Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ciencias y Sistemas

Seminario de Sistemas 2 Sección A

Primer Semestre 2023



**Práctica 2**

**Apache - Hadoop**

Nombre: William Alejandro Borrayo Alarcón

Carné: 201909103

Contenido

[Procedimiento realizado 2](#_Toc131080413)

[Preparación de Hadoop en contenedor 2](#_Toc131080414)

[Procedimiento para el conteo de palabras en el archivo Correos.txt 7](#_Toc131080415)

[Procedimiento para el conteo de palabras en el archivo Punteos.txt 14](#_Toc131080416)

[Archivo de comandos utilizados 20](#_Toc131080417)

[Análisis acerca de los resultados de cada archivo 20](#_Toc131080418)

[Conclusiones acerca de los resultados de cada archivo 22](#_Toc131080419)

[Conclusiones acerca del uso de Hadoop en BigData 22](#_Toc131080420)

[Bibliografía 22](#_Toc131080421)

# Procedimiento realizado

Para realizar este procedimiento es necesario tener Docker instalado.

## Preparación de Hadoop en contenedor

1. Descarga de imagen de Hadoop: Si vemos el mensaje ‘Already exists’ es porque ya la tenemos descargada.

Texto

Descripción generada automáticamente

1. Creación del contenedor (ejecución de la imagen descargada): En el comando se utiliza un parámetro para darle un nombre específico al contenedor: ‘hadoop’. Si se creó correctamente veremos la consola de esta manera:

Texto

Descripción generada automáticamente

También en localhost:50070 veremos esto:

Captura de pantalla de computadora

Descripción generada automáticamente

En utilities, browser directory se mostrará así:

Captura de pantalla de un celular

Descripción generada automáticamente

1. Luego de esto se crea una carpeta con el nombre ‘201909103\_Practica2’ en el contenedor, aquí se guardarán algunos archivos necesarios (archivos de entrada y el WordCount.java). Para la copia de los archivos se utiliza el comando: docker cp <Ruta en computadora> hadoop:/201909103\_Practica2 en una consola aparte.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

1. Se exportan las variables HADOOP\_HOME y CLASSPATH de la siguiente forma:

Texto

Descripción generada automáticamente

1. Ejecutamos el archivo .java, si se ejecuta bien deberá crear 3 archivos .class como se muestra en la imagen, la advertencia que se muestra no representa mayor problema:

