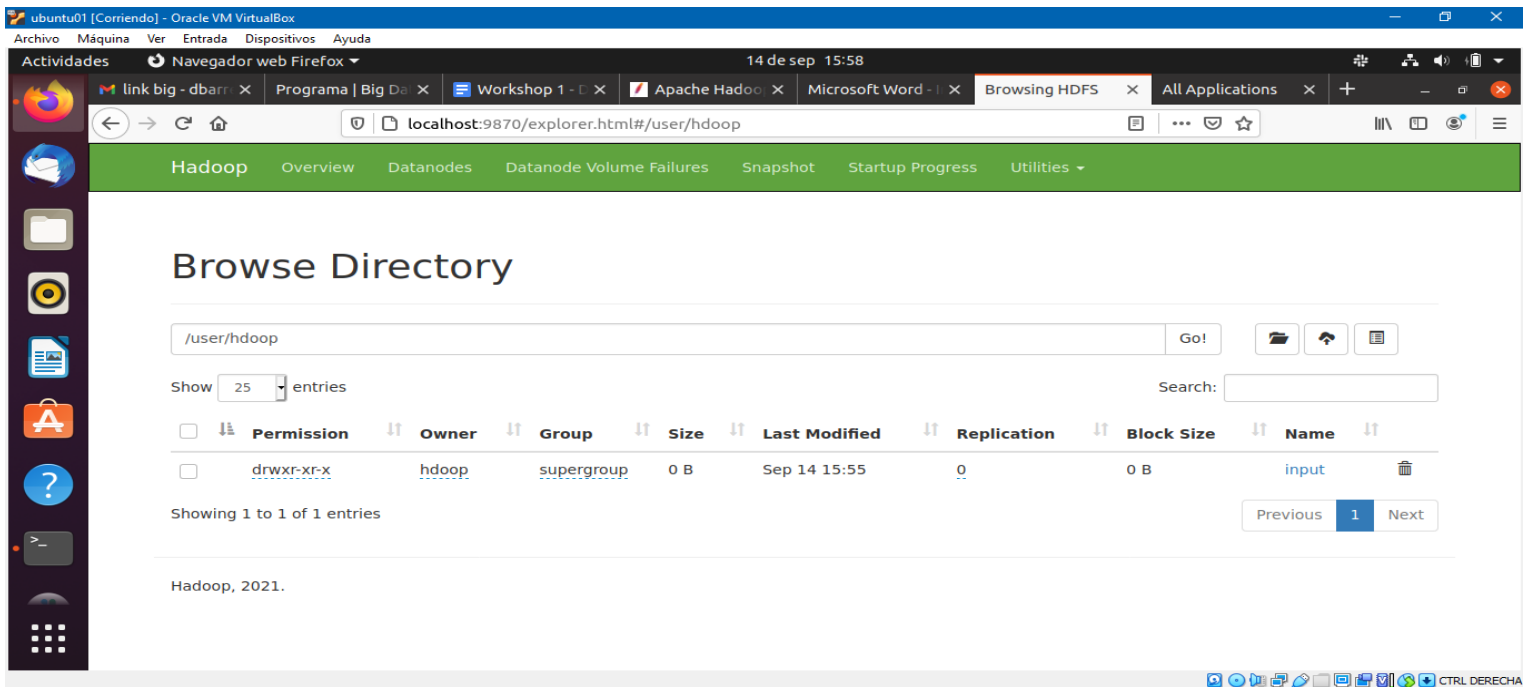
Showing 0 to 0 of 0 entries



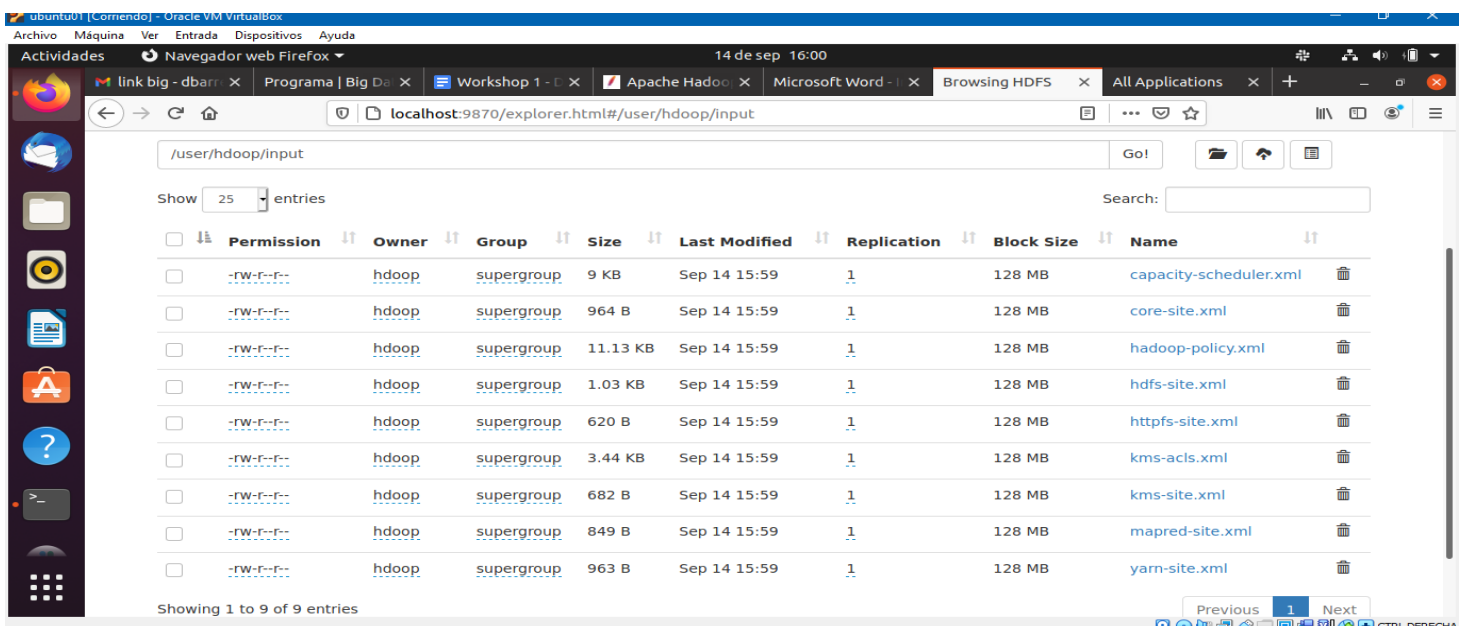
➤ Parte 2 - MapReduce componente de Hadoop

1. Se siguió la guía oficial de Apache, en su parte de “Execution” planteada en el taller.

Se realizó la creación la carpeta de salida llamada input en el usuario de hdoop, proceso que podemos evidenciar en el screenshot.



➤ Verificación de los archivos que se crean en el entorno de Hadoop.





- ¿Qué resultados generó el programa y cuáles son los pasos MapReduce que implementa?

