

Clase 8

Consigna: Por cada ejercicio, escribir el código y agregar una captura de pantalla del resultado obtenido.

Diccionario de datos:

<https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?select=results.csv>

1. Crear las siguientes tablas externas en la base de datos f1 en hive:
 - a. driver_results (driver_forename, driver_surname, driver_nationality, points)
 - b. constructor_results (constructorRef, cons_name, cons_nationality, url, points)
2. En Hive, mostrar el esquema de driver_results y constructor_results
3. Crear un archivo .bash que permita descargar los archivos mencionados abajo e ingestarlos en HDFS:
 - results.csv
<https://data-engineer-edvai.s3.amazonaws.com/f1/results.csv>
 - drivers.csv
<https://data-engineer-edvai.s3.amazonaws.com/f1/drivers.csv>
 - constructors.csv
<https://data-engineer-edvai.s3.amazonaws.com/f1/constructors.csv>
4. Generar un archivo .py que permita, mediante Spark:
 - a. insertar en la tabla driver_results los corredores con mayor cantidad de puntos en la historia.
 - b. insertar en la tabla constructor_result quienes obtuvieron más puntos en el Spanish Grand Prix en el año 1991
5. Realizar un proceso automático en Airflow que orqueste los archivos creados en los puntos 3 y 4. Correrlo y mostrar una captura de pantalla (del DAG y del resultado en la base de datos)