USO DE LA PLATAFORMA CLOUDERA CDH COMO UNA HERRAMIENTA DE BIG DATA Y CIENCIA DE DATOS

Ing. Rodrigo Vázquez López[1]
Ing. Esther Viridiana Vázquez Carmona[1]
Dr. Juan Carlos Herrera Lozada[1]
Dr. Jacobo Sandoval Gutiérrez[2]

[1]Instituto Politécnico Nacional
Centro de Innovación y Desarrollo Tecnológico en Cómputo (IPN-CIDETEC)

[2]Universidad Autónoma Metropolitana Unidad Lerma
Departamento de Procesos Productivos

rodrigo_em2@hotmail.com
ev.vazquezc@gmail.com
jcrls.ipn@gmail.com
j.sandoval@correo.ler.uam.mx

Resumen

La ciencia de datos y big data son dos conceptos que han surgido en la actualidad gracias a la cantidad de información oculta que se puede extraer de los grandes conjuntos de datos.

Este trabajo busca incentivar el uso de la plataforma Cloudera CDH como herramienta para el desarrollo de aplicaciones encaminadas a big data y ciencia de datos, por lo cual, se ejemplifica el uso de plataforma en el análisis de datos utilizando como ejemplo el texto del libro “Cien años de Soledad” de Gabriel García Márquez.

Introducción

En la actualidad, la cantidad de datos que generan los dispositivos electrónicos contiene información valiosa que puede aprovecharte en distintos campos.

La ciencia de datos es el proceso en el cual se descubren nuevos conocimientos extrayendo información oculta de los conjuntos de datos combinando técnicas de aprendizaje automático, estadística y programación [1].

Por otra parte, big data se utiliza para describir grandes volúmenes de datos crecientes cuyo tamaño no permite un análisis con herramientas tradicionales [2].

MapReduce

MapReduce es un modelo asociativo de programación diseñado por Google con el objetivo de procesar grandes conjuntos de datos. En este modelo el usuario especifica una función de mapeo que procesa un par clave-valor para generar un conjunto intermedio, y una función de reducción que combina todos los valores intermedios. Los programas escritos utilizando el modelo MapReduce se paralelizan automáticamente y se ejecutan en un clúster de máquinas que no requieren características técnicas especializadas [3].

Los principales componentes del modelo son las funciones Map y la función Reduce, las cuales son definidas por el usuario. La función Map, toma un par clave-valor de entrada y produce un conjunto intermedio de pares clave-valor. La función se define de la siguiente forma:

$$\mathrm{map}(\mathrm{k1}, \mathrm{v1}) \rightarrow \mathrm{List}(\mathrm{k2}, \mathrm{v2})$$

La función Reduce acepta una clave intermedia y un conjunto de valores para esa clave. La función combina estos valores para formar un conjunto de valores más pequeño. La definición de la función es de la siguiente forma:

$$\mathrm{reduce}(\mathrm{k2}, \mathrm{List}(\mathrm{v2})) \rightarrow \mathrm{List}(\mathrm{v2})$$

La figura 1 describe el funcionamiento del modelo MapReduce aplicado en el problema de conteo de palabras, donde se observa que el conjunto de datos de entrada se particiona en M subconjuntos cuyo tamaño puede variar entre 16 y 64 MB. Una vez obtenidos los subconjuntos, estos se copian a diferentes nodos del clúster, los cuales ejecutan la función Map sobre el respectivo subconjunto. Finalmente, los resultados parciales de cada nodo se combinan con la función Reduce para obtener el conjunto de salida.

