Desafío - Sqoop y Hive

- Para realizar este desafío debes haber estudiado previamente todo el material disponibilizado correspondiente a la unidad.
- Una vez terminado el desafío, comprime la carpeta que contiene el desarrollo de los requerimientos solicitados y sube el `.zip` en el LMS.
- Desarrollo desafío:
    - El desafío se debe desarrollar de manera Individual/Grupal.
    - Debes tener precaución con las rutas y el sistema operativo que utilices.

Preliminares: Sobre los datos del ejercicio

Para esta sesión vamos a trabajar con los datos de la aplicación `last_fm`. Los datos se encuentran en un motor de bases de datos `Postgresql` con la siguiente especificación:

- **Url del motor:**

```
ls-07d09717f869f2ab37ff6e2b5aa45a1441ebfef3.cll3q1dgryg1.us-east-1.rds.amazonaws.com
```

- **Database:** `last_fm`.
- **Tablas:** `user_profile y artists`.
- **Usuario:** `alumnos_big_data`.
- La **contraseña** será administrada por el profesor.
- **Puerto de conexión:** 5432.

Estructura de las tablas

- **Tabla `user_profile`:** Cada registro representa un usuario único que utilizó la aplicación.

user_mboxsha1	gender	age	country	signup
000063d3fe1cf2ba248b9e3c3f0334845a27a6bf	m	19	Mexico	Apr 28, 2008

Donde `user_mboxsha1` es el identificador único del usuario en la aplicación. `gender` es el género (masculino, femenino y 99999 para identificar los campos vacíos). `age` es la edad del usuario. country es el país de origen declarado por el usuario y `signup` es la fecha de ingreso a la aplicación.

- **Tabla `artists`:** Cada registro representa la cantidad de reproducciones de un artista específico por un usuario específico.

user_mboxsha1	musicbrainz_artist_id	artist_name	plays
000063d3fe1cf2ba248b9e3c3f0334845a27a6bf	af8e4cc5-ef54-458d-a194-7b210acf638f	cannibal corpse	48
000063d3fe1cf2ba248b9e3c3f0334845a27a6bf	eaaee2c2-0851-43a2-84c8-0198135bc3a8	elis	31

Donde `user_mboxsha1` es el identificador único del usuario en la aplicación que escuchó a X artista. `musicbrainz_artist_id` es un identificador único del artista en la base de datos de MusicBrainz. artist_name es el nombre del artista y `plays` es el número de reproducciones del usuario para X artista.

Las tablas presentan una relación $user\_profile \xrightarrow{1:N} artists$.

Requerimientos

Ejercicio 1 - Preparación del ambiente de trabajo

- Genere una instancia EMR en AWS con una configuración avanzada que incluya Hue, Hive, Sqoop, Pig y Hadoop.
- Verifique la existencia de los binarios de los programas mencionados con `which`. Genere un pantallazo de este procedimiento.
- Habilite la interfaz gráfica de Hue que se presenta en AWS EMR utilizando un puerto dinámico descrito en la lectura. Genere un pantallazo de este procedimiento.

Ejercicio 2 - Importación de los datos con Sqoop

- Desde la instancia AWS EMR, importe ambas tablas en la base de datos `last_fm` utilizando `sqoop` a `hdfs` de su instancia AWS EMR.
- Por defecto las bases de datos migradas estarán en la ubicación del home. Liste los archivos y sus contenidos. Genere un pantallazo de este procedimiento.
- Desde Hue, genere las tablas `user_profile` y `artists`. Tenga en consideración el tipo de dato de cada columna, así como el hecho de que cada campo estará separado por `,`. Guarde los comandos en un archivo con la siguiente extensión: `create_last_fm_tables.hql`.
- Desde Hue, cargue los datos de cada tabla en HDFS con su respectivo esquema en Hive. Tenga en consideración el hecho que si tiene un error en la escritura de los datos, deberá modificar los permisos de usuario. Refiérase a la lectura para lograr este punto. Guarde los comandos en un archivo con la siguiente extensión: `load_last_fm_data_to_table.hql`.
- Desde Hue, muestre las primeras 10 observaciones para la tabla `user_profile` y `artists`. Genere un pantallazo de cada una. Guarde los comandos en un archivo con la siguiente extensión: `select_10_first_ from table.hql`.

Ejercicio 3 - Queries desde Hive

Desde la gerencia de Last.fm tienen una serie de dudas respecto al comportamiento y composición de sus usuarios. En su calidad de analista de datos, le entregan una minuta con las siguientes preguntas:

- Desde gerencia buscan identificar en qué países es más prevalente el uso de last.fm. Guarde los comandos en un archivo con la siguiente extensión: `count_users_by_country.hql`.

- Una de las tendencias más preocupantes desde la gerencias es el hecho que algunos usuarios optan por no declarar su edad ni sexo. Reporte cuántos usuarios han optado por esto. Guarde los comandos en un archivo con la siguiente extensión: `count_null_users_by_gender_and_age.hql`.

- De una manera similar, desde gerencia le piden identificar en qué países es más prevalente este escenario donde por lo menos una de las dos columnas presente datos perdidos. Guarde los comandos en un archivo con la siguiente extensión: `count_nulls_by_country.hql`.

- Lamentablemente el equipo TI de Last.fm tuvo un desperfecto con las máquinas y se perdieron algunos registros de la tabla `artists`. Desde gerencia le piden contar la cantidad de usuarios que estando registrados en la tabla `user_profile`, no figuran en los registros de `artists`. Guarde los comandos en un archivo con la siguiente extensión: `identify_unmatched_users.hql`.

- Desde una compañía discográfica le piden información sobre el consumo asociado a `metallica`. Específicamente buscan tres respuestas:
    - ¿En qué tipo de usuario (a nivel de país, género y edad) hay más individuos que escuchan Metallica? Guarde los comandos en un archivo con la siguiente extensión: `crosstabulate_at_user_level_metallica.hql`.
    - ¿Qué países presentan una mayor y menor cantidad de reproducciones promedio de Metallica? Guarde los comandos en un archivo con la siguiente extensión: `metallica_average_plays_by_country.hql`.
    - ¿Cuál es la correlación entre edad y cantidad de reproducciones de Metallica a nivel de país? ¿En qué países existe una correlación negativa? Guarde los comandos en un archivo con la siguiente extensión: `correlate_metallica_age_and_play_by_country.hql`.

Ejercicio 4 - Descarga de resultados y eliminación de Instancia

- Saque un screenshot del término de la instancia AWS EMR.

- En un `.zip`, adjunte los siguientes archivos:
  - La declaración de la tabla, la importación de los datos y queries en Hive en un archivo con extensión `hql`. Cada archivo debe seguir el nombre indicado en el enunciado.
  - El screenshot del listado de archivos en HDFS.
  - El screenshot de la identificación de los binarios.
  - El screenshot de la habilitación de la interfaz gráfica.
  - El screenshot de la terminación de la instancia.