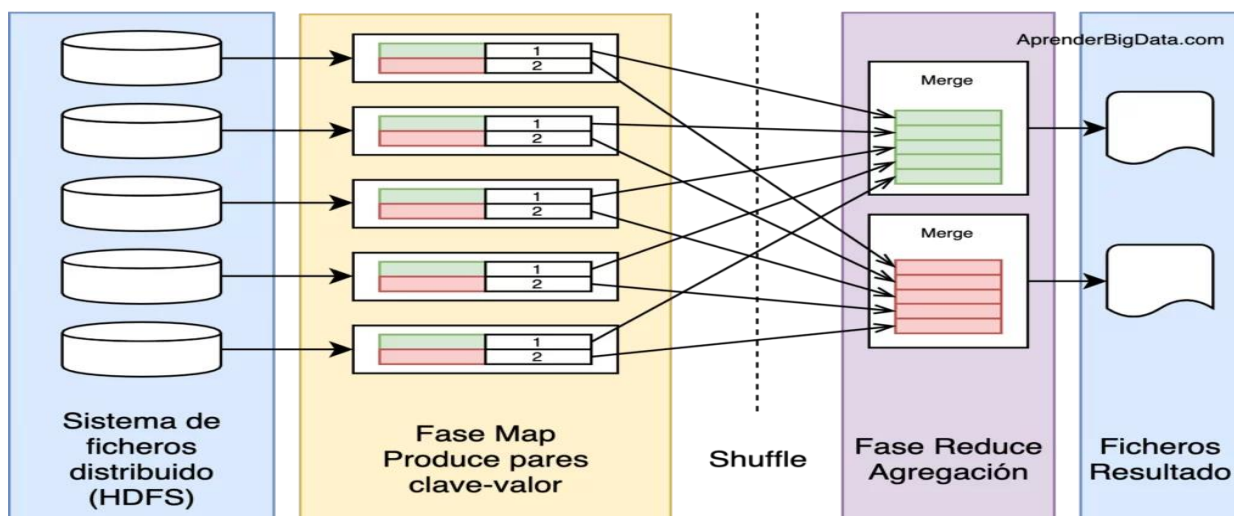
I. **Resultados:** Posterior a esto se procedió a realizar la ejecución del mapReduce con un archivo que trae por defecto Hadoop, procedimiento que siempre tiende a ser un poco demorado y para este se realizó el conteo de las palabras que empiecen con las iniciales “dfs”. Obteniendo los resultados que se ven Screenshot.

```
hadoop@david-VirtualBox: ~/hadoop-3.2.2
Peak Reduce Virtual memory (bytes)=2488565760
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=196
File Output Format Counters
Bytes Written=54
hadoop@david-VirtualBox:~/hadoop-3.2.2$ bin/hdfs dfs -cat output/*
2 dfsdata
2 dfs.data.dir
1 dfsadmin
1 dfs.replication
hadoop@david-VirtualBox:~/hadoop-3.2.2$
```

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Sep 14 16:12	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	54 B	Sep 14 16:12	1	128 MB	part-r-00000

II. Pasos MapReduce:

Para empezar el proceso del MapReduce, se reduce en realizar el Map y el reduce para el procesamiento de datos. Estos subprocesos asociados a la tarea se ejecutan de manera distribuida, en diferentes nodos de procesamiento o esclavos. Siguiendo los siguientes pasos.





- En el trabajo de Hadoop MapReduce, se dividen los datos de entrada en fragmentos independientes que son procesados por los mappers en paralelo. A continuación, se ordenan los resultados del map, que son la entrada para los reducers. Generalmente, las entradas y salidas de los trabajos se almacenan en un sistema de ficheros, siendo los nodos de almacenamiento y de cómputo los mismos. También es muy común que la lógica de la aplicación no se pueda descomponer en una única ejecución de MapReduce, por lo que se encadenan varias de estas fases, tratando los resultados de una como entrada para los mappers de la siguiente fase.
- A. La fase Map se ejecuta en subtarefas llamadas mappers. Estos componentes son los responsables de generar pares clave-valor filtrando, agrupando, ordenando o transformando los datos originales. Los pares de datos intermedios, no se almacenan en HDFS.
 - B. La fase Shuffle (sort) puede no ser necesaria. Es el paso intermedio entre Map y reduce que ayuda a recoger los datos y ordenarlos de manera conveniente para el procesamiento. Con esta fase, se pretende agregar las ocurrencias repetidas en cada uno de los mappers.
 - C. La fase Reduce gestiona la agregación de los valores producidos por todos los mappers del sistema (o por la fase shuffle) de tipo clave-valor en función de su clave. Por último, cada reducer genera su fichero de salida de forma independiente, generalmente escrito en HDFS.

2.

Se realizó el mismo procedimiento del numeral dos, en donde se crea la carpeta receptora para guardar y procesar los datos.