Texto

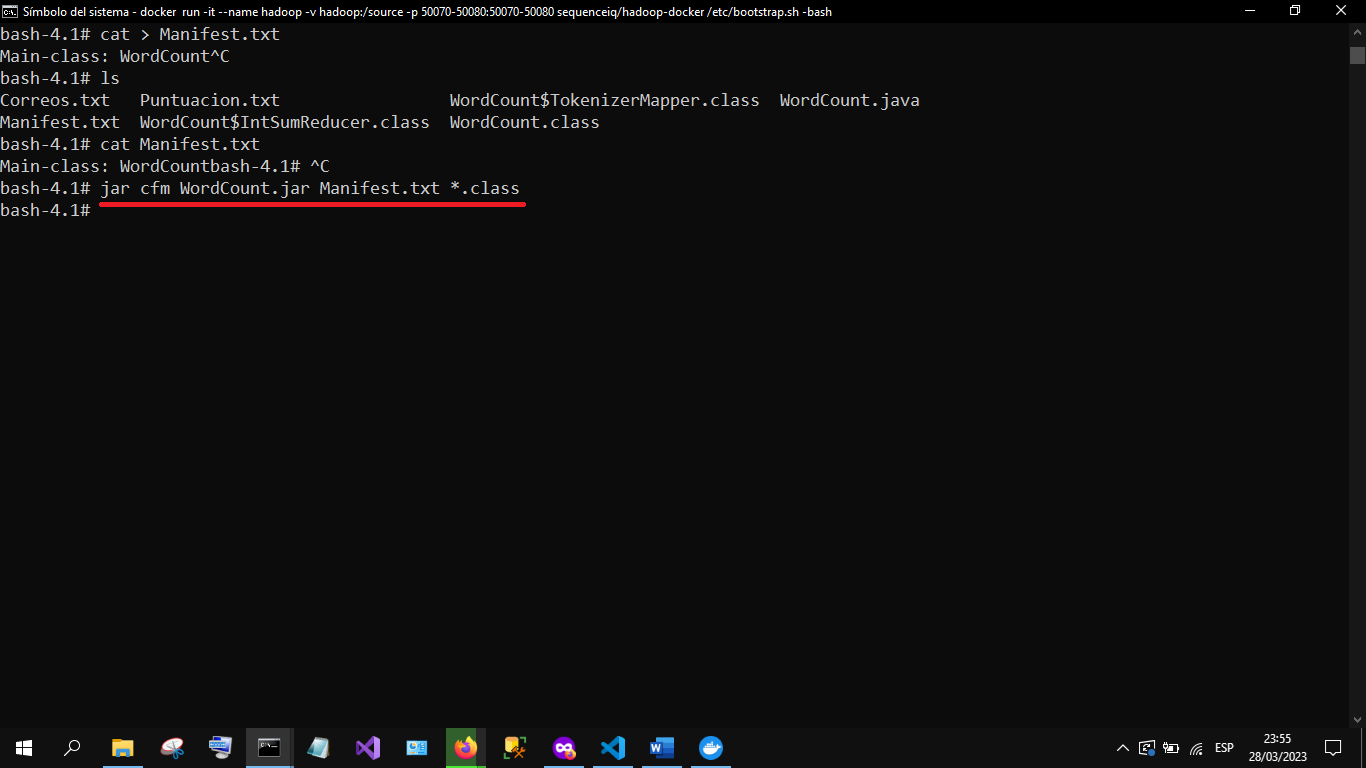
Descripción generada automáticamente

1. Se crea el archivo Manifest.txt con el contenido “Main-class: WordCount”. Se puede utilizar cat y luego de escribir el contenido la combinación Ctrl + D.

