Effects of Boro Taxis on the New York City Taxi Market

Israel Malkin, Alex Pine, Tian Wang
Big Data, Spring 2015

In the summer of 2013, New York City created a new taxi program called the “Five Borough Taxi Plan” that was designed improve the taxi service in areas of the city that had been historically underserved by the traditional yellow medallion taxicabs. This program created a new type of taxi, the green-colored “Boro” taxi, that is only permitted to to pick up riders in Brooklyn, the Bronx, Queens, and northern Manhattan (see Fig 1 and 2 below).

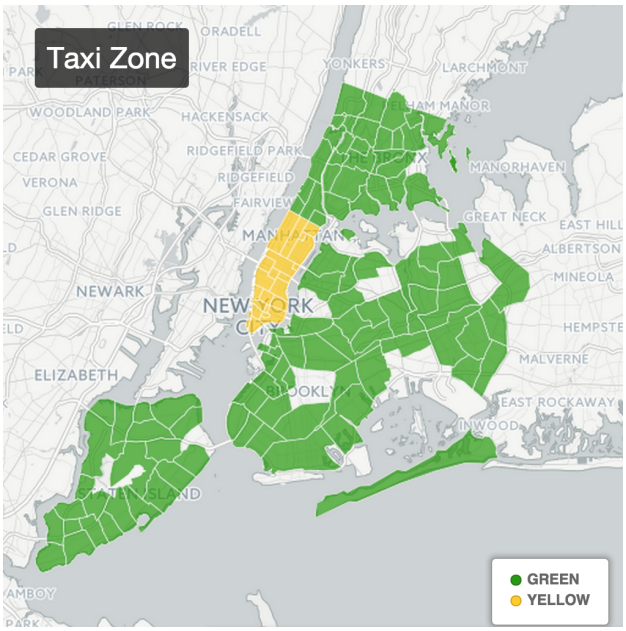

Figure 1: Boro Taxi Regions.
The blank spots in the New York City are a result in gaps in our data. These are explained in the “Data Issues” section of the paper.

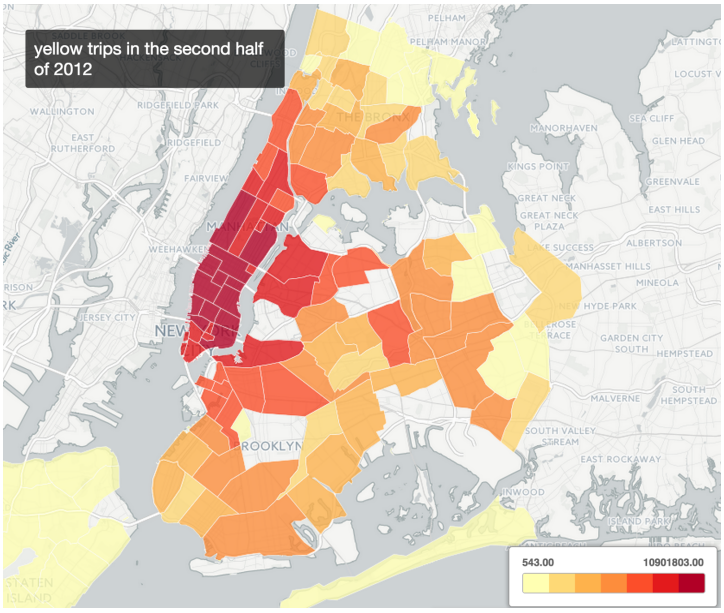

Figure 2: Number of trips by medallion taxis grouped by the originating neighborhood, August-December 2012. This demonstrates motivation behind creating the Boro taxis: Medallion taxis serve Manhattan far better than the outer Boros.

When this plan was initially proposed, it was met with protests[1] and a lawsuit[2] from medallion taxi drivers, who claimed the increased competition from the new taxis would drive many of them out of business. Our group has tried to determine if their prediction was accurate--did the introduction of Boro taxis negatively impact the business of the medallion taxis?