The screenshot shows the Hadoop web interface for the path `/user/hadoop`. The interface includes a navigation bar with tabs like 'Overview', 'Datanodes', and 'Snapshot'. The main content area is titled 'Browse Directory' and shows a table of files and directories. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The files listed are 'input', 'inputPunto2', and 'output'.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	Sep 14 15:59	0	0 B	input
drwxr-xr-x	hadoop	supergroup	0 B	Sep 14 17:34	0	0 B	inputPunto2
drwxr-xr-x	hadoop	supergroup	0 B	Sep 14 16:12	0	0 B	output



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	9 KB	Sep 14 17:34	1	128 MB	capacity-scheduler.xml
-rw-r--r--	hadoop	supergroup	964 B	Sep 14 17:34	1	128 MB	core-site.xml
-rw-r--r--	hadoop	supergroup	11.13 KB	Sep 14 17:34	1	128 MB	hadoop-policy.xml
-rw-r--r--	hadoop	supergroup	1.03 KB	Sep 14 17:34	1	128 MB	hdfs-site.xml
-rw-r--r--	hadoop	supergroup	620 B	Sep 14 17:34	1	128 MB	https-site.xml
-rw-r--r--	hadoop	supergroup	3.44 KB	Sep 14 17:34	1	128 MB	kms-acls.xml
-rw-r--r--	hadoop	supergroup	682 B	Sep 14 17:34	1	128 MB	kms-site.xml
-rw-r--r--	hadoop	supergroup	849 B	Sep 14 17:34	1	128 MB	mapred-site.xml
-rw-r--r--	hadoop	supergroup	963 B	Sep 14 17:34	1	128 MB	yarn-site.xml

- Posterior a esto se hizo la carga del poema de pedro, el cual consta de 220 líneas de texto.

```

GNU nano 4.8 my_wordcount.txt Modificado
y siguiendo el carril de la carreta
un boyero se extingue con la tarde.

Después no quiero más que paz.
Un nido de constructiva paz en cada palma.
Y quizás a propósito del alma
el enjambre de besos y el olvido.

¿Guardar el búfer modificado?
S Si
N No
AC Cancelar
  
```




- Verificación de la creación de manera exitosa en la ubicación requerida para el procesamiento de Hadoop.

The screenshot shows a virtual machine window titled 'ubuntu01 [Corriendo] - Oracle VM VirtualBox'. The main window displays a web browser at 'localhost:9870/explorer.html#/data/input' showing the 'Browse Directory' page for Hadoop. The page lists a file 'my_wordcount.txt' with a size of 7.62 KB, last modified on Sep 15 19:28, and a replication of 1. The file is owned by 'hadoop' and 'supergroup'. The browser also shows a terminal window with the following commands and output:

```
hadoop@david-VirtualBox: ~/hadoop-3.2.2
-bash: cd: /data/input: No existe el archivo o el directorio
hadoop@david-VirtualBox:~/hadoop-3.2.2$ cd bin/hdfs dfs -ls /data/input
-bash: cd: demasiados argumentos
hadoop@david-VirtualBox:~/hadoop-3.2.2$ cd /bin/hdfs df --ls /data/input
-bash: cd: demasiados argumentos
hadoop@david-VirtualBox:~/hadoop-3.2.2$ cd bin/hdfs
hadoop@david-VirtualBox:~/hadoop-3.2.2$ bin/hdfs dfs -ls /data/input
Found 1 items
-rw-r--r-- 1 hadoop supergroup 7801 2021-09-15 19:28 /data/input/my_wordcount.txt
hadoop@david-VirtualBox:~/hadoop-3.2.2$
```

- Finalmente se realizó la ejecución, para el procesamiento del archivo previamente cargado. Cabe resaltar que este procedimiento no culminó con éxito, entendemos que fue por el limitante que se otorga a la máquina virtual, además del tamaño del archivo cargado, porque se dejó trabajando la ejecución por un extenso tiempo, para finalmente no obtener la salida esperada, además de que el procesamiento de Hadoop consume bastante máquina debido a sus procesos.

The screenshot shows a terminal window in the same virtual machine. The user runs the command: `bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.2.jar wordcount /data/input/my_wordcount.txt /data/out/my_wordcount`. The output shows the job submission and execution details, including the job ID 'job_1631750816412_0001'. The job failed with the state 'FAILED' due to 'Application application_1631750816412_0001 failed 2 times due to ApplicationMaster for attempt appattempt_1631750816412_0001_000002 timed out. Failing the application.'



Parte 3 – Configuración ambiente Spark

- Inicialmente se realizó la instalación y configuración del ambiente Spark. Posterior a esto se realizó la ejecución del puerto 127.0.0.1:8080, Obteniendo la respuesta esperada para el entorno de ejecución de Spark.