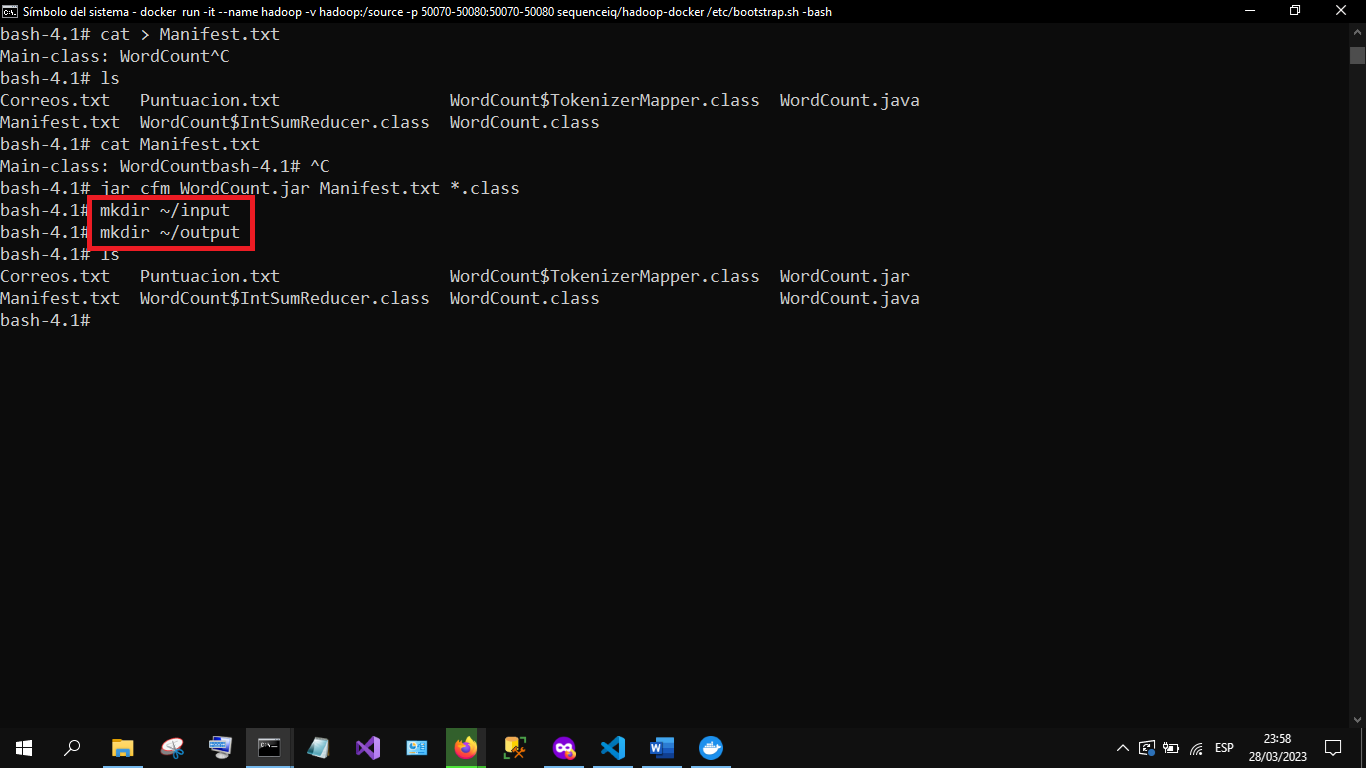
Texto

Descripción generada automáticamente

1. Se debe crear un archivo .jar utilizando el archivo Manifest.txt previamente creado:



1. Se crean las carpetas ‘input’ y ‘output’ en el home del usuario root:



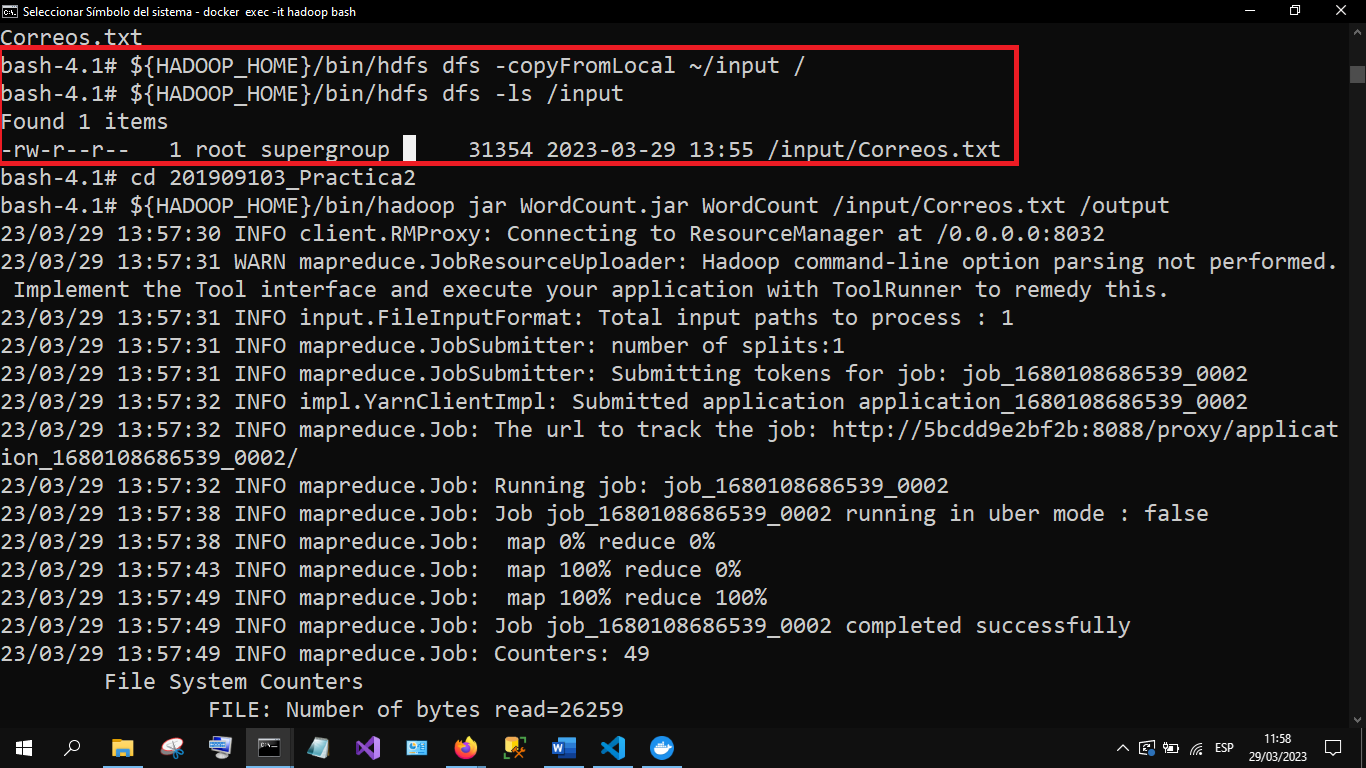
## Procedimiento para el conteo de palabras en el archivo Correos.txt