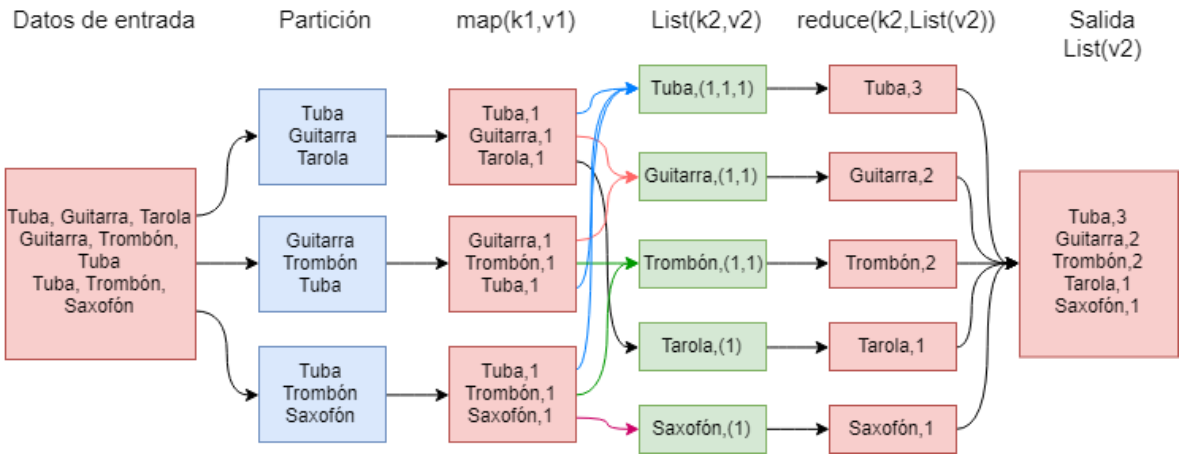

Figura 1. Modelo de MapReduce aplicado al problema de conteo de palabras.

Apache Hadoop

Apache Hadoop es un sistema de archivos distribuido y un framework de código abierto desarrollado por la Fundación Apache cuyo enfoque principal es el procesamiento de grandes conjuntos de datos utilizando el modelo de programación de MapReduce. Hadoop permite particionar los datos y la ejecutar en paralelo de aplicaciones utilizando N hosts. Hadoop es la implementación en código abierto del modelo MapReduce siendo el sistema de archivos (Hadoop File System) el componente principal [4,5].

Desarrollo

Se utilizó la plataforma Cloudera CDH, la cual es una distribución de Apache Hadoop que incluye herramientas propias para el análisis de datos [6]. La plataforma ofrece clústers en forma de máquinas virtuales que se pueden descargar de forma gratuita y cuyo propósito es de desarrollo de aplicaciones y la enseñanza. Las máquinas virtuales se pueden descargar desde [7].

Para ejemplificar el funcionamiento de hadoop se utilizó el ejemplo wordcount que provee Cloudera y se encuentra disponible en [8], además, como conjunto de entrada se utilizó el texto del libro “Cien años de soledad”, el cual se encuentra disponible en [9] para su consulta de forma gratuita.

Para comenzar a trabajar con hadoop, es necesario crear los directorios de entrada y salida dentro del sistema HDFS, para ello ejecutamos los siguientes comandos:

```
$ sudo su hdfs
$ hadoop fs -mkdir /user/cloudera
$ hadoop fs -chown cloudera /user/cloudera
$ exit
$ sudo su cloudera
$ hadoop fs -mkdir /user/cloudera/wordcount /user/cloudera/wordcount/input
```

Una vez generados los directorios, es necesario copiar el archivo de entrada al sistema HDFS (figura 2), para ello utilizamos el siguiente comando:

$ hadoop fs -put /ruta_archivo_entrada /user/Cloudera/input

Figura 2. Proceso de copia del archivo de entrada.

Procedemos a ejecutar el ejemplo utilizado el siguiente comando (figura 3):

$ hadoop jar wordcount.jar org.myorg.WordCount /user/Cloudera/wordcount/input /user/Cloudera/wordcont/output