Spark Master at spark://david-VirtualBox:7077

URL: spark://david-VirtualBox:7077

Alive Workers: 1

Cores in use: 1 Total, 0 Used

Memory in use: 1894.0 MiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20210914165116-10.0.2.15-45843	10.0.2.15:45843	ALIVE	1 (0 Used)	1894.0 MiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

- Se procedió a realizar la comprobación del correcto funcionamiento del ambiente Spark, para lenguajes como scala y Python, obteniendo las respuestas esperadas.

```
scala> :q
david@david-VirtualBox:~$ pyspark
Python 3.8.5 (default, Jul 28 2020, 12:59:40)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
21/09/14 16:54:40 WARN Utils: Your hostname, david-VirtualBox resolves to a loopback address: 127.0.0.1; using 10.0.2.15 instead (on interface enp0s3)
21/09/14 16:54:40 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

scala>
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1631656486587).
SparkSession available as 'spark'.
```



```
ubuntui01 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades Terminal 14 de sep 17:09
david@david-VirtualBox: ~/words

[1]+ Detenido pyspark
david@david-VirtualBox:~$ pyspark
Python 3.8.5 (default, Jul 28 2020, 12:59:40)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
21/09/14 17:03:16 WARN Utils: Your hostname, david-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
21/09/14 17:03:16 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
21/09/14 17:03:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

      ____      __
     / __ )____/ /  __
    / __ /_____/ /  / /
   /_/ /_____/ /  /_/
  /___/_____/ /  /___/
 version 3.1.2

Using Python version 3.8.5 (default, Jul 28 2020 12:59:40)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1631657015109).
SparkSession available as 'spark'.
>>> text_file = sc.textFile("poemaPedro.txt")
>>> counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> counts.saveAsTextFile("words")
>>> quit()
david@david-VirtualBox:~$ cd words/
david@david-VirtualBox:~/words$ ls
part-000000 _SUCCESS
david@david-VirtualBox:~/words$
```

- Finalmente se realizó el conteo de palabras del archivo que spark python genera de respuesta al archivo cargado, y se obtuvo el conteo de palabras del poema.

```
ubuntui01 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades Terminal 14 de sep 17:11
david@david-VirtualBox: ~/words

Spark context available as 'sc' (master = local[*], app id = local-1631657015109).
SparkSession available as 'spark'.
>>> text_file = sc.textFile("poemaPedro.txt")
>>> counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> counts.saveAsTextFile("words")
>>> quit()
david@david-VirtualBox:~$ cd words/
david@david-VirtualBox:~/words$ ls
part-000000 _SUCCESS
david@david-VirtualBox:~/words$ cat part-000000
cat: part-000000: No existe el archivo o el directorio
david@david-VirtualBox:~/words$ cat part-000000
('HAY', 1)
('UN', 1)
('PAIS', 1)
('EN', 1)
('EL', 1)
('MUNDO', 1)
(' ', 20)
('Hay', 3)
('un', 10)
('pais', 6)
('en', 39)
('el', 45)
('mundo', 3)
('colocado', 1)
('mismo', 1)
('trayecto', 1)
('del', 40)
('sol', 1)
('Ortundo', 1)
('de', 58)
('anoche', 1)
```