Data scientist Chris Whong[3] recently acquired the Boro taxi trip and fare history from their inception in 2013 through the end of 2014 through a “Freedom of Information Law” request to the New York City government, and published it on his website[4].

We compared this data set with the corresponding data from the traditional medallion taxis, and found little evidence that the medallion taxis were negatively affected by the creation of the Boro taxis. Boro taxis largely serve areas of New York City that were largely unserved by medallion taxis, such as the Bronx and Queens, as their creators intended.

Results

The trip and fare history for Boro (a.k.a “green”) taxis in 2013 clearly demonstrates that the green taxi program was a success in its original intention: providing taxi service to previously underserved outer boroughs of New York City. In its first half-year, green taxis provided over

[1] http://www.newsday.com/news/new-york/cabbies-to-protest-outer-borough-taxi-proposal-1.2969177
[2] http://www.wnyc.org/story/285364-breaking-judge-halts-mayor-bloombergs-taxi-plan/
[3] http://chriswhong.com/
[4] http://chriswhong.com/open-data/foiling-nycs-boro-taxi-trip-data/

half a million rides. Figure 3, shown below, demonstrates that many of these trips originated in neighborhoods that the medallion (a.k.a “yellow”) taxis do not serve well. It also demonstrates that green taxis rarely break the rules by picking up fares in areas reserved for yellow taxis.

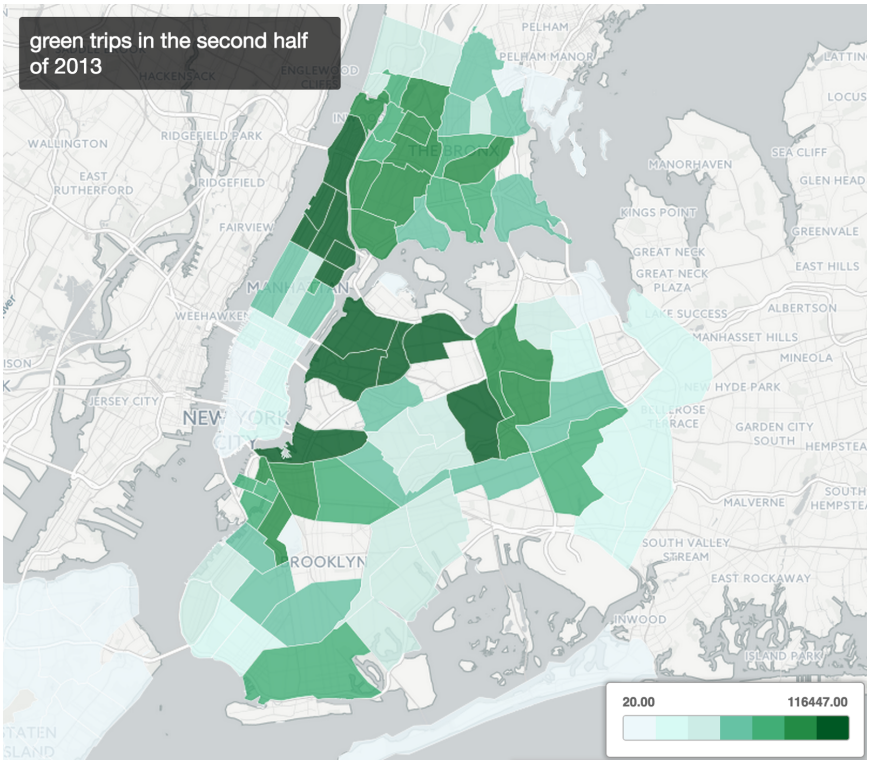

Figure 3: The number of green taxi rides, grouped by originating neighborhood, August to December 2013

Even if the green cabs were a success on their own right, the question of whether they negatively affected the business of the yellow cabs still remains. To answer this question, we compared the rides provided by green taxis during the second half of 2013, when they first appeared, to those made by yellow taxis during the same time period. We also compared the trips provided by yellow taxis in 2013 to those from 2012, before green cabs existed.