1. Se copia el archivo de entrada hacia la carpeta ‘input’ creada en el paso anterior:

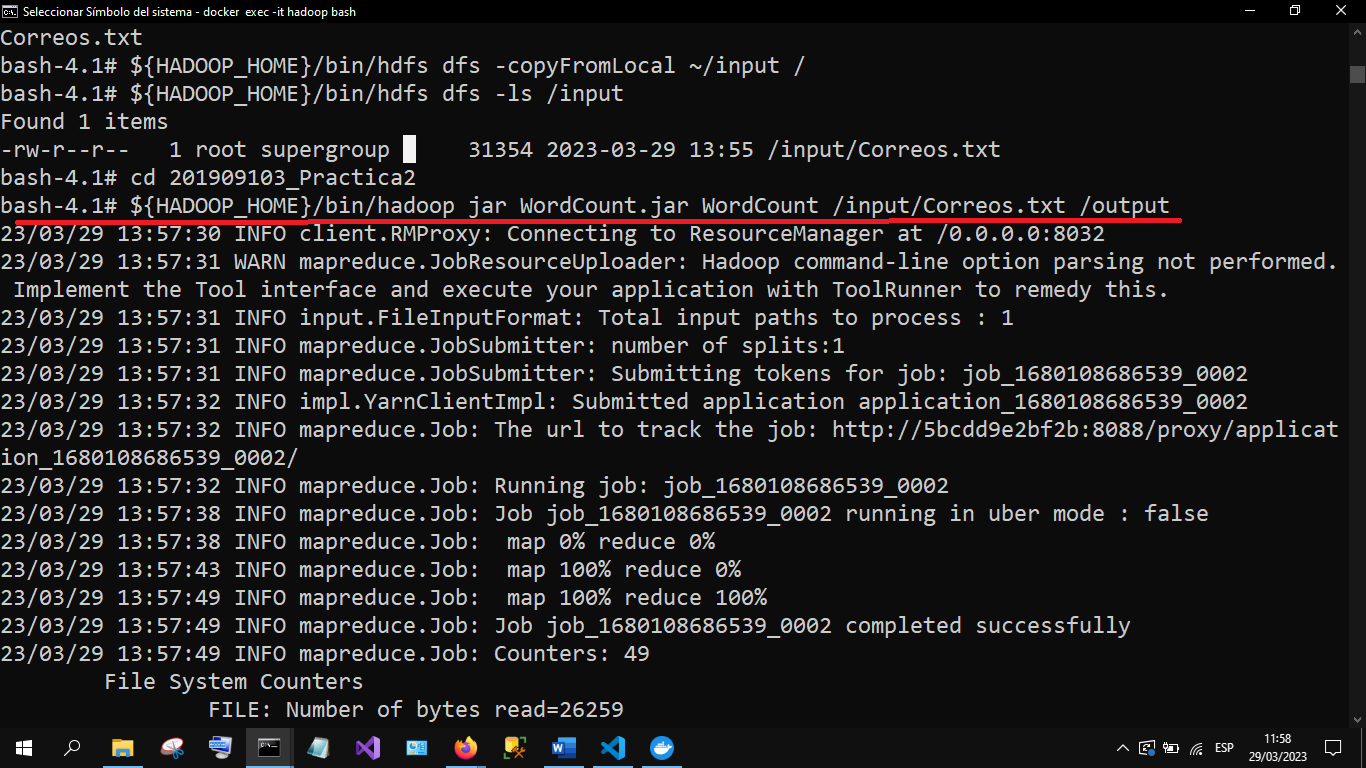
Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