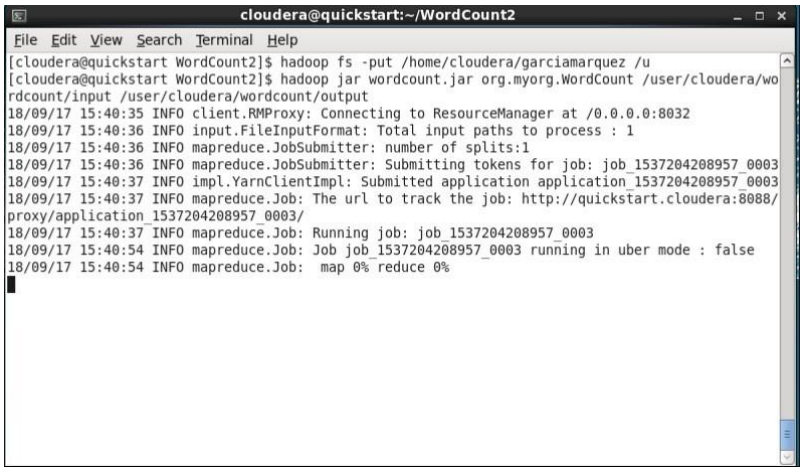

Figura 3. Ejecución del programa.

Con lo cual se procederán a ejecutar las tareas de MapReduce, tal y como se observa en la figura 4.

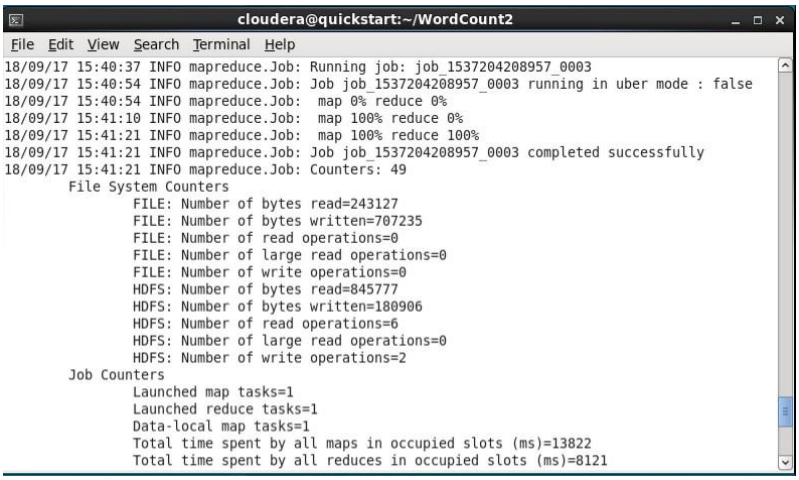

Figura 4. Resultados de la ejecución.

Finalmente, para extraer los resultados del sistema de archivos de hadoop utilizamos el comando (figura 5):

$ hadoop fs -get /user/Cloudera/wordcount/output /directorio_salida

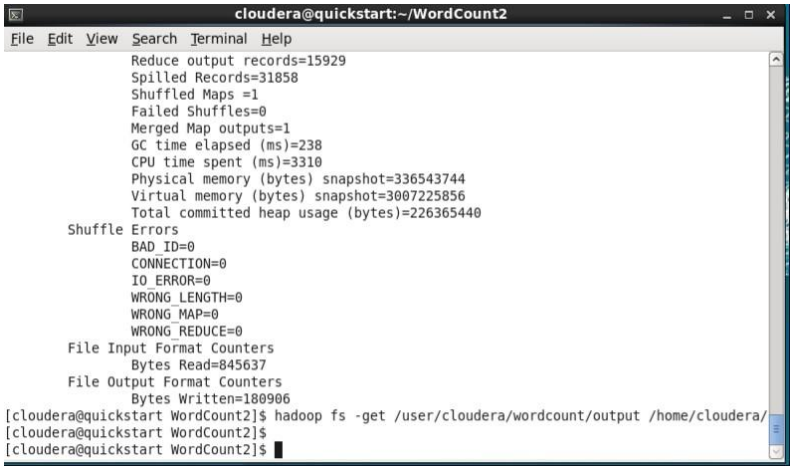

Figura 5. Extrayendo los resultados del sistema HDFS.

Resultados

El programa procesó 125702 palabras. El texto hace uso de 15928 palabras diferentes y 15533 signos de puntuación. La palabra más utilizada es la preposición “de”, la cual se repite 8861 veces, seguida del artículo “la” (se repite 6117 veces). Los nombres de los personajes que más se mencionaron fueron los siguientes: Aureliano (794 veces), Úrsula (514 veces) y Arcadio (480 veces).