Figures 4 and 5 display the number of trips provided by yellow and green taxis and the amount of money they made from 2012 to 2013, respectively. These graphs show that the number of trips and the amount of money made by the green taxis grew sharply after their conception. It also shows that the number of trips and the amount of money made yellow cabs does not seem to decrease at all after the introduction of the green taxis. There are large fluctuations in this yellow cab data, but there is no apparent relationship, positive or negative, to the corresponding green taxi data.

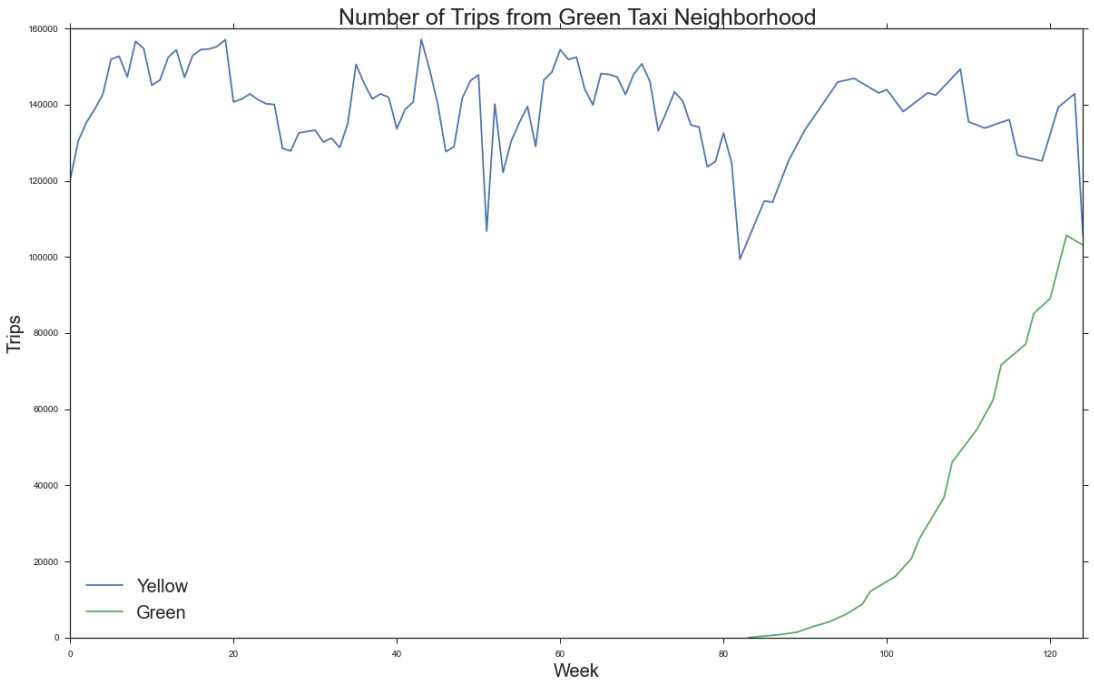

Figure 5: The number of trips per week originating in green neighborhoods from 2012 to 2013.

