NYC taxi data ingestion

Purpose of a taxi dataset

One of the key objectives of this project is to investigate the influence of the taxi ridership in NYC under various factors. The factors under consideration are different pickup locations, different times of the day, and different types of weather conditions, etc.. From this analysis, we can provide NYC taxi drivers when and where to find their customers much easier.

Description of Data sources

Taxi trip records contains the historical data from 1/1/2009 - 7/30/2021. The record is issued monthly under four different taxi categories, yellow taxi, green taxi, for-hire vehicle[1] and high volume for-fire vehicle, respectively. Since the data quality (comprehensiveness and wholeness of the data) of the yellow taxi is the best among these four, we choose yellow taxi data to conduct the analysis. In selected time periods from January 2019 to July 2021 the yellow taxi trip records' dataset size is approximately 11.65 GB.

Design and Implementation

Raw data

```
rayichiu@rayis-mbp yellow_taxi_data % cat yellow_tripdata_2021-01.csv | head -n 5
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare_amount
,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount,congestion_surcharge
1,2021-01-01 00:30:10,2021-01-01 00:36:12,1,2.10,1,N,142,43,2,8,3,0.5,0,0,0.3,11.8,2.5
1,2021-01-01 00:51:20,2021-01-01 00:52:19,1,.20,1,N,238,151,2,3,0.5,0.5,0,0,0.3,4.3,0
1,2021-01-01 00:43:30,2021-01-01 01:11:06,1,14.70,1,N,132,165,1,42,0.5,0.5,8.65,0,0.3,51.95,0
1,2021-01-01 00:15:48,2021-01-01 00:31:01,0,10.60,1,N,138,132,1,29,0.5,0.5,6.05,0,0.3,36.35,0
```

Each monthly published data is in CSV format. The first line of each file is the schema/ field name[2], and the following lines are line-based records.

Data cleansing and data Profiling by MapReduce

I use a single mapper to filter out the inaccurate data, such as the arrival time is earlier than the departure time, the driving duration is longer than 24 hours, or the driving distance is zero.

---

[1] FHV data includes trip data from high-volume for-hire vehicle bases (bases for companies dispatching 10,000+ trip per day, meaning Uber, Lyft, Via, and Juno). See Trip Record User Guide for more descriptions.

[2] Yellow Trips Data Dictionary describes yellow taxi trip data.

The mapper output is as follows:

```
[pc3095@hlog-2 pc3095]$ hdfs dfs -cat /user/pc3095/projectMR/output/part-m-00000 | head
2020-03-07 18:39:07,9.70,132
2020-03-07 18:27:02,1.01,170
2020-03-07 18:40:08,1.31,163
2020-03-07 14:57:54,15.60,138
2020-03-07 18:19:51,1.12,48
2020-03-07 18:34:23,3.04,142
2020-03-07 18:58:30,1.38,234
2020-03-07 18:45:58,.71,79
2020-03-07 18:09:42,.72,163
2020-03-07 18:28:36,1.92,170
```

The fields of the mapper outputs are pick-up date and time, the distance of the trip and the pick-up location ID.

I further count the pick-up instance according to the hour of the day and the day of the week, by using the dynamic counter in the mapper.

