Instituto Tecnológico de Costa Rica,
Área Académica de Ingeniería en Computadoras,
Bases de Datos : CE3101
Reporte de Investigación
Angelo Ortiz Vega

Descripción: El presente documento corresponde al Reporte de Investigación sobre el tema de Big Data el mismo consiste en un documento con evidencia del desarrollo del taller o descripción de los pasos solicitados por el equipo que presenta el taller en caso de tener problemas para poder llevar a cabo el taller.

Desarrollo:

¿Qué es Big Data? Se refiere a una acumulación de datos de tal extensión y complejidad que no se puede analizar por medios tradicionales, por ejemplo información de redes sociales, información de teléfonos celulares, compras y ventas, entre otros. Gran parte de los datos son obtenidos por data mining el cual es procesos utilizados por compañías para transformar datos crudos en información útil, se realiza de tal forma que se encuentren patrones entre los datos relevantes.

Usos: Dentro de los principales usos que posee Big Data se centraliza en Publicidad y Ventas, ejemplo de ello es cuando aparecen anuncios de grandes compañías en redes sociales.

Pasos del Taller:

Se consideraron importantes los archivos a descargar:

- https://www.oracle.com/virtualization/technologies/vm/downloads/virtualbox-downloads.html
- https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html

1. Abrir máquina virtual en el computador.
2. Seleccionar *File -> Import* y abrir el archivo de máquina virtual que se descargó previamente.
3. Click en *START* para inicializar la máquina virtual.
4. Al cargar completamente aparece una terminal con la siguiente leyenda:

   HDP 2.5
   http://hortonworks.com
   To initiate your Hortonworks Sandbox session, please open a browser and enter this address in the browser's address field:
   http://127.0.0.1:8888/

5. Después de esto, se abre su navegador favoritos y abre la dirección http://localhost:8888
6. Se presiona el botón en el contenedor izquierdo de *LAUNCH DASHBOARD*
7. Se presiona el botón en el contenedor derecho de *QUICK LINKS*
8. Se selecciona AMBARI y se copia el nombre de usuario y contraseña del CLUSTER OPERATOR, en este caso sería **raj_ops**
9. Se inicia sesión con las credenciales de: username: **raj_ops** password: **raj_ops**
10. Se cambia la contraseña desde el puerto 4200

    ambari-admin-password-reset
    Se escribe la contraseña actual.
    Se escribe la nueva contraseña. En este caso es **admin**.
    Espera a que se ejecute la actualización de directorios.

11. Inicia sesión con los nuevos credenciales en puerto 8080: login
12. Se redirige al Dashboard de Ambari.
13. En la barra superior, se selecciona el botón de cuadrícula, y el subitem *File View*.
14. Dentro de *File View* se selecciona la carpeta maria_dev y se suben los archivos enviados por telegram con los nombres:

    drivers.csv
    timesheet.csv

    Cabe destacar que tiene que ser un archivo a la vez.

15. Se redirige al puerto 4200 y se copian las sentencias:

    su maria_dev
    cd
    pig

    Se inicializa pig grunt

16. Se deben copiar las siguientes sentencias en terminal

    ```
    touch sum_of_hours_miles
    vi sum_of_hours_miles
    drivers = LOAD 'drivers.csv' USING PigStorage(',');
    raw_drivers = FILTER drivers BY $0>1;
    drivers_details = FOREACH raw_drivers GENERATE $0 AS driverId, $1 AS name;
    timesheet = LOAD 'timesheet.csv' USING PigStorage(',');
    ```

raw_timesheet = FILTER timesheet by $0>1;

timesheet_logged = FOREACH raw_timesheet GENERATE $0 AS driverId, $2 AS hours_logged, $3 AS miles_logged;

grp_logged = GROUP timesheet_logged by driverId;

sum_logged = FOREACH grp_logged GENERATE group as driverId,

SUM(timesheet_logged.hours_logged) as sum_hourslogged,

SUM(timesheet_logged.miles_logged) as sum_mileslogged;

drivers = LOAD 'drivers.csv' USING PigStorage(',');

pig -x mr -f sum_of_hours_miles

join_sum_logged = JOIN sum_logged by driverId, drivers_details by driverId;

join_data = FOREACH join_sum_logged GENERATE $0 as driverId, $4 as name, $1 as hours_logged, $2 as miles_logged;

dump join_data;

17. Después de esto te aparece el resultado de correr el script, el cual su función inicial era observar el promedio de horas conducidas por un chofer de camión.

No pude realizar el taller en vivo ya que no tengo espacio en el Disco de mi computadora, y a pesar de los inconvenientes que se presentaron en la explicación.

Adjunto ScreenShot de descarga.