PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE | la clase ejecutiva

Instrucciones Miniproyecto 1

Introducción

En este miniproyecto, deberá procesar datos que contienen las preferencias musicales de un conjunto de usuarios. En particular, trabajará con un gran volumen de registros provenientes de la red social de servicios musicales *Last.fm* [1], los cuales contienen información acerca de las canciones escuchadas por cada usuario durante un período de tiempo. Utilizando este conjunto de datos, deberá obtener información estadística de las preferencias de los usuarios, y como éstos se distribuyen de acuerdo a su género y edad para un determinado artista. En esta experiencia se formula un grupo de preguntas relativas a los datos, las cuales deberán ser respondidas utilizando las herramientas *Apache Hive* y *Apache Pig*, pertenecientes al ecosistema Hadoop.

Análisis de Datos

El conjunto de datos que utilizaremos en esta oportunidad, corresponde a la base de datos generada por la red social de servicios musicales *Last.fm* [1]. Esta red social permite a sus usuarios la creación de un perfil musical basado en las canciones que estos van escuchando, ya sea desde su colección personal, o bien, desde el servicio de radio por internet que ofrece el portal. Las canciones escuchadas son almacenadas dentro de un registro del sistema, que permite determinar los artistas y canciones favoritas por cada usuario, y en base a sus gustos, hacerles recomendaciones. En esta oportunidad, los datos serán utilizados para obtener informaci ón estadística mediante las herramientas de análisis del ecosistema Hadoop vistas hasta ahora.

Los datos de esta colección vienen almacenados dentro de dos archivos, cuyo contenido es descrito a continuación:

- `userid-profile.tsv`: archivo de texto con información acerca de 992 usuarios registrados en el sistema. Cada línea contiene 5 campos separados por el caracter `tab` (`\t`), los que son listados a continuación:

  ```
  userid | gender | age | country | signup
  ```

  Es importante destacar que los campos de este archivo pueden venir vacíos, a excepción del campo *userid* que siempre estará presente.

- `userid-timestamp-artid-artname-traid-traname.tsv`: archivo de texto con información de las canciones escuchadas por cada usuario. Posee un total de 19.150.868 de líneas, en donde cada una de ellas, contiene 6 campos separados por el caracter `tab` (`\t`), los que son listados a continuación:

  ```
  userid | timestamp | musicbrainz-artist-id | artist-name |
  musicbrainz-track-id | track-name
  ```

Para comenzar a trabajar, deberá descargar y descomprimir el archivo que contiene los datos. Para facilitar esta tarea, junto a este enunciado entregaremos unas plantillas ipython notebook, en cuyas celdas se ha dispuesto el código necesario para instalar el software base de Hadoop, y para la descarga y descompresión de datos dentro de la máquina virtual de Colab.

Actividades

A continuación se presentan 3 actividades que deberá realizar utilizando las herramientas Apache Hive y Apache Pig del ecosistema Hadoop. Para ello, deberá programar las rutinas que estime conveniente en cada lenguaje, y responder las preguntas que aparecen en el notebook, dentro de una celda de texto.

A continuación se especifican las actividades que deberá realizar para poder responder las preguntas:

1. Elabore un ranking con los 10 artistas más populares, según el número de reproducciones de sus canciones. Como salida deberá generar una tabla con dos columnas que contengan el nombre del artista y la cantidad de reproducciones. Esta lista deberá estar ordenada de forma descendente y SOLO deberá contener los 10 artistas más populares.

2. Para el artista más popular encontrado en la actividad anterior, obtenga la distribución de sus auditores según su género. Es decir, del subconjunto de usuarios que han escuchado al artista más popular, deberá determinar cuántos son hombres y cuántos son mujeres. Para los casos en que el campo *gender* no esté definido, deberá omitir el dato en cuestión.

3. Para el artista más popular, obtenga la distribución de sus auditores según su edad. Es decir, del subconjunto de usuarios que han escuchado al artista más popular, deberá determinar cuántos hay por cada edad. Para los casos en que el campo *age* no esté definido, deberá omitir el dato en cuestión. Además, la tabla generada deberá estar ordenada de forma ascendente según edad.

Requisitos de entrega

Para la entrega de este miniproyecto, deberá enviar el ipython notebook con todos sus códigos ejecutados, de manera tal que las salidas y/o resultados de cada comando estén visibles en el archivo, y con las respuestas a las preguntas planteadas en celdas de texto.

Para esta entrega se proporcionarán las plantillas hadoop-apache-hive.ipynb y hadoop-apache-pig.ipynb. En ellas vendrán los códigos necesarios para la instalación del software base, para la descarga de datos y las preguntas a responder.

Es requisito de la entrega que tanto los comandos *HiveQL* como *Pig Latin* sean generados dentro de la máquina virtual de Colab como scripts con extensión *.sql* y *.pig*. Para ellos, deberá escribir sus scripts dentro de una celda del notebook colab, y almacenarla dentro de un archivo utilizando el magic cell *writefile*, como se muestra en el siguiente ejemplo:

```
% %writefile load_data.pig

raw = LOAD 'dataset.tsv' AS (field_1, field_2, field_3)
```

Referencias

[1] https://
www.last.fm