Taxi Data Visualization of NYC

Yu Yang
Mentor: Prof. James Abello
April 25th, 2016

Goals

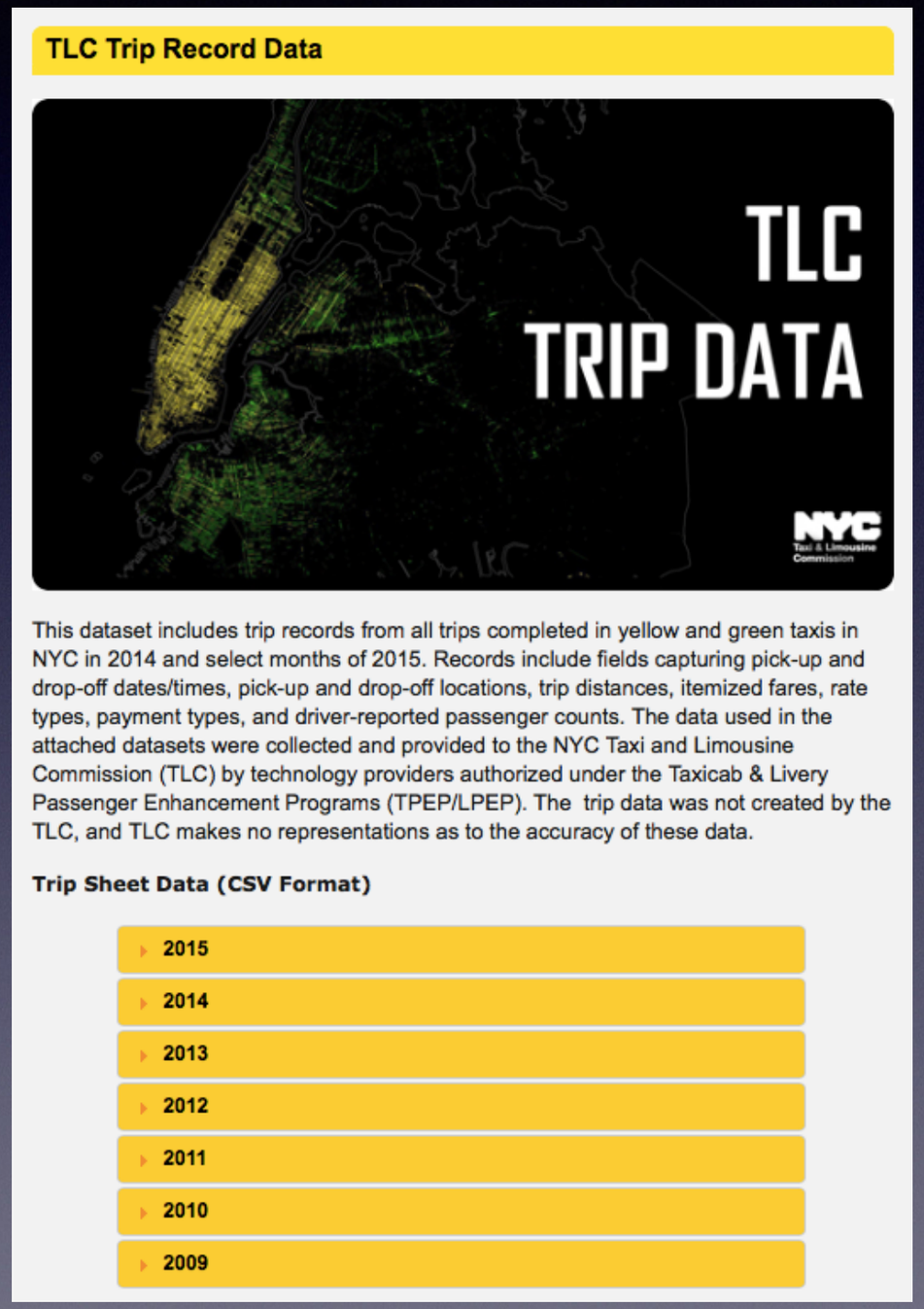
- Develop a taxi data visualization system for millions records level datasets
- Use Google BigQuery platform for offline data analysis
- Analyze traffic patterns of New York City

Outline

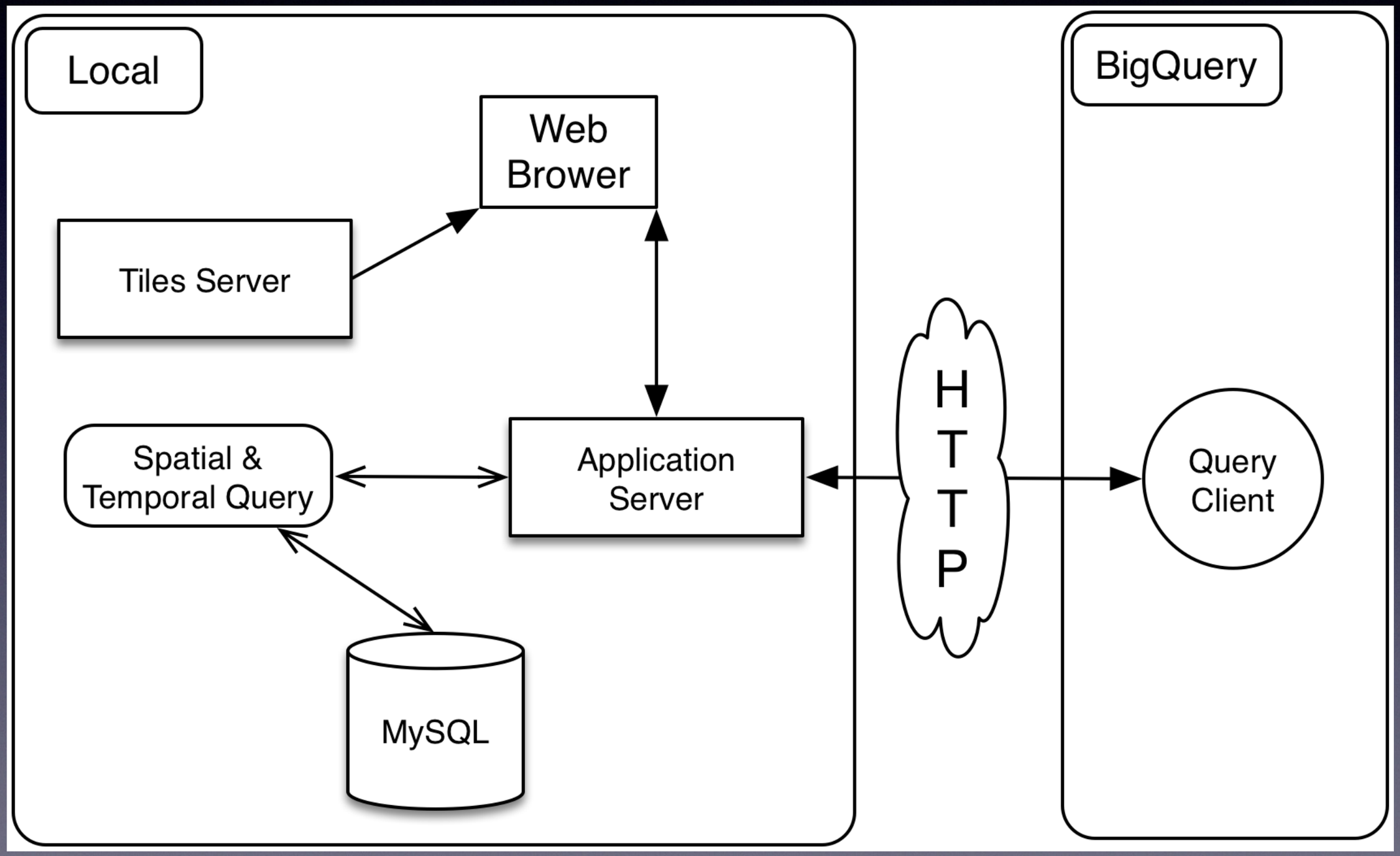
- Data Overview
- System Architecture
- Google BigQuery VS Amazon EMR
- Data Analysis
- Demo

Data Overview

- Taxi trips records of New York City in 2014
- http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- 165,114,361 records a year
  Average 18848/hour
- Include pick-up and drop-off locations, time, distance, passenger number, fare, payment type, etc