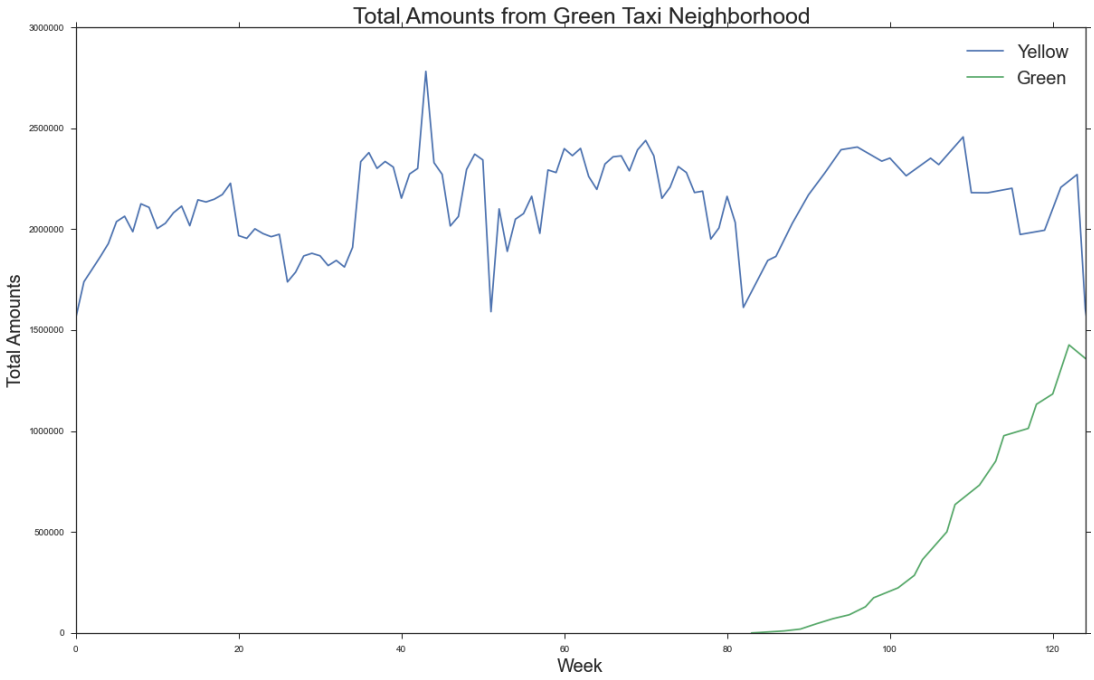

Figure 4: The total amount of money made per week from trips originating in green neighborhoods from 2012 to 2013.

If we break down the trips by neighborhood, we find that the green taxis only dominate neighborhoods far away from the yellow zones, with northern Manhattan as a notable exception. As figure 6 shows, green cabs handled up to as much as 17 times as many pickups relative to yellow cabs in the northern neighborhoods of the Bronx. Note that these are neighborhoods that had virtually no taxi pick-ups prior to the introduction of Boro cabs.

