NYC TAXI DATA ANALYSIS

Jiaming Zhang Mengdi Wang

New York University

Introduction

Taxi data is more than a log for transportation, because of its flexible nature and lack of rigid routine, taxis can go to places where public transportation cannot, hence its trip data are way more complex than that of subways or buses. Also, taxi data tells stories about the city where they are operated: not every neighborhood gets equal amount of taxi visit, nor does every neighborhood share the same amount of pickups. This may serve as evidences for gentrification and social-economic studies.

Using our knowledge in MapReduce and relational databases, we aim to unravel the stories and myths in these data by developing algorithms to clean up the large raw dataset, serving it efficiently through a web server API, and creating an intuitive graphic user interface.

Data Source

Taxi fact book:

- Publisher, Year: http://www.nyc.gov/ , 2015
- Records: 19,233,777 records for green taxi, 141,620,046 records for yellow taxi
- Size: 30GB

The definition in the columns falls into the following sections:

- Date time: pickup datetime, drop off datetime
- Location information: pickup longitude, pickup latitude, dropoff longitude, pickup latitude.
- Categories: store_and_fwd_flag, VenderID, RateCodeID, Trip_Type, Payment_Type
- Scalar: the remaining columns are all scalars.

New York Neighborhood Data:

- Publisher, Year: Zillow, 2015
- File format: shapefile, we extracted 128 NY objects.

The location data consists of the following columns:

- Name, example: ['NY', 'Greenwich Village' ,'New York City-Mahattan']
- Bound Box, the outer box of the neighborhoods
- Polygon coordinates list

MapReduce (Factbook)

Input of each map and reduce are the data and map output

- **Map**

Output: ((Date_time, taxi_type, rateID), [val1, val2, val3…])

- **Reduce**

Output: ((Date_time, taxi_type, rateID), [total_val1, total_val2, total_val3…])

MapReduce (Spatial Analysis)

Input of each map and reduce are the data and map output

- **Map**

Output: ((Date_time, taxi_type, rateID), (pickup_poly_id, dropoff_poly_id))

- **Reduce**

Output: (Date_time, taxi_type, rateID, action), [value1, value2, …]

- **Reduce (Origin-Destination Anaylysis)**

Output: (origin_poly_id, dest_poly_id, taxi_type, rateID), count

NYC Taxi Factbook

Based on existing NYC Taxi Factbook, we extracted the content that our current dataset allows to generate, and developed a web service API that returns data matching a given query, and created data tables and interactive graphs for visualization.

In terms of how each of these queries' results are computed. The Total Traveling Distance was the sum of all data row's individual traveling distances for each taxi type; Average Trips was the sum of all data row's individual total count, divided by the interval, 1 for yearly, 12 for monthly, 52 for weekly, and 365 for daily; Passenger Counts were calculated in the same fashion as trip count; Trip distributions were calculated by aggregating trip counts in a hash table, where key is the integer encoded timerange (January is 1, February is 2, etc. Similarly, Monday is 1, Tuesday is 2, etc.).

- NYC Factbook gives out the mileage a typical taxi travels within a year, but unfortunately, we do not have the medallion info of each taxi, therefore we calculated the total distance all taxis travelled within this year. For yellow taxi, this number is over 750 million miles, which is about 4 round trips from earth to the sun; For green taxi, this number is over 55 million miles, which is about halfway from earth to the sun.
- Some other results about average trips, passenger counts, and trip distribution are represented in figure 1, figure 2, figure 3

Average Trips

Time Interval	Yellow	Green
Year	119579474	19164472
Month	9964956	1597039
Week	2299605	368547
Day	327614	52505

Table 1, Average Trips

Passenger Counts

Time Interval	Yellow	Green
Year	201035209	26346814
Month	16752934	2195567
Week	3866061	506669
Day	550781	72183

Table 2, Passenger Counts

Trip Distribution

By Month

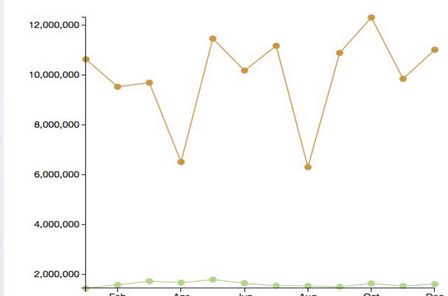

Table 3, Trip Distribution

Spatial Analysis

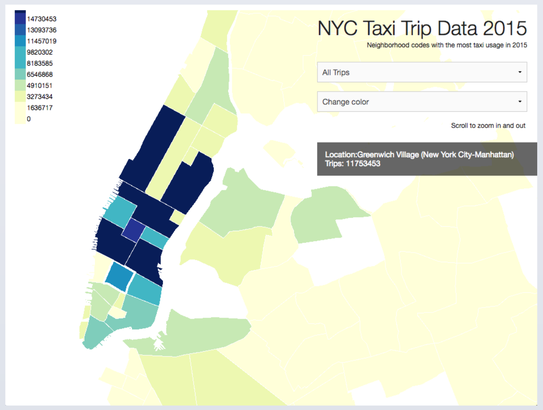

Figure 1 All Taxi Usage Heat Map Screenshot

Our first task is to find statistic results for taxi usage in each neighborhood. And we built a website to visualize these taxi usage in heat map.

