Universidad el Bosque

Ingeniería de Sistemas

Big data

Daniel Pedraza

Juan Forero

**Taller 1**

Apache Hadoop es una plataforma de código abierto que permite el procesamiento distribuido y escalable de grandes conjuntos de datos, basándose en un grupo de computadores básicos. Proporciona una solución rentable para almacenar y procesa cantidades masivas de datos estructurados, semiestructurados y no estructurados sin importar el formato. Hadoop utiliza los datos localmente para procesarla en nodos en vez de enviar la información a una red externa, lo que reduce el tráfico y la velocidad.

La arquitectura master-slave tiene dos componentes principales: name node y data node. El primero tiene la tarea de saber dónde está cada bloque de un archivo; solo existe uno por cluster. El segundo guarda los bloques de datos y puede haber varios. Ambos se comunican por pulsos llamados Heartbeats.

Para el análisis de grandes cantidades de información, hadoop se basa en el algoritmo MapReduce que permite que un mismo trabajo se corra en paralelo. Básicamente, divide una misma tarea en muchas que se ejecutan en diferentes equipos. La fase de Map filtra, agrupa y organiza la información. Los datos de entrada se dividen en splits y cada tarea trabaja en paralelo en diferentes máquinas y su resultado es un par de llave-valor. Reduce junta todos los resultados de los diferentes nodos, resume y guarda los resultados en HDFS.

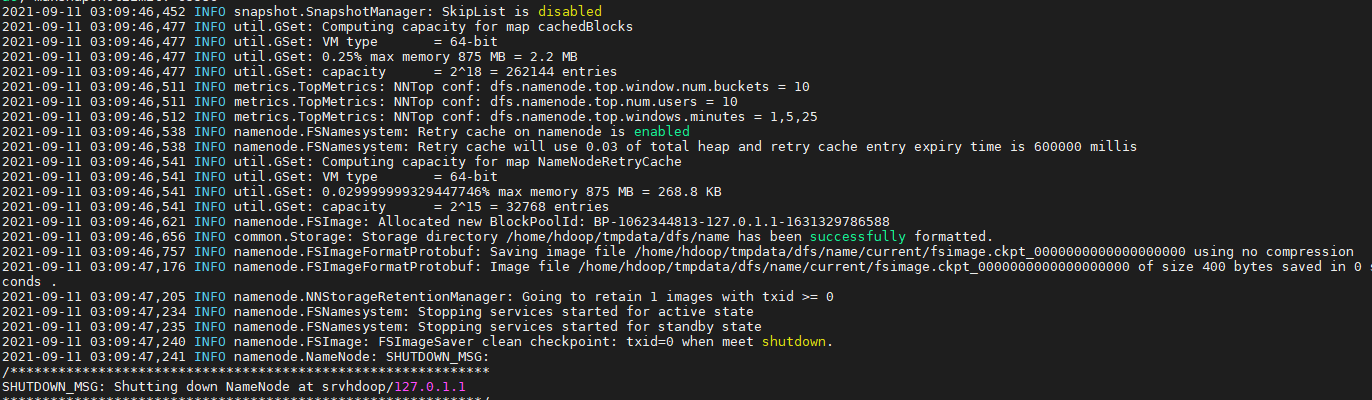
**Figura 1. Algoritmo MapReduce**Diagram

Description automatically generated

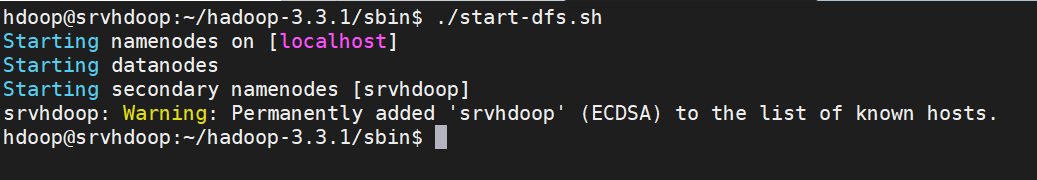
***Nota.*** Extraído de [1].

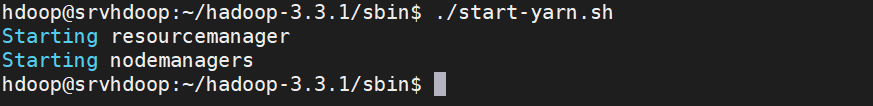
Se realizó la instalación de Hadoop en una máquina virtual Ubuntu 20. Como se puede ver a continuación, se formateó el name node y se inició el cluster y el Yarn (Yet Another Resource Negotiator) que se encarga de gestionar los recursos y cómo los datos almacenados en HDFS son procesados.

