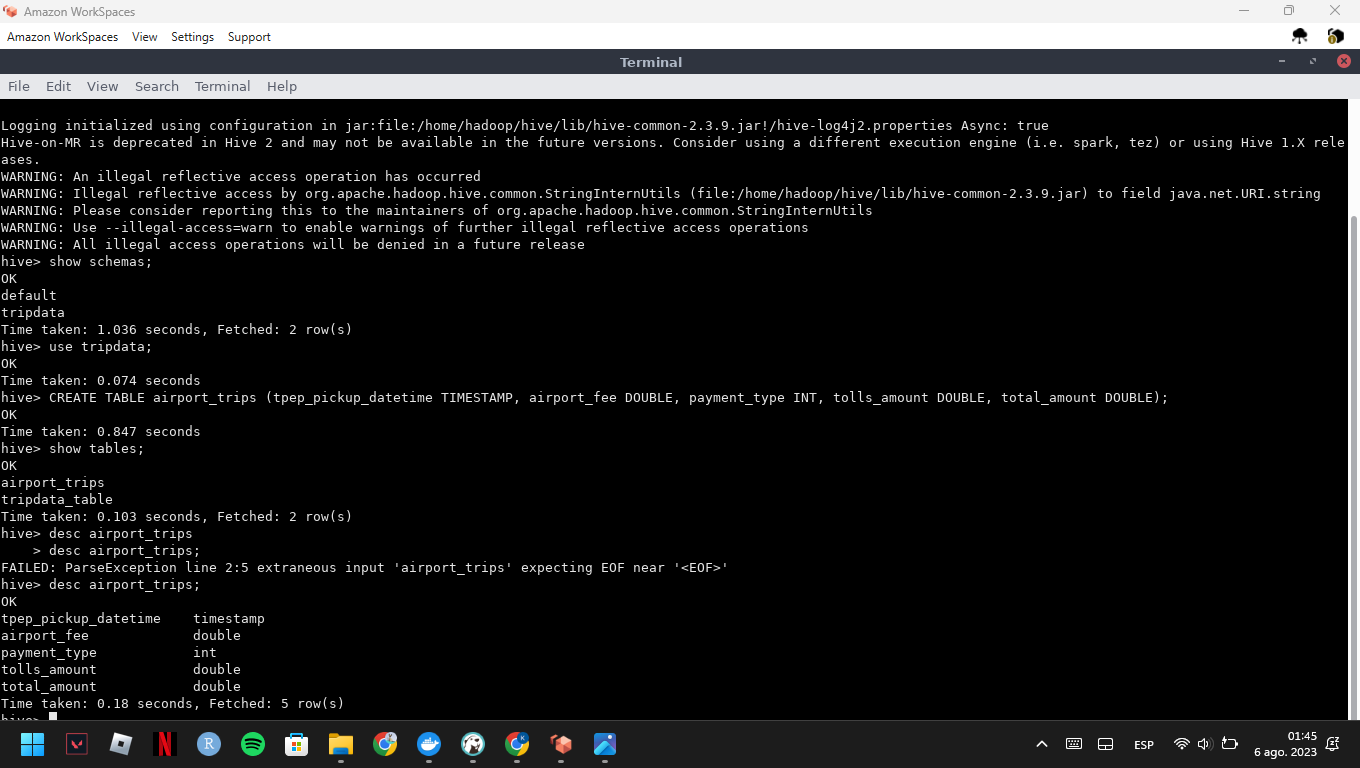
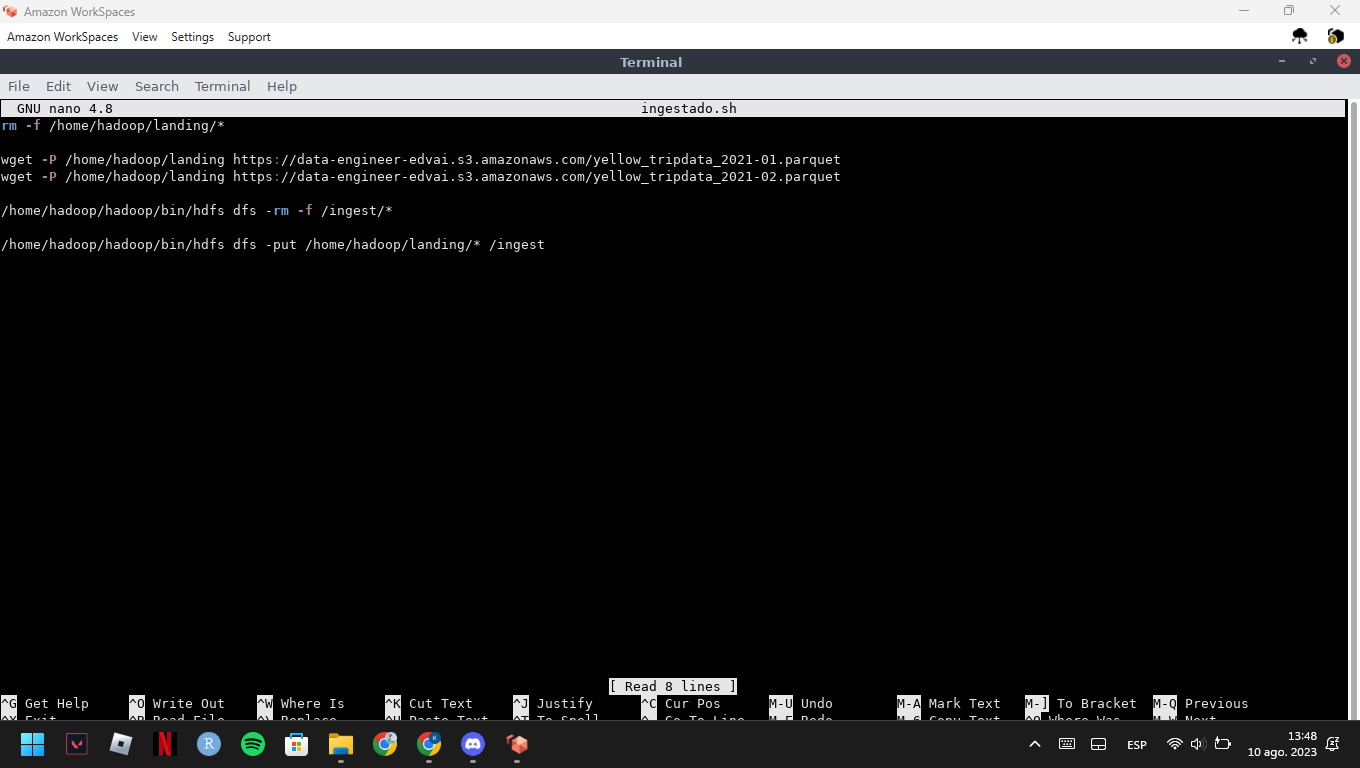
**Clase 7**

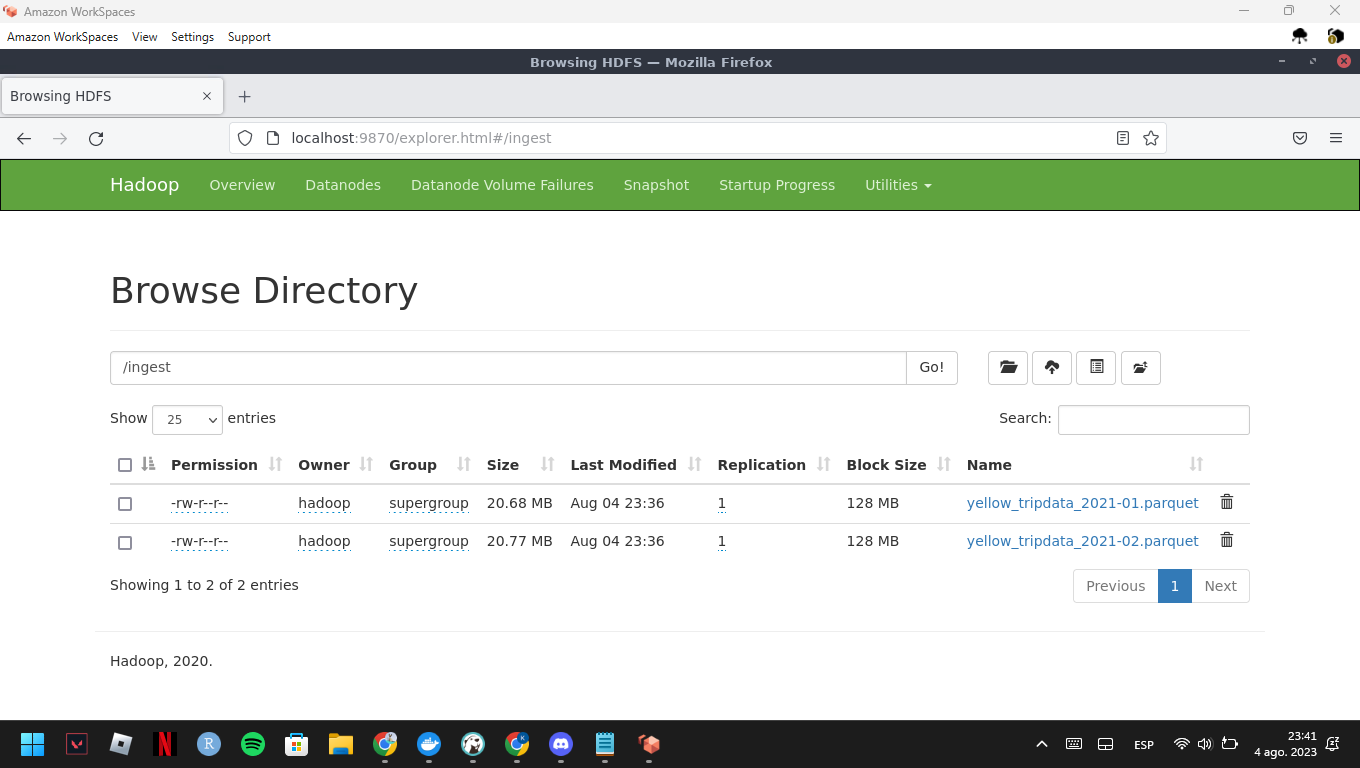
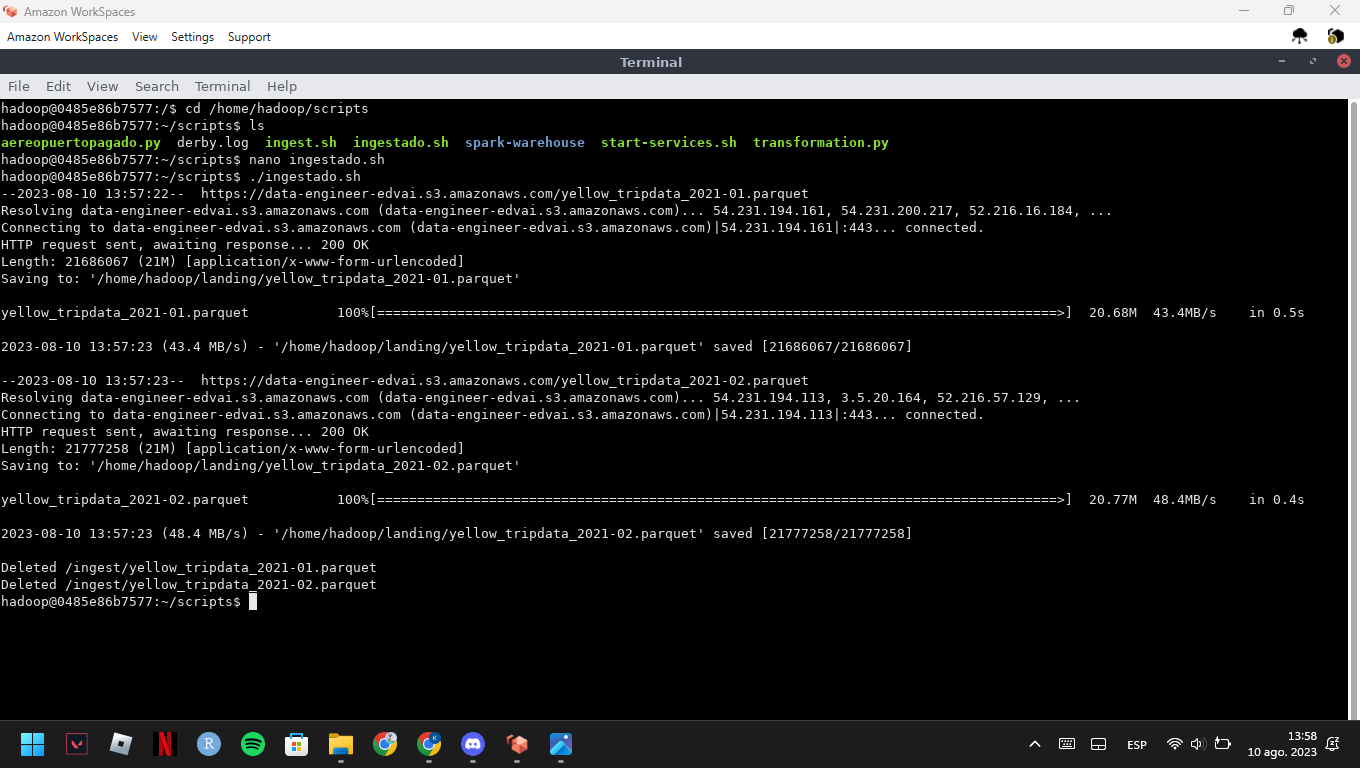