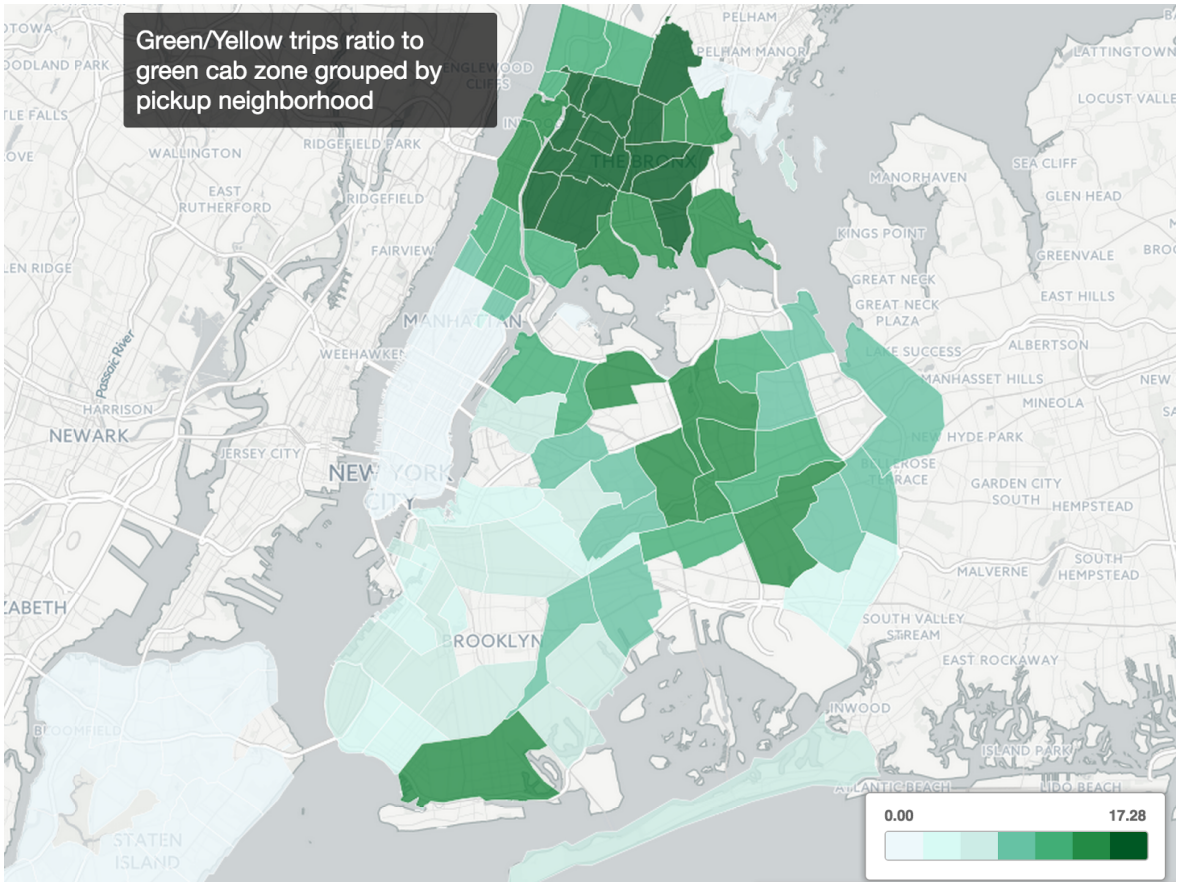

Figure 6: The ratio of the number of green trips originating in green neighborhoods, to the number of yellow trips originating in green neighborhoods, Aug-Dec 2013.

Lastly, we wanted to see if green taxis met pent-up demand or if they actually took business away from yellow taxis. Figure 7 shows the change in the number pickups between Aug-Dec of 2012 and Aug-Dec of 2013. The values are normalized by the total number of trips during the Aug-Dec 2012 period, which only consisted of yellow cabs. Note that if we had pure substitution from yellow to green, then the orange and green bars would be of equal length. Since the green bars are in general much greater than the orange ones, this graph implies that the green taxis grew the taxi market much more than they replaced yellow taxis.

