Ejercicio Clase 5

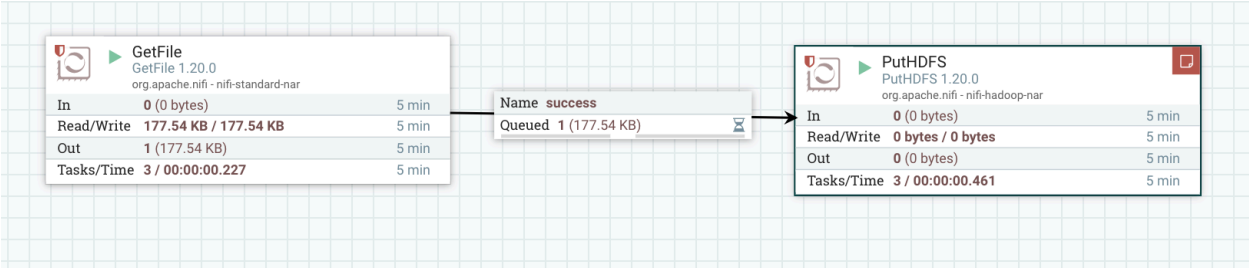
1. En el container de Nifi, crear un .sh que permita descargar el archivo yellow_tripdata_2021-01.parquet desde

   wget -O /home/fpineyro/test/yellow_tripdata_2021-01.parquet https://dataengineerpublic.blob.core.windows.net/data-engineer/yellow_tripdata_2021-01.parquet

   y lo guarde en /home/nifi/ingest.

