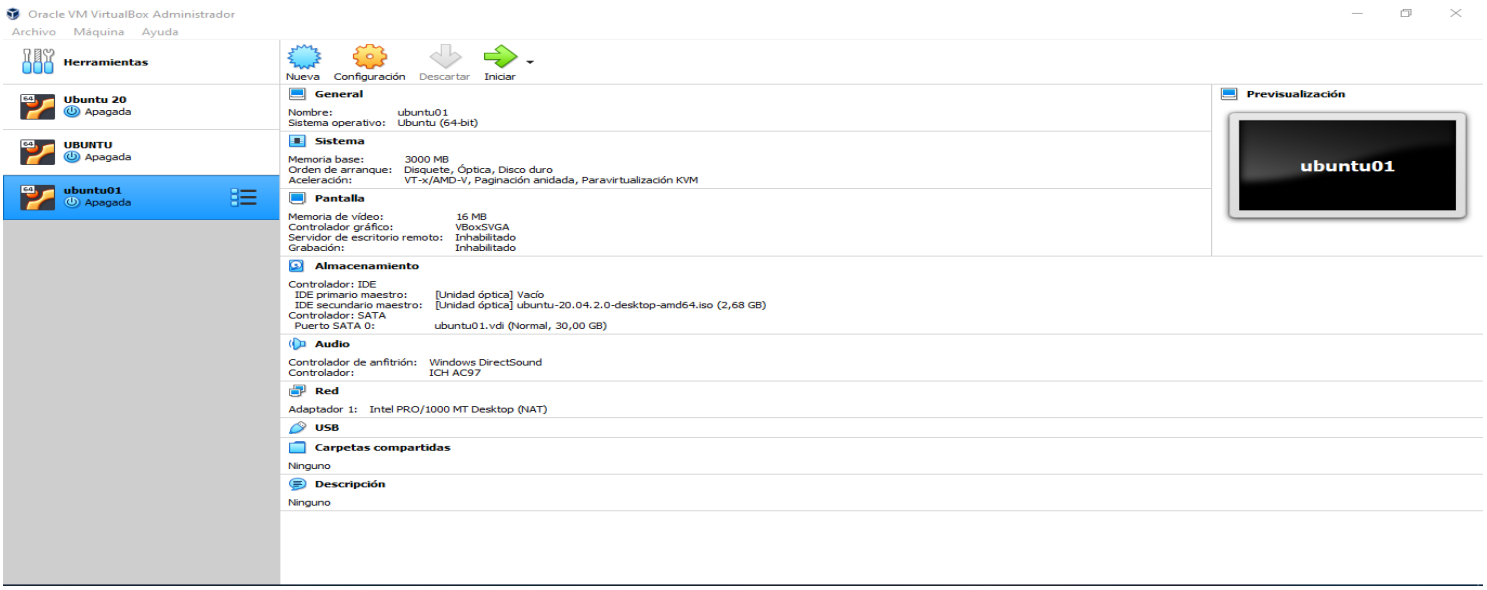
David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

Informe Taller 1 – Ecosistema de Hadoop

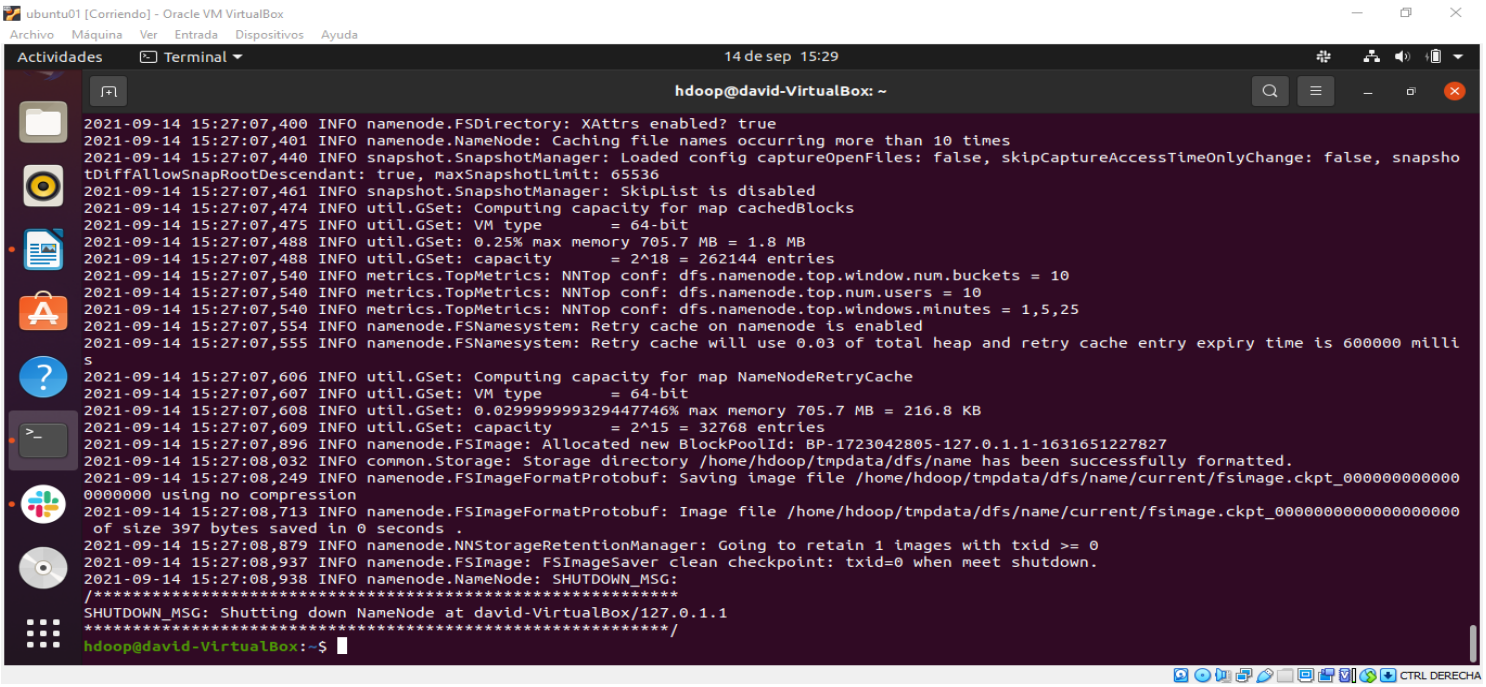
- ***Parte 1 – Configuración Ecosistema Hadoop***

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



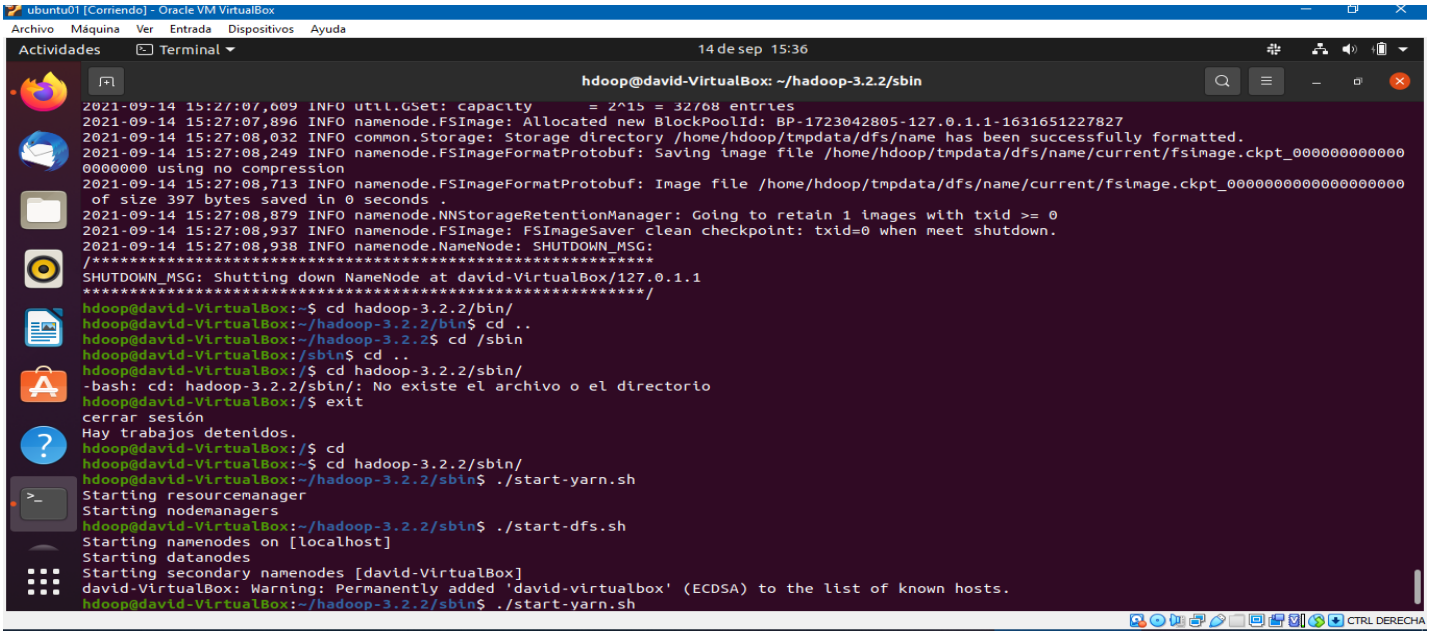
- ***Posterior a esto se realizó la configuración del Hadoop***

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.



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

- 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.