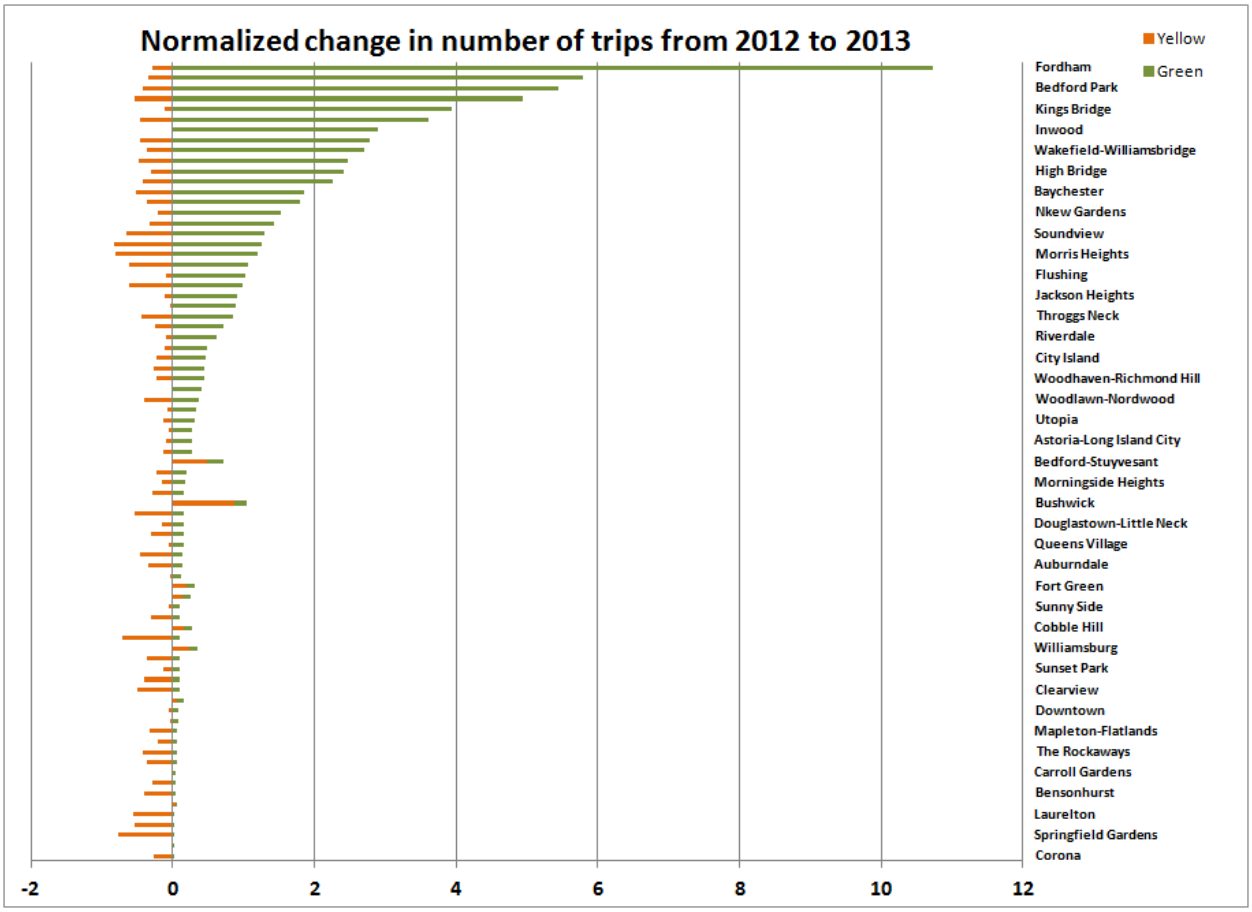

Figure 7: Percent change in the number of trips Aug-Dec 2013 relative to previous year

Methods

We created a data processing pipeline that transforms the raw data files provided by the New York City government about every taxi trip into a single MySQL table that could be easily queried. The results from these queries are what we used to generate the graphs above and to reach our conclusions. All of our code and methodology can be found in our project's Github page[5].

The initial input to the pipeline is the medallion and Boro taxi data provided by the New York City government. This data contains the details of every taxi trip taken in New York City: the spatial coordinates of where the trip began and ended, the fare amount, the number of passengers taken, and so on. This data was split into several files, divided by time, type of taxi (medallion or Boro), and by attribute groups.

[5] https://github.com/pinesol/bigdata

The first step of the pipeline was to merge these files with different attributes into a single set of files with common attributes. This transformation was achieved by storing the input files on Amazon's Simple Storage Service ("S3"), and processing them with a map-reduce program executed on Amazon's Elastic Map-Reduce ("EMR") system.

Once the data was merged into a set of files with a common format, we were faced with the problem of transforming the data so that it could be easily queried. Initially, we thought these would be most easily achieved by using a tool designed to process structured queries on very large amounts of data, such as Apache Hive, Apache Pig, or Apache Spark. However, we realized that the data could be summarized in such a way that all the information required for our analysis could be preserved, while reducing its size enough so that it could handled by a traditional relational database.

The question of whether or not Boro taxis have taken some of the market away from medallion taxis was not going to be determined by analyzing individual trips--it was going to be answered by analyzing trips aggregated over days, weeks, and months. Additionally, Boro and medallion taxis only differ by the neighborhoods in which they are allowed to pick up fares. This implies that spatial coordinates are too finely grained for our analysis, and that trips taking place within the same neighborhoods can be considered equivalent for our purposes. Given these insights, we decided to average together the attributes of trips that took place within the same hour, and began and ended in the same neighborhoods. This summarization reduced the size size of the data by an order of magnitude, which allowed us to load the data into a MySQL database.

Once the summarized data was loaded into MySQL, the final step in our data pipeline was simply to run SQL queries we had composed through it, and to plot their results using Python's matplotlib library, and CartoDB.