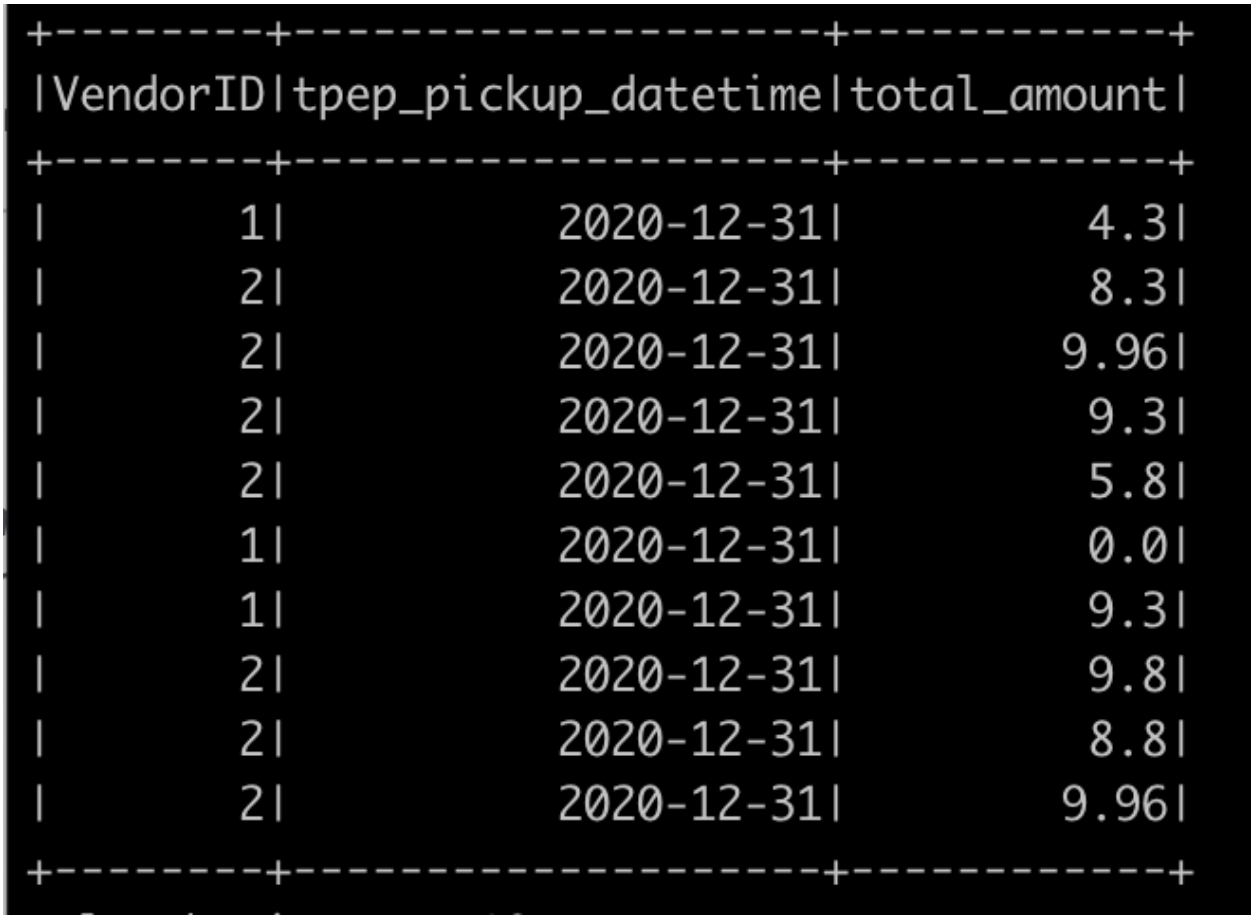
   Ejecutarlo

2. Por medio de la interfaz gráfica de Nifi, crear un job que tenga dos procesos.
   a) GetFile para obtener el archivo del punto 1 (/home/nifi/ingest)
   b) putHDFS para ingestarlo a HDFS (directorio nifi)



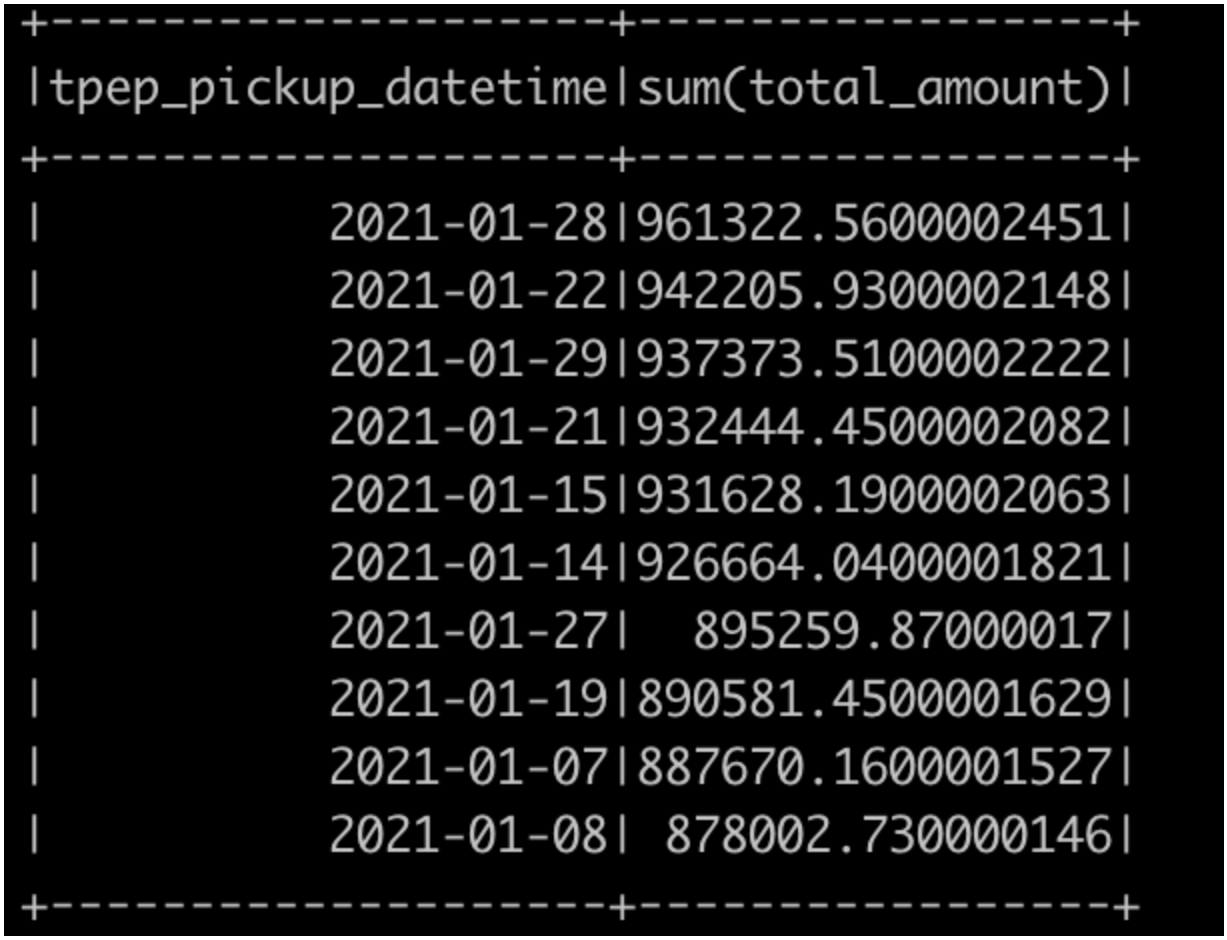
3. Con el archivo ya ingestado en HDFS/nifi, escribir las consultas y agregar captura de pantalla del resultado. Para los ejercicios puedes usar SQL mediante la creación de una vista llamada yellow_tripdata.
   También debes chequear el diccionario de datos por cualquier duda que tengas respecto a las columnas del archivo
   https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