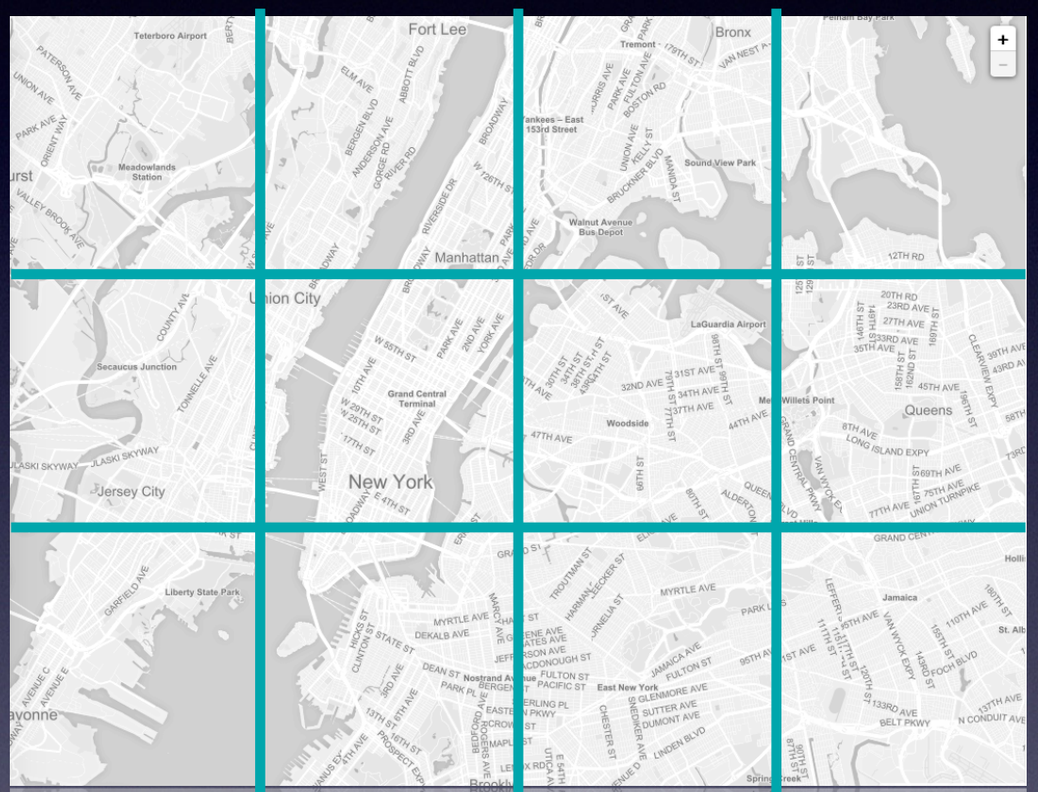
System Architecture

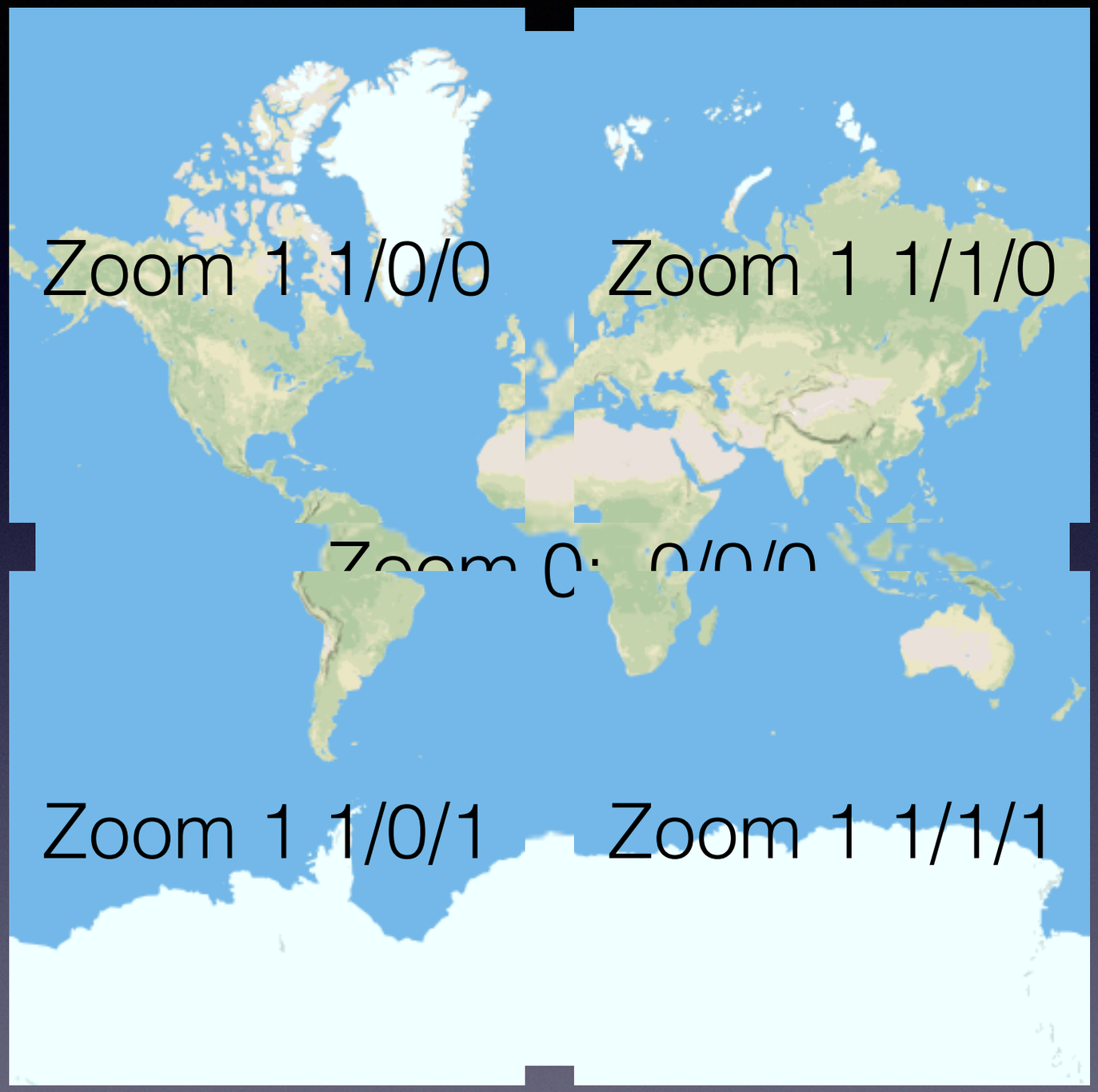


Tiles Server

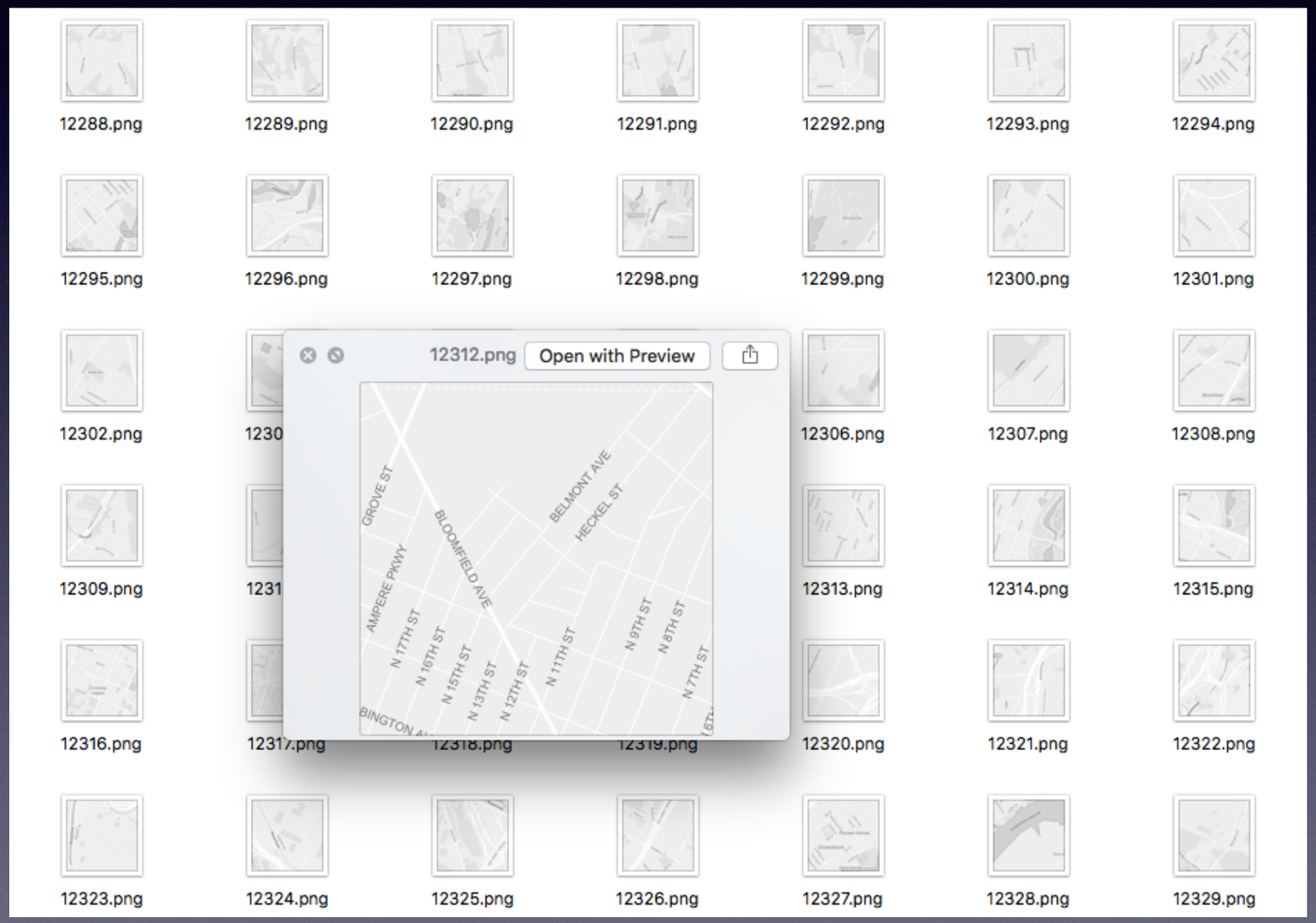
Tiles and Zoom

- Web maps are made up of many small, square images called tiles
- Maps of different zoom levels consist of different numbers of tiles
- http://hostname/Z/X/Y.png