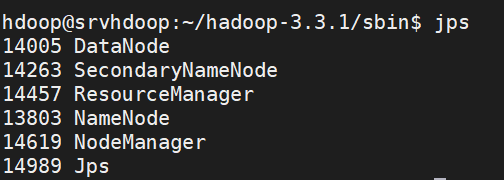
**Figura 2. Format HDFs name node**



**Figura 3. Inicio Hadoop Cluster y Yarn**

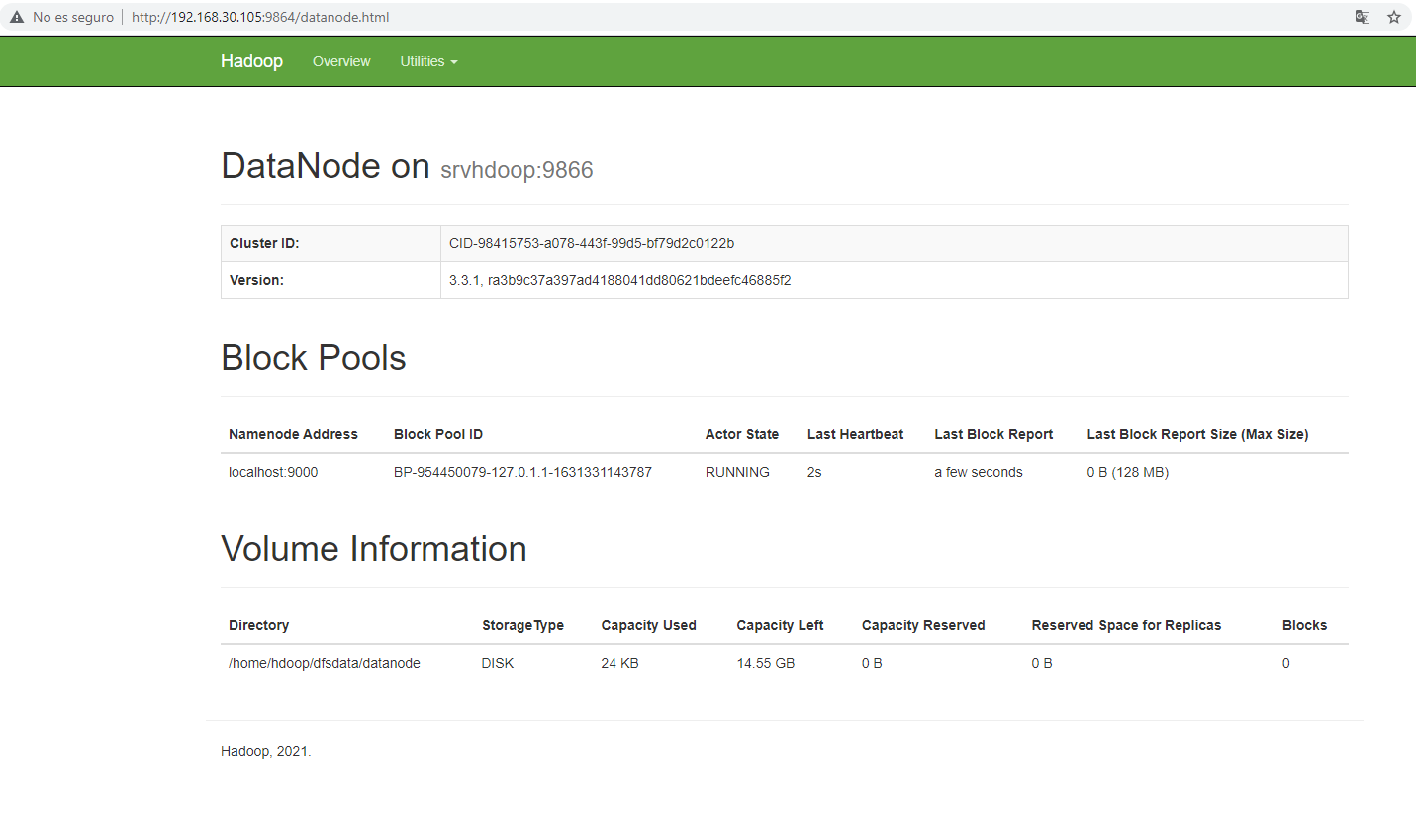


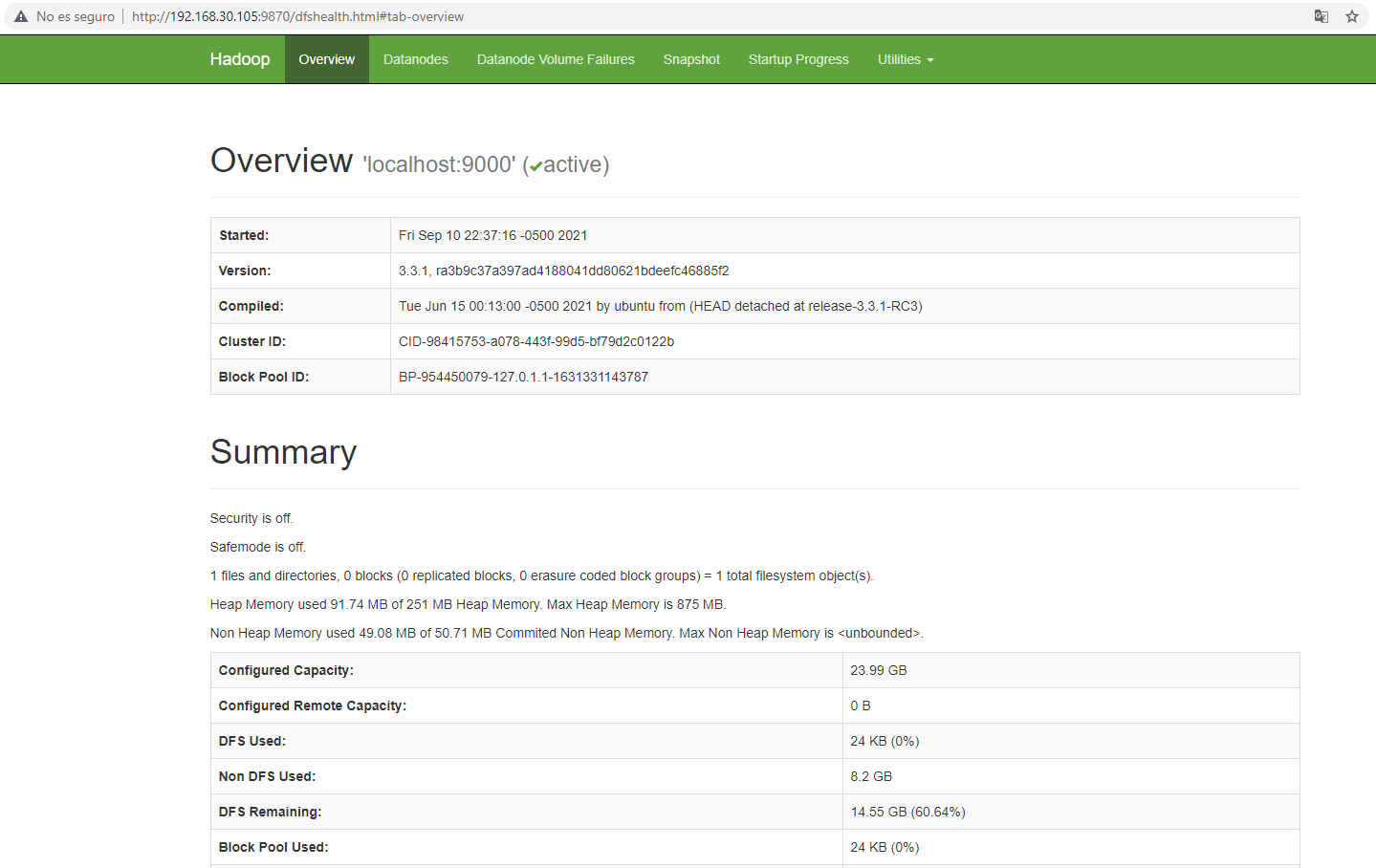