Form the trip spatial analysis, we basically get a rank for the usage of taxi between different neighborhoods. Table 7 shows the top 10 neighborhoods in NYC that are better served by taxi.

Midtown is the most popular neighborhood, and the neighborhoods on the Top 10 list have high median household income among NYC.

Secondly, we did some origin-destination analysis. The trips between Midtown, Upper East Side, Upper West Side, Garment District, Gramercy are of the highest frequency (Table 8).

We created some confusion matrices for trips. The confusion matrix gives a direct idea about the taxi usage between neighborhood. The following figure is the confusion matrix for taxi usage in Manhattan, each row stands for the origin neighborhood, and each column is the destination neighborhood. In the following figure, we found that the diagonal is significant, which implies for most trips, the origin and destination are in the same neighborhood. For these trips, the number of the trips in midtown, upper east, upper west, gramercy ranked top. For the trips between different neighborhoods, both row and column for midtown are the most significant. This implies that people either in midtown or go to midtown would more likely to use taxi.

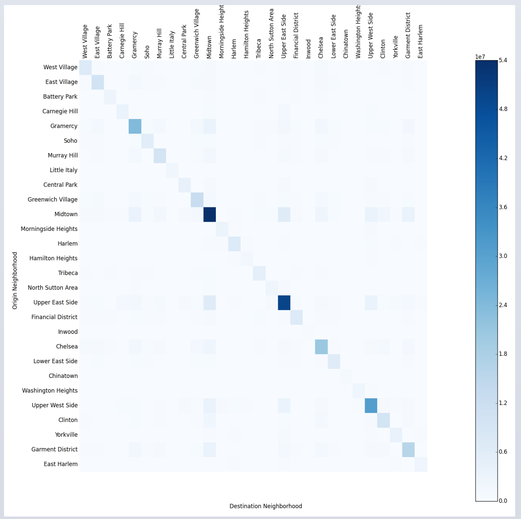

Figure 2, Confusion matrix for trips in Manhattan

Conclusion

In our recreation of NYC Taxi Factbook, we obtained valuable insights in taxi trips' volume, type, and distribution through our REST API server and responsive charting front end. We also developed a conceptual framework for automatically generating NYC taxi data and efficiently serving them on the fly.

During the data processing steps, we used Hadoop streaming for data statistics and polygon detection. All the Hadoop tasks can be completed for about 15 minutes.

For the spatial data analysis, we found the most popular neighborhoods and most popular taxi trips. Besides, We built an heat map website for visualizing the taxi usage in NYC and confusion matrices for visualizing the trips between neighborhoods.

Future Work

- For next steps of our Factbook implementations, we would streamline the data import process, ideally having a web app that allows the Factbook publisher to select raw data file (can be uploaded or linked to a public data storage), then the system automatically performs data processing, database entry, and front end UI generation.
- We could also enable user to pick historical data for comparison, and allow better navigation in our data visualization such as panning and zooming.

Results Table

tripsStats
id (Integer())
datetime (DateTime())
taxi_type (Integer(6))
rate_code (Integer(6))
total_cnt_vendorID_1 (Integer(10))
total_cnt_vendorID_2 (Integer(10))
total_trip_time (Decimal(12,)
.....
total_payment_6 (Integer(10))
total_trip_type_0 (Integer(10))
total_trip_type_1 (Integer(10))
total_trip_type_2 (Integer(10))
total_tolls_amount (Integer(10))
total_record_cnt (Integer(10))

Table 4, Schema of trips statistics

tripsPolygonStats
id(Integer())
datetime(DateTime())
taxi_type(Integer(6))
rate_code(Integer(6))
action(Integer(6))
PolygonId(Integer(6))
Count(Integer(10))

Table 5, Schema of Trip Spatial Query

tripsSpatialQuery
id(Integer())
datetime(DateTime())
taxi_type(Integer(6))
rate_code(Integer(6))
action(Integer(6))
total_record_cnt(String(2048)

Table 6, Schema of Spatial statistics

rank	Neighbourhood	Amount
1	Midtown, New York City-Manhattan	45009718
2	Upper East Side, New York City-Manhattan	37231528
3	Upper West Side, New York City-Manhattan	23548857
4	Gramercy, New York City-Manhattan	21513593
5	Chelsea, New York City-Manhattan	18138800
6	Garment District, New York City-Manhattan	15162592
7	Greenwich Village, New York City-Manhattan	11753453
8	Murray Hill, New York City-Manhattan	9522897
9	East Village, New York City-Manhattan	9423773
10	Clinton, New York City-Manhattan	9031628

Table 7, Top 10 Taxi Usage (Pickup, Drop off)

rank	Origin	Destination	Amount
1	Upper East Side	Midtown	6516892
2	Midtown	Upper East Side	6516892
3	Midtown	Garment District	3764784
4	Garment District	Midtown	3764784
5	Midtown	Upper West Side	3678148
6	Upper West Side	Midtown	3678148
7	Midtown	Gramercy	3560040
8	Gramercy	Midtown	3560040
9	Upper East Side	Upper West Side	3392189
10	Upper West Side	Upper East Side	3392189

Table 8, Trip Rank