Zoom 1 1/0/0
Zoom 1 1/1/0
Zoom 0: 0/0/0
Zoom 1 1/0/1
Zoom 1 1/1/1

Tiles Server



Z12: 54 pieces 1.1 MB

Z13: 211 pieces 3 MB

Z14: 813 pieces 9.9MB

Z15: 3193 pieces 25MB

Z16: 12657 pieces 67.3MB

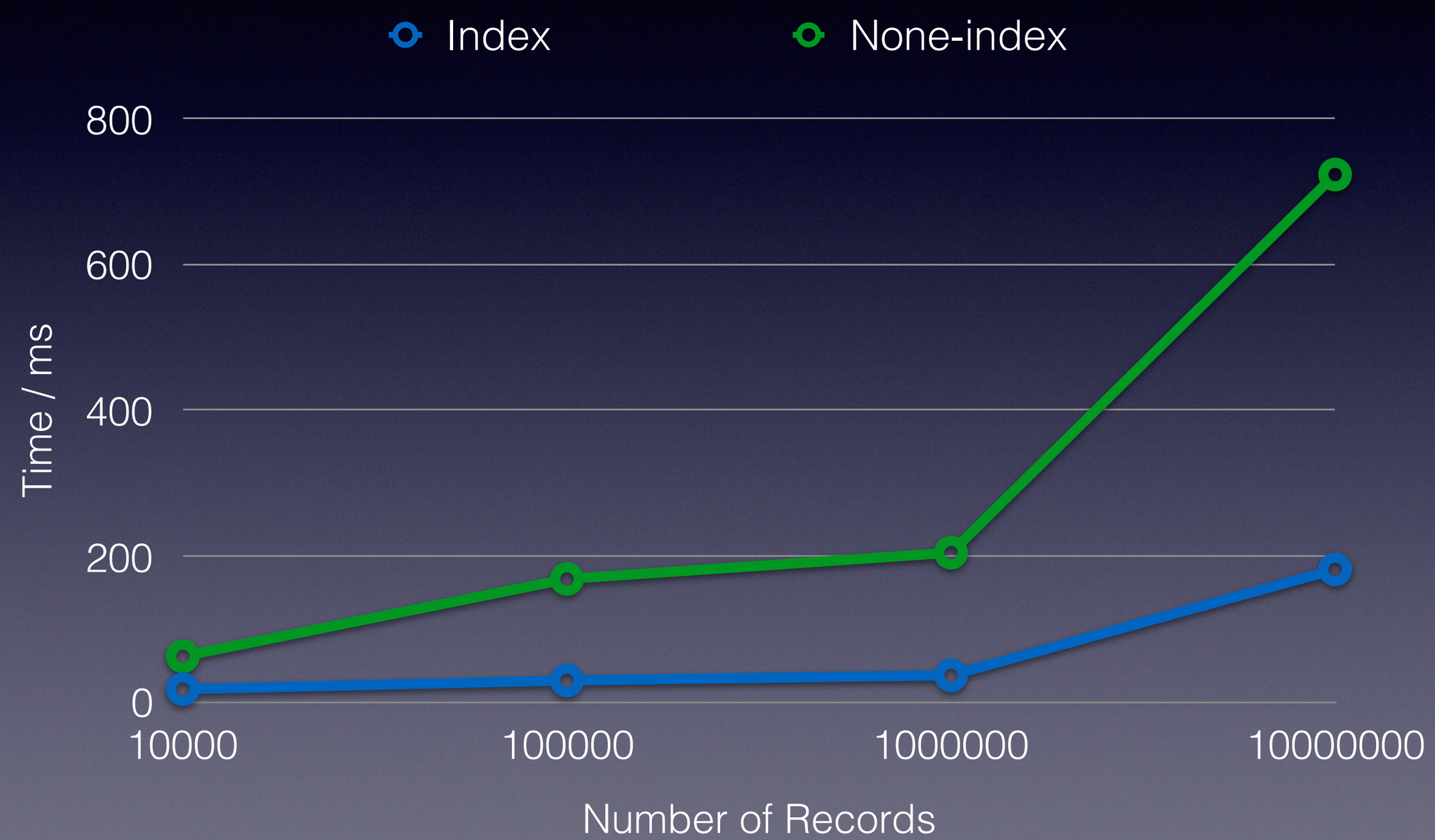
Web Interface - Leaflet.js

- An open-source JavaScript library for mobile-friendly interactive maps
- Developed by Vladimir Agafonkin
- Provides plenty of features for maps eg: tile layers, markers. interactions, plugins(draw rectangle, heatmap)

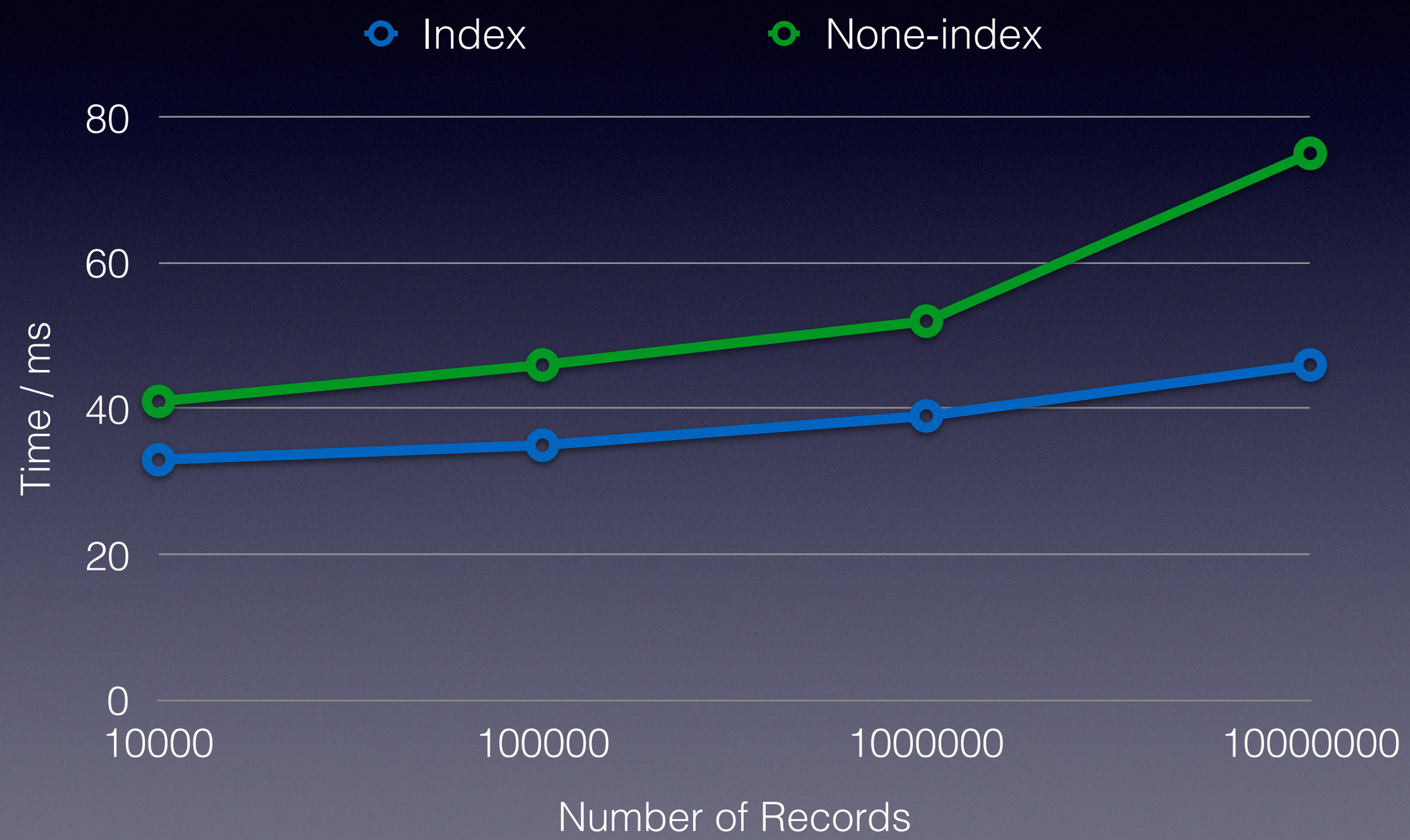
Spatio-Temporal Query

- Index is one of the most important features in DBMS to speed up query
- For range query, use B+ tree index
  eg: pick-up time between 10:00 and 11:00
- For equality query, use Hash index
  eg: trips count where passengers number is 4

Range Query


Index
None-index
800
600
400
200
0
Time / ms
10000
100000
1000000
10000000
Number of Records