   3.1) Mostrar los resultados siguientes
   a. VendorId Integer
   b. Tpep_pickup_datetime date
   c. Total_amount double
   d. Donde el total (total_amount sea menor a 10 dólares)



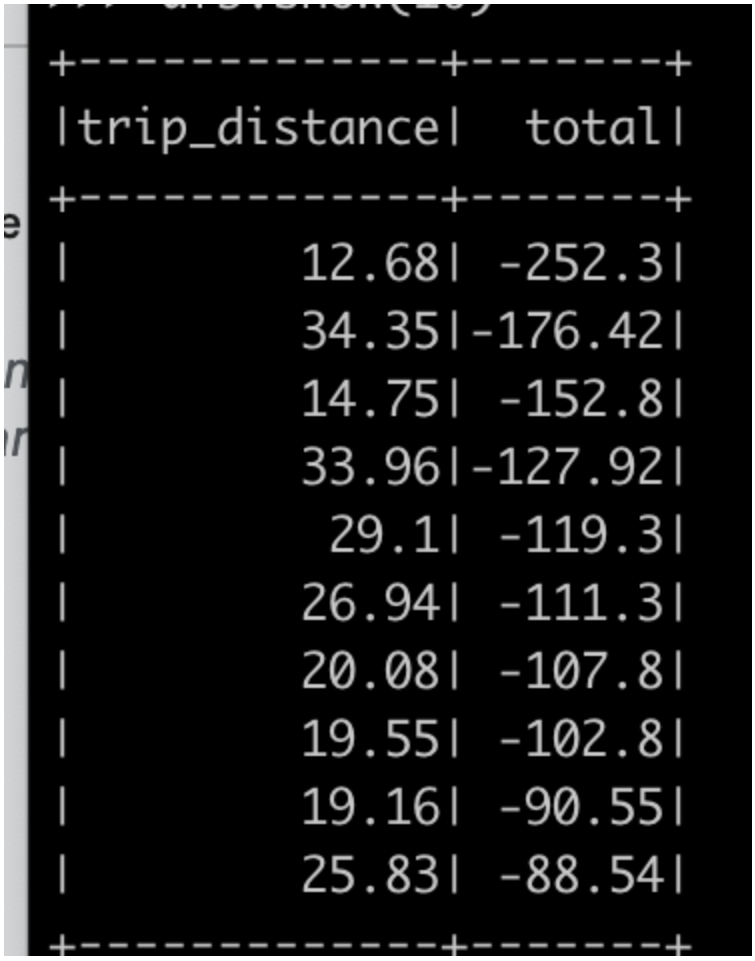
VendorID	tpep_pickup_datetime	total_amount
1	2020-12-31	4.3
2	2020-12-31	8.3
2	2020-12-31	9.96
2	2020-12-31	9.3
2	2020-12-31	5.8
1	2020-12-31	0.0
1	2020-12-31	9.3
2	2020-12-31	9.8
2	2020-12-31	8.8
2	2020-12-31	9.96