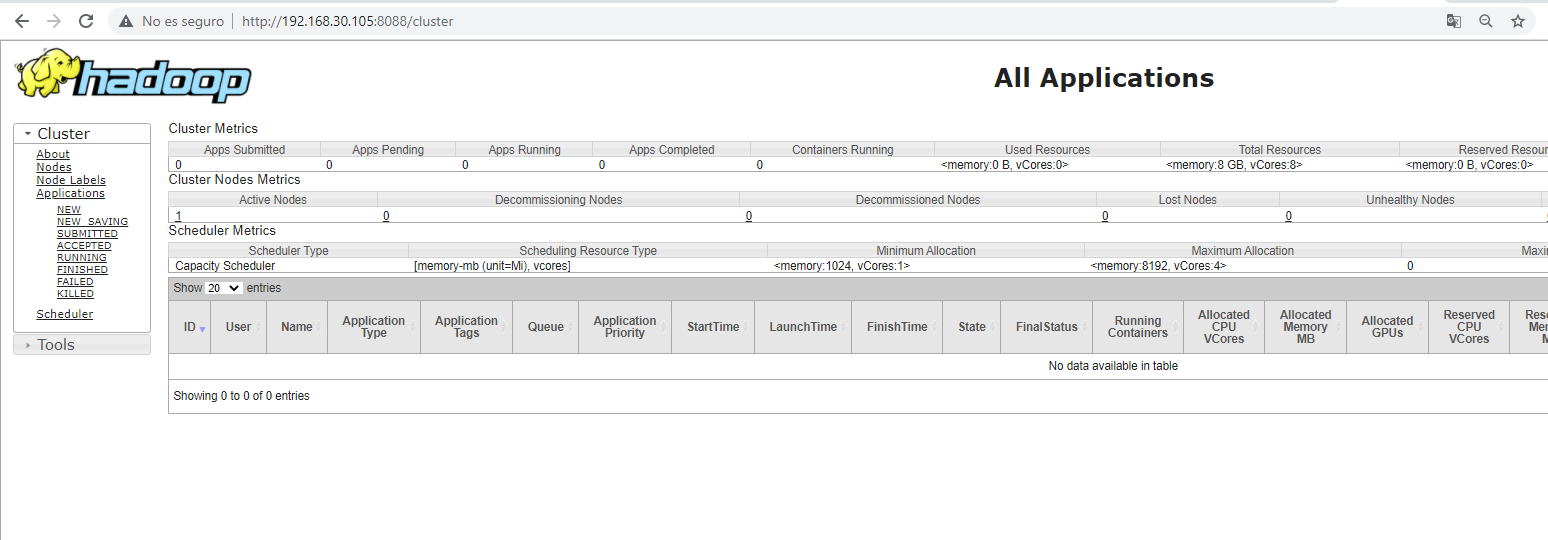


Se puede acceder a Hadoop desde el navegador. Con el puerto 9870 se accede al Hadoop name node. El puerto 9864 permite ver data nodes individuales y el puerto 8088 permite visualizar el Yarn.

