Tipo : Guía de laboratorio
Capítulo : Componentes de Hadoop
Duración : 30 minutos

I. OBJETIVO

Conocer cómo implementar los componentes de Hadoop.

II. REQUISITOS

Los siguientes elementos de software son necesarios para la realización del laboratorio:

- Virtual Box
- Hortonworks Sandbox for HDP
- Putty
- Winscp

III. EJECUCIÓN DEL LABORATORIO

1. Ingresar al servidor Hortonworks, para esto deberá ingresar a la siguiente dirección:

   http://127.0.0.1:8080

2. Ir al menú de Canvas

3. Seleccionar la opción File View

4. Luego de esto navegar hasta seleccionar la carpeta /user/maria_dev, luego de esto dar click en el botón Upload, seleccionar los archivos drivers.csv y timesheet.csv de la unidad compartida.

5. Ir al menú Canvas, seleccionar la opción Hive View 2.0

6. Deberá crear una base de datos con el nombre clase01, para esto escribir el comando:

7. Seleccionar la base de datos clase01, ir a la opción Browse

8. Se deberá crear la tabla temp_drivers, para esto escribir el siguiente comando:

9. Luego de esto realizar una consulta para cargar la información del archivo drivers.csv hacia la tabla temp_drivers.

10. ¿Qué paso con el archivo drivers.csv, existe todavía?, Ir a la opción File View

11. Crear la tabla DRIVERS, para esto deberá crear una tabla con seis columnas, indicándoles los tipos de datos, por cada columna, para esto deberá escribir el comando:

12. Crear una consulta para extraer datos de TEMP_DRIVERS y guardarlos en DRIVERS, en cada columna.

13. Realizar una consulta para visualizar el contenido de los primeros 10 registros de la tabla DRIVERS

14. Realizar la carga de la siguiente tabla: TIMESHEET. Para esto repetir el paso 7 hasta el paso 13.

15. Crear una consulta de la tabla timeshett, para agrupar los datos por el campo driverId, con el fin de encontrar la suma de horas (hours_logged) y millas de puntaje(miles_logged) registrado.

16. Crear una consulta para realizar un JOIN de las tablas DRIVERS y TIMESHEET (DRIVERID, NAME, HOURS_LOGGED, MILES_LOGGED)