3.2) Mostrar los 10 días que más se recaudó dinero (tpep_pickup_datetime, total amount)



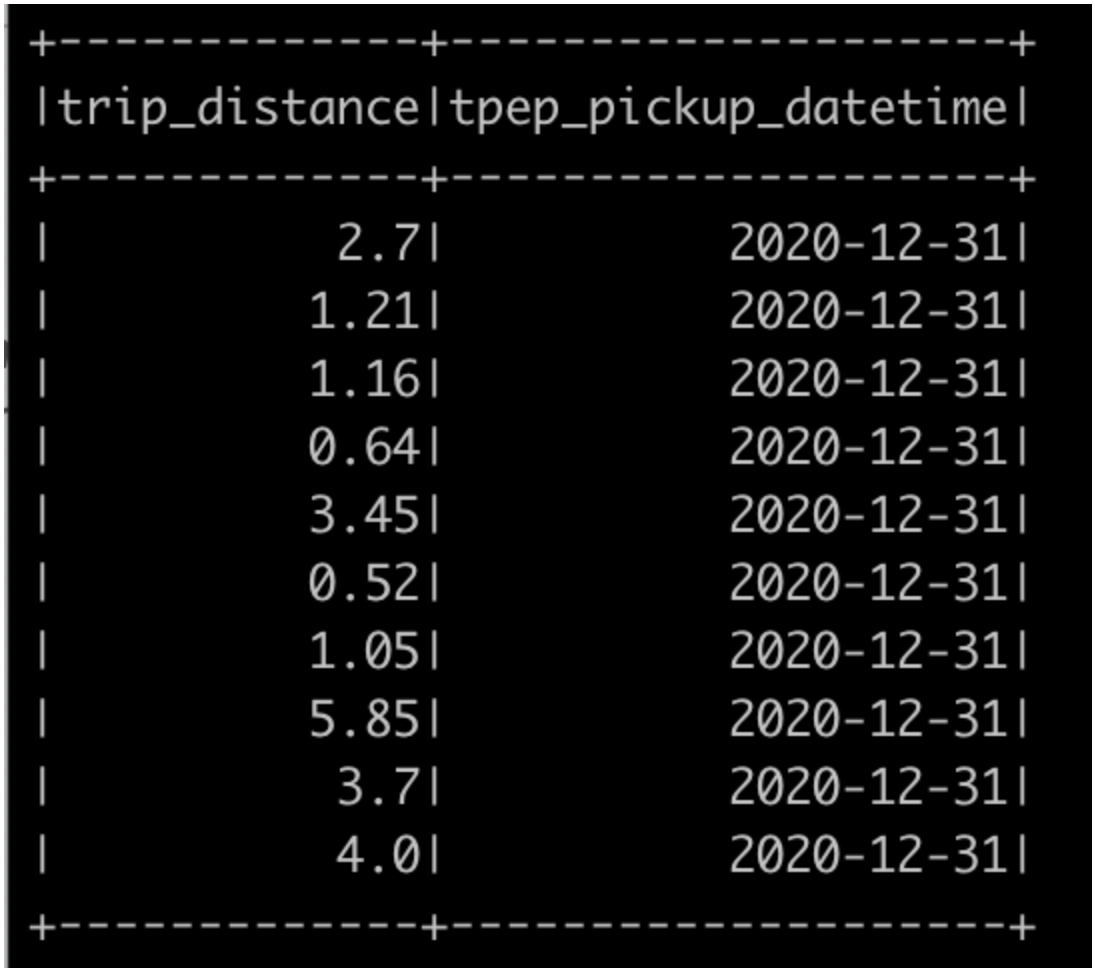
tpep_pickup_datetime	sum(total_amount)
2021-01-28	961322.5600002451
2021-01-22	942205.9300002148
2021-01-29	937373.5100002222
2021-01-21	932444.4500002082
2021-01-15	931628.1900002063
2021-01-14	926664.0400001821
2021-01-27	895259.87000017
2021-01-19	890581.4500001629
2021-01-07	887670.1600001527
2021-01-08	878002.730000146