**Figura 4. Acceso a hadoop desde el browser**



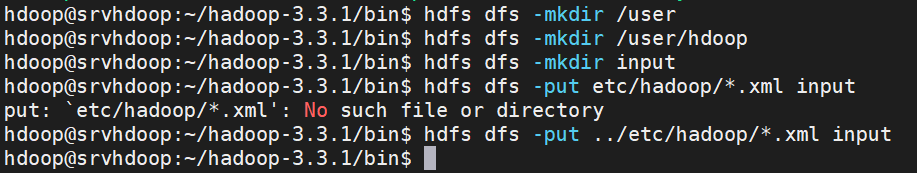


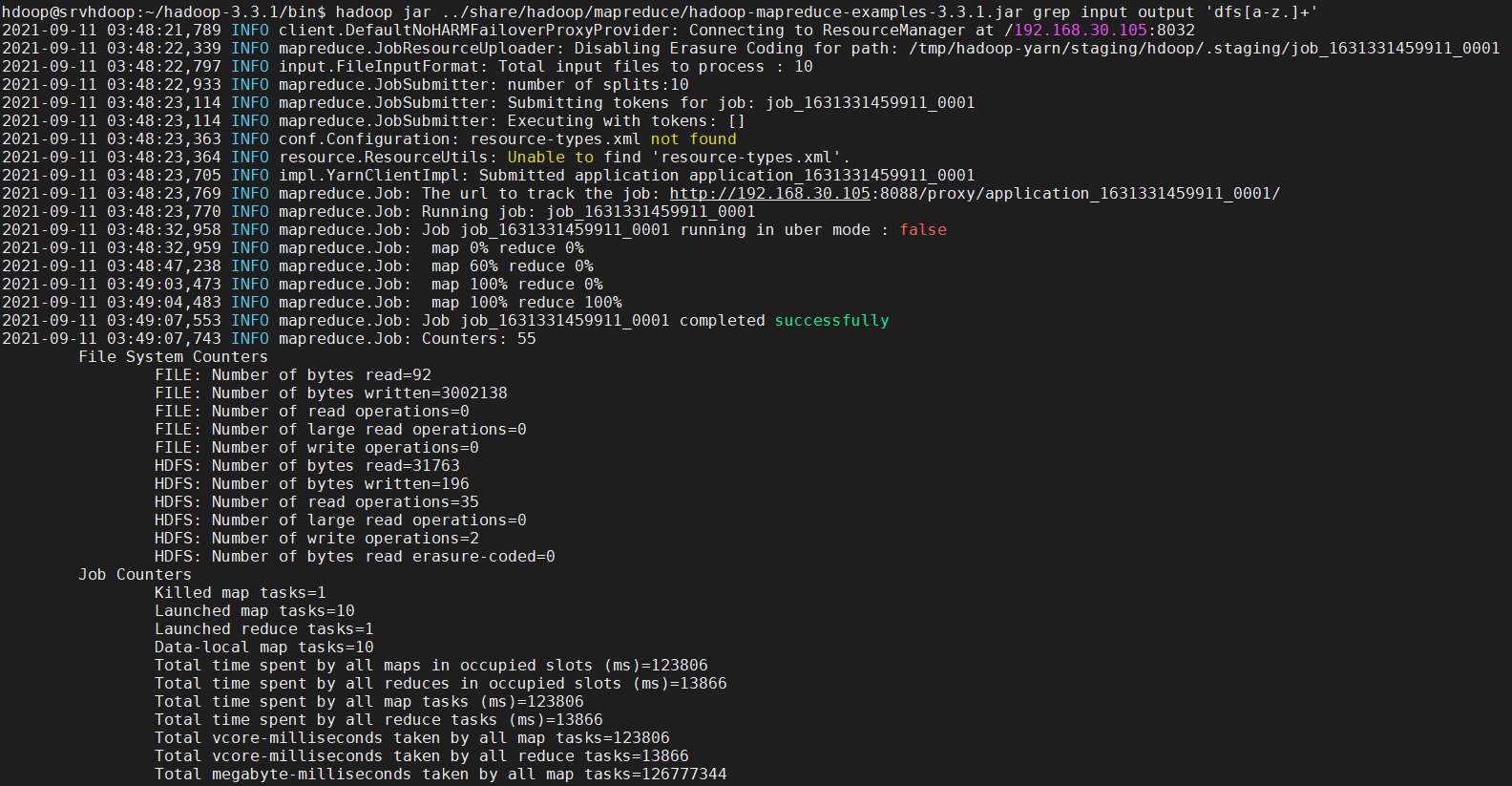


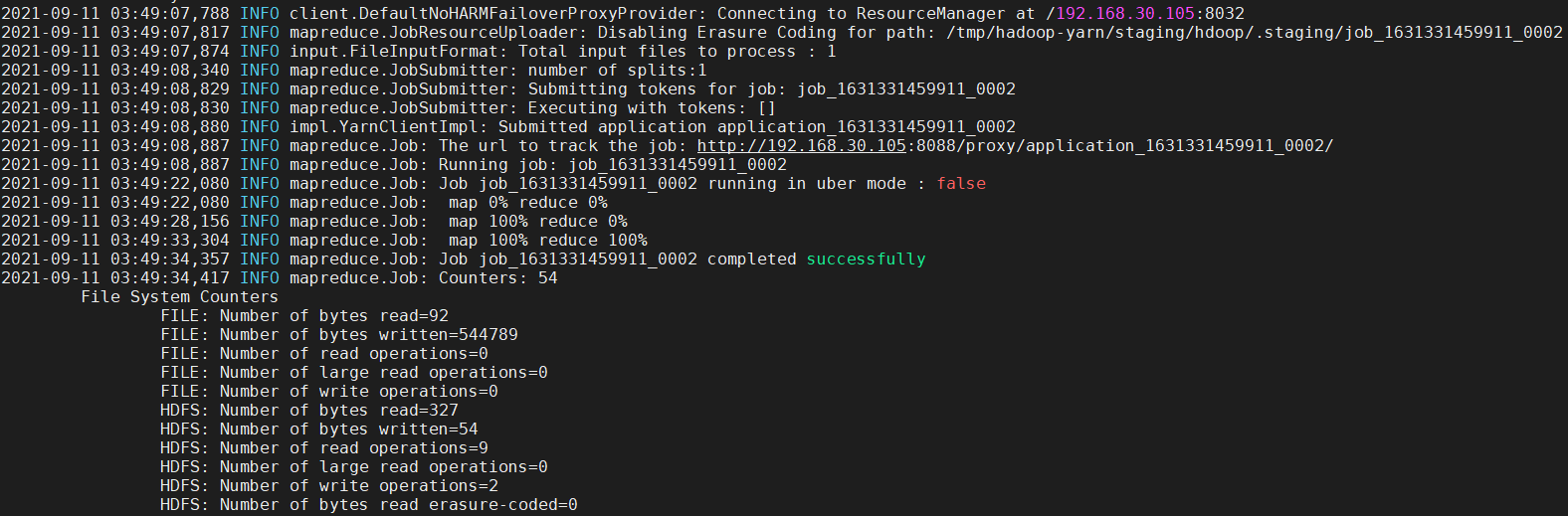
**Ejemplos Hadoop**

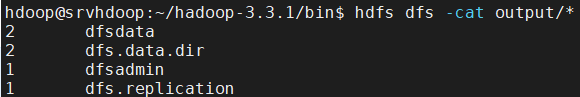
Se implementaron grep y wordcount. En el primero se guardaron todos los archivos terminados en \*.xml de la carpeta etc/hadoop en la carpeta input. Luego se dividen todos los archivos en Split que generan un par llave-valor de todos los archivos, seleccionando aquellos que cumplen con la expresión regular ‘dfs[a-z]+’. Finalmente, el Reduce junta todos los resultados, realiza el resumen y guarda los archivos en HDFS.