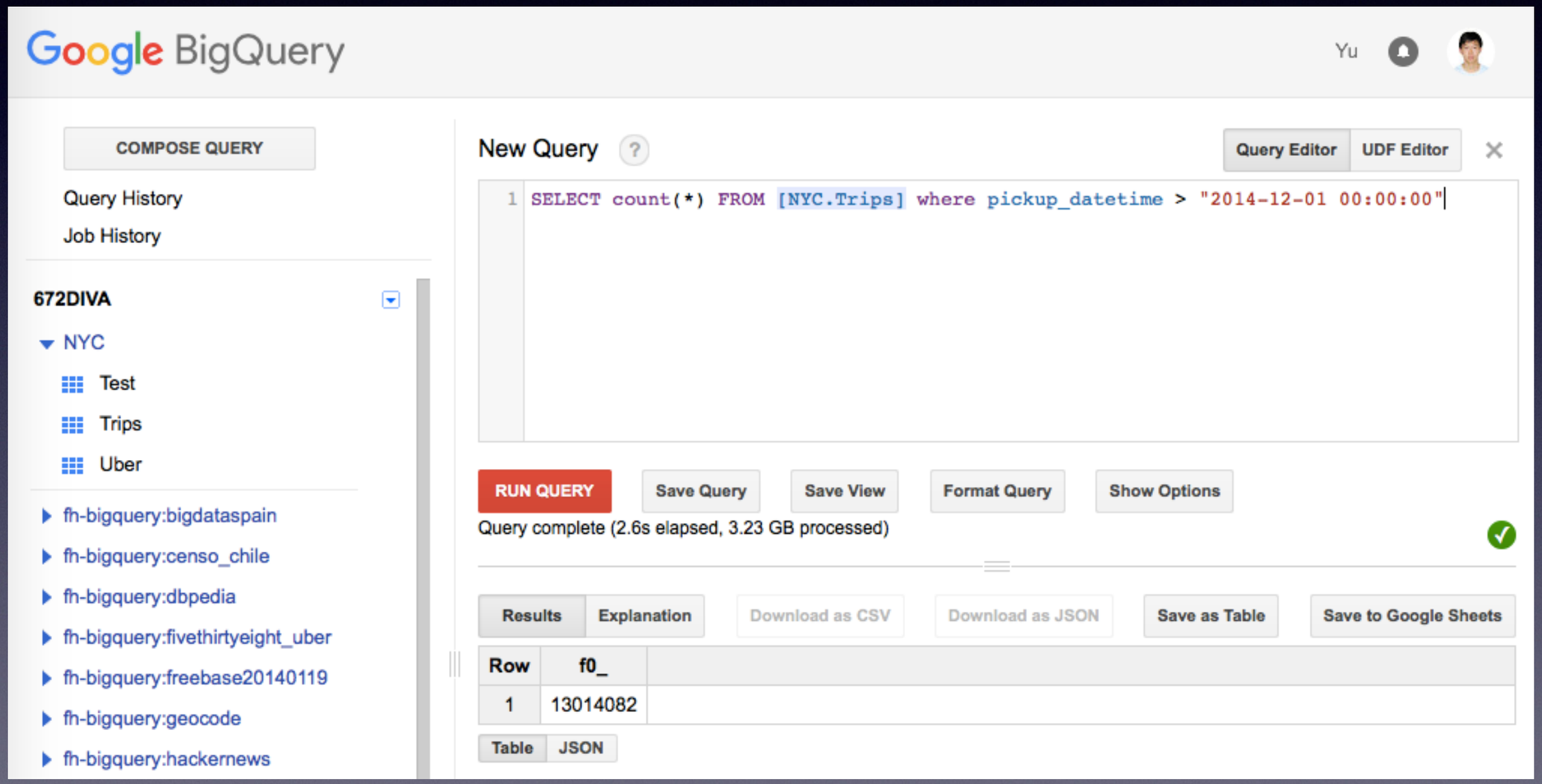
Equality Query



Google BigQuery

- Cloud-based interactive query service for massive datasets
- Implementation of Dremel
- Dremel can scan 35 billion rows without an index in tens of seconds
- SQL-Like query language
- http://dl.acm.org/citation.cfm?id=1920886

Google BigQuery

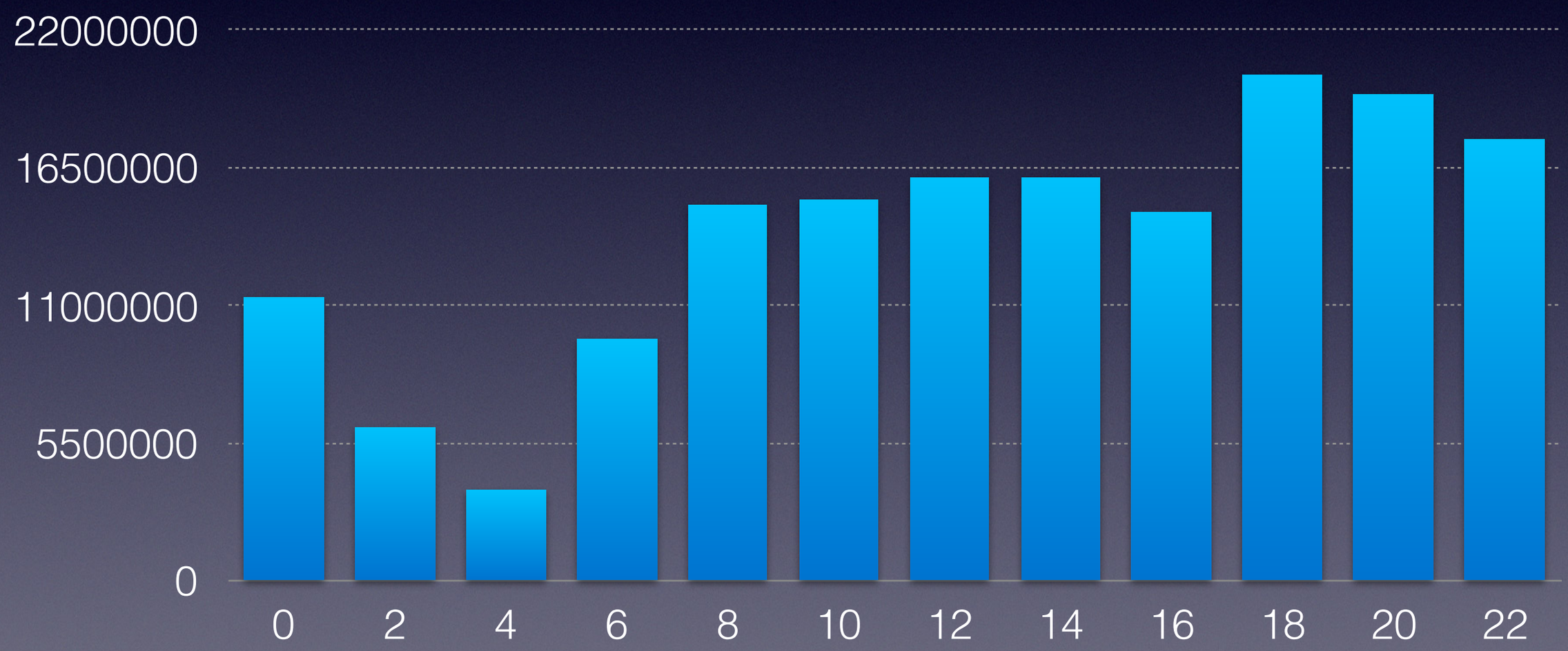


BigQuery VS EMR

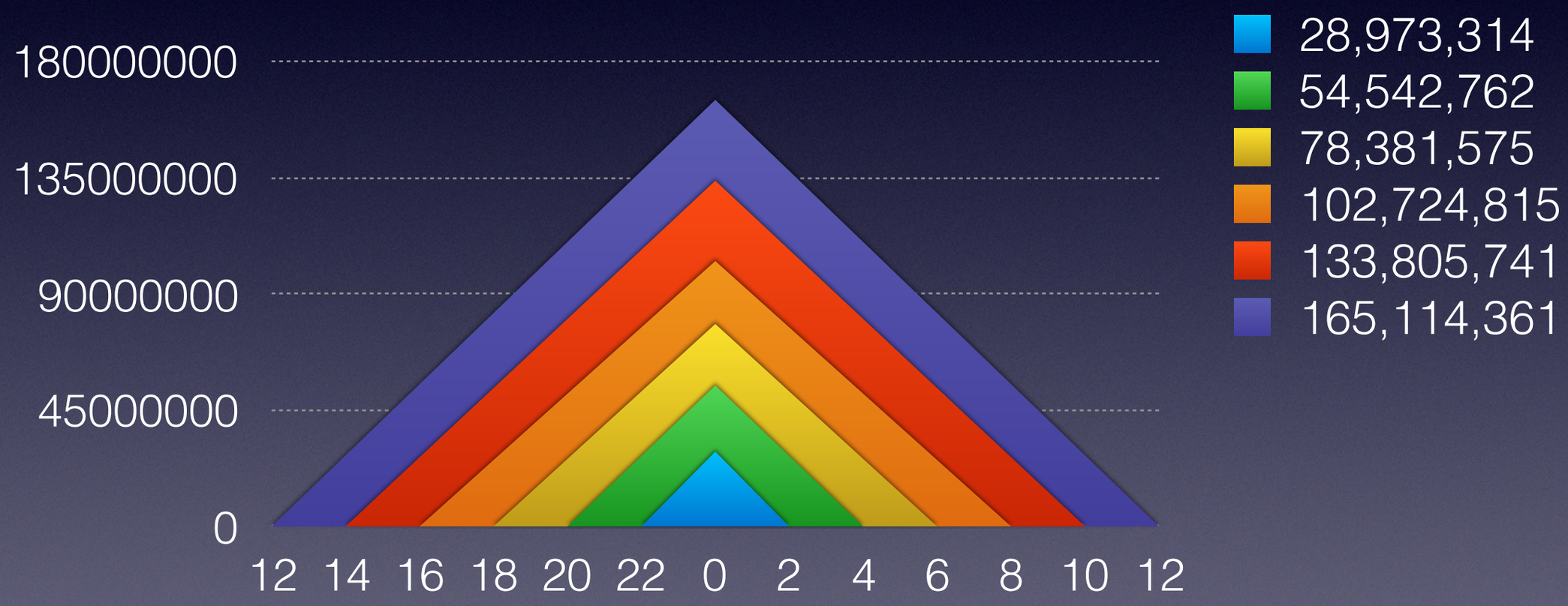
	BigQuery	Elastic Map Reduce
What is it	Query service for large datasets	Programming model for processing large datasets
Common use cases	Ad hoc interactive query of large dataset	Batch processing of large dataset for time-consuming data aggregation
Fast response	Yes	No
Non-programmers friendly	Yes	No
Complex data processing logic	No	Yes
Price	$5 per TB (free 1 TB)	$0.07 per hour (single instance)