Input Data

Yellow cab data from 2012 and 2013 was obtained from the University of Illinois (https://uofi.app.box.com/NYCtaxidata), and green cab data from August through December 2013 was obtained from Chris Wong (http://chriswhong.com/nycborotaxidata/). The raw yellow cab data was approximately 70 GB for both years, whereas the 4 months of green cab data amounted to only 200 MB. Once the raw data was downloaded and hosted in S3, we set up a github repo and started processing the data for analysis. The green cab data contains fewer columns than the yellow data, so we only kept the intersecting columns, listed below:

- 'pickup_datetime'
- 'dropoff_datetime'
- 'store_and_fwd_flag'
- 'rate_code'

- 'pickup_longitude'
- 'pickup_latitude'
- 'dropoff_longitude'
- 'dropoff_latitude'
- 'passenger_count'
- 'trip_distance'
- 'fare_amount'
- 'surcharge'
- 'mta_tax'
- 'tip_amount'
- 'tolls_amount'
- 'total_amount'
- 'payment_type'

Map-Reduce workflow

Map-Reduce to merge yellow taxi's trips and fares

We created two map-reduce jobs to merge and clean the datasets. The first map-reduce task read in all raw datasets from yellow taxis and output a dataset containing both trip and fare information for yellow taxis.

Map-Reduce to merge green and yellow and geocode locations

The second map-reduce task took in the joined dataset from the first task and added the "neighborhood", "borough" and "taxi service zone" attributes. It accomplished this by using a "shape" file from Zillow (www.zillow.com), and an R-Tree library that can map spatial coordinates to a neighborhood name, provided to us by the class's teaching assistant Tuan-Anh Hoang-Vu.

The EMR cluster used to run the map-reduce was set to have 1 master node (m3.xlarge) and 2 slave nodes (m3.xlarge). Joining trip and fare data sets utilized 506 mappers, 7 reducers and 528 mappers, 7 reducers for the 2013 and 2012 datasets, respectively. The process took 144 minutes to complete. The task of mapping spatial coordinates to neighborhoods utilized 381 mappers, 7 reducers and 392 mappers, 7 reducers for 2013 and 2012, respectively. It took 365 minutes to complete.

Map-Reduce to aggregate trips by type, hour, and location

After obtaining merged and cleaned data, we decided to aggregate the trips into hour-pickup neighborhood-dropoff neighborhood buckets. This task was accomplished by a relatively simple map-reduce job. Its mapper emitted keys of the tuple (taxi color, date, hour, pickup-location, dropoff-location), with the rest of the trip's data as the value. The reducer

simply summed the numeric values in the “data” tuple. This task was repeated for 2012 and 2013 data respectively. In both years, 6 slave nodes were used. The aggregation was completed in just 17 minutes, and 496 mapper and 23 reducer tasks were utilized. Each year's data was compressed from approximately 30GB to around 1.7GB, a compression factor of over 15. For example, the raw 2012 data (which only contains yellow cab trips) was reduced from 175 million rows to 10 million by aggregating trips to the hourly neighborhood-pair level.

This more manageable data size allowed us to perform our analysis in MySQL. After loading the data, we created an index on (color, pickup_zone, dropoff_zone, pickup_borough, pickup_neighborhood, dropoff_neighborhood), which took about 45 minutes on a desktop machine with 24GB memory. All of the results and plots are the results of relatively simple SQL queries.