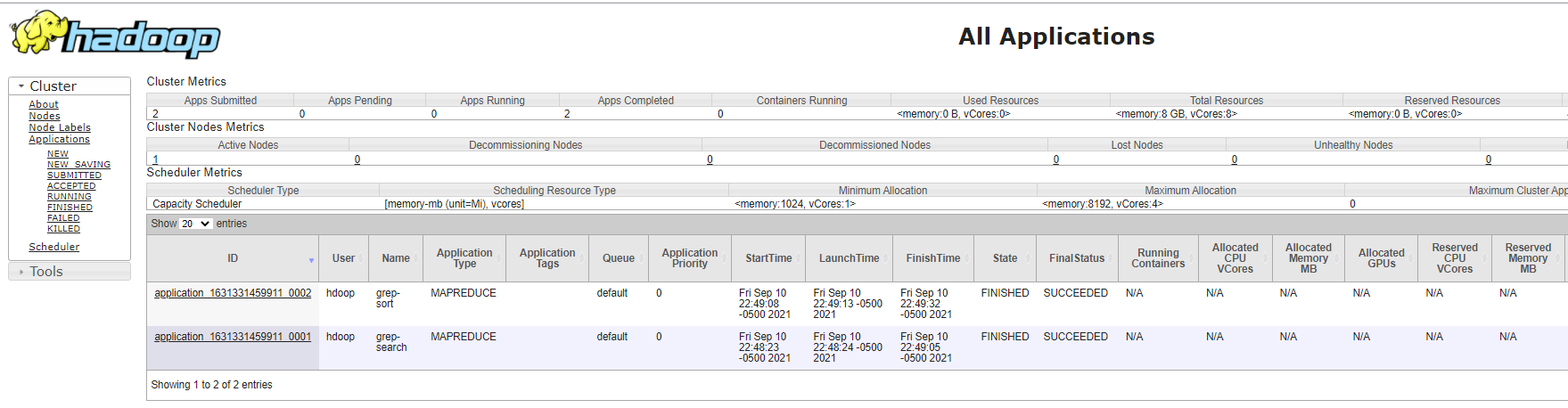
**Figura 5. Ejemplo Grep**

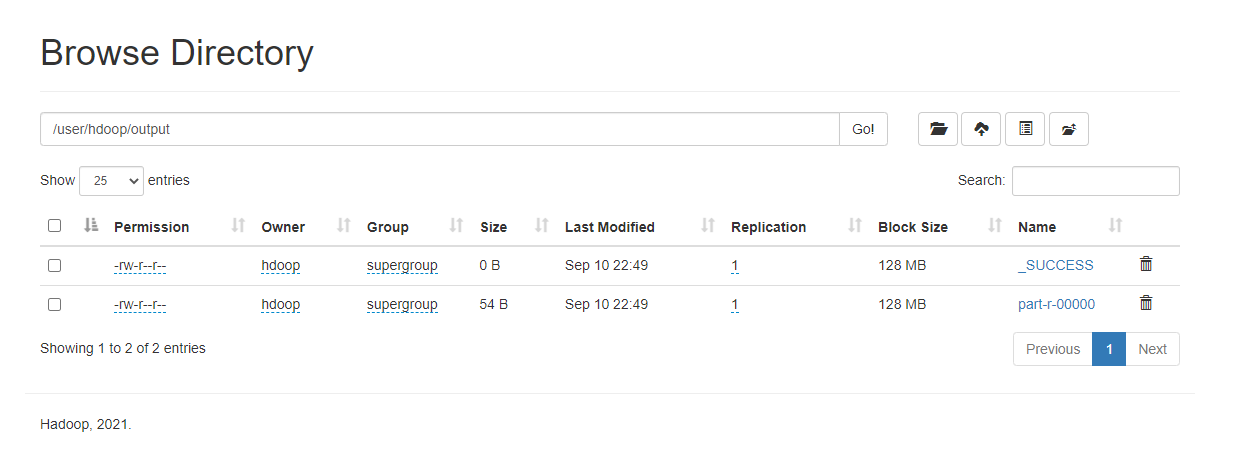






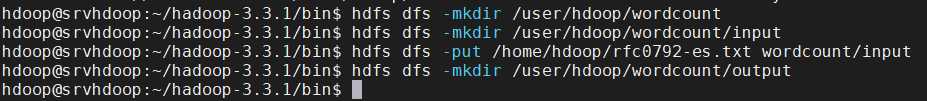


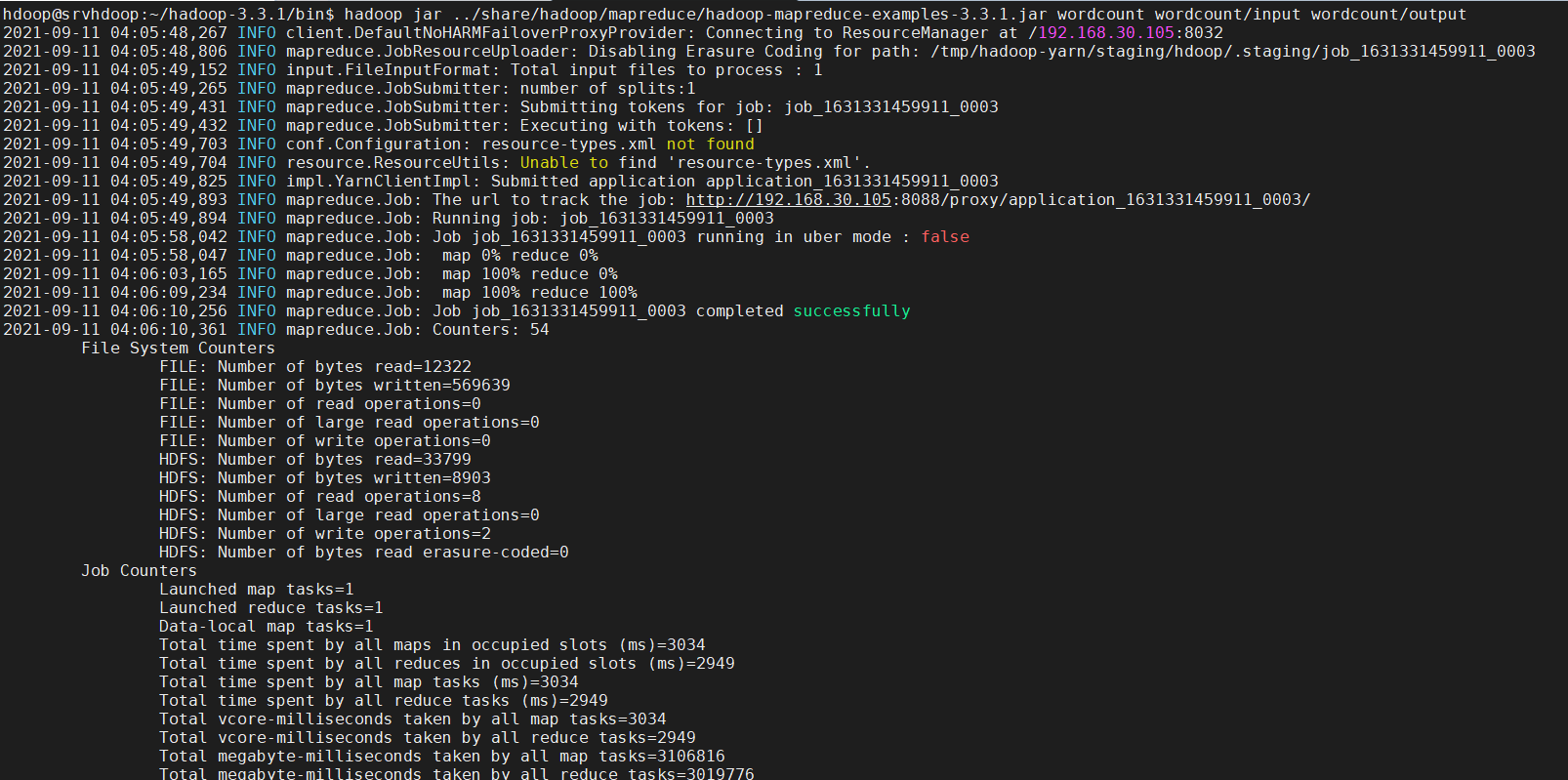


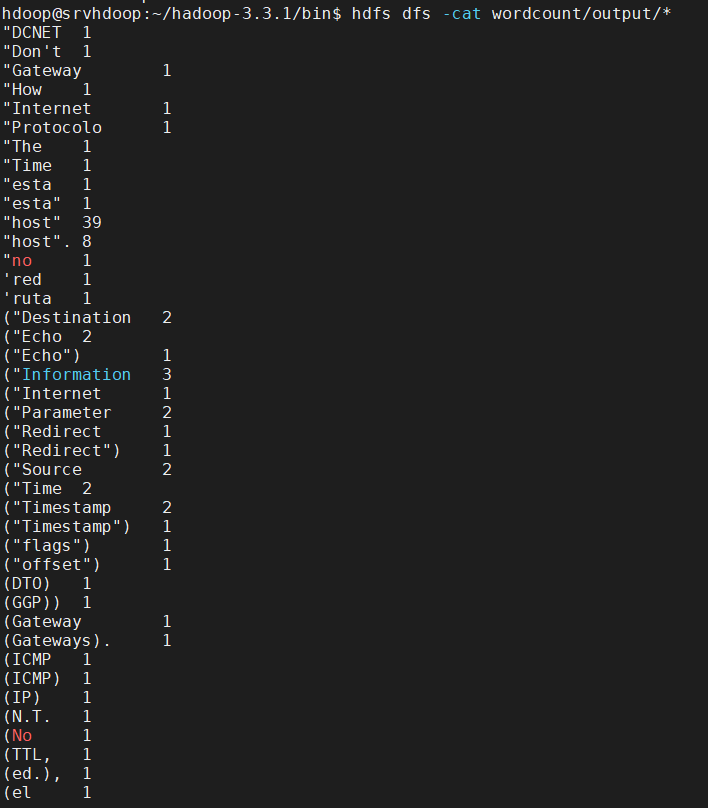


En el ejemplo WordCount, se carga un archivo de texto plano rcf0792 en la carpeta wordcount/input y el Map divide el trabajo de nuevo en Split que se procesan de manera paralela para contar las palabras. Se genera un resultado llave-valor de palabra-cantidad. El Reduce reúne todos los resultados y los guarda en la carpeta output.