Data Analysis

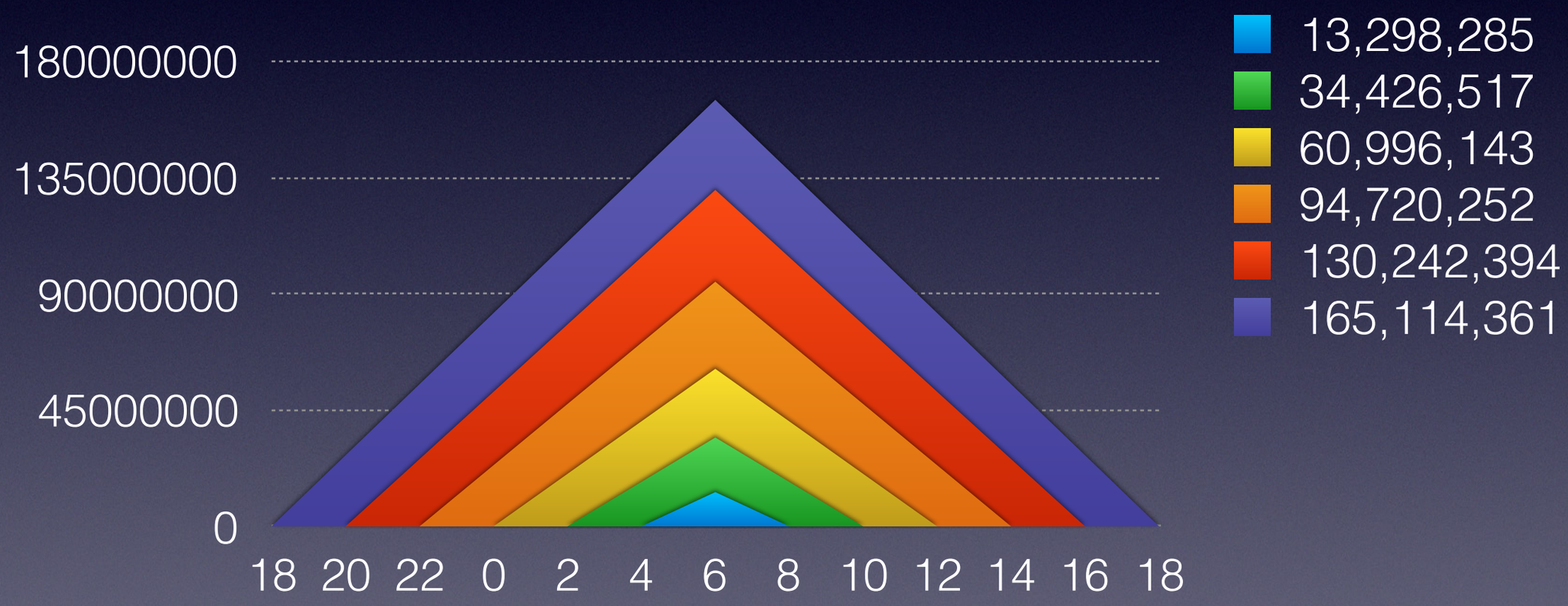
Taxi Demanded at different times



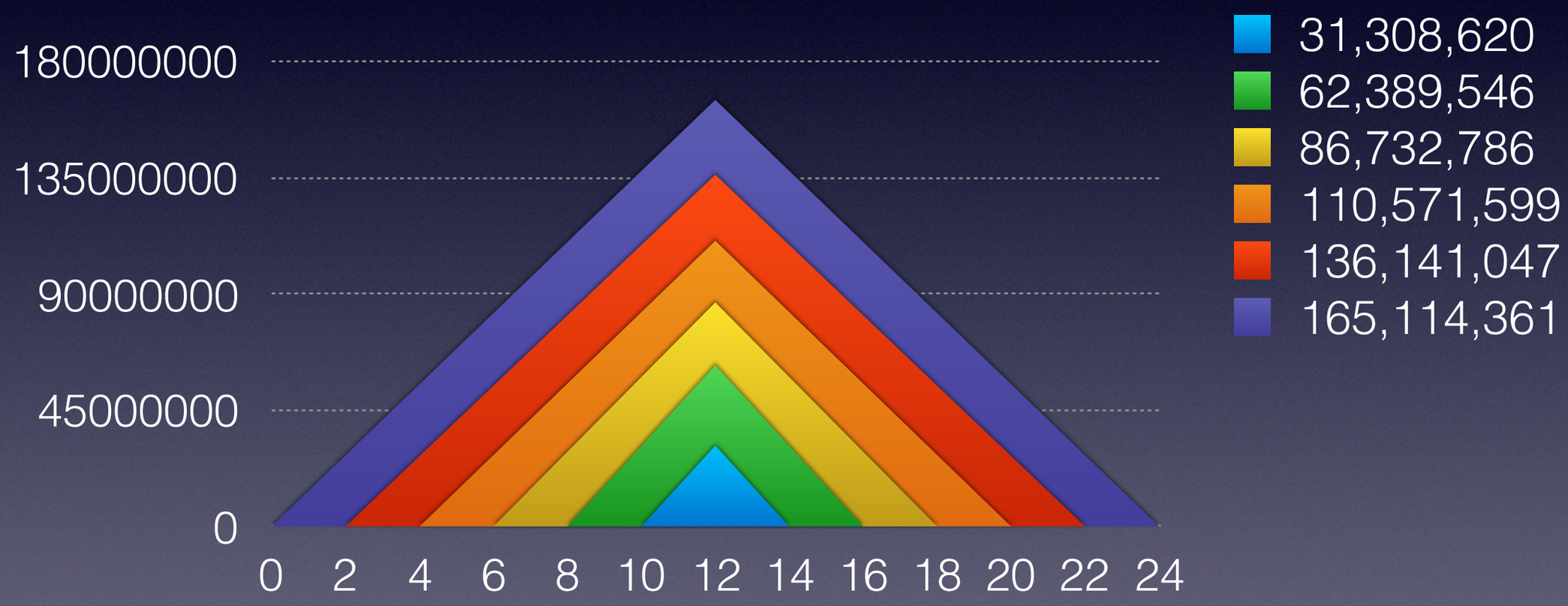
Trips Distribution



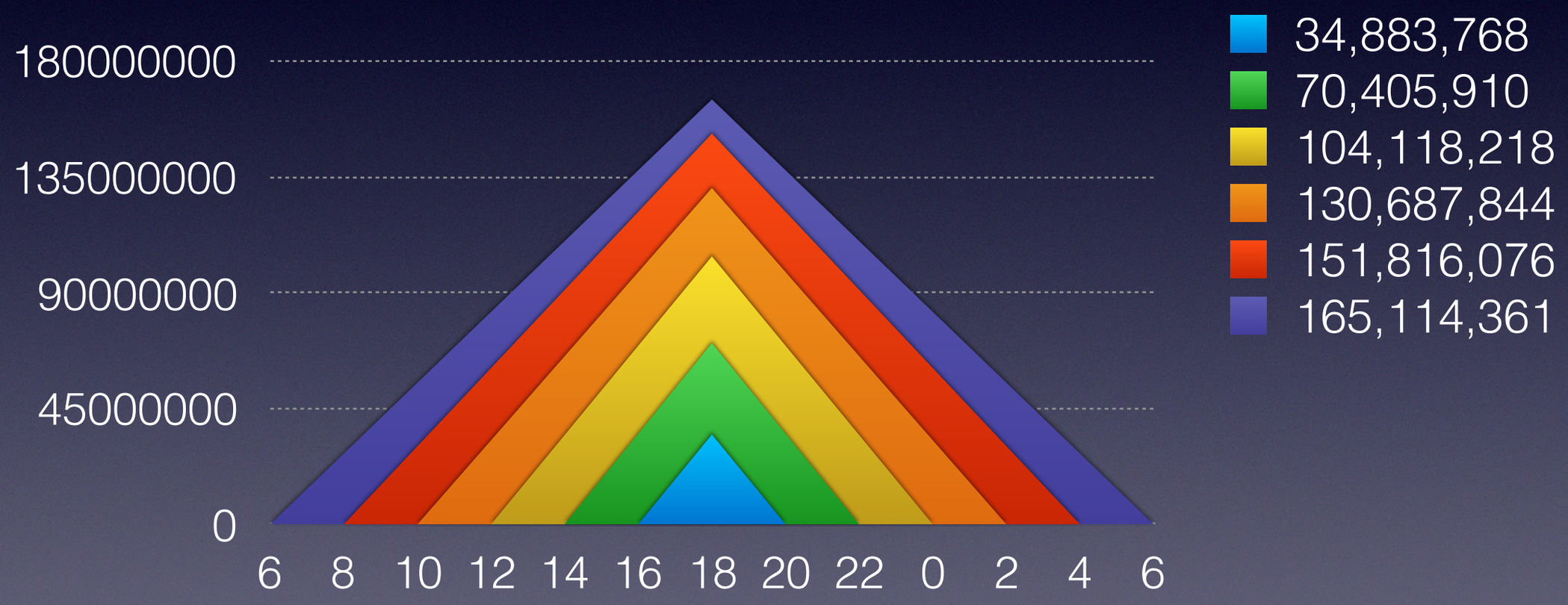
Trips Distribution



Trips Distribution