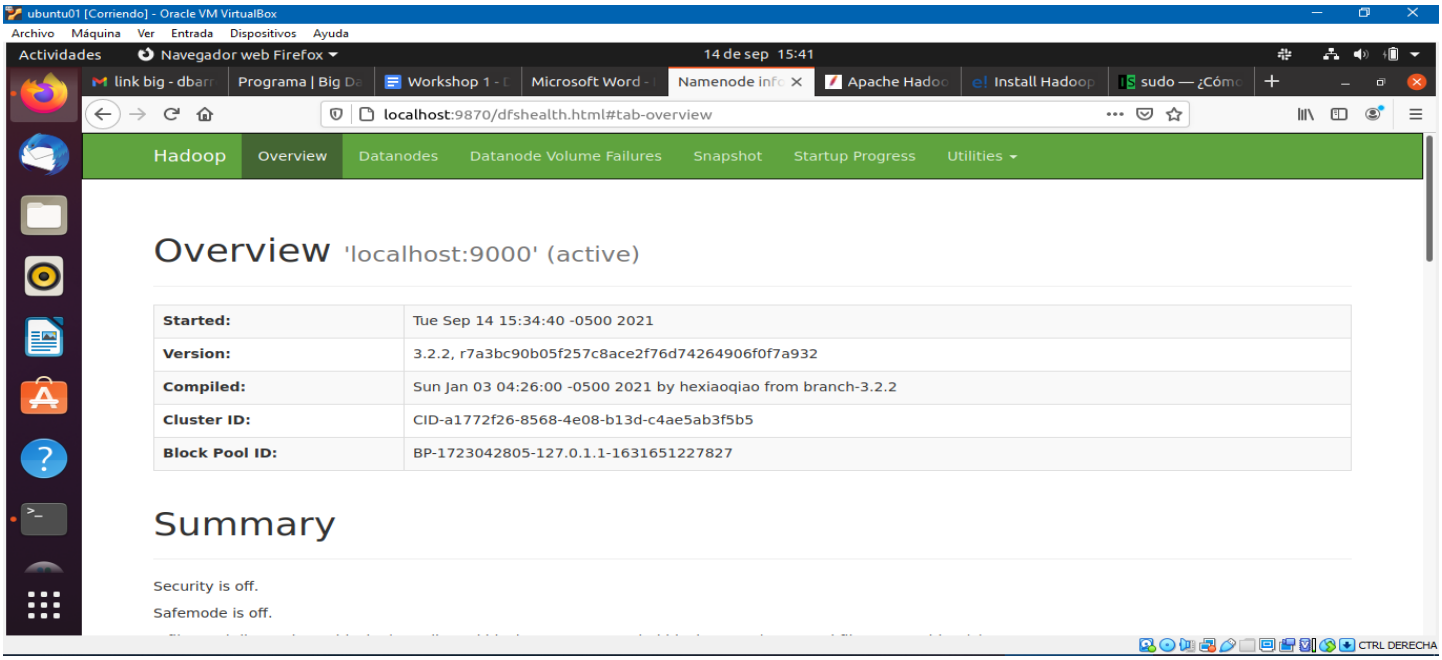


- Finalmente ejecutamos el comando jps, para verificar la existencia de los archivos que vienen con el Hadoop.

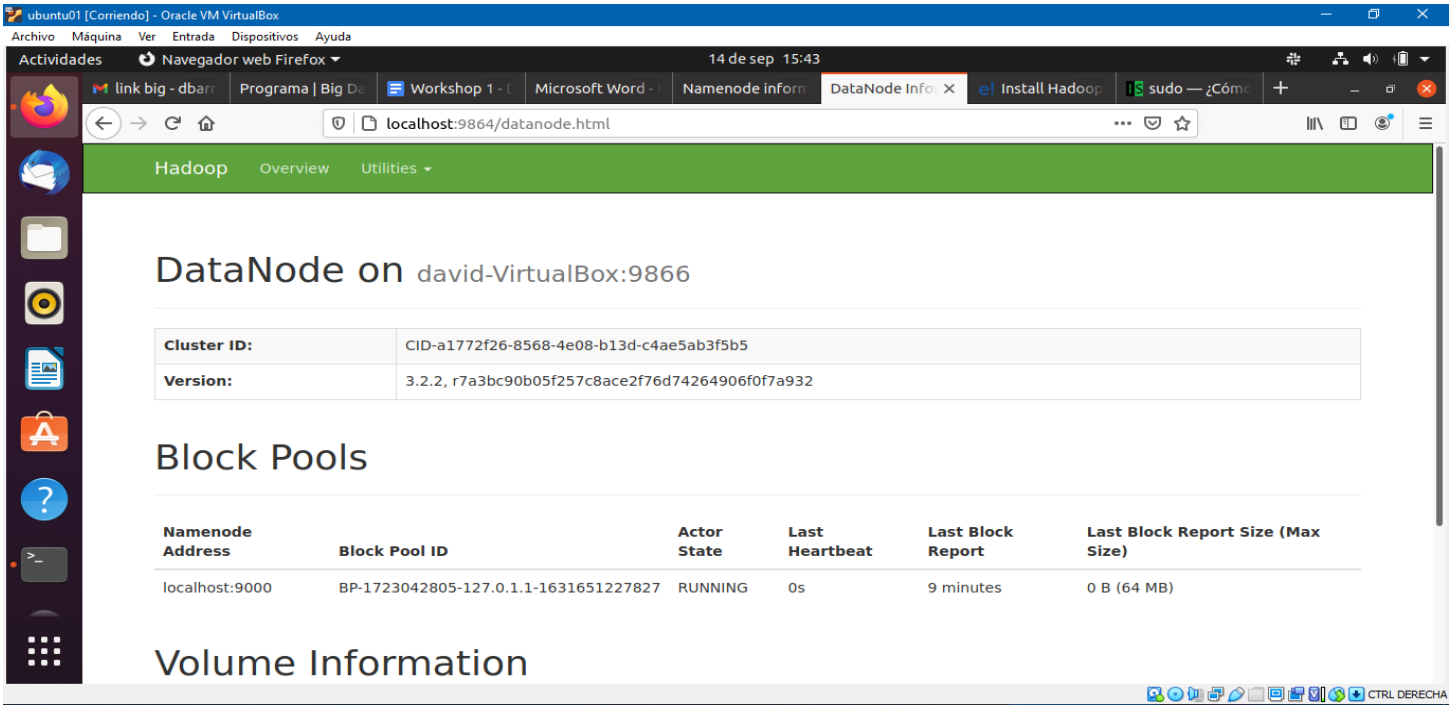
David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

UNIVERSIDAD EL BOSQUE

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

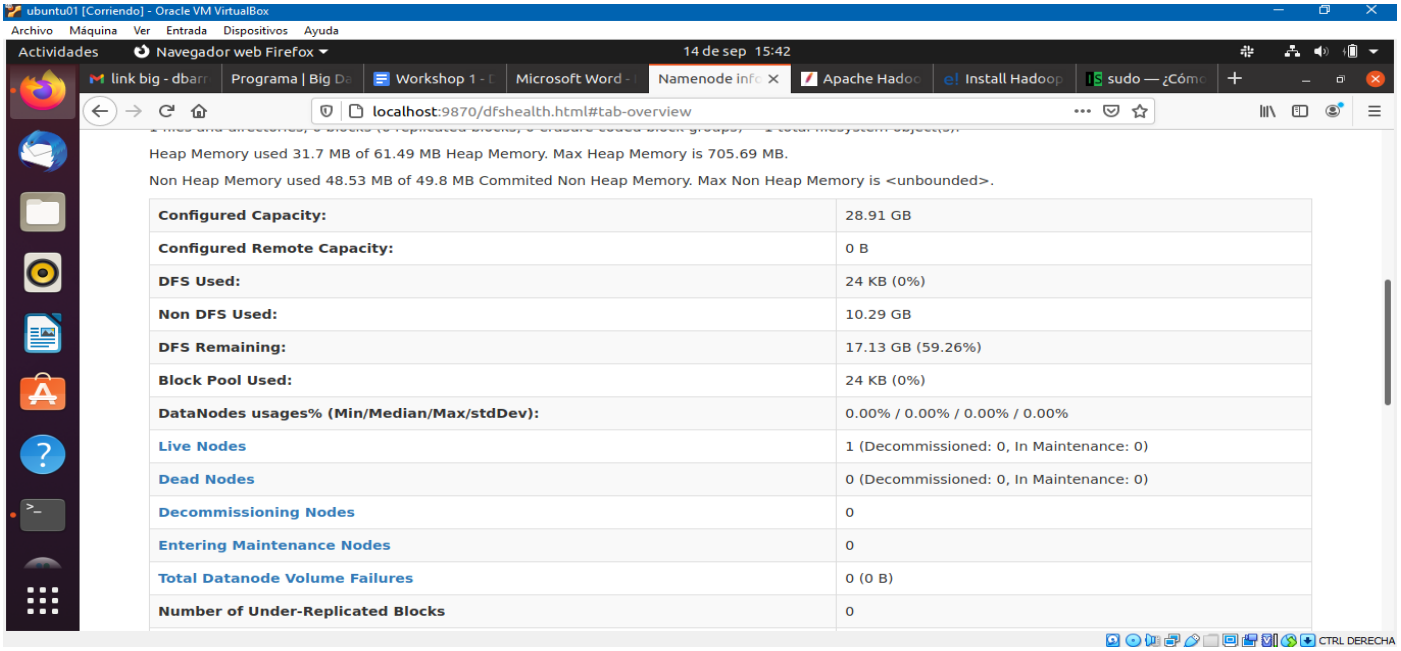


- Se procede a ejecutar el puerto localhost en el puerto 9866 para la ejecución de los entornos de Hadoop.