1. Se copia el archivo de entrada hacia el sistema de archivos de hadoop. Y se verifica la correcta copia del archivo:



1. Para iniciar con el conteo de palabras es necesario estar en la carpeta que tiene el archivo WordCount.java (201909103\_Practica2). Desde ahí se ejecuta el comando: ${HADOOP\_HOME}/bin/hadoop jar WordCount.jar WordCount /input/Correos.txt /output:



Se iniciará el proceso de conteo, al final veremos detalles como la cantidad de bytes que fueron escritos, leídos, si hubieron errores u otras cosas:

Texto

Descripción generada automáticamente

1. Podremos ver los archivos que fueron generados con el primer comando, el segundo hará que se muestre el contenido del archivo de resultado:

Texto

Descripción generada automáticamente

1. Renombramos el archivo a ‘Resultado\_Correos.txt’ y verificamos que tenga el mismo contenido:

Texto

Descripción generada automáticamente

1. Copiamos el archivo de resultado a la carpeta output del home del usuario root, luego hacia la carpeta Salidas (es necesario crearla en el directorio inicial del contenedor), para su fácil extracción del contenedor.

Texto

Descripción generada automáticamente

1. Finalmente pasamos el archivo de resultado desde el contenedor hacia nuestra computadora, esto desde una nueva consola y hacia el directorio de nuestra preferencia.

Captura de pantalla de un celular

Descripción generada automáticamente

Podremos comprobar que el archivo ya está en nuestra computadora y con el mismo contenido:

Captura de pantalla de computadora

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente

1. Desde el navegador en localhost:50070, utilities, browse directory se pueden ver los archivos de esta manera:

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

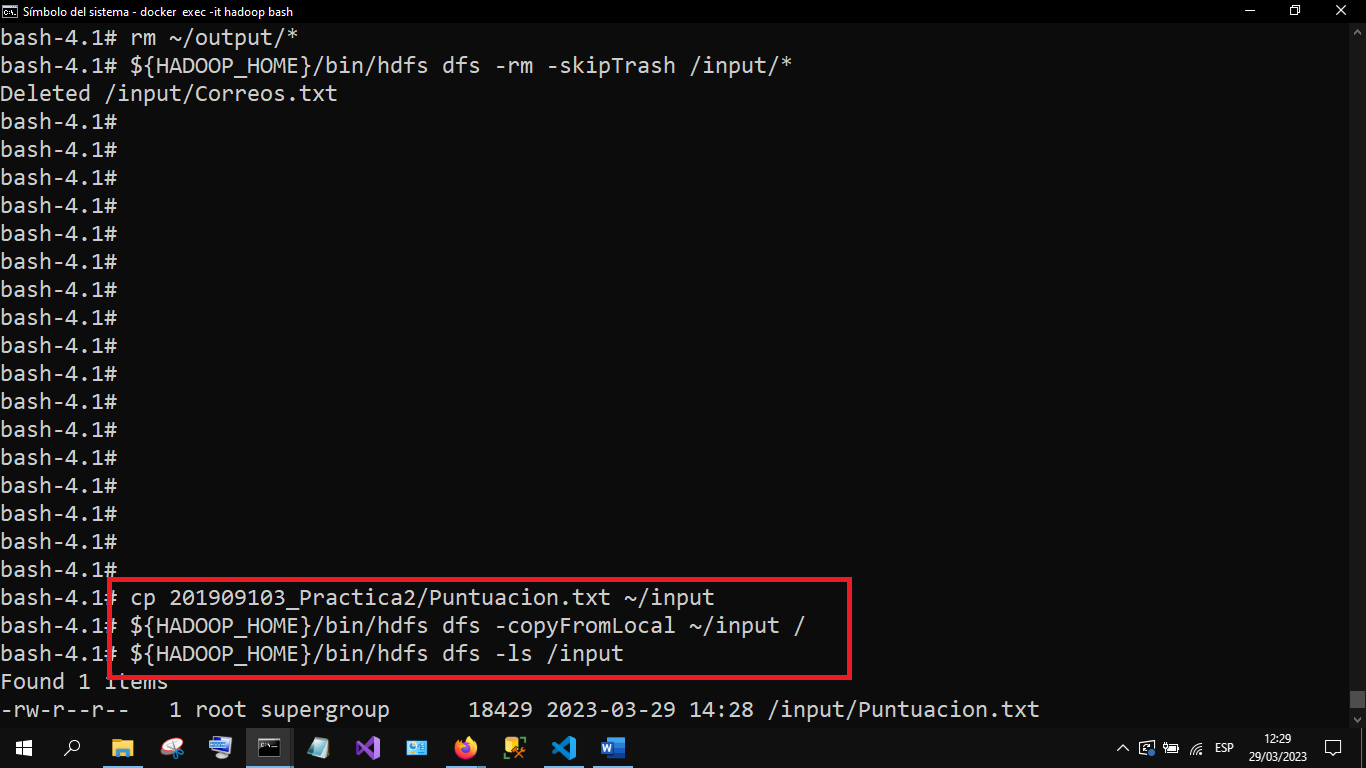
## Procedimiento para el conteo de palabras en el archivo Punteos.txt

1. Es necesario eliminar los archivos en las carpetas output e input, tanto del sistema de archivos de Hadoop como del home del usuario root de esta manera:

Texto

Descripción generada automáticamente

1. Se copia el archivo de entrada hacia la carpeta ‘input’ en el home del usuario root y luego hacia el sistema de archivos de hadoop:



1. Se inicia el conteo de palabras de la misma manera que en el archivo anterior:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

1. El archivo generado como resultado se verá de esta manera:

Texto

Descripción generada automáticamente

1. Renombramos el archivo y verificamos que el contenido sea el mismo:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

1. Copiamos el archivo de resultado a la carpeta output del home del usuario root, luego hacia la carpeta Salidas (es necesario crearla en el directorio inicial del contenedor), para su fácil extracción del contenedor.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

1. Finalmente pasamos el archivo de resultado desde el contenedor hacia nuestra computadora, esto desde una nueva consola y hacia el directorio de nuestra preferencia.