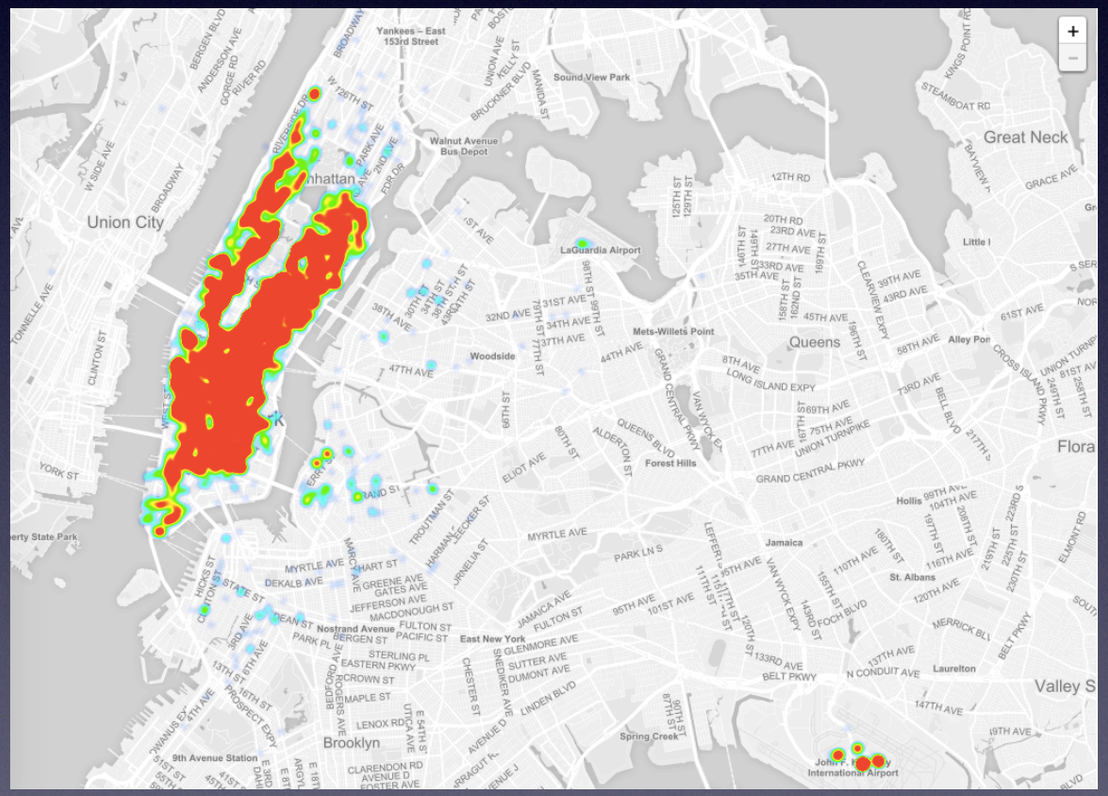


Trips Distribution

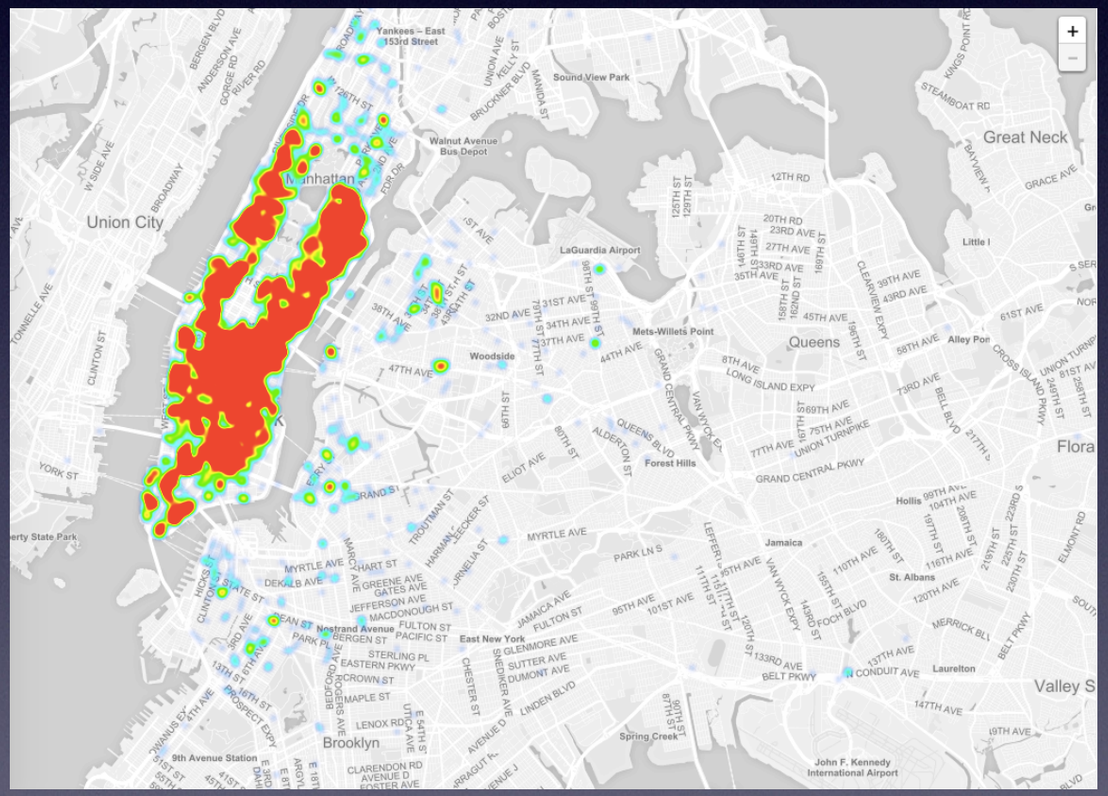


Pick Up / Drop Off Regions

00:00

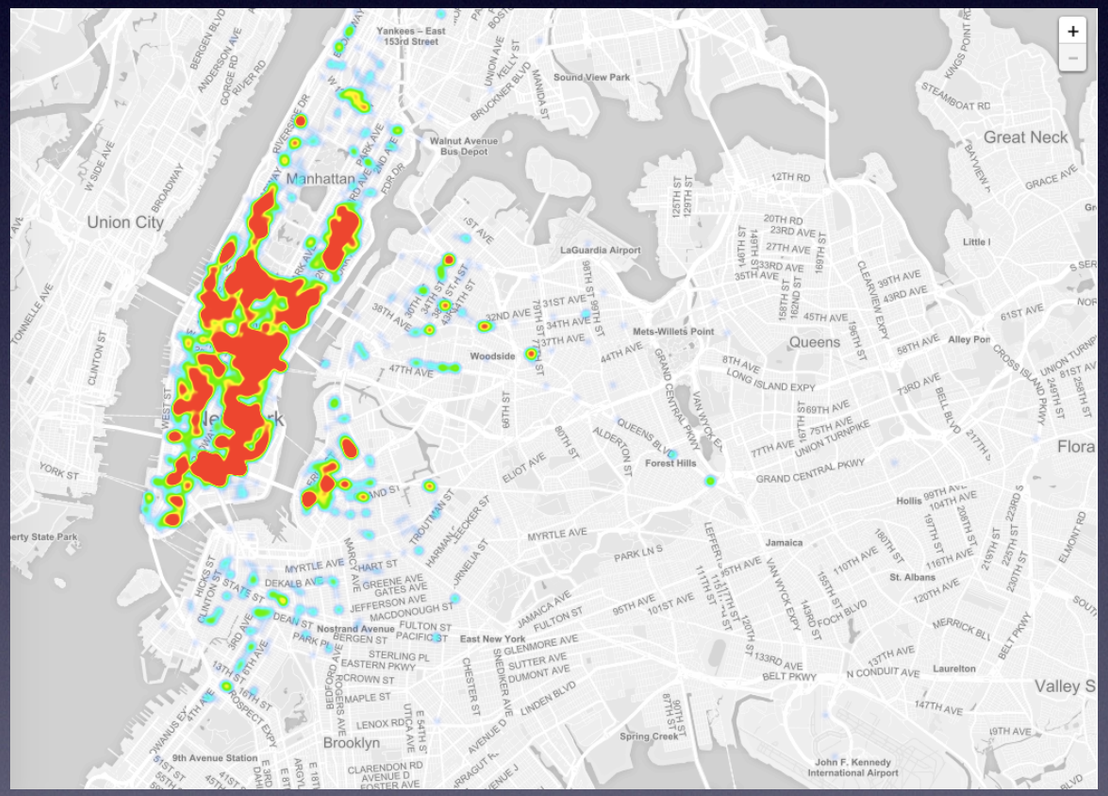


Pick

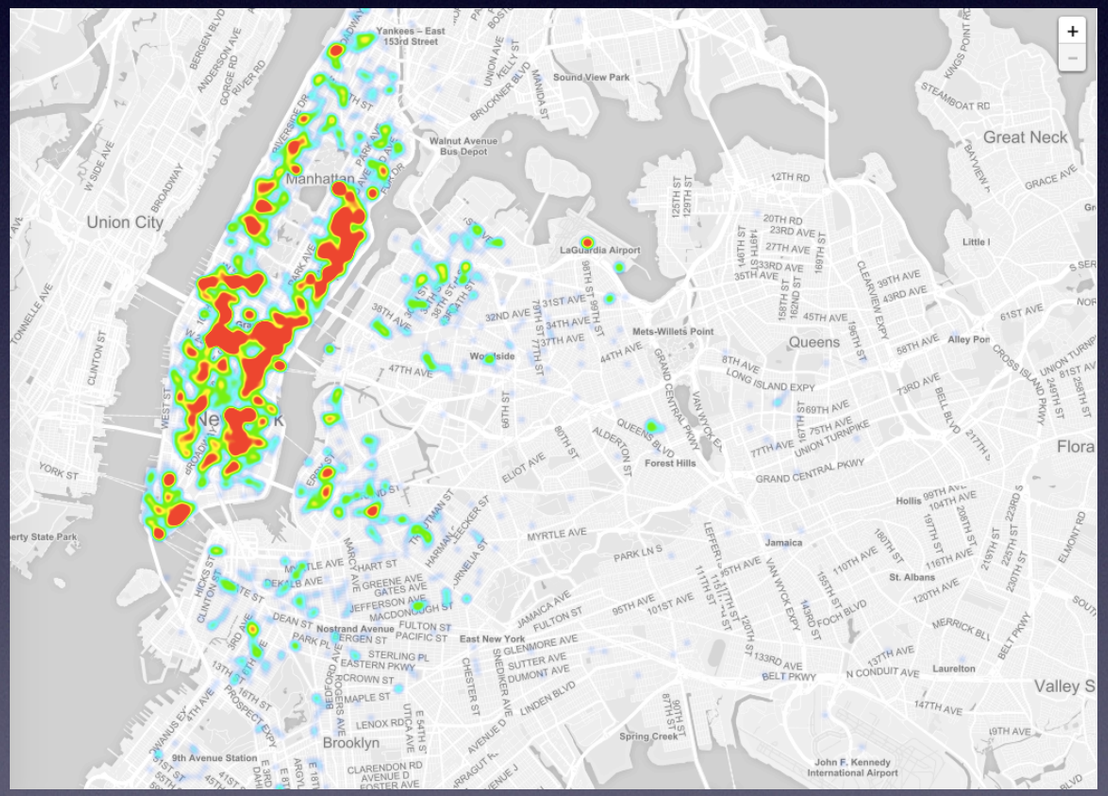


Drop