En el gráfico de la figura 6 se observa la frecuencia de repetición de palabras en valores porcentuales, de las cuales se observa lo siguiente:

- 54% de las palabras no se repiten en el texto (8142 palabras).
- 13% de las palabras aparecen dos veces en el texto (1966 palabras).
- 8% de las palabras se repiten 3 veces (1273 palabras).
- 5% de las palabras se repiten 4 veces (810 palabras).
- 16% de las palabras se repiten más de 5 veces (2861 palabras).

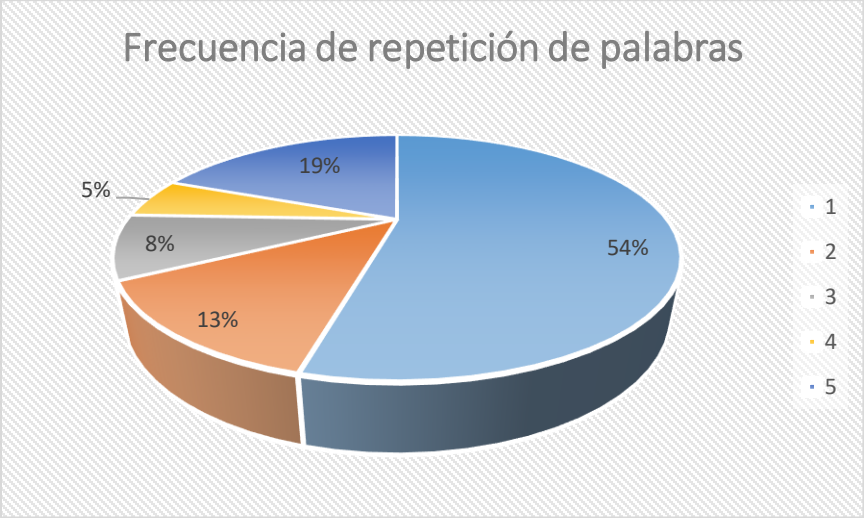

Figura 6. Gráfica que muestra el porcentaje de repetición de las palabras en el texto.

Conclusiones

Cloudera CDH es una herramienta poderosa que permite realizar el análisis de datos utilizando el ecosistema de Apache Hadoop y el paradigma de programación de MapReduce. Cloudera cuenta con sus propias herramientas que permiten realizar análisis más complejos.

Los resultados obtenidos, aunque son muy básicos, pueden servir para el desarrollo de técnicas de análisis de calidad de textos o para el desarrollo de criterios de calidad de los autores.

Debido a la importancia y actualidad de estos temas es importante utilizar herramientas que acerquen el enfoque de big data y ciencia de datos a los estudiantes por lo que el objetivo principal era demostrar la facilidad de uso de las plataformas para el análisis de datos.

Referencias

[1] IBM Analitycs, (2019), “IBM – Ciencia de los Datos – IBM Analytics - México”, https://www.ibm.com/analytics/mx/es/technology/data-science/

[2] IBM Developer, (2019), “¿Qué es big data?”, https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html

[3] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[4] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In MSST (Vol. 10, pp. 1-10).

[5] Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. Hadoop Project Website, 11(2007), 21.

[6] Cloudera Inc., (2019), "CDH Overview | 5.13.x | Cloudera Documentation", https://www.cloudera.com/documentation/enterprise/5-13-x/topics/cdh_intro.html#xd_583c10bfdbd326ba--5a52cca-1476e7473cd--7f59

[7] Cloudera Inc., (2019), "Download QuickStarts for CDH 5.13 | Cloudera" https://www.cloudera.com/downloads/quickstart_vms/5-13.html

[8] Cloudera Inc., (2019), "Hadoop Tutorial", http://tiny.cloudera.com/hadoopTutorialSample

[9] Internet Archive, "Garcia Marquez Gabriel Cien Años De Soledad I", https://archive.org/stream/GarciaMarquezGabrielCienAnosDeSoledad1/garcia-marquez-gabriel-cien-anos-de-soledad1_djvu.txt