**Figura 7. Ejemplo WordCount**

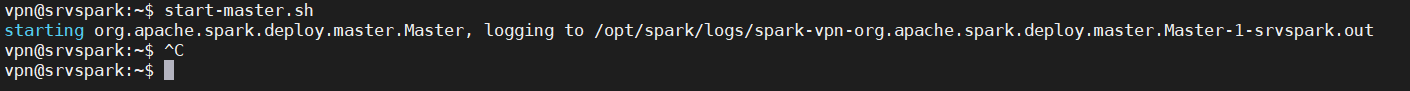


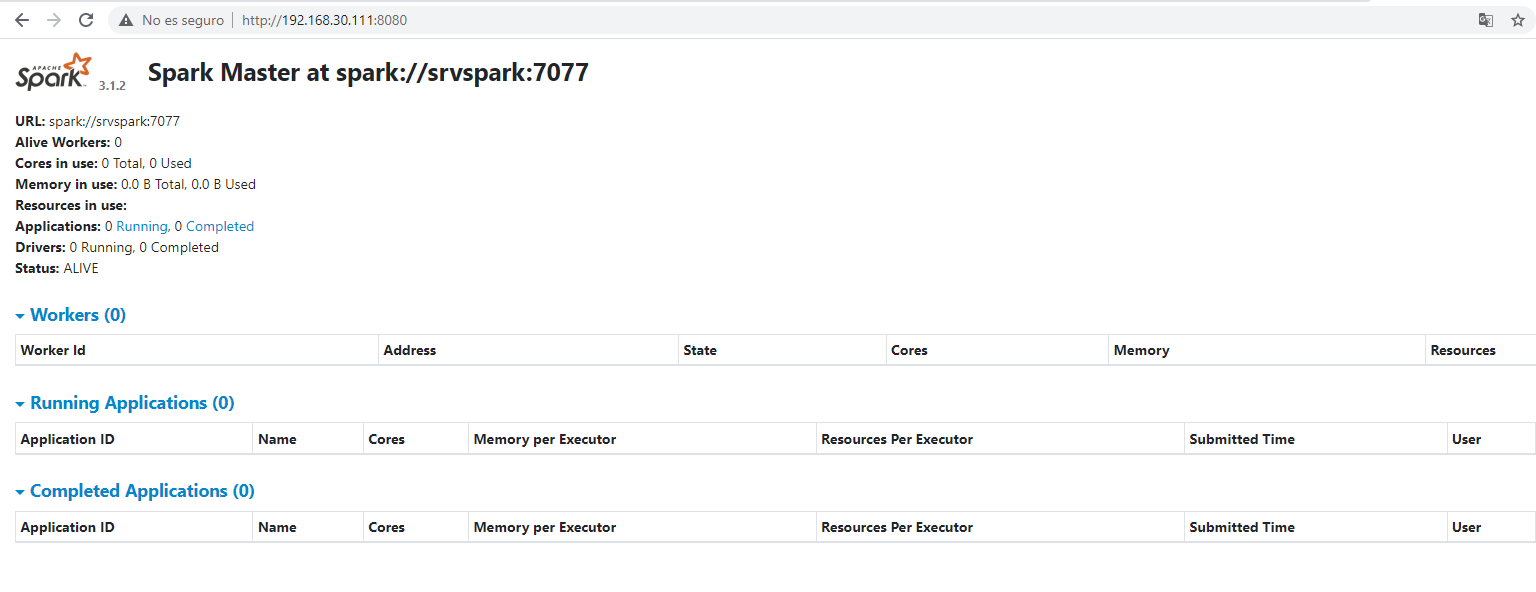


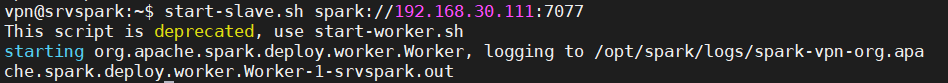


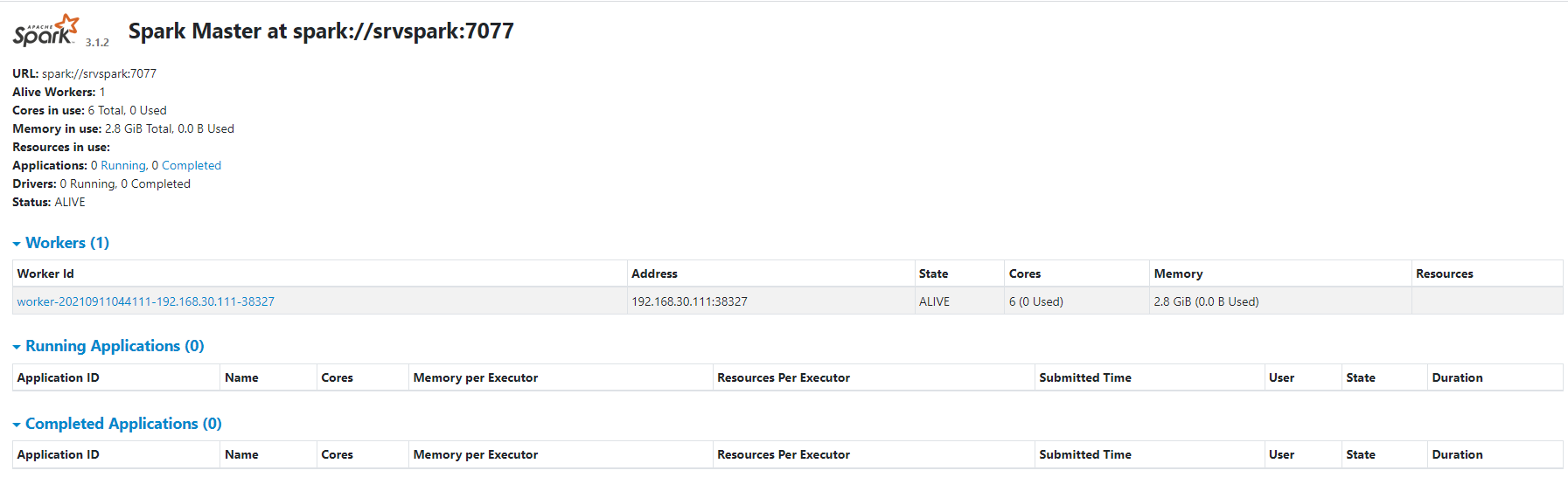
Por otra parte, se instaló Spark, que es la evolución natural de hadoop, cuya funcionalidad es más rígida y limitada. Algunas de sus ventajas es el procesamiento en memoria, lo que disminuye las operaciones de lectura/escritura, la posibilidad de análisis interactivo con SQL y la facilidad para interactuar con múltiples sistemas de almacenamiento persistente [2]. Se inicializó el servidor maestro y el servidor esclavo como se muestra en la figura 8.

**Figura 8. Instalación de Spark**



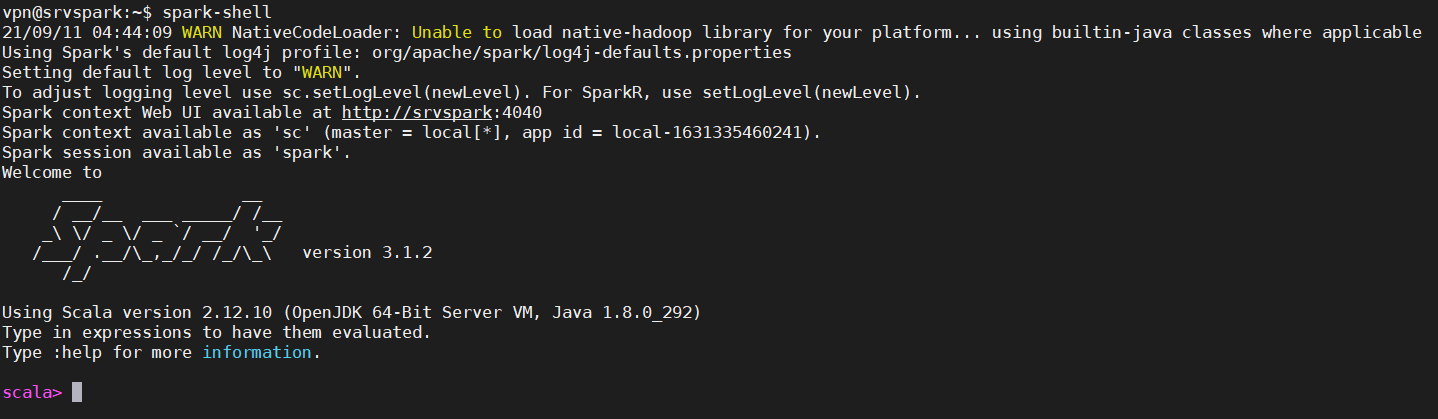


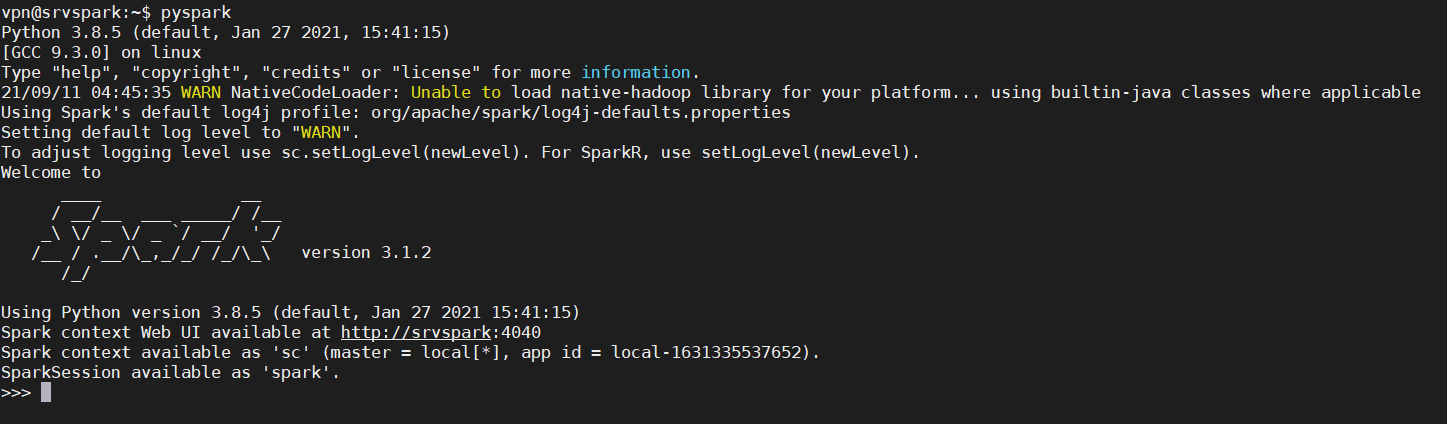




Se comprobó que spark funciona ingresando la interface por defecto de Scala (spark-shell) y con la interface de Python pyspark.