Consigna: Por cada ejercicio, escribir el código y agregar una captura de pantalla del resultado obtenido. Diccionario de datos: https://www.nyc.gov/assets/tlc/downloads/pdf/data\_dictionary\_trip\_records\_yellow.pdf

1. En Hive, crear la siguiente tabla (externa) en la base de datos tripdata: a. airport\_trips(tpep\_pickup\_datetetime, airport\_fee, payment\_type, tolls\_amount, total\_amount)

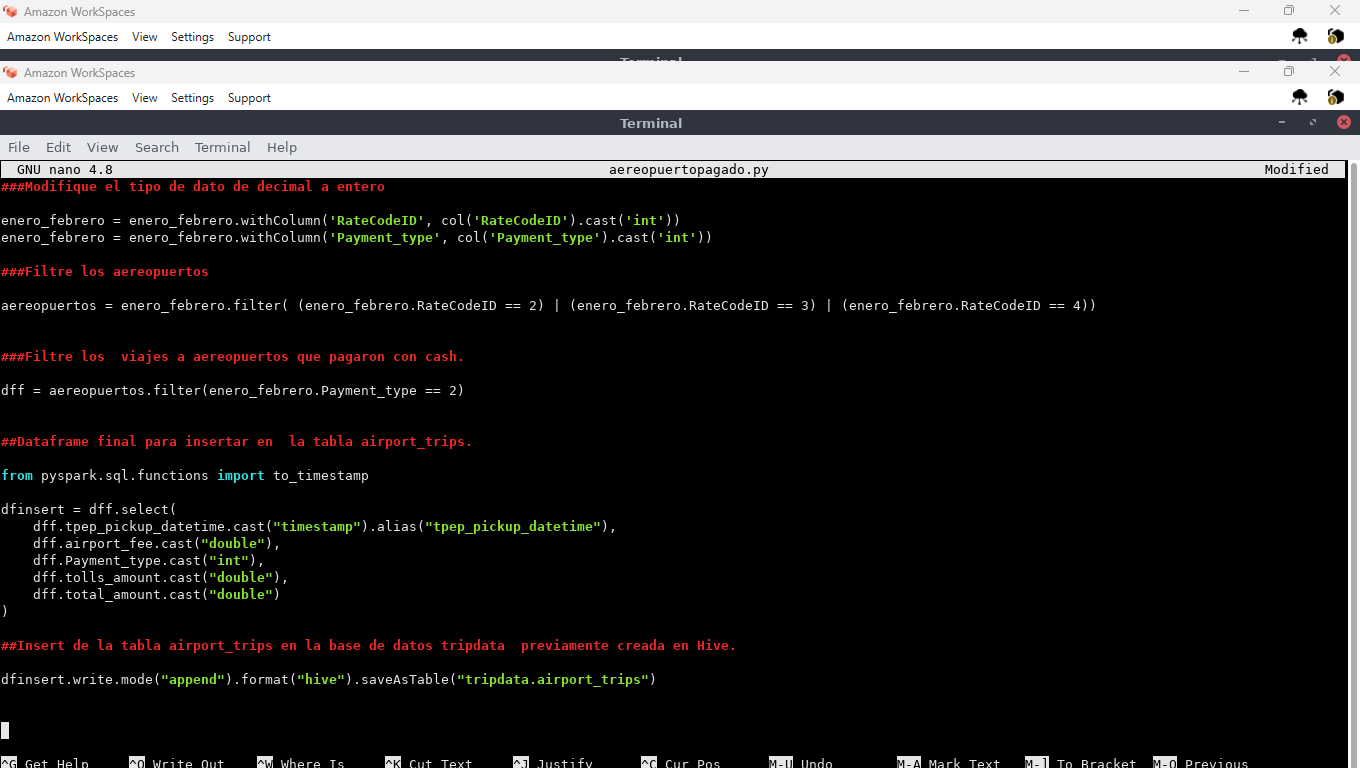
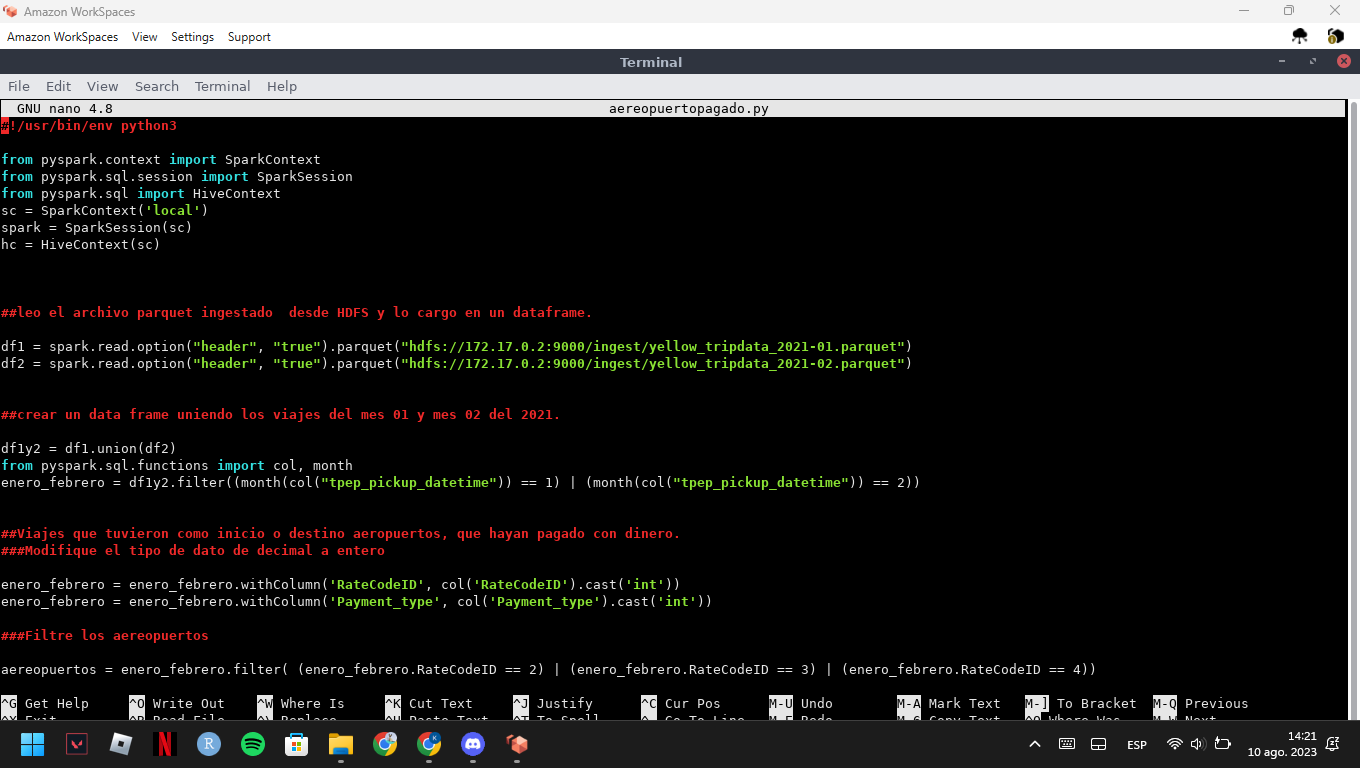
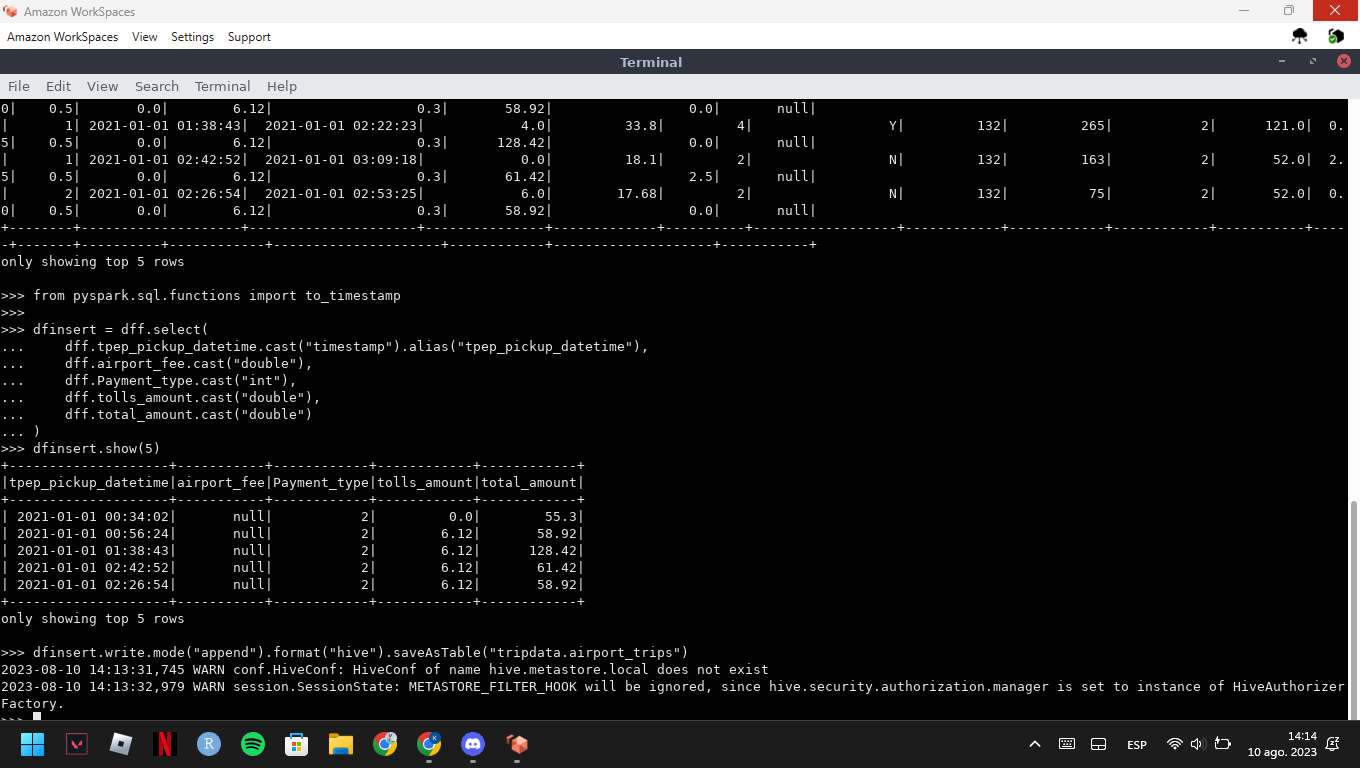
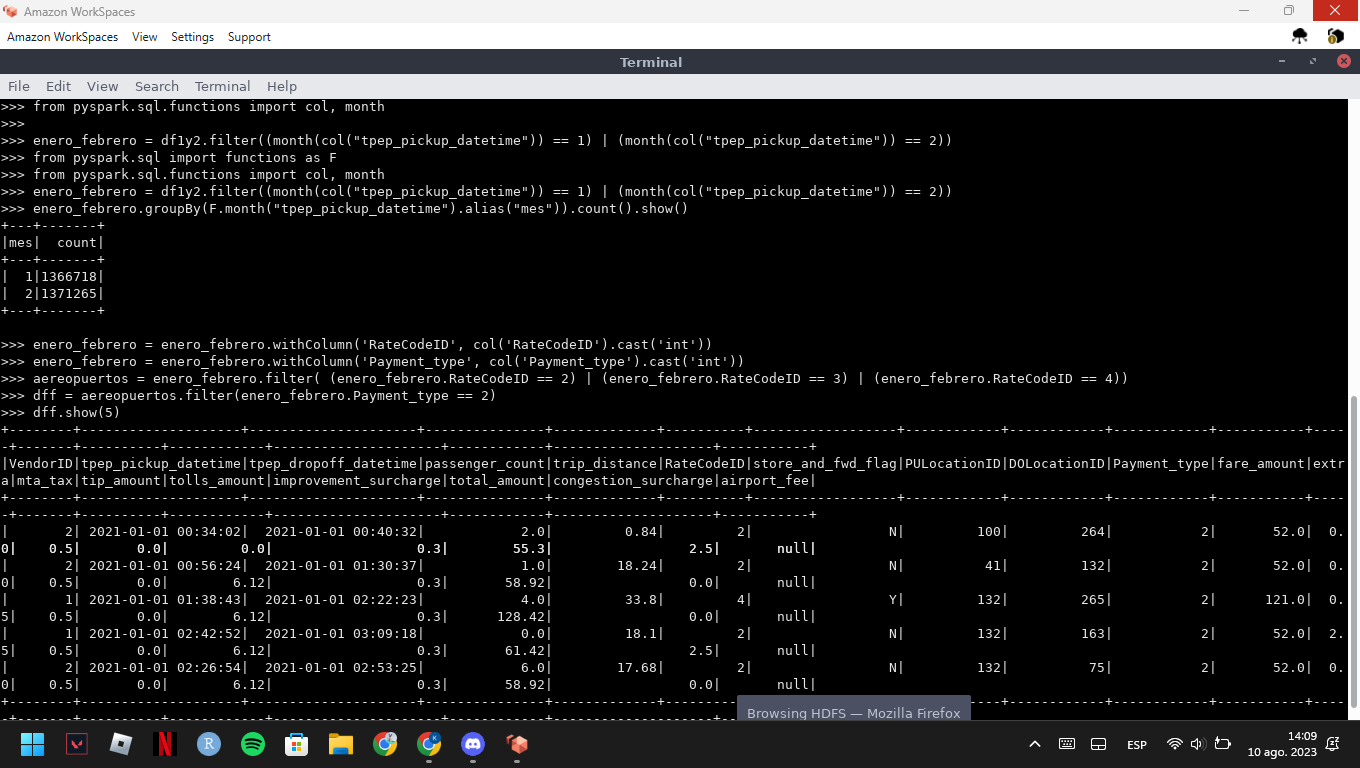