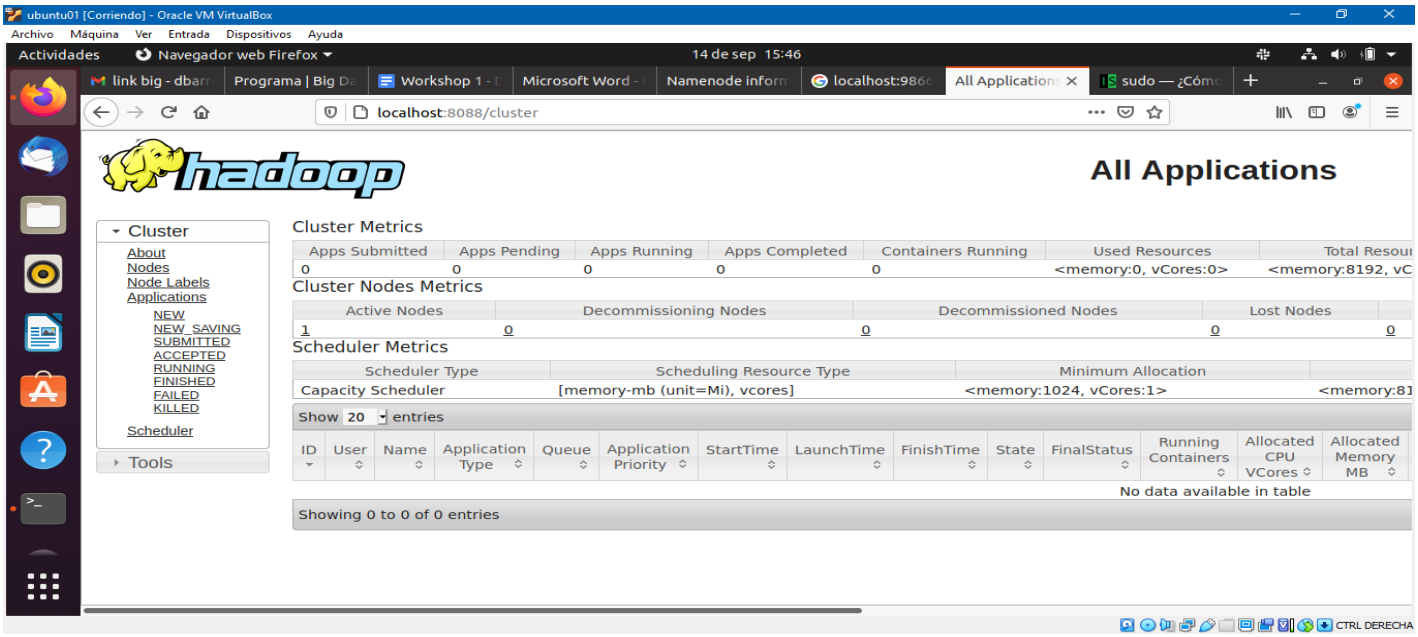


David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

- Características del Hadoop en la maquina virtual.



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

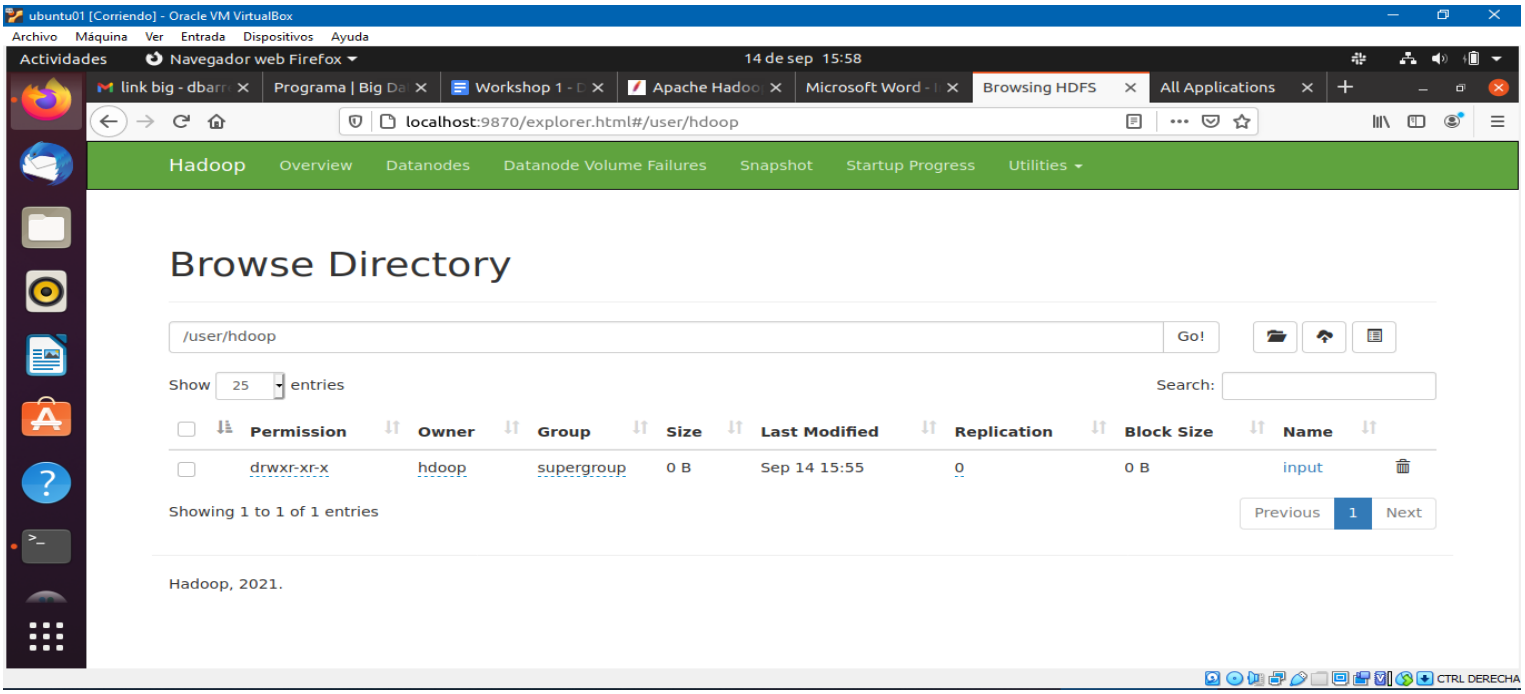


David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

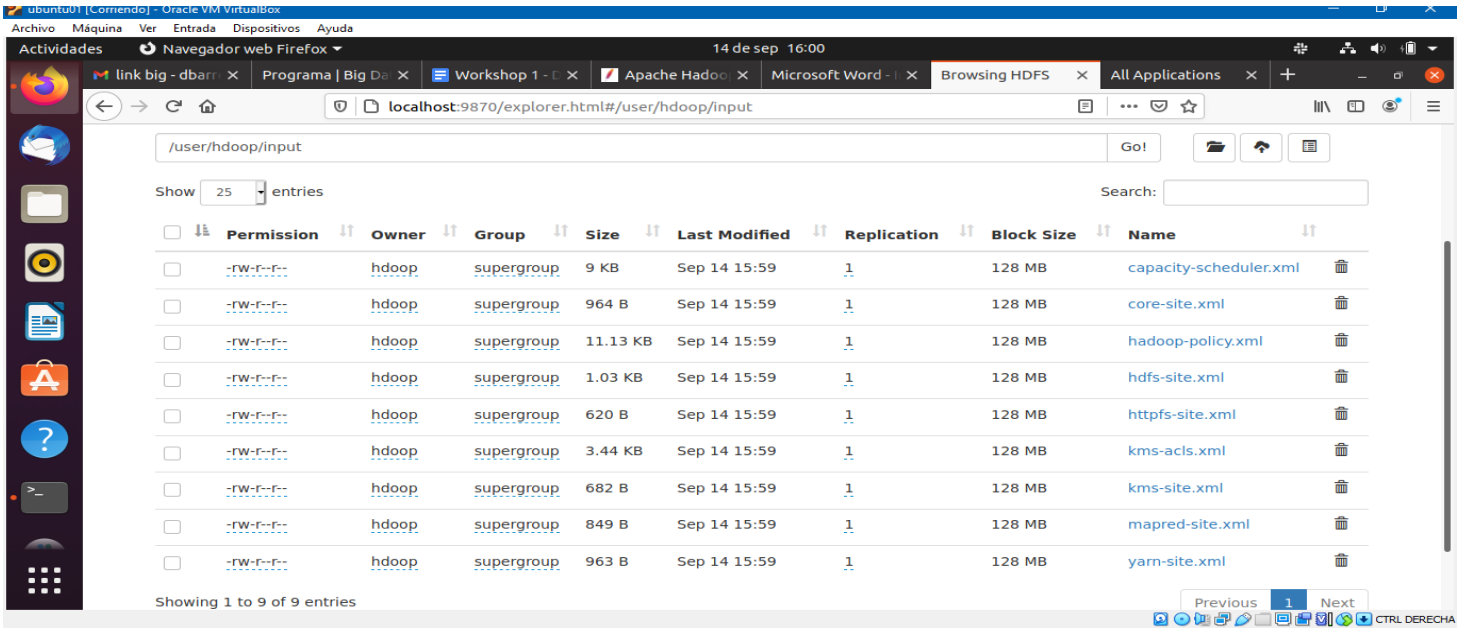
➢ Parte 2 - MapReduce componente de Hadoop

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

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



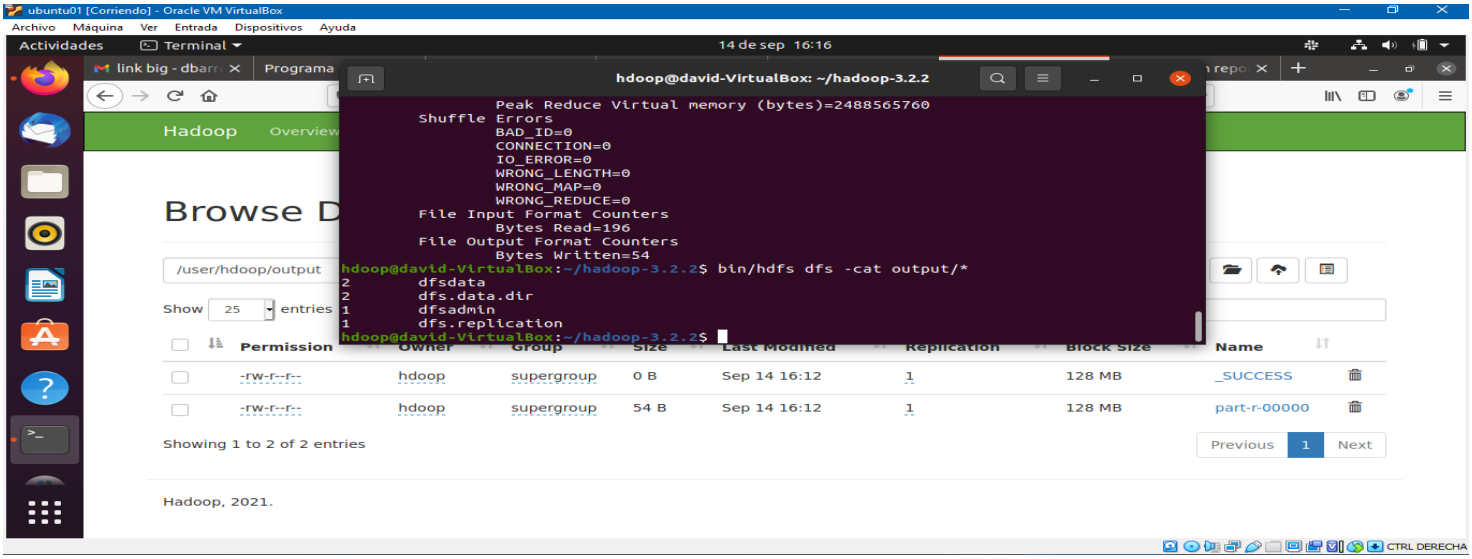
➢ Verificación de los archivos que se crean en el entorno de Hadoop.