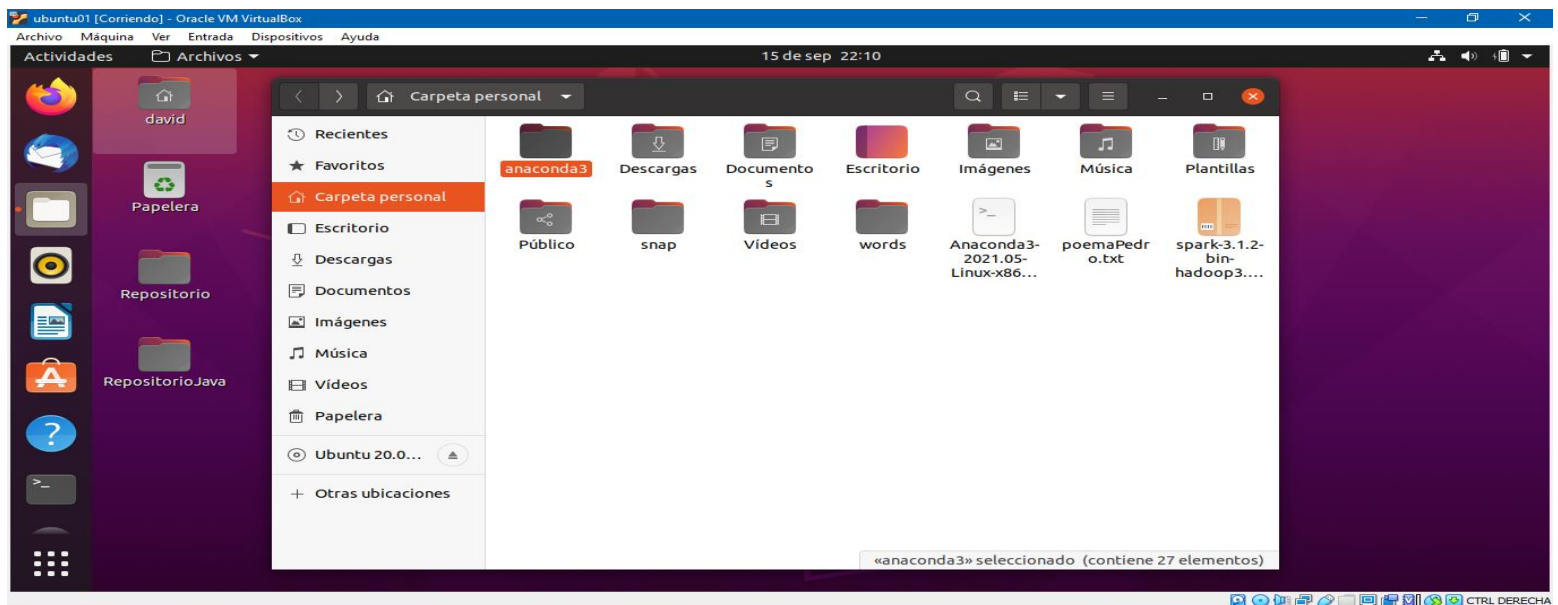
- Cabe resaltar que el procesamiento y análisis de datos de Spark es mucho más rápido a comparación con el mapReduce, que, por su procedimiento se demora en términos de tiempos y consumo de máquina más. Esto también



se debió al fuerte de procesamiento de datos que Python realiza en sus algoritmos y métodos.

```
('llenas', 1)
('abogados', 1)
('placas', 1)
('silencio', 1)
('poetas', 1)
('nieblas', 1)
('silencio', 1)
('jueces', 1)
('silenciosos.', 1)
('Sube', 1)
('salta', 1)
('delira', 1)
('esquinas', 1)
('resuelve', 1)
('dólar', 1)
('inminente.', 1)
('¡Un', 1)
('dólar!', 1)
('He', 1)
('aquí', 1)
('Un', 2)
('borbotón', 1)
('sangre.', 2)
('Silenciosa', 1)
('terminante.', 1)
('Sangre', 2)
('viento', 1)
('efectivo', 1)
('amargura.', 1)
('merece', 1)
('nombre', 1)
('Sino', 1)
('tumba', 1)
```

- **Parte 4 (IDE) Anaconda**
- Se realizó la instalación del IDE de anaconda, con ello se habilita el ambiente de desarrollo proporcionado por este IDE, el cual incluye varias aplicaciones que permiten trabajar diferentes lenguajes y frameworks de desarrollo.