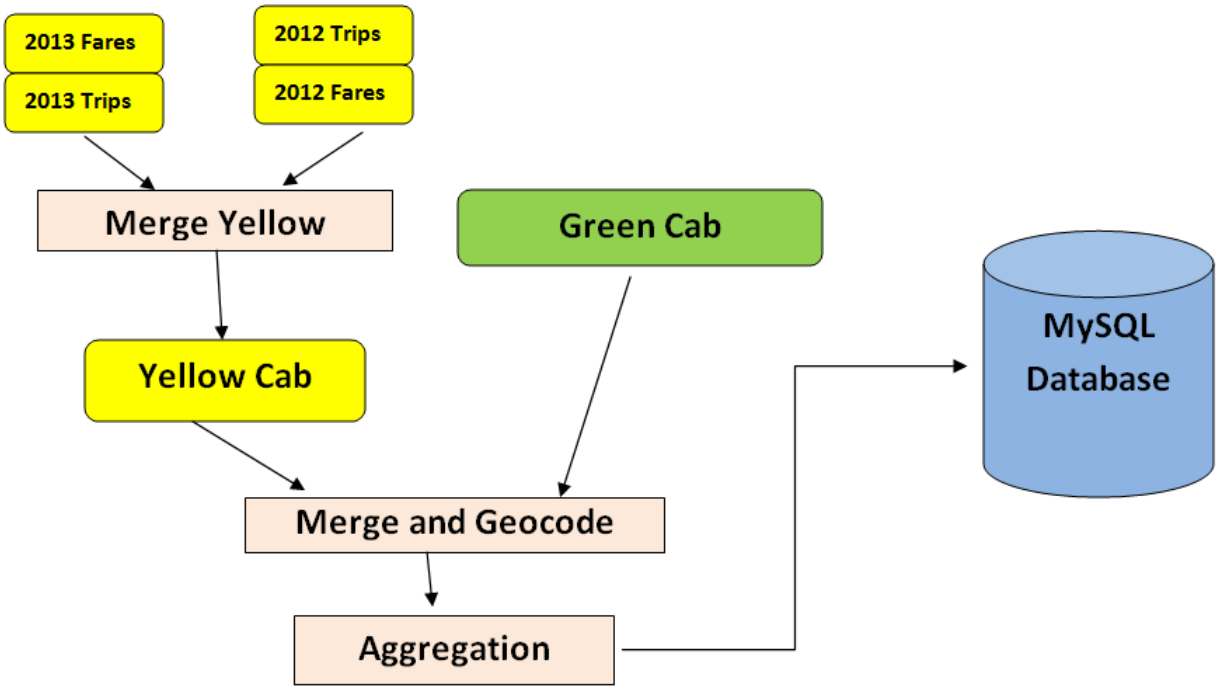

Figure 8: The data processing pipeline.

Visualization

We used matplotlib package in ipython notebook and CartoDB to do visualization. Matplotlib is a python plotting library, and we mainly used it to do 2-D time-series analysis. CartoDB (http://cartodb.com) is an online platform to visualize data geographically. We merged the shape file with datasets and created interactive maps on this website.

Data Issues

We encountered a few data issues, the major one being that 2014 yellow cab data is not yet available. This prevented us from seeing if Boro taxis continued to gain market share within the designated green zones. We also had an issue with the Zillow NYC neighborhoods shape file. Initially, we noticed that some of the trip pickup/drop-off locations locations were not being captured by the Zillow shapefile. After sampling a few of the unmatched trips we realized that all of them were trip either to or from airports. This turned out to be false, because only after plotting some of our results we realized that the shapefile was missing about half a dozen neighborhoods in Queens and Brooklyn. Although the data involving these neighborhoods is not included in our analysis, the evidence seems to be strong and consistent enough for us to be confident in our claim that the Boro Taxi project succeeded in its goal and did not adversely affect yellow cab revenues.

The green cab data does have a lot of trips with distance equal to zero, especially early in the sample, but this is not particularly relevant for our analysis.

Additionally, our workflow structure generates a fairly long pipeline which had be rerun every a bug was encountered. We spent a lot of time considering which big data analysis tools to use (MR, Pig, Hive, Spark, etc) and settled on using SQL which all the team members were familiar with. We had to invest a more time in compressing the data, but this paid off in terms of faster and simpler queries using MySQL. AWS and S3 are not straightforward to use and we found their UI's to be lacking, especially when collaborating and sharing files across S3 buckets.

Work breakdown

Israel: Four-merging map-reduce, SQL queries, paper, presentation.
Alex: Geen taxi data, github, zoning code, MySQL setup, MySQL queries, paper, presentation.
Tim: Parsing input data, merging map-reduce, graphs, paper, presentation.