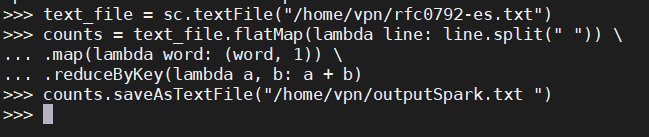
**Figura 9. Spark Shell y pyspark**

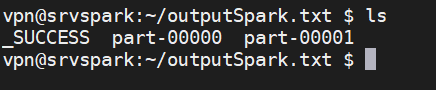


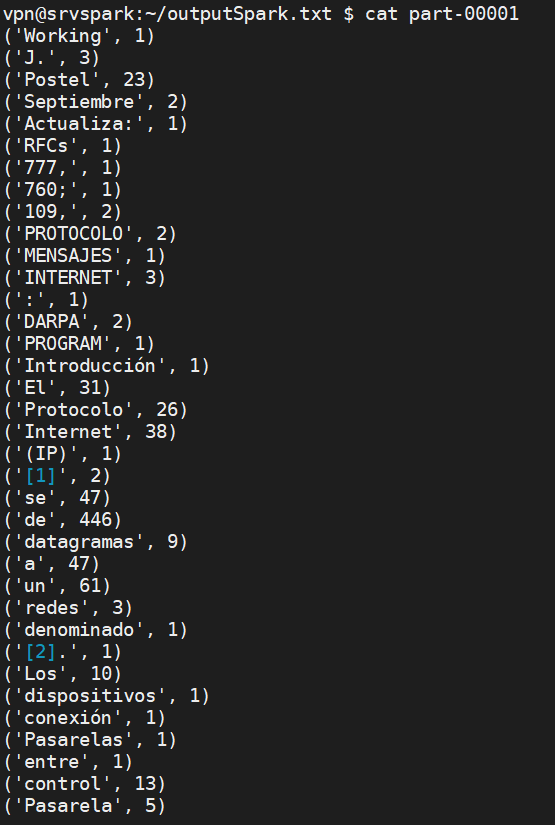
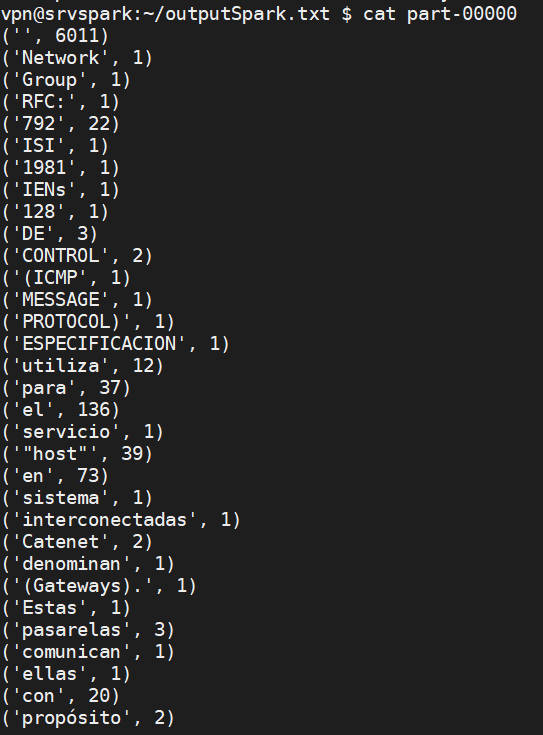


Se implementó el ejemplo WordCount para que realizara los cálculos de conteo sin tener en cuenta mayúsculas/minúsculas y signos de puntuación. El mapeo genera el par llave-valor de las palabras a partir de los espacios en blancos y luego el Reducer integra los resultados en el archivo outputSpark.txt.

**Figura 10. WordCount**

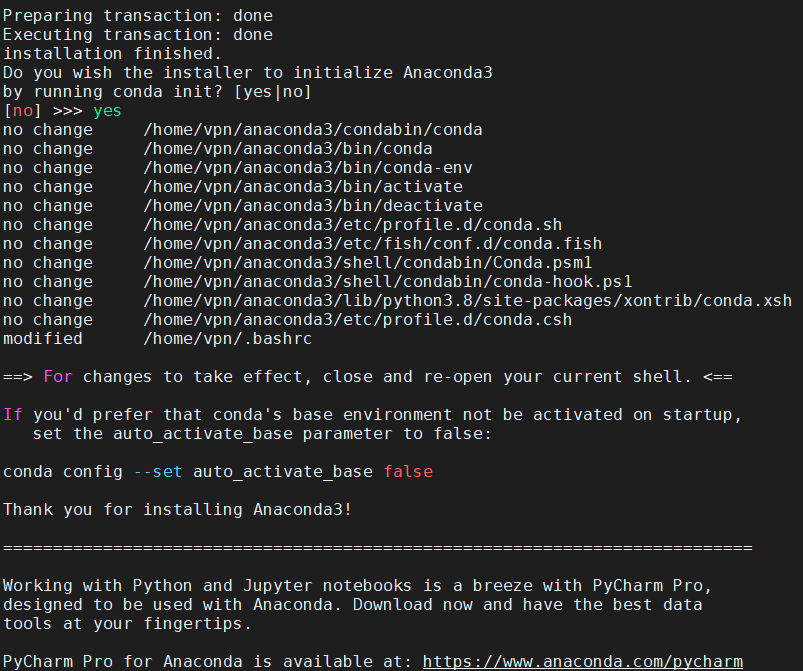




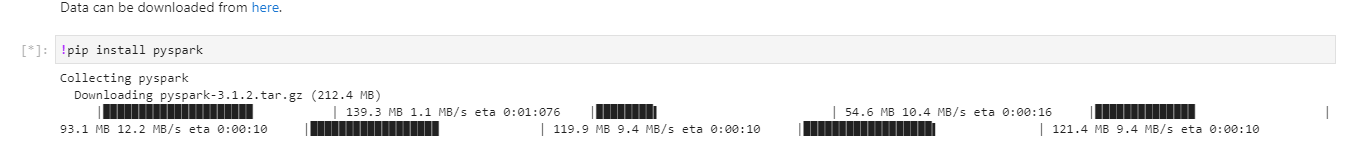


Finalmente, se clonó el repositorio sparkTutorial y se instaló Anaconda, la plataforma de distribución de Python más popular con más de 25 millones de usuarios.

**Figura 11. Instalación de Anaconda**

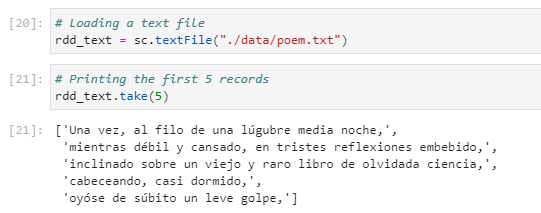


Se ejecutaron los scripts de de pyspark-basics como se muestra a continuación.





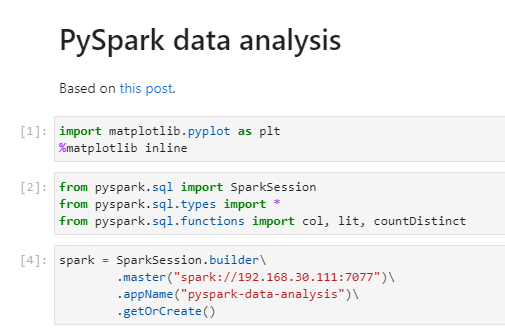
Se conecta primero al cluster, luego se carga el poem.txt y se obtienen los primeros cinco registros.