Pick Up / Drop Off Regions 04:00

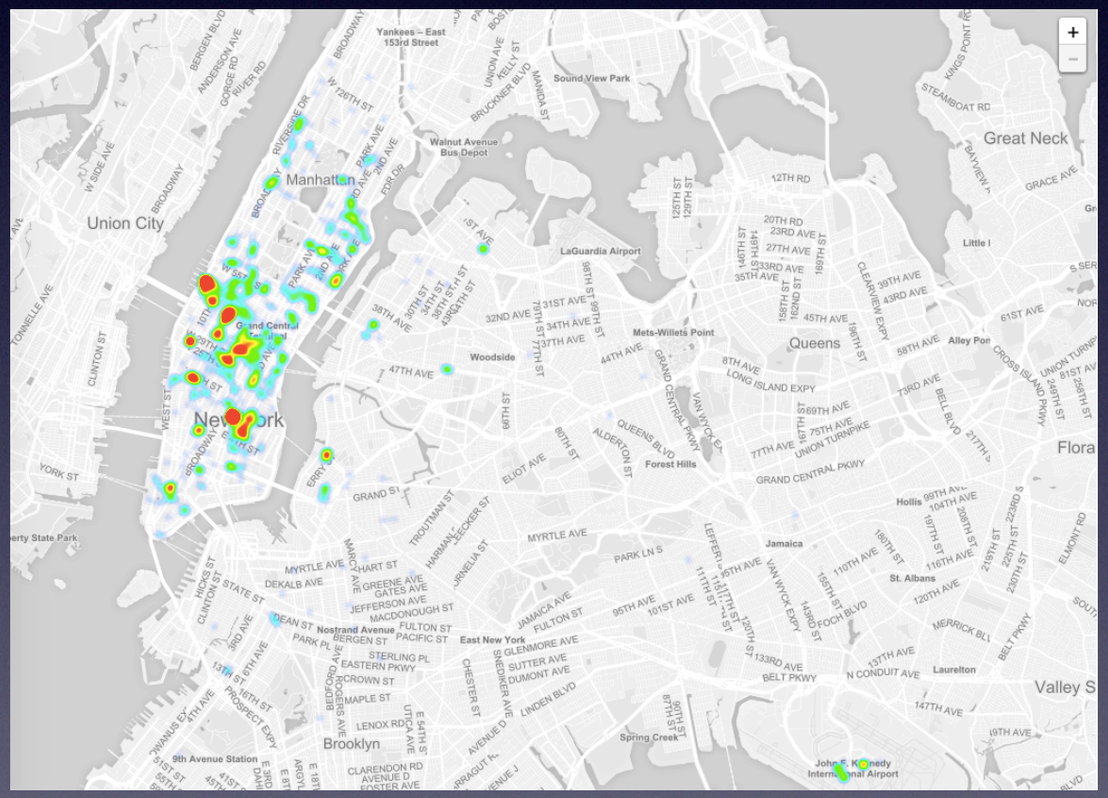


Pick

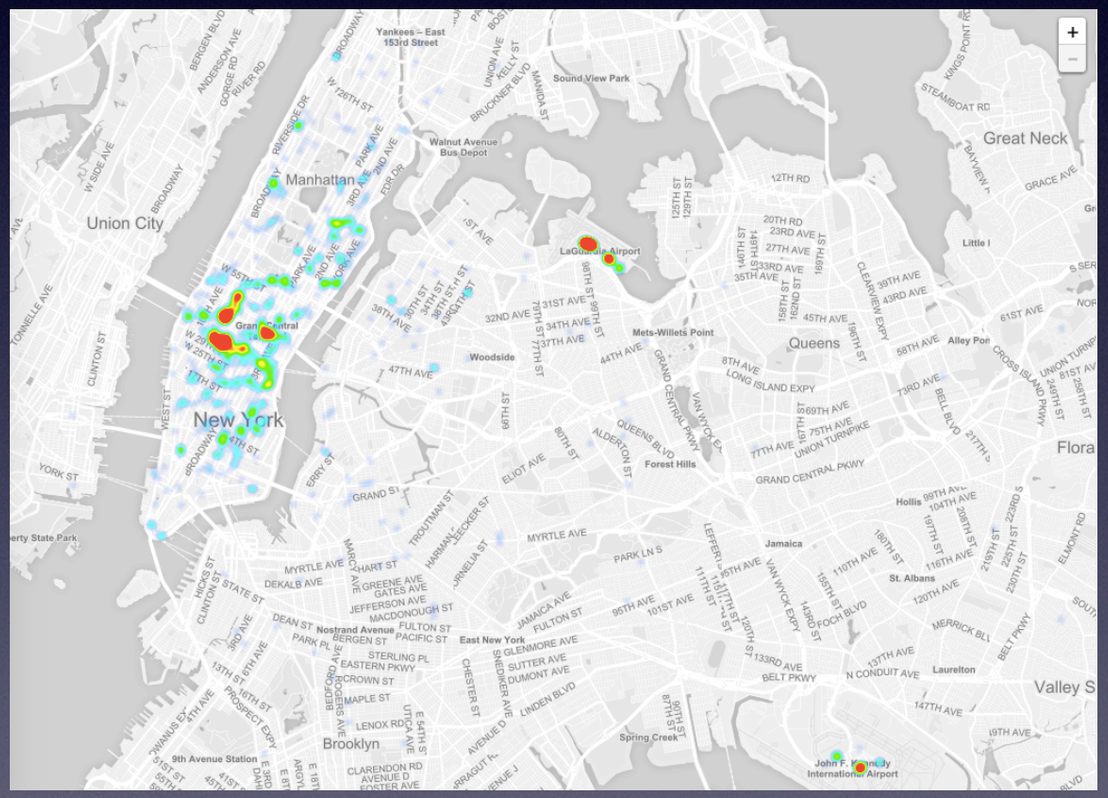


Drop

Pick Up / Drop Off Regions 08:00

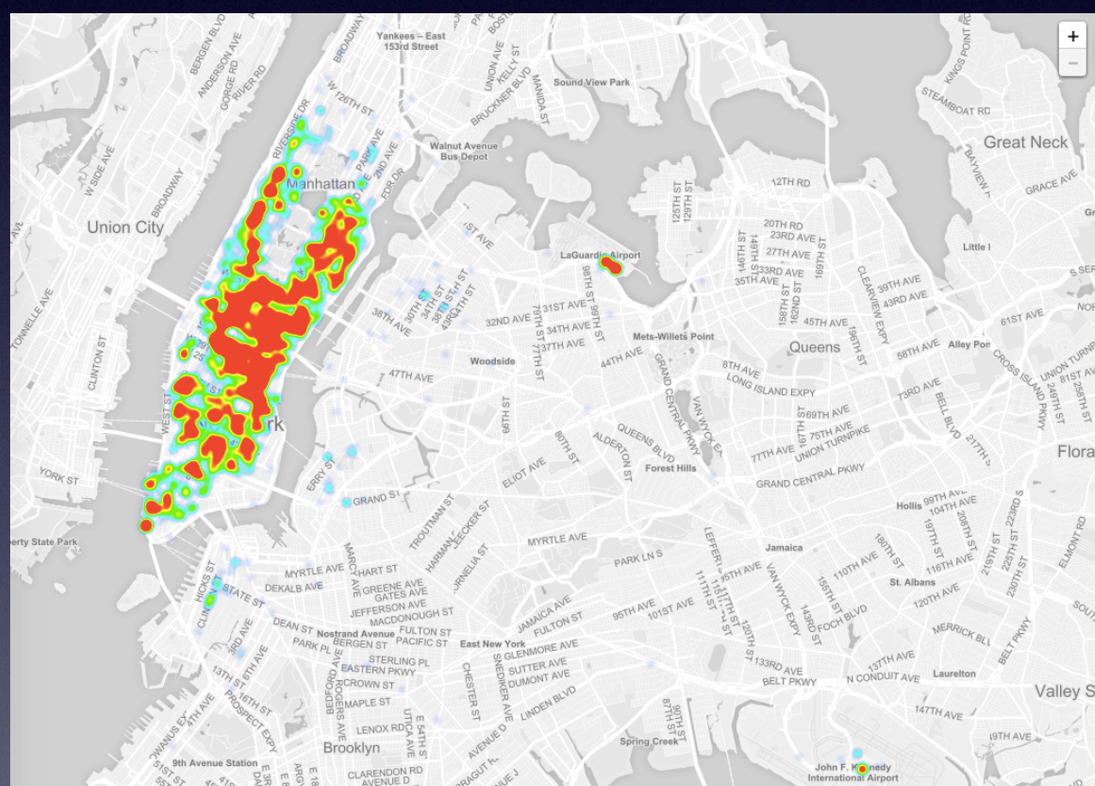


Pick

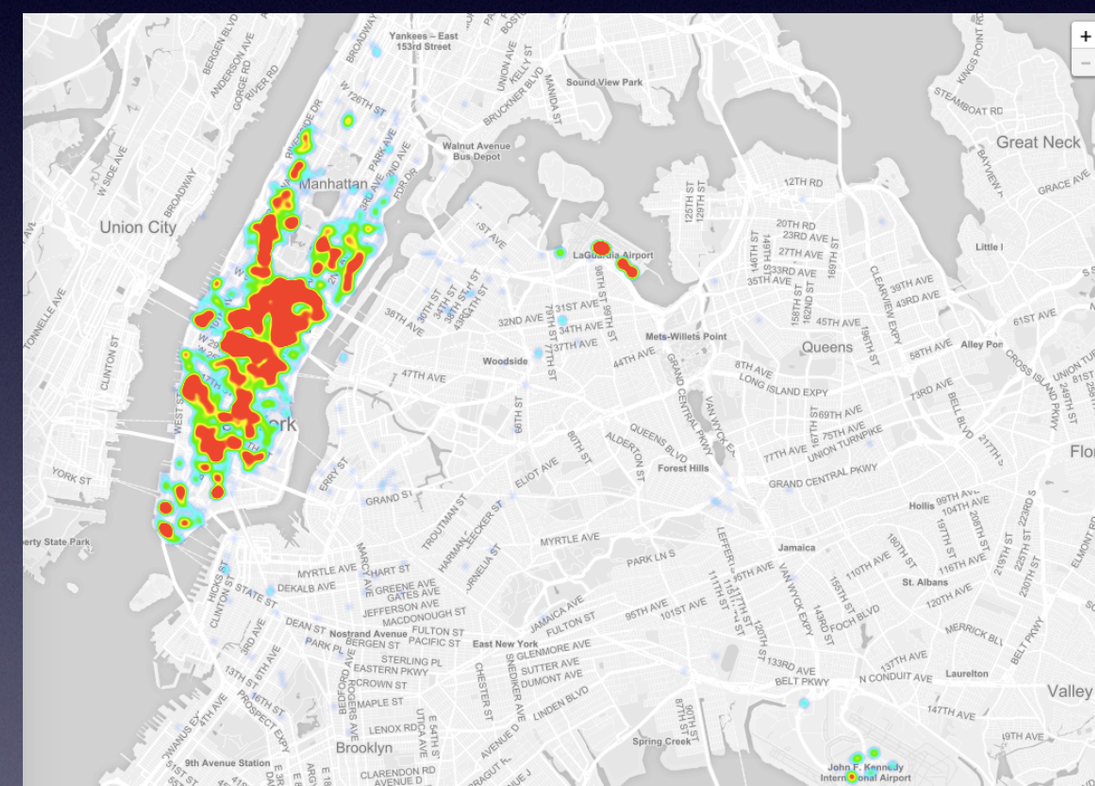