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

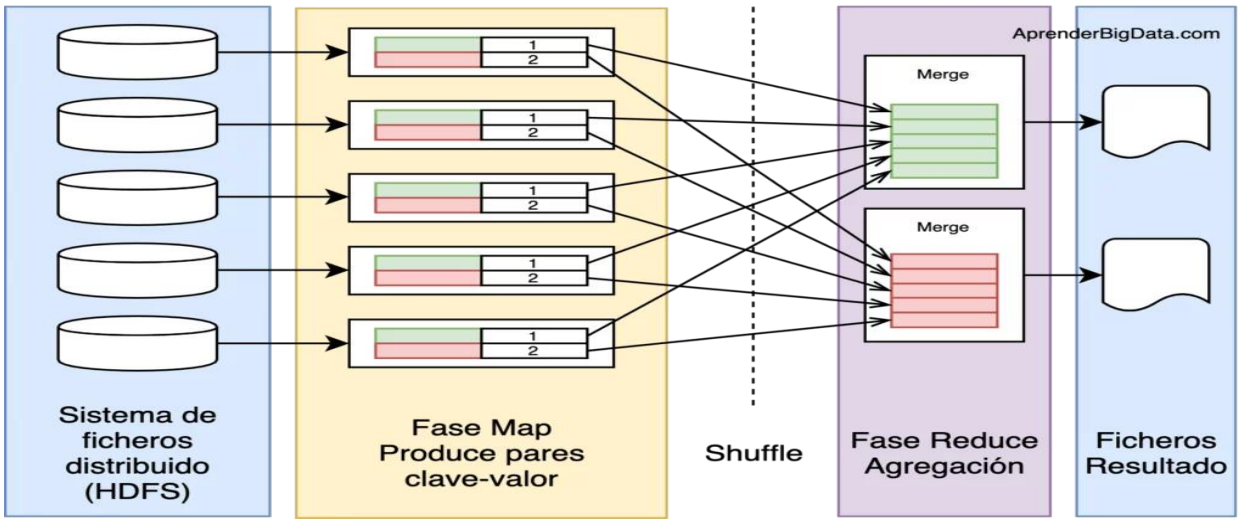
- ¿Qué resultados generó el programa y cuáles son los pasos MapReduce que implementa?

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.



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.



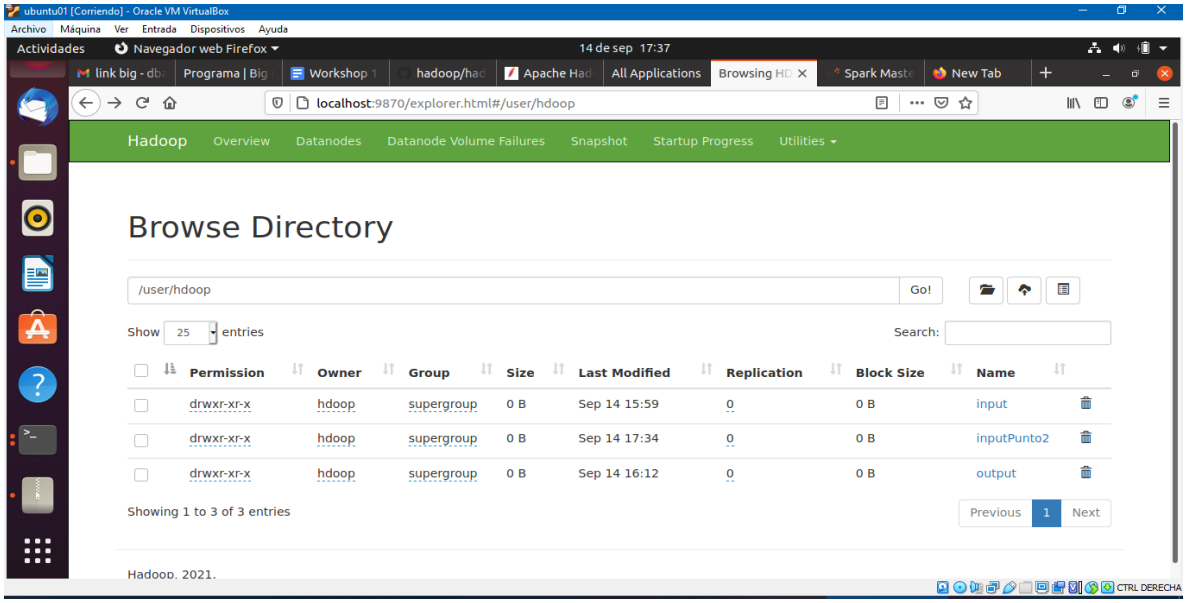
David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

- 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.

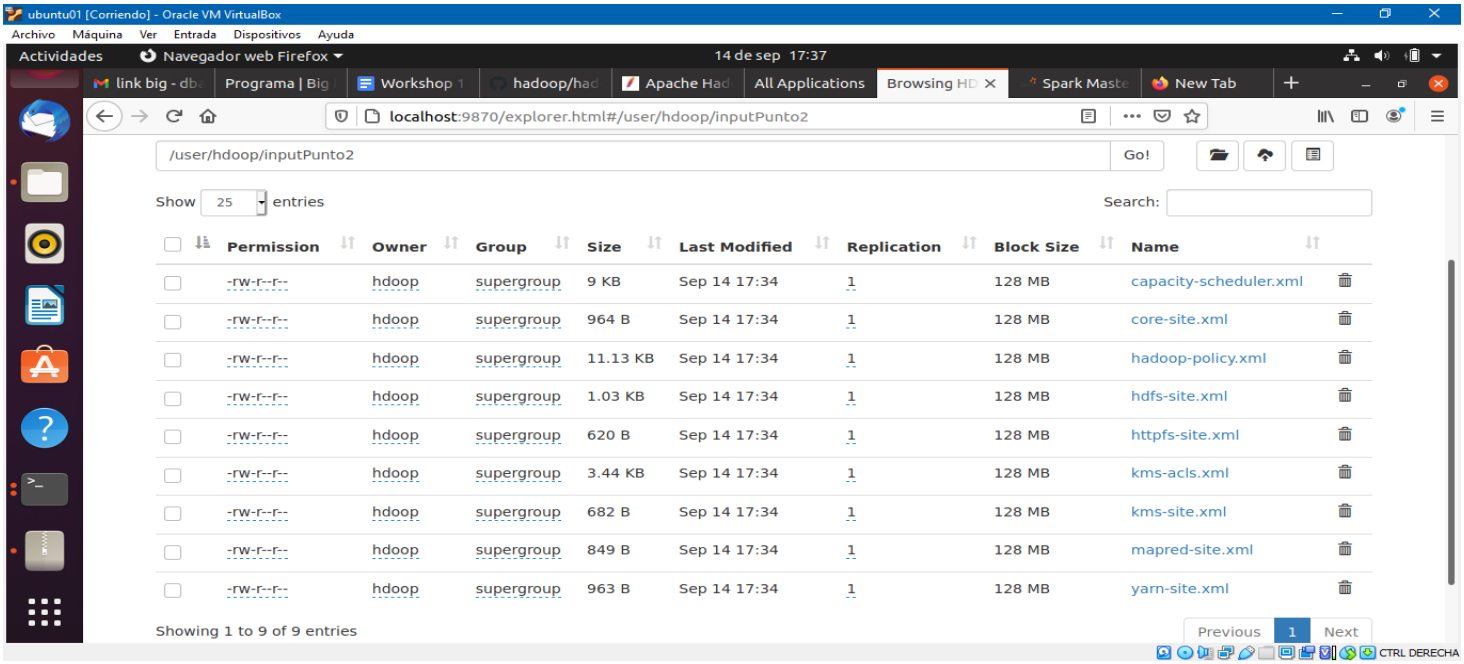
A. La fase Map se ejecuta en subtareas llamadas mappers. Estos componentes son los responsables de generar pares clave-valor filtrando, agrupando, ordenando o transformando los datos originales. Los pares de datos intermedios, no se almacenan en HDFS.
B. La fase Shuffle (sort) puede no ser necesaria. 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.