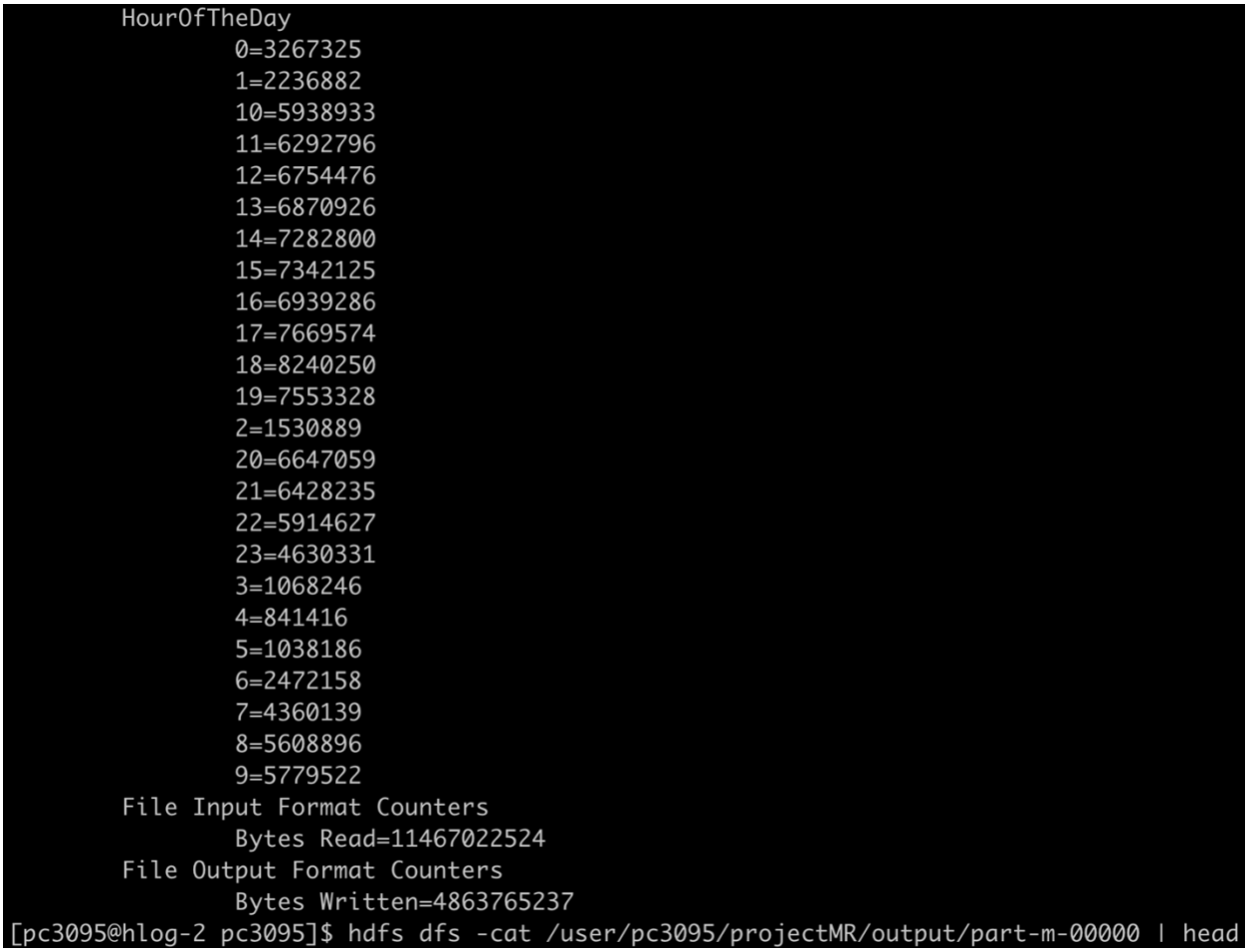
Counter “DayOfTheWeek” result:

```
DayOfTheWeek
        0=14547883
        1=15885027
        2=17941870
        3=18653845
        4=19201616
        5=19095377
        6=17382787
```

The returned value (0 = Sunday, 1 = Monday, 2 = Tuesday, 3 = Wednesday, 4 = Thursday, 5 = Friday, 6 = Saturday)

The counter result shows the ride count of Sunday and Monday are the lowest of the week. We can then use Hive to count the daily ride number in a time period (e.g. two consecutive month) to see if there is a weekly pattern in taxi ride.

Counter “HourOfTheDay” result:



When getting the summarization data of pick-up instance according to the time of the day, the day of week and the pick-up locations we can predict the ride-hailing peak time/ place.

Furthermore, we can join with weather and crime dataset, to study the taxi ride corresponding to different weather conditions/ criminal rate.