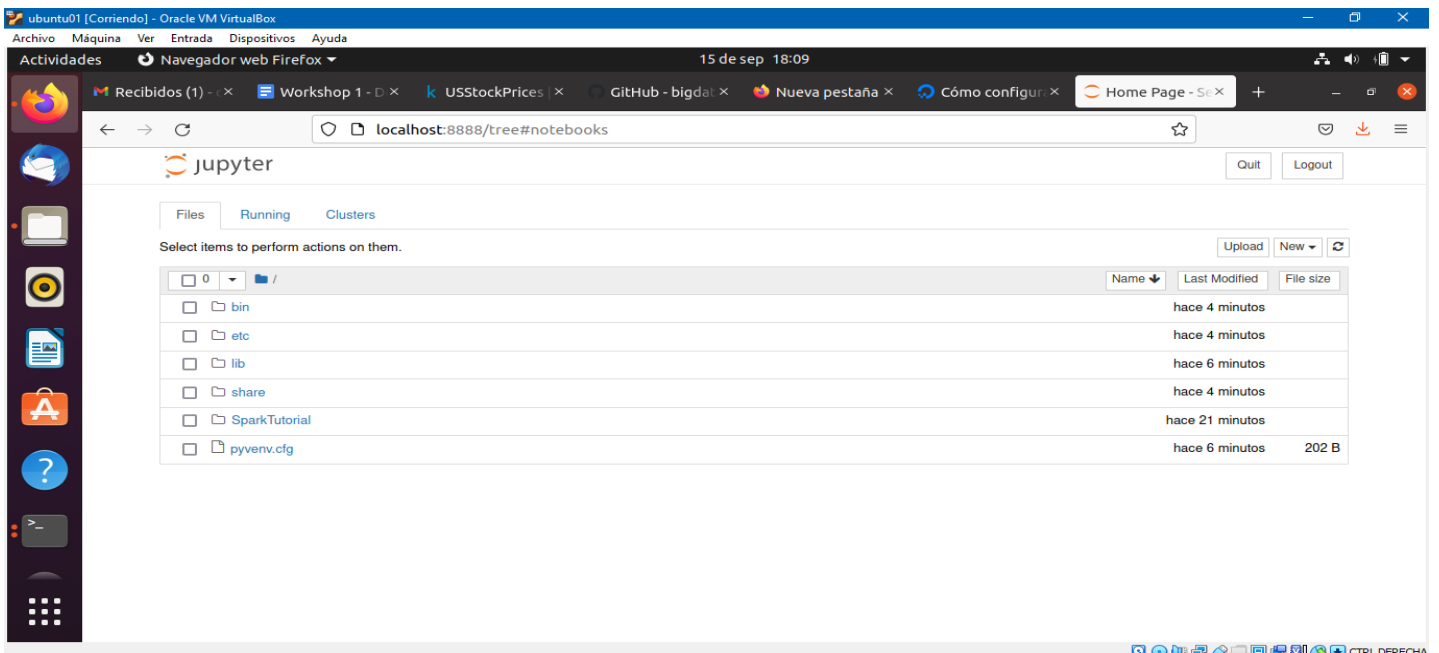




- Posterior a la configuración de anaconda se realizó la configuración del entorno virtual para realizar la conexión entre jupyter y VirtualBox, ya que sin esto no es posible la conexión, en el Screenshot se puede observar la correcta y exitosa conexión.

```
20 pygments-2.10.0 pyparsing-2.4.7 pyrsistent-0.18.0 python-dateutil-2.8.2 pyzmq-22.2.1 qtconsole-5.1.1 qtpy-1.11.1 six-1.16.0 terminado-0.12
.1 testpath-0.5.0 tornado-6.1 trattlets-5.1.0 wcwidth-0.2.5 webencodings-0.5.1 widgetsnbextension-3.5.1
WARNING: You are using pip version 21.2.3; however, version 21.2.4 is available.
You should consider upgrading via the '/home/david/Escritorio/Repositorio/bin/python -m pip install --upgrade pip' command.
(Repositorio) david@david-VirtualBox:~/Escritorio/Repositorio$ cd /home/david/Escritorio/Repositorio/
(Repositorio) david@david-VirtualBox:~/Escritorio/Repositorio$ jupyter lab
Traceback (most recent call last):
  File "/home/david/Escritorio/Repositorio/bin/jupyter", line 8, in <module>
    sys.exit(main())
  File "/home/david/Escritorio/Repositorio/lib/python3.8/site-packages/jupyter_core/command.py", line 285, in main
    command = _jupyter_abspath(subcommand)
  File "/home/david/Escritorio/Repositorio/lib/python3.8/site-packages/jupyter_core/command.py", line 124, in _jupyter_abspath
    raise Exception(
Exception: Jupyter command 'jupyter-lab' not found.
(Repositorio) david@david-VirtualBox:~/Escritorio/Repositorio$ jupyter notebook
[I 18:06:24.852 NotebookApp] Writing notebook server cookie secret to /home/david/.local/share/jupyter/runtime/notebook_cookie_secret
[I 18:06:25.626 NotebookApp] Serving notebooks from local directory: /home/david/Escritorio/Repositorio
[I 18:06:25.627 NotebookApp] Jupyter Notebook 6.4.0 is running at:
[I 18:06:25.627 NotebookApp] http://localhost:8888/?token=361647be99ca81eeae868d6d9692dec2a3ce4d8353874f86
[I 18:06:25.627 NotebookApp] or http://127.0.0.1:8888/?token=361647be99ca81eeae868d6d9692dec2a3ce4d8353874f86
[I 18:06:25.627 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 18:06:25.896 NotebookApp]

To access the notebook, open this file in a browser:
    file:///home/david/.local/share/jupyter/runtime/nbserver-11566-open.html
Or copy and paste one of these URLs:
    http://localhost:8888/?token=361647be99ca81eeae868d6d9692dec2a3ce4d8353874f86
    or http://127.0.0.1:8888/?token=361647be99ca81eeae868d6d9692dec2a3ce4d8353874f86
/usr/lib/python3.8/json/encoder.py:257: UserWarning: date_default is deprecated since jupyter_client 7.0.0. Use jupyter_client.jsonutil.json_
default.
    return _iterencode(o, 0)
```