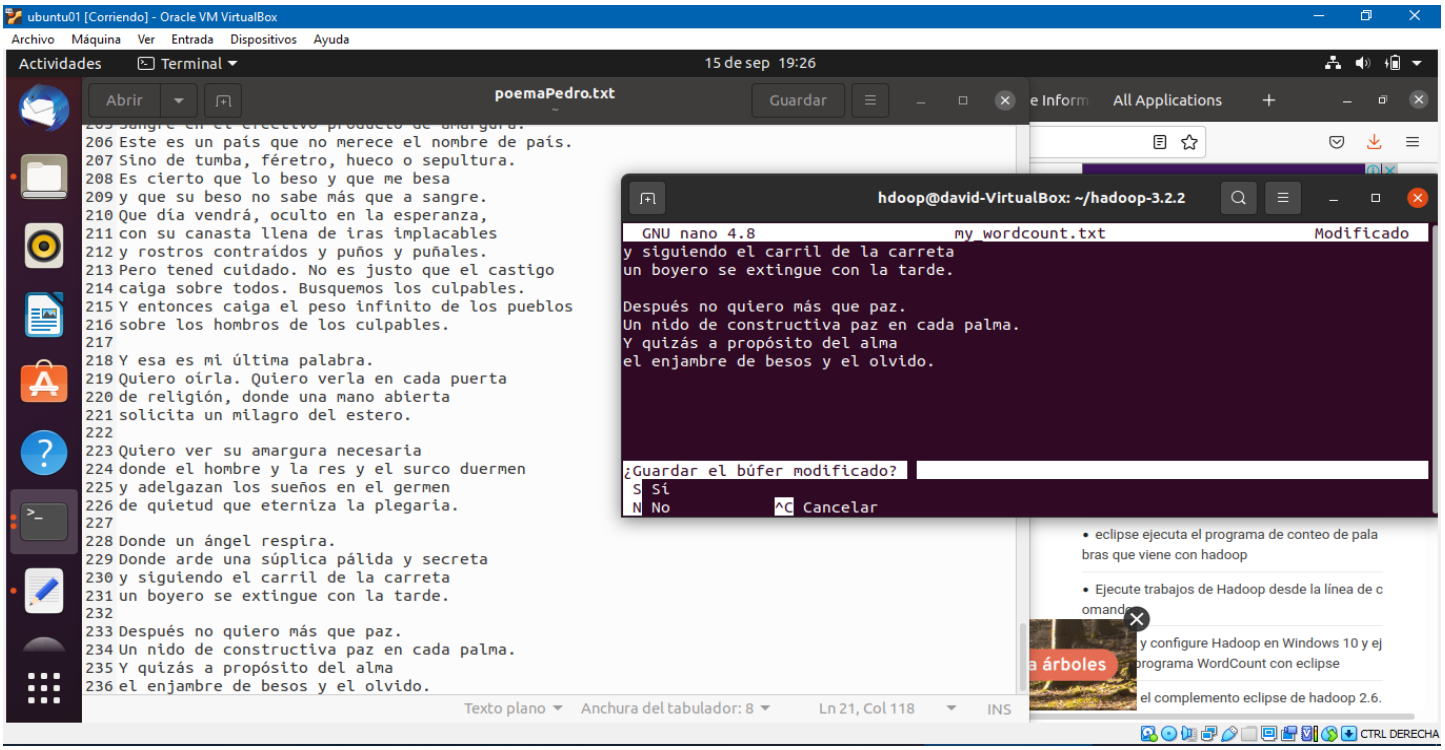
Se realizo el mismo procedimiento del numeral dos, en donde se crea la carpeta receptora para guardar y procesar los datos.



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

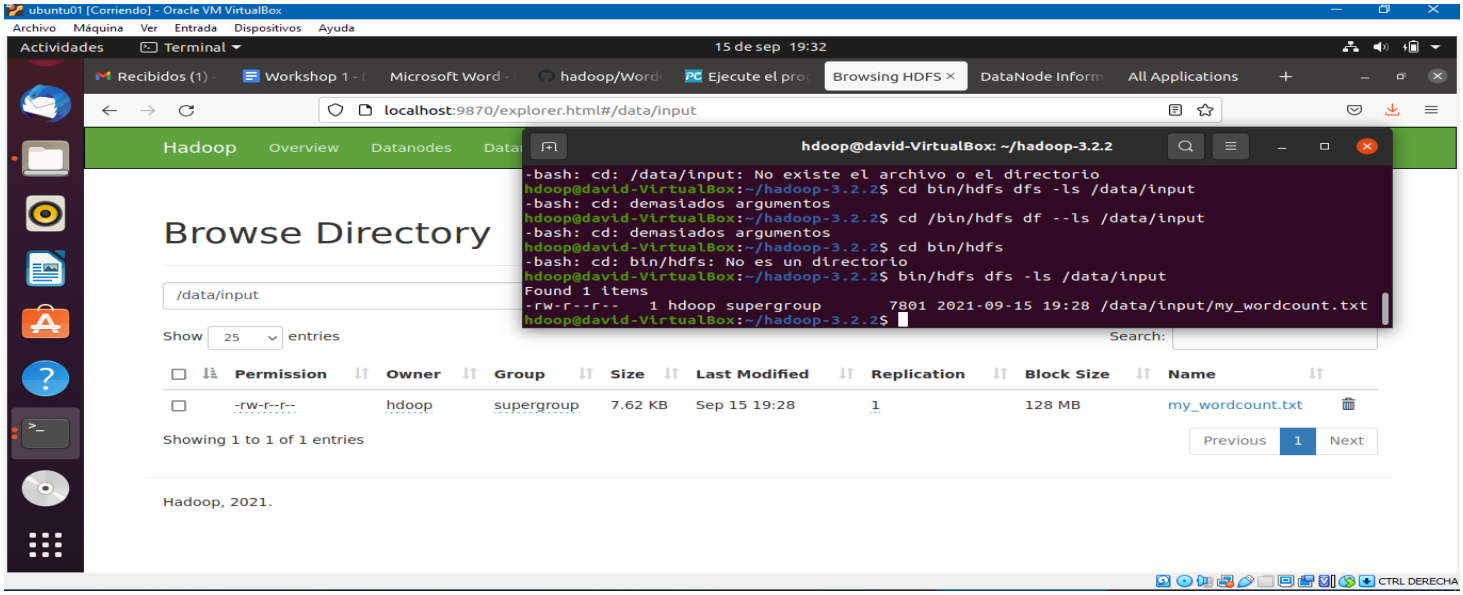


- Posterior a esto se hizo la carga del poema de pedro, el cual consta de 220 líneas de texto.

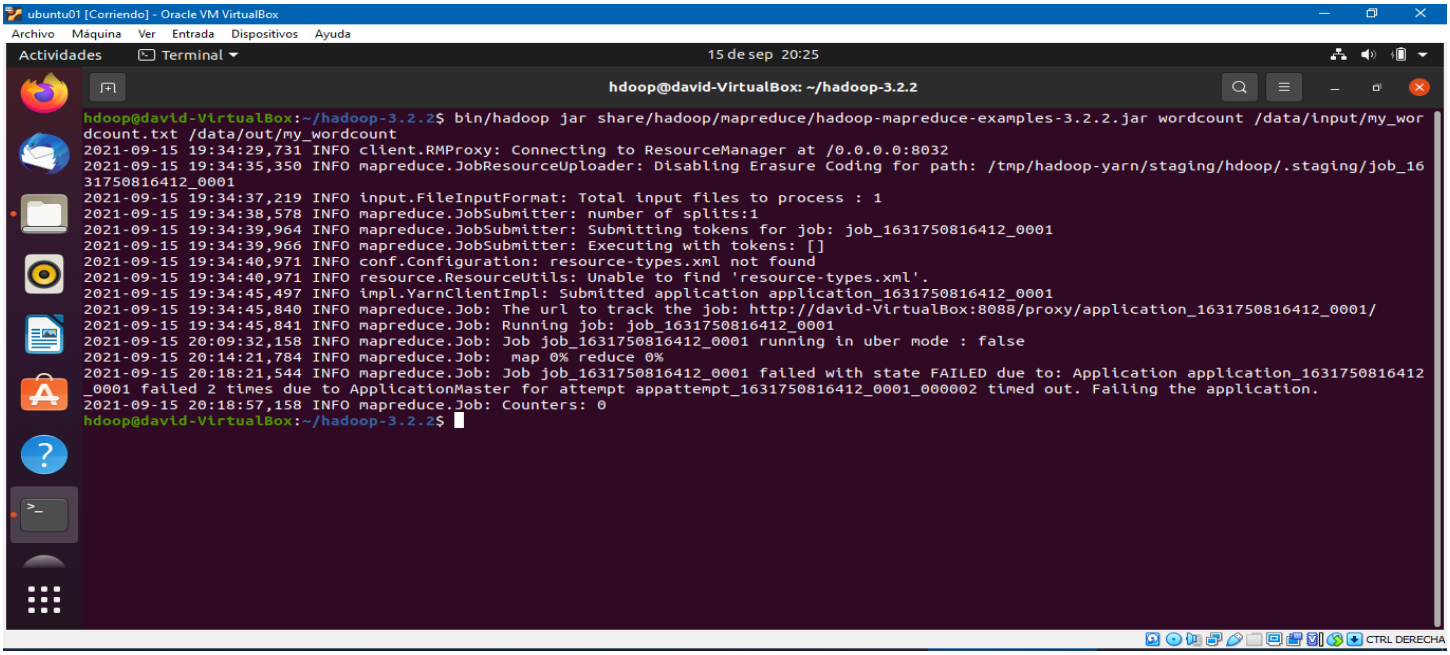


David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

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



- Finalmente se realizó la ejecución, para el procesamiento del archivo previamente cargado. Cabe resaltar que este procedimiento no culmino 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 maquina debido a sus procesos.

