75.06/95.58 Organización de Datos
Finger 3: Apache Spark SQL y Catalyst Optimizer

Puntos: Se otorgarán los siguientes puntos en orden de entrega 5,3,2,1,1,1.
Tiempo Estimado de Resolución: 3 hs.

Utilizando el API Apache Spark SQL, trabajaremos sobre el set de datos '2015 Flight Delays and Cancellations'
Los datos del set están disponibles en el siguiente link de kaggle:

https://www.kaggle.com/usdot/flight-delays

Considerando la siguiente información del set de datos

- airlines.csv
- airports.csv
- flights.csv

Se plantean los siguientes puntos a realizar:

1. Mostrar los 5 aeropuertos de origen que tienen mayor cantidad de cancelaciones.
2. Mostrar el nombre de aerolíneas y la cantidad de vuelos desde Atlanta (ATL) a Los Ángeles (LAX) ordenadas cantidad de vuelos
3. Mostrar y Analizar el Query Plan del punto 2 para entender las optimizaciones que realiza Catalyst Optimizer, contestando las siguientes preguntas:
    a. ¿Se realiza alguna optimización lógica, como filter pushdown? ¿En qué etapa?
    b. ¿Que tipo de Join Físico se realiza? ¿En qué etapa?

En https://spark.apache.org/docs/latest/sql-programming-guide.html puede encontrar la documentación de Spark para apoyar lo dado en clase.

Recomendamos utilizar para la resolución un Notebook de Jupyter Notebook que permita organizar claramente el output y comentar los hallazgos, de tal forma que puedan compartirse, por ejemplo en Github.
Los ejercicios deben resolverse utilizando Apache Spark SQL a partir de los visto en clase.