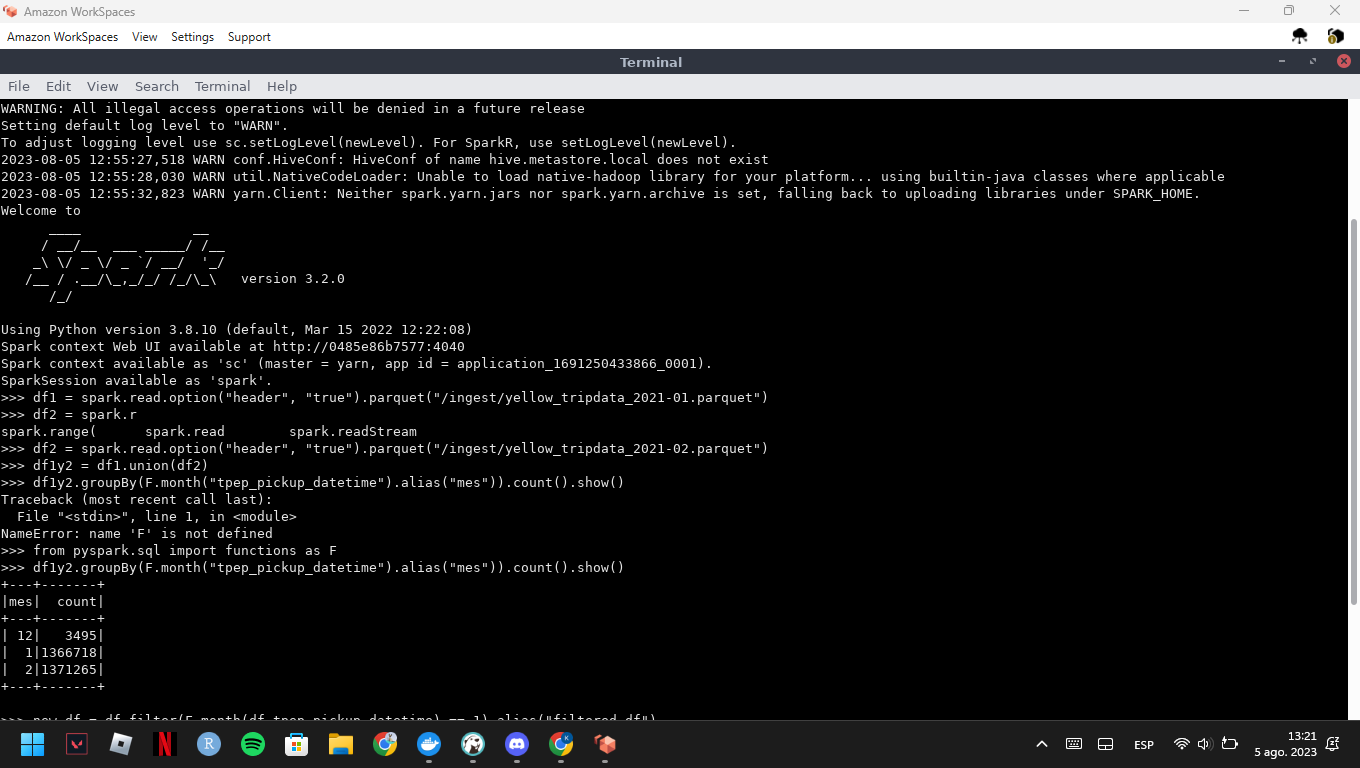
2. En Hive, mostrar el esquema de airport\_trips.



3. Crear un archivo .bash que permita descargar los archivos mencionados abajo e ingestarlos en HDFS: Yellow\_tripdata\_2021-01.parquet (https://data-engineer-edvai.s3.amazonaws.com/yellow\_tripdata\_2021-01.parquet) Yellow\_tripdata\_2021-02.parquet (https://data-engineer-edvai.s3.amazonaws.com/yellow\_tripdata\_2021-02.parquet) 



4. Crear un archivo .py que permita, mediante Spark, crear un data frame uniendo los viajes del mes 01 y mes 02 del año 2021 y luego Insertar en la tabla airport\_trips los viajes que tuvieron como inicio o destino aeropuertos, que hayan pagado con dinero.



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)