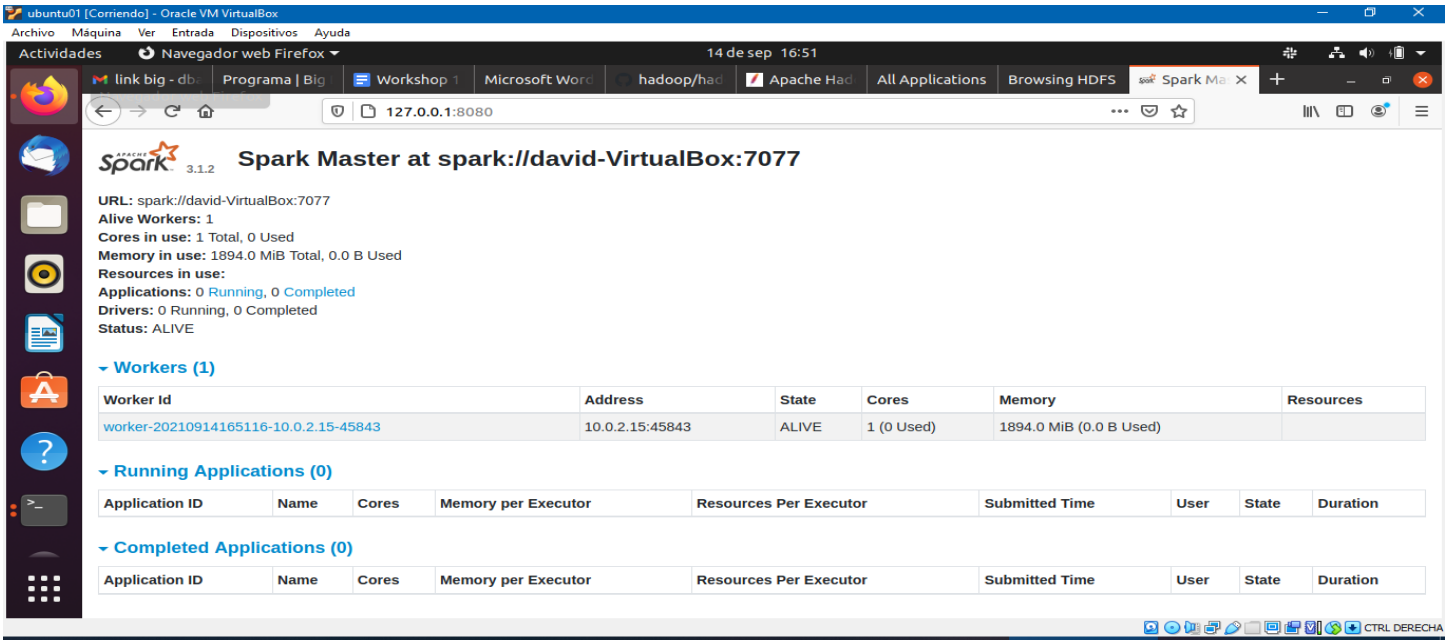


David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

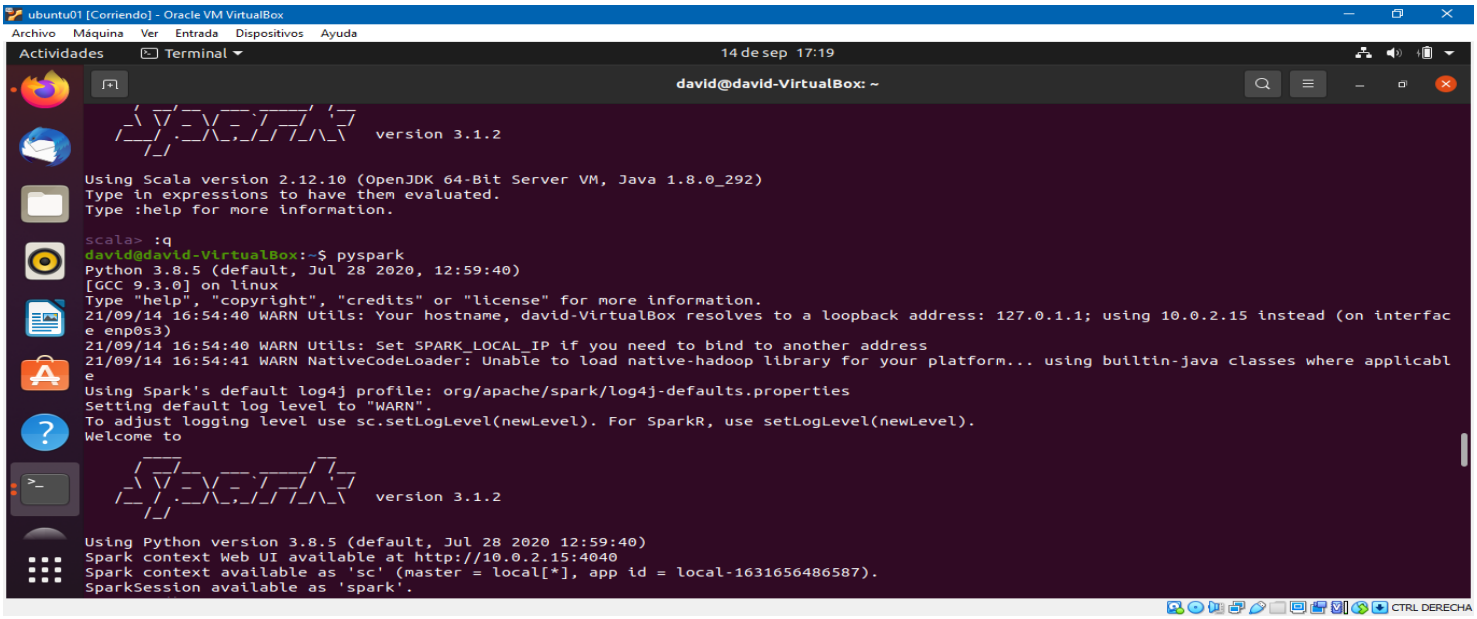
UNIVERSIDAD EL BOSQUE

Parte 3 – Configuración ambiente Spark

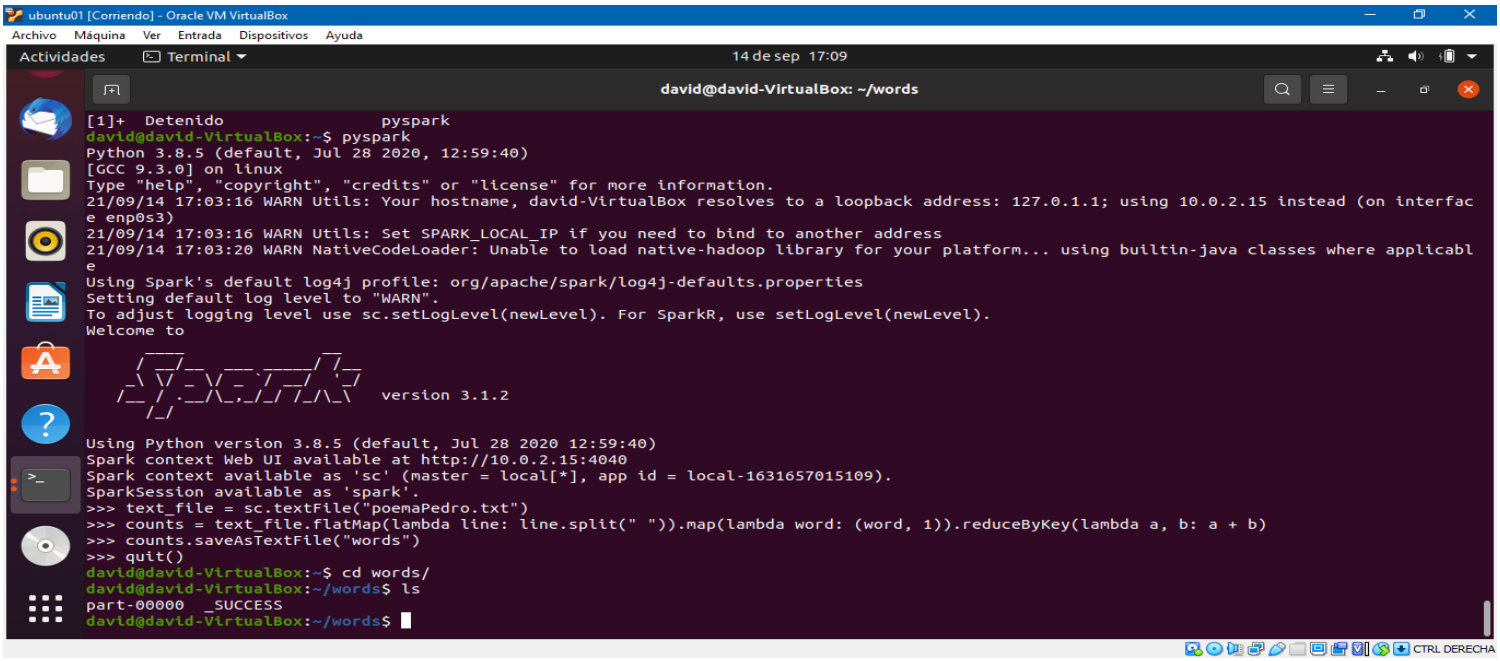
- Inicialmente se realizó la instalación y configuración del ambiente Spark. Posterior a esto se realizó la ejecución del puerto 127.0.0.1.:8080, Obteniendo la respuesta esperada para el entorno de ejecución de Spark.



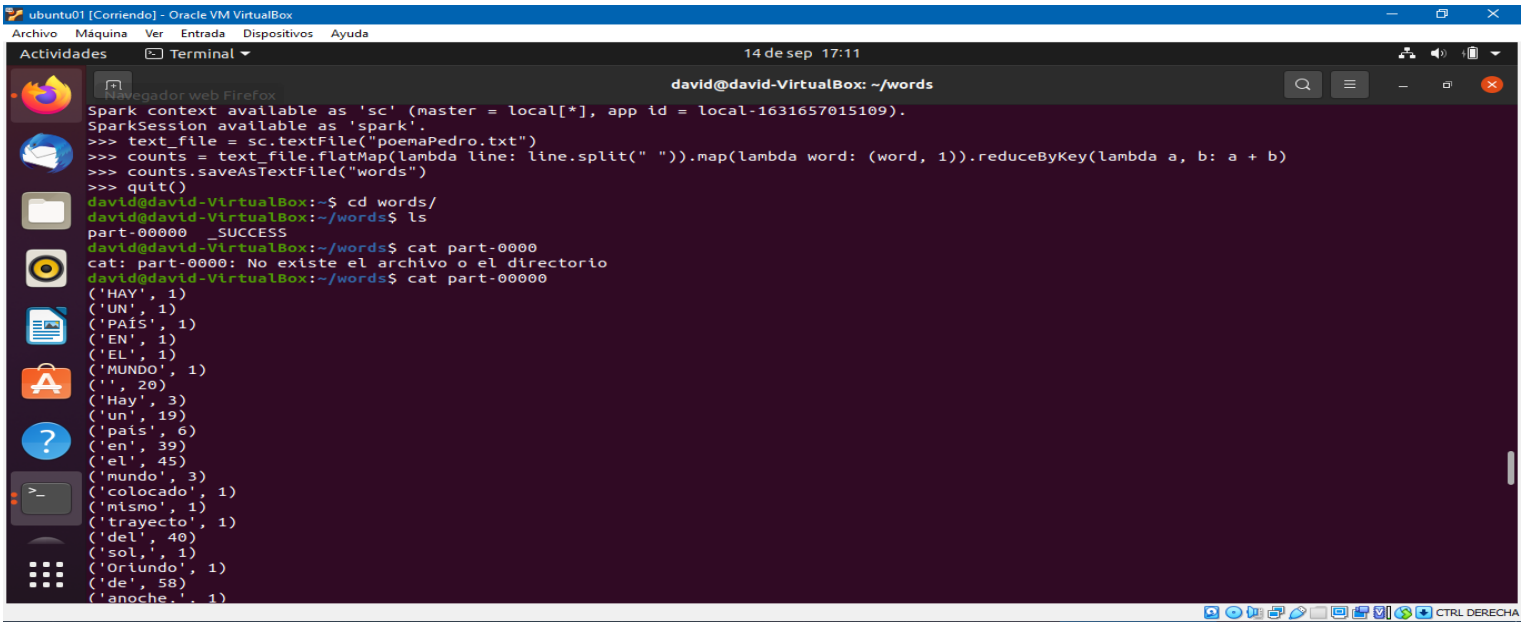
- Se procedió a realizar la comprobación del correcto funcionamiento del ambiente Spark, para lenguajes como scala y Python, obteniendo las respuestas esperadas.