- Posterior a esto podemos ver la visualización de los archivos del repositorio previamente clonados del Git.

ubuntu01 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda

Actividades Navegador web Firefox 15 de sep 18:11

Recibidos (1) Workshop 1 - USStockPrices GitHub - bigda Nueva pestaña SparkTutorial/ pyspark-dat x pyspark-basi x

localhost:8888/notebooks/SparkTutorial/pyspark-data-analysis.ipynb

jupyter pyspark-data-analysis (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

PySpark data analysis

Based on [this post](#).

```
In [ ]: import matplotlib.pyplot as plt
      %matplotlib inline

In [ ]: from pyspark.sql import SparkSession
      from pyspark.sql.types import *
      from pyspark.sql.functions import col, lit, countDistinct

In [ ]: spark = SparkSession.builder\
      .master("spark://localhost:7078")\
      .appName("pyspark-data-analysis")\
      .getOrCreate()
```

Loading and analysing data structure

ubuntu01 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda

Actividades Navegador web Firefox 15 de sep 18:12

Recibidos (1) Workshop 1 - USStockPrices GitHub - bigda Nueva pestaña SparkTutorial/ pyspark-dat x pyspark-basi x

localhost:8888/notebooks/SparkTutorial/pyspark-basics.ipynb

jupyter pyspark-basics (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

PySpark basics

Based on [this post](#).

Data can be downloaded from [here](#).

```
In [ ]: !pip install pyspark

In [ ]: from pyspark import SparkContext
      from pyspark.sql import SparkSession
```

Connecting to Spark cluster

```
In [ ]: sc = SparkContext("spark://localhost:7078", appName = "pyspark-basics")

In [ ]: spark = SparkSession.builder\
      .master("spark://localhost:7078")\
      .appName("pyspark-basics")\
      .getOrCreate()
```




Se realizó la ejecución de los scripts:

1) ***spark-basics.ipynb***

Donde inicial se realiza la importación de la librería de pyspark la cual nos permite realizar procesamiento de datos de manera más analítica y eficaz. Teniendo como ventajas principales:

- Procesamiento en memoria de los resultados parciales.
- Soporte para múltiples lenguajes.
- Tolerancia a fallos implícita.
- 100% Open Source.

2) ***spark-data-analysis.ipynb***.

En este ejercicio se diferencia con el primero en cuanto, a que se realiza un procesamiento de graficas y datos con ayuda de la librería de matplotlib, esto ayuda a entender mejor los datos, que sirve de complemento al procesamiento de la data con pyspark.

- Hasta 100 veces más rápido que Hadoop MapReduce.

