Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ciencias y Sistemas
Seminario de Sistemas 2 Sección A Ing. Luis Alberto Vettorazzi España Aux. Escarleth Andrea Velasco Campos

Práctica 2
Apache - Hadoop

OBJETIVOS:

- Que el estudiante aprenda cómo trabaja la herramienta Hadoop para el manejo de Big Data.
- Que el estudiante experimente y aprenda cómo trabajar con datos no estructurados.
- Que el estudiante experimente el análisis de datos y cómo llevarlo a una forma entendible.

DESCRIPCIÓN:

Debido a la buena implementación que se realizó con el proyecto de la empresa SG-Food, la empresa ha quedado satisfecha. Debido a esto lo ha contactado nuevamente pero esta vez para el procesamiento de grandes cantidades de datos que más adelante se detallarán.

Para el análisis que se le solicita se le proveerán distintos archivos con ***datos no estructurados*** en los cuales se necesita de un entorno de trabajo para software como Hadoop para el procesamiento de estos debido a la cantidad y estructura de estos.

IMPLEMENTACIÓN SUGERIDA

Para un mejor control podrá utilizar Docker para tener un contenedor con Hadoop (como se vio en clase) o instalar Hadoop de manera local.

- Se podrá utilizar el sistema operativo que usted desee.
- Instale Docker.
- Descargue la imagen de Hadoop.
- Cree sus archivos .java y junto con los archivos de entrada proceda a copiarlos en su imagen de Hadoop.
- Corra la imagen de Hadoop.

FLUJO DE DATOS

Diagrama 1.

ARCHIVOS

Se tendrán 2 archivos en los cuales se detalla la siguiente información:

- Archivo 1 ***Correos.txt***: este será un archivo que contendrá el texto de correos sobre la opinión de personas acerca de un hotel y se necesita darle un procesamiento correcto.

- Archivo 2 ***Puntuacion.txt***: este será un archivo que contendrá las puntuaciones en un rango de 1-5 que dan personas acerca de un hotel y se necesita darle un procesamiento y análisis correcto.

Para los dos archivos se requiere que se haga un conteo de palabras con más repitencia en cada uno de ellos para llegar a interpretar el porqué de estas repitencias en cada uno de sus ámbitos.

RESTRICCIONES

- Se debe utilizar ***Hadoop*** como entorno de trabajo.
- Los pasos del entregable que se le solicita deben de ir bien detallados y en el caso de las capturas **todas** deben **contener la fecha y hora** de su computadora en todo momento.
- Los resultados deben tener el siguiente formato ***Resultado_[nombre del archivo].txt*** donde ***#n*** es el número del archivo.

ENTREGABLES

- Manual con los pasos más importantes que realizaron para llegar al resultado final del procesamiento, este debe contener:
    - Capturas de pantalla sobre el procedimiento realizado y su descripción detallada.
      Obligatorio captura donde se pueda ver el Browse HDFS (i*magen 1*).
      *---Todo esto solo es **necesario para cada** archivo.---*

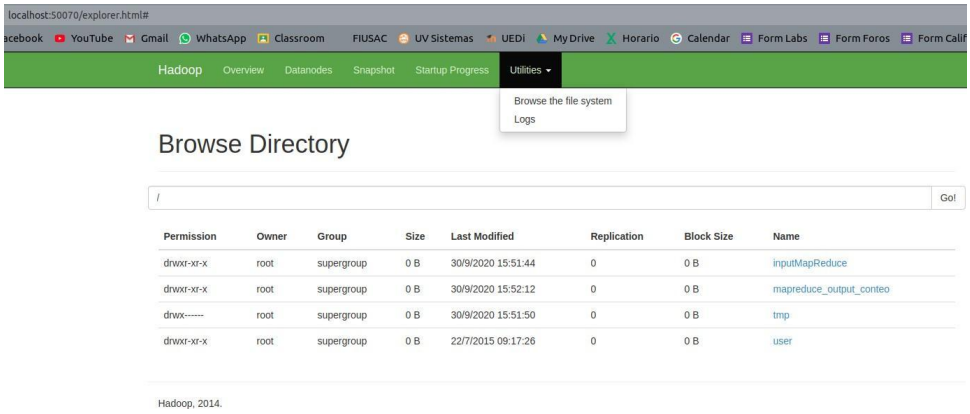

Imagen 1.

    - Conclusiones acerca de los resultados de cada archivo.
    - Conclusiones acerca del uso de Hadoop en BigData.

- Resultados de cada archivo con el formato especificado ***Resultado_[nombre del archivo].txt***

CONSIDERACIONES

- La entrega es individual. **No habrá prórroga.**
- Todas las dudas con respecto a esta Fase deberán ser planteadas en los foros creados en la plataforma UEDI.
- Enviar la práctica vía UEDI en un rar|zip con el nombre: **[SS2]Practica2_carne.rar el día Domingo 02 de abril de 2023 a las 23:59 horas.**
- Entregas tarde no se calificarán.
- De encontrar copias se tendrá una nota de 0 y el reporte a la escuela de sistemas.