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

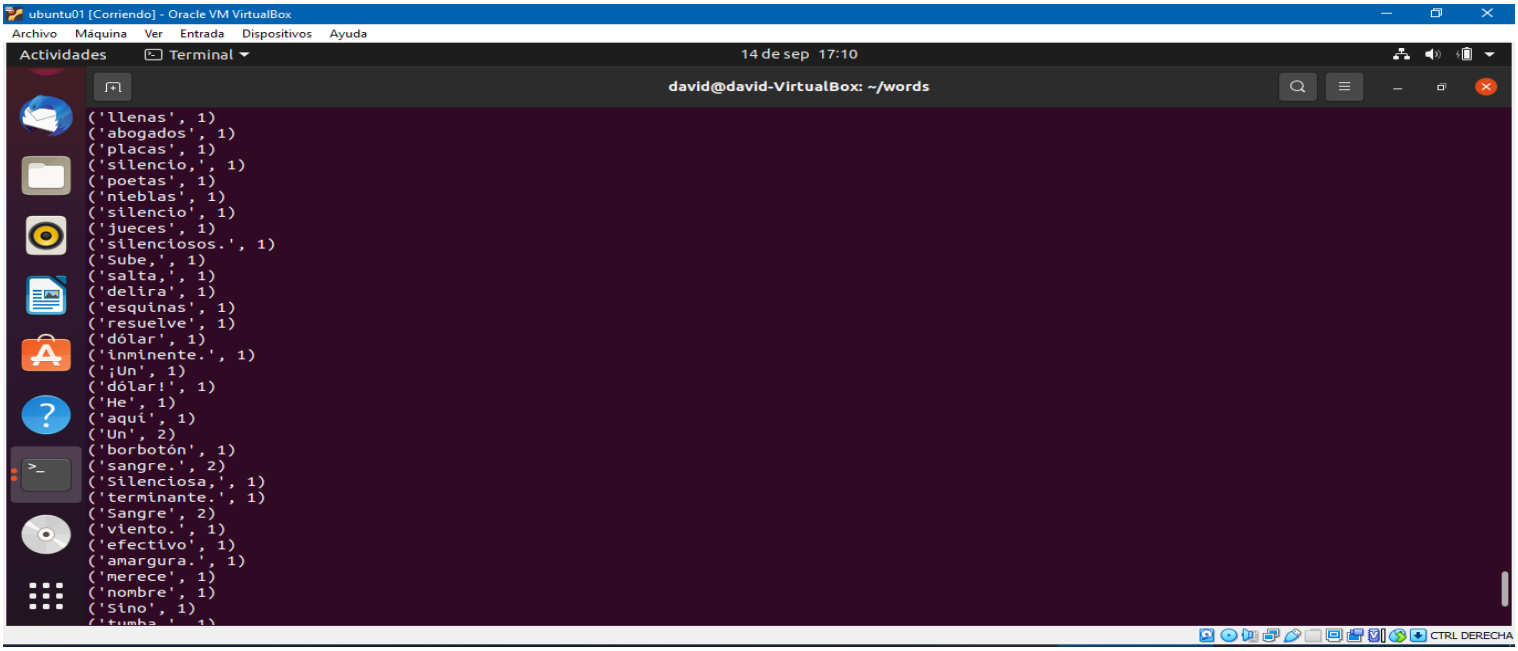


- Finalmente se realizo el conteo de palabras del archivo que spark python genera de respuesta al archivo cargado, y se obtuvo el conteo de palabras del poema.



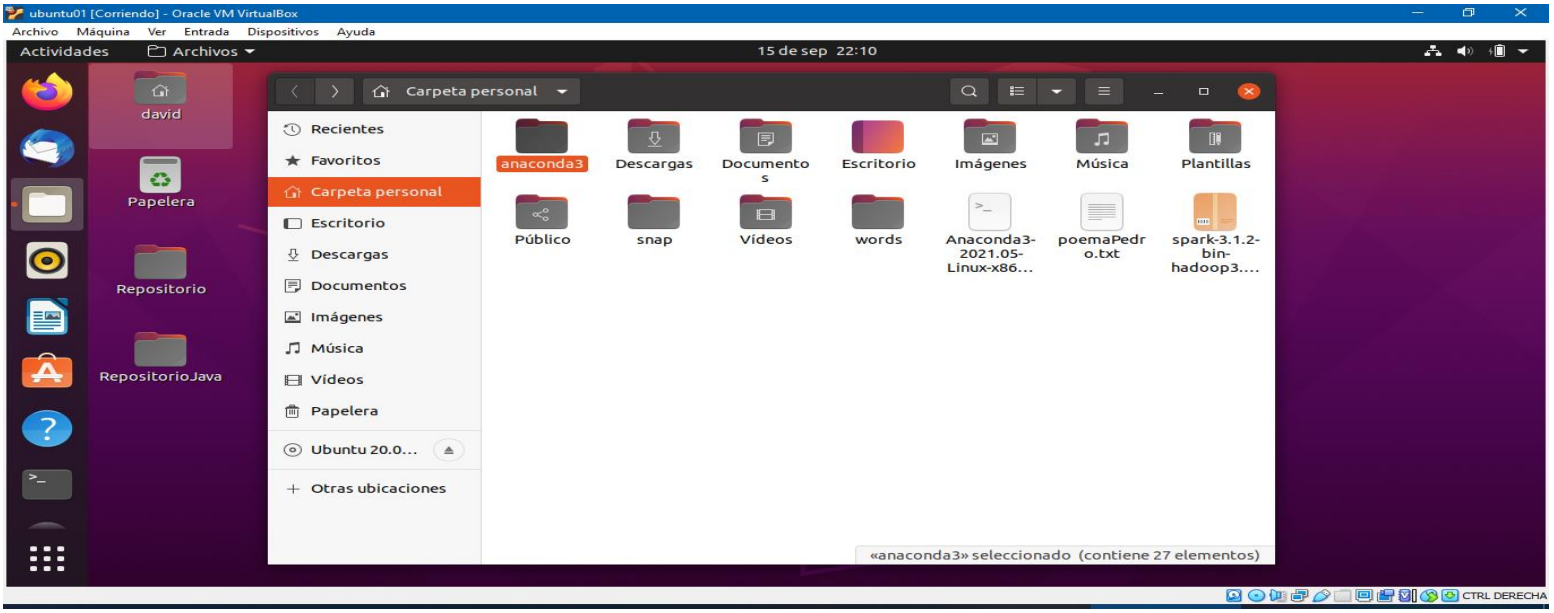
- Cabe resaltar que el procesamiento y análisis de datos de Spark es mucho más rápido a comparación con el mapReduce, que, por su procedimiento se demora en términos de tiempos y consumo de maquina más. Esto tambien