3.3) Mostrar los 10 viajes que menos dinero recaudó en viajes mayores a 10 millas (trip_distance, total_amount)



trip_distance	total
12.68	-252.3
34.35	-176.42
14.75	-152.8
33.96	-127.92
29.1	-119.3
26.94	-111.3
20.08	-107.8
19.55	-102.8
19.16	-90.55
25.83	-88.54

3.4) Mostrar los viajes de más de dos pasajeros que hayan pagado con tarjeta de crédito (mostrar solo las columnas trip_distance y tpep_pickup_datetime)

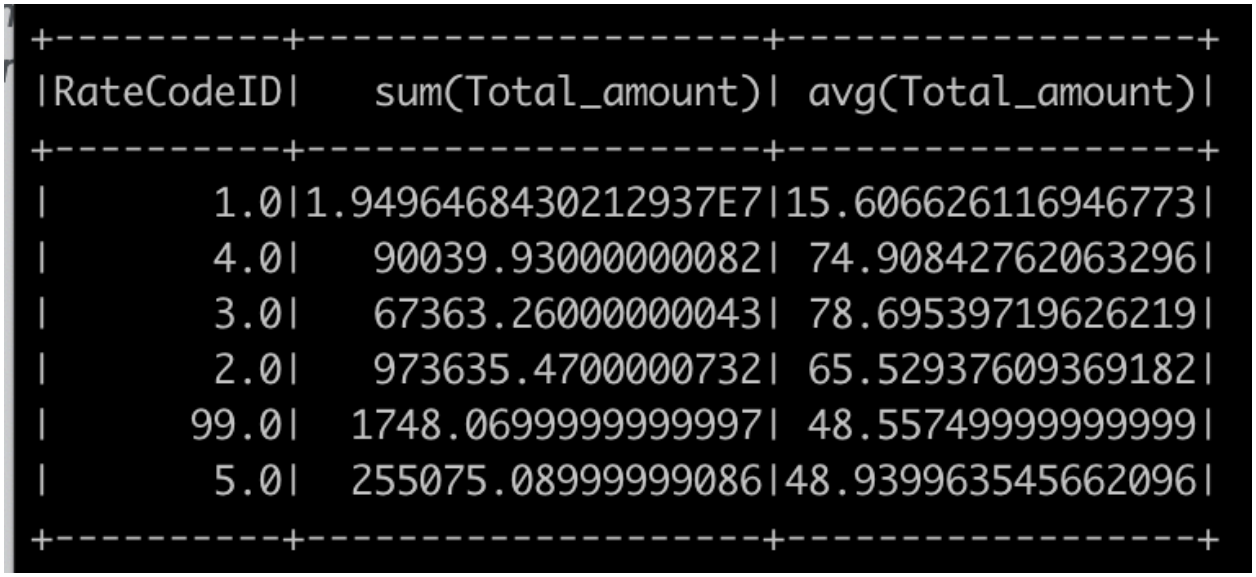


trip_distance	tpep_pickup_datetime
2.7	2020-12-31
1.21	2020-12-31
1.16	2020-12-31
0.64	2020-12-31
3.45	2020-12-31
0.52	2020-12-31
1.05	2020-12-31
5.85	2020-12-31
3.7	2020-12-31
4.0	2020-12-31

3.5) Mostrar los 7 viajes con mayor propina en distancias mayores a 10 millas (mostrar campos tpep_pickup_datetime, trip_distance, passenger_count, tip_amount)

trip_distance	tpep_pickup_datetime	passenger_count	tip_amount
427.7	2021-01-20	1	1140.44
267.7	2021-01-03	1	369.4
326.1	2021-01-12	0	192.61
260.5	2021-01-19	1	149.03
11.1	2021-01-31	0	100.0
14.86	2021-01-01	2	99.0
13.0	2021-01-18	0	90.0

3.6) Mostrar para cada uno de los valores de RateCodeID, el monto total y el monto promedio. Excluir los viajes en donde RateCodeID es ‘Group Ride’



RateCodeID	sum(Total_amount)	avg(Total_amount)
1.0	1.9496468430212937E7	15.606626116946773
4.0	90039.93000000082	74.90842762063296
3.0	67363.26000000043	78.69539719626219
2.0	973635.4700000732	65.52937609369182
99.0	1748.0699999999997	48.55749999999999
5.0	255075.08999999086	48.939963545662096