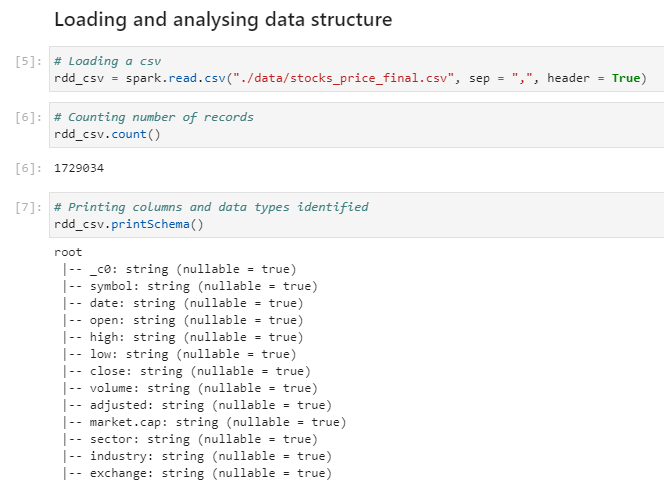
Se multiplicó cada valor de la colección collect con el lambda x: x\*2 y se filtraron los mayores a cinco. Luego se separaron las palabras del texto y se tomaron las primeras dos filas.



A continuación, se ejecutó el script data análisis.

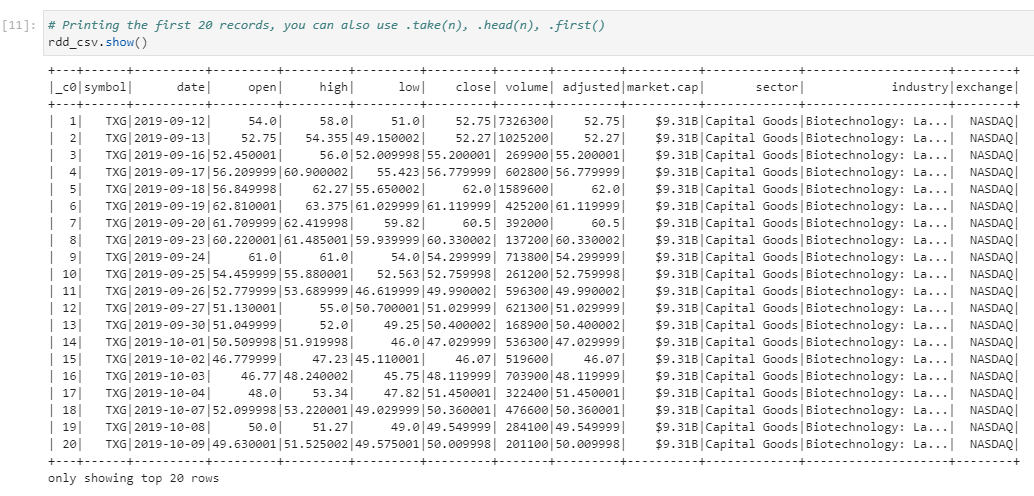


Se cargó el archivo CSV stocks\_price\_final, se contó el número de palabras y se imprimió las columnas con su tipo de dato.



Se creó un esquema con los campos y sus respectivos tipos de datos para el archivo cvs y luego se imprimieron los primeros 20 registros.

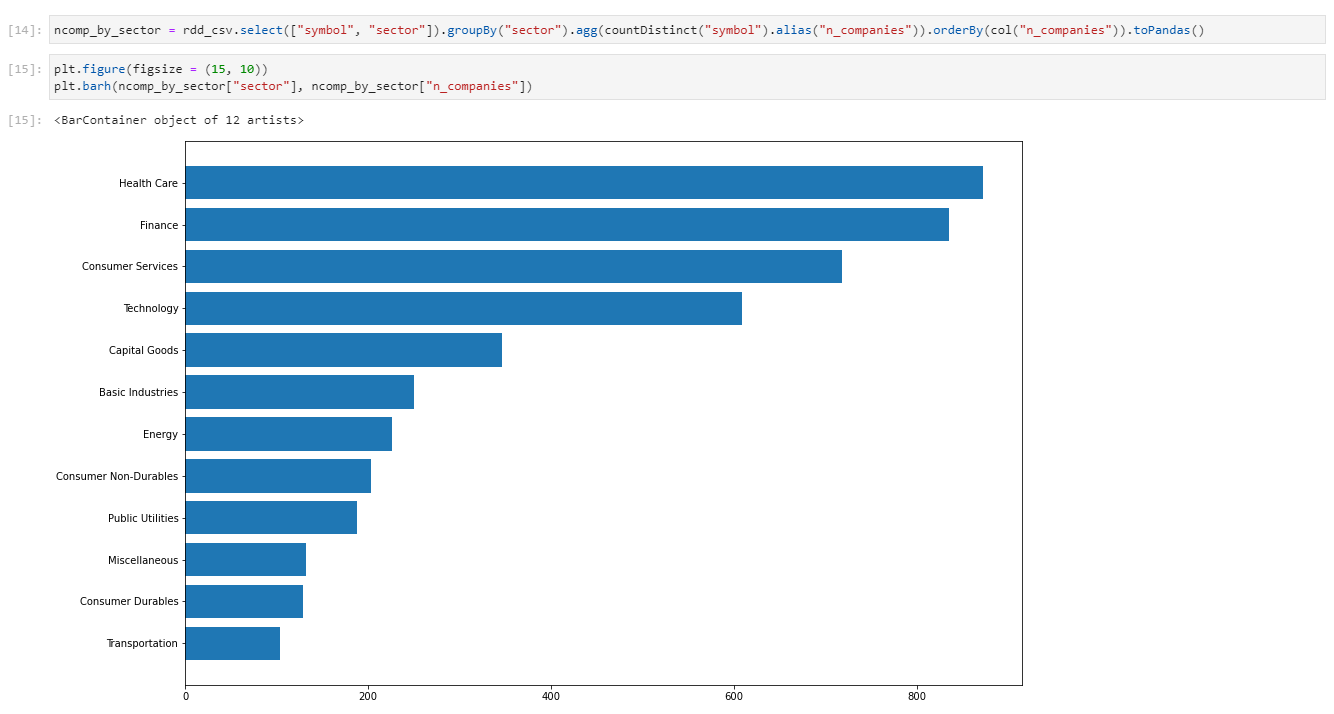


2

Se filtró por columna “symbol” y fila “tsla” y “symbol” y “gme”. Luego se transformó a pandas para mejorar la visualización.



Por último, se selecciono “symbol” y “sector”, se agrupó por “sector”, se agregó el countDistinct y se ordenó por “n\_companies” para luego transformarlo a Panda y visualizar.



**Referencias**

[1] IBM. Apache Hadoop [online]. https://www.ibm.com/co-es/analytics/hadoop

[2]. Aniruddha Bhandari. Introduction to the Hadoop Ecosystem for Big Data and Data Engineering. October 23, 2020. [online.] https://www.analyticsvidhya.com/blog/2020/10/introduction-hadoop-ecosystem/

[3]. ESIC. Apache Spark: Introducción, qué es y cómo funciona. Octubre 2018. https://www.esic.edu/rethink/tecnologia/apache-spark-introduccion-que-es-y-como-funciona