Texto

Descripción generada automáticamente

Podremos comprobar que el archivo ya está en nuestra computadora y con el mismo contenido:

Captura de pantalla de un celular

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente

1. Desde el navegador en localhost:50070, utilities, browse directory se pueden ver los archivos de esta manera:

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Captura de pantalla de un celular

Descripción generada automáticamente

## Archivo de comandos utilizados

<https://drive.google.com/file/d/1LnGh4jACTDktksCw8ECojQKIctKYfFOD/view?usp=sharing>

# Análisis acerca de los resultados de cada archivo

**Respecto al archivo Correos.txt:**

Estas son las 5 palabras encontradas en los correos con mayor frecuencia:



Son palabras que se relacionan mucho a la administración de hoteles.

Dado que la empresa SG-Food es una megaempresa destinada a la compra, distribución y comercialización de productos de diferentes marcas y categorías, y que en estos años debido a la crisis que atraviesa el país por el COVID-19 la industria alimentaria o de productos de uso personal han sido las más demandadas, se puede hacer la hipótesis que los compradores están abasteciendo sus hoteles con este tipo de productos, muy probablemente porque las personas no han podido viajar mucho y hospedarse en estos hoteles, entonces los dueños están buscando formas de mantener la seguridad de sus visitantes y que así las personas puedan hospedarse nuevamente.



Las siguientes 5 palabras más frecuentes también se relacionan a la idea anterior, algunas pueden tener contextos un poco abiertos y variados, pero tampoco representan cantidades demasiado considerables en los conteos ya que según el archivo de Puntuaciones.txt se tienen 9215 opiniones de clientes.

El resto de las palabras de igual forma pueden tener contextos muy variados y son menos utilizadas, aunque se tiene el caso también que Hadoop distinguió aquellas palabras acompañadas de comas o mal escritas. En este caso esas palabras tampoco representan una cantidad muy significativa, pero es conveniente que se realice un agrupamiento de estas palabras con la ayuda de otras herramientas y programas de software.

Para saber si los comentarios son positivos o negativos podemos considerar también el archivo de puntuaciones.

**Respecto al archivo Puntuacion.txt:**



Las puntuaciones con más frecuencia son las de 5 y 4, esto da a entender que la empresa está brindando un buen servicio a sus clientes. En la tercera posición se encuentra la valoración de 1, esto también indica que una cantidad considerable no está satisfecha con los servicios de la empresa, para analizar estos motivos puede ser necesario acudir a los comentarios de los clientes y buscar aquellos que utilicen palabras negativas como: n’t, not, aggravated, bad, entre otras, ya que son este tipo de palabras las que generalmente se emplean al expresar disgusto. Con una base de datos que relacione esta información puede ser más sencillo, pero en este caso los datos no están estructurados.

El último orden de las puntuaciones es 3 y 2, por lo que se sigue considerando que en su mayoría los clientes quedan satisfechos.

# Conclusiones acerca de los resultados de cada archivo

**Correos.txt**

* Se deben utilizar herramientas y programas para agrupar palabras similares (mal escritas o con comas), porque con mayores cantidades de datos sí puede perjudicar la toma correcta de decisiones.
* La identificación de palabras negativas ayuda a descubrir motivos de disgusto en los clientes.
* La mayoría de los consumidores son personas con trabajos relacionados a los hoteles.

**Puntuación.txt**

* La mayoría de las clientes está satisfecha con el servicio que la empresa les brinda.
* Una cantidad considerable de personas no está satisfecha con el servicio que les provee la empresa.
* Estos datos ayudan a tener conclusiones generales de manera más rápida gracias a que son numéricos y más directos. Para su profunda investigación se acude a archivos como el de Correos.txt

# Conclusiones acerca del uso de Hadoop en BigData

* Es una herramienta capaz de realizar diferentes tipos de análisis gracias a su funcionamiento con códigos de java.
* Ayuda a visualizar de una mejor forma los datos no estructurados para generar hipótesis y/o brindar conclusiones.
* Si el contexto y la complejidad de los datos es muy amplia para realizar análisis específicos, Hadoop funciona bien como guía para saber por dónde empezar.

# Bibliografía

* <https://hadoop.apache.org/>
* <https://www.docker.com/>