Drop

Pick Up / Drop Off Regions 12:00

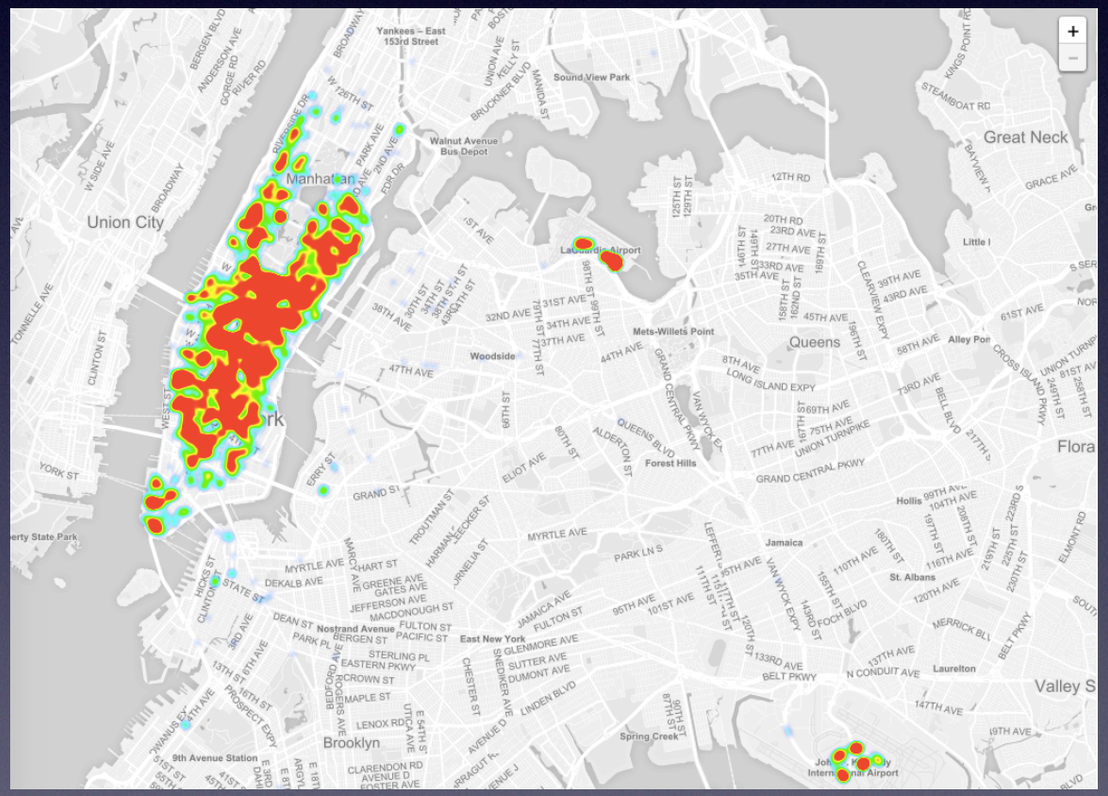


Pick

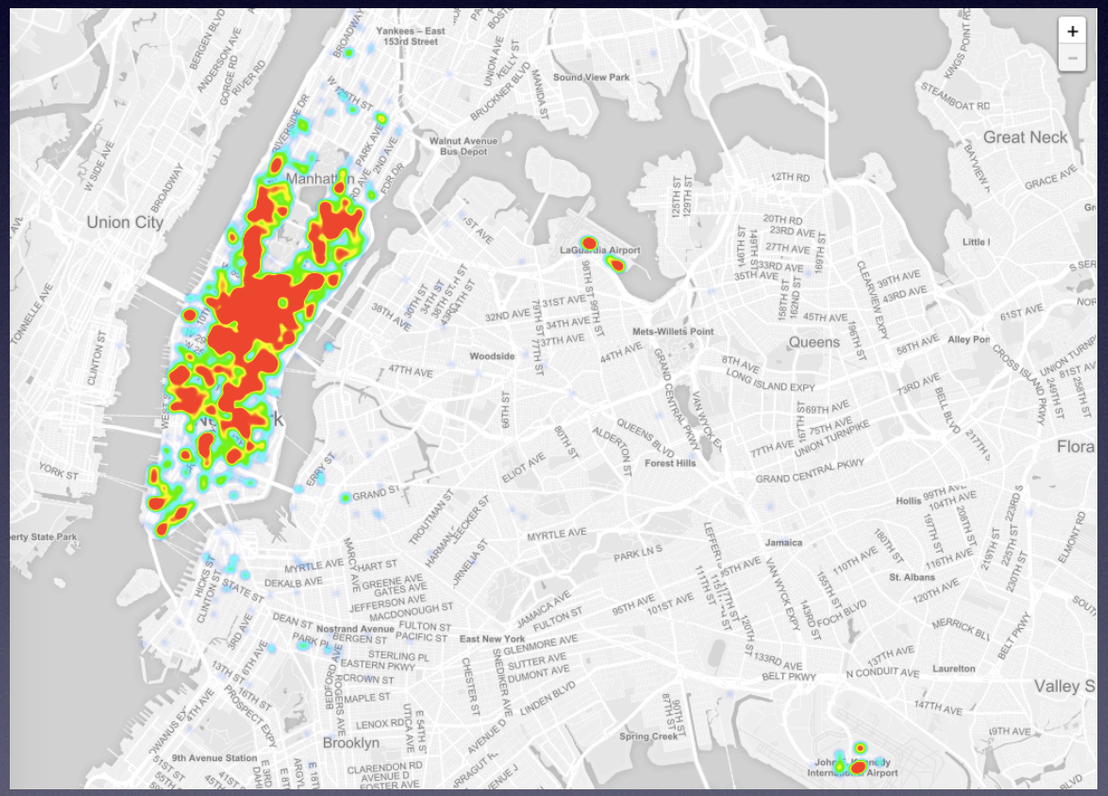


Drop

Pick Up / Drop Off Regions 16:00

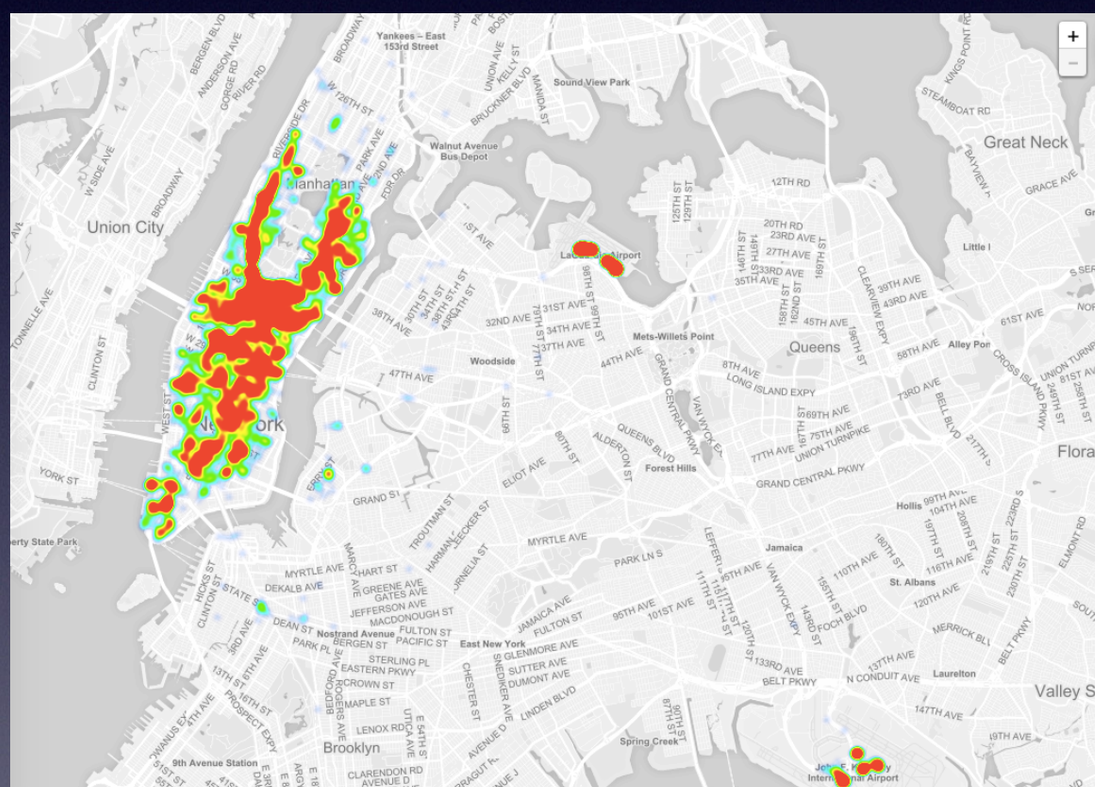


Pick