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



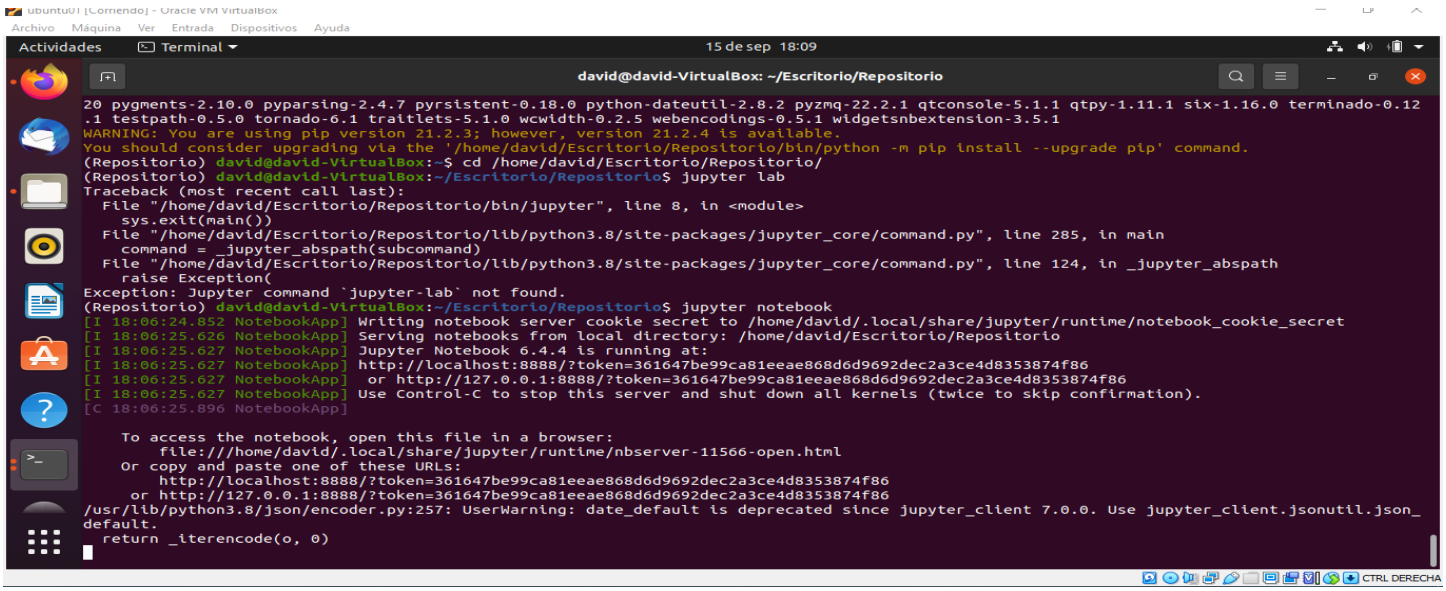
- ***Parte 4 (IDE) Anaconda***

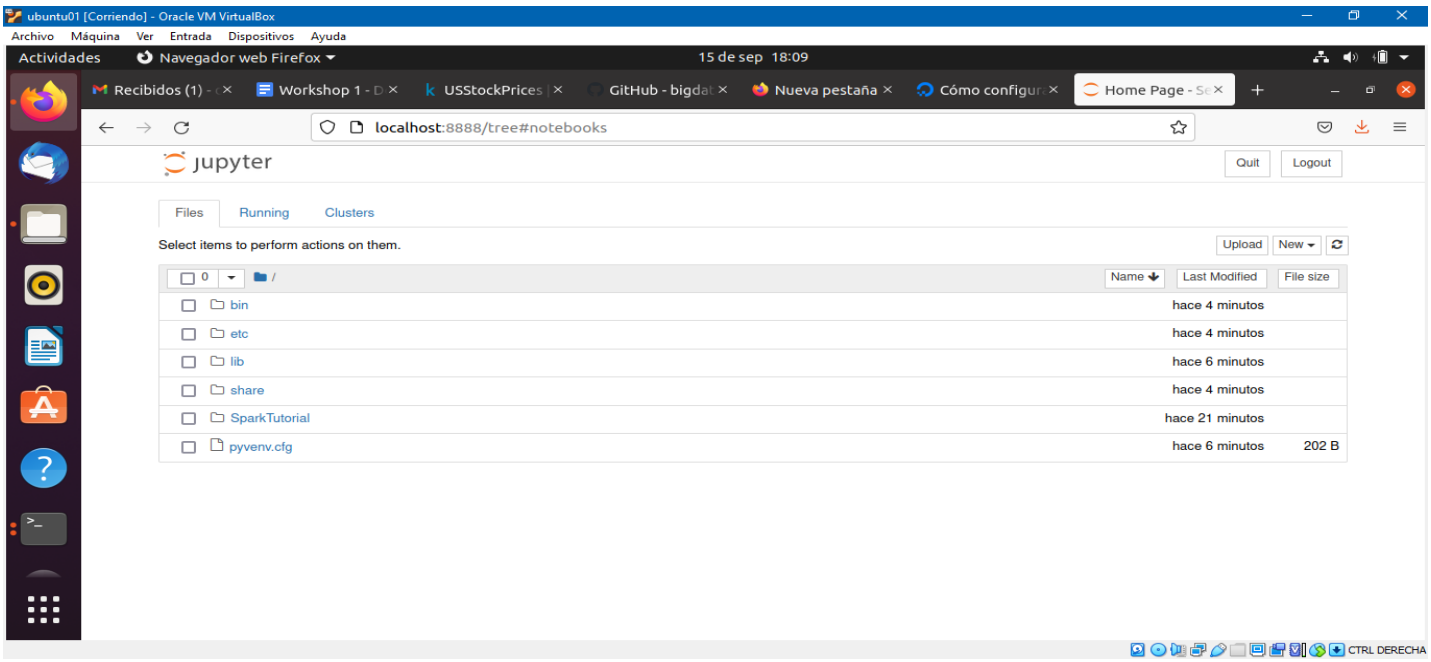
- Se realizo la instalación del IDE de anaconda, con ello se habilitada el ambiente de desarrollo proporcionado por este IDE, el cual incluye varias aplicaciones que permite trabajar diferentes lenguajes y frameworks de desarrollo.



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

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

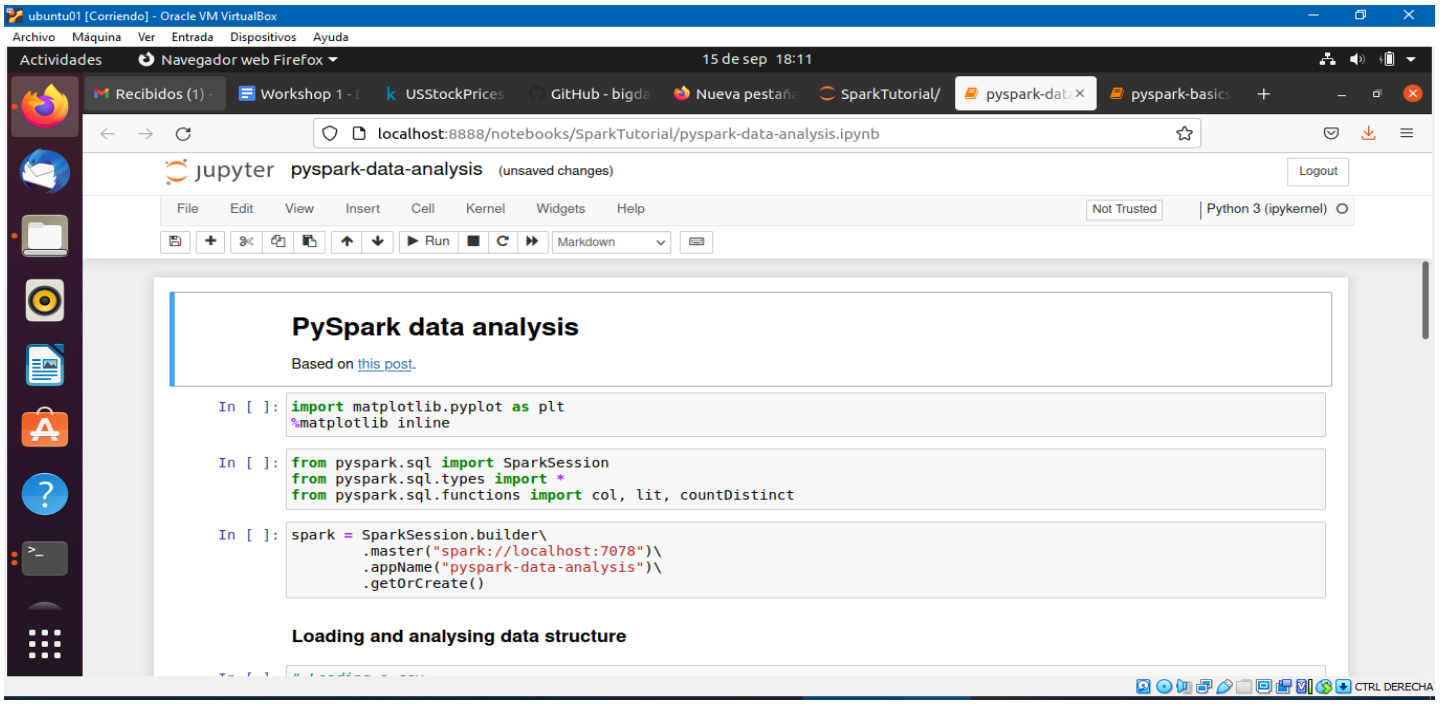


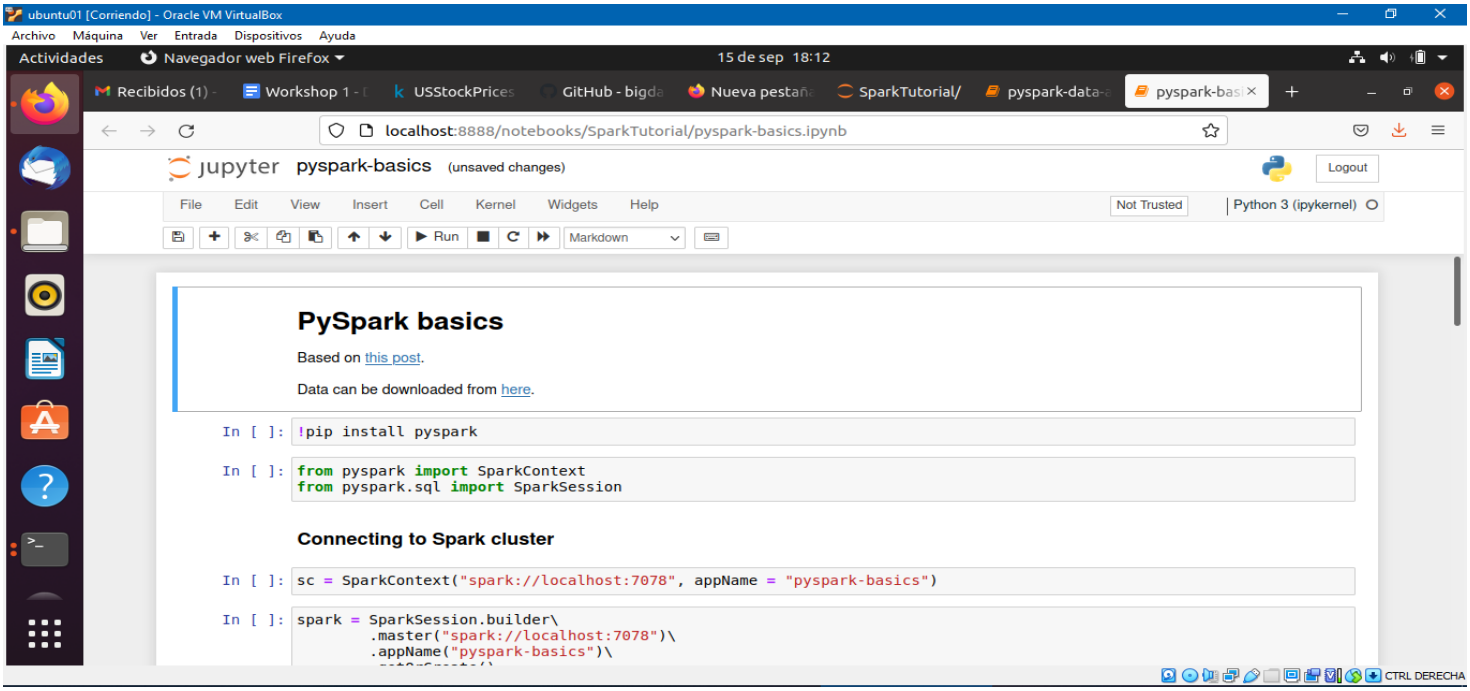


David Leonardo Barrera
David Esteban Zamora
Materia – Big Data and Analytics
Docente Fabian Peña
2021 - II

UNIVERSIDAD EL BOSQUE

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





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

Se realizo 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 libre 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.