Taller 1 - Big Data

Diplomado de Análisis y modelamiento de datos - Databricks
José Miguel Tobar - jtobarr@utem.cl

- Fecha: 11 de Mayo 2022
- Fecha de entrega: 18 de Mayo 2022
- Formato de entrega: Link de notebook compartido desde databricks, se debe entregar en Canvas.
- Importante: Escribir su nombre en el notebook.

Dataset

Todos los datasets se encuentran en el siguiente link y deben usar su cuenta utem para acceder:

https://drive.google.com/drive/folders/1cOfvk6wQVxyGmdgc5fNmZIkIPbaovWRB?usp=sharing

El primer dataset que se utilizará en este taller contiene los ganadores de los premios Nobel entre el año 1901 a 2016. El archivo a cargar en databricks se llama `nobel.csv`.

El segundo dataset corresponde a una lista de canciones de spotify con sus correspondientes atributos como nombre, artista, popularidad, año, que tan bailable es, entre los principales.

Para hacer el análisis de estos datos se debe cargar un nuevo notebook de tipo SQL. Recuerde usar Markdown para describir las distintas etapas del proceso.

Carga de archivo

Abrir el espacio de trabajo en https://community.cloud.databricks.com/login.html y loguearse con las credenciales utilizadas durante la clase.

1.1.- Cargar en el cluster el archivo `nobel.csv` en una nueva tabla de nombre nobel. Es importante inferir los datos al crear la tabla y

Exploración los datos (50 pts)

Responda a través de consultas SQL:

2.1.- Describa los tipos de datos de la tabla creada. (Comando para describir tablas)

2.2.- ¿Cuál es el total de los datos contenidos en la tabla?

2.3.- Muestre los primero 10 registros

2.4. ¿Qué género ha ganado más premios Nobel?

2.5. ¿En qué año ganó Albert Einstein el premio Nobel?

2.6. ¿Cuál es el país donde han nacido más premios Nobel?

2.7. ¿Cuál fue la primera mujer en ganar un premio Nobel?

Visualización de los datos (30 pts)

3.1- Generar un gráfico de líneas que represente la evolución de premios nobel ganados.

3.2.- Genere un gráfico de barras que muestre los 10 países que hayan ganado más premios nobel de forma descendente.

3.3.- Genere dos gráficos Pie que muestren la distribución de cuántos premios nobel hay por las distintas categorías. El primer gráfico debe mostrar la distribución de antes de 1970 y el segundo después de 1970. ¿Qué diferencias existen?

Archivo de mayor tamaño (20 pts)

4.1.- Cargar en el cluster el archivo `spotify.csv` en una nueva tabla de nombre spotify. Es importante inferir los datos al crear la tabla.

4.2.- Realice una exploración del archivo y muestre 3 insights encontrados en los datos.

Compartir resultados

5.1.- Publicar los resultados en databricks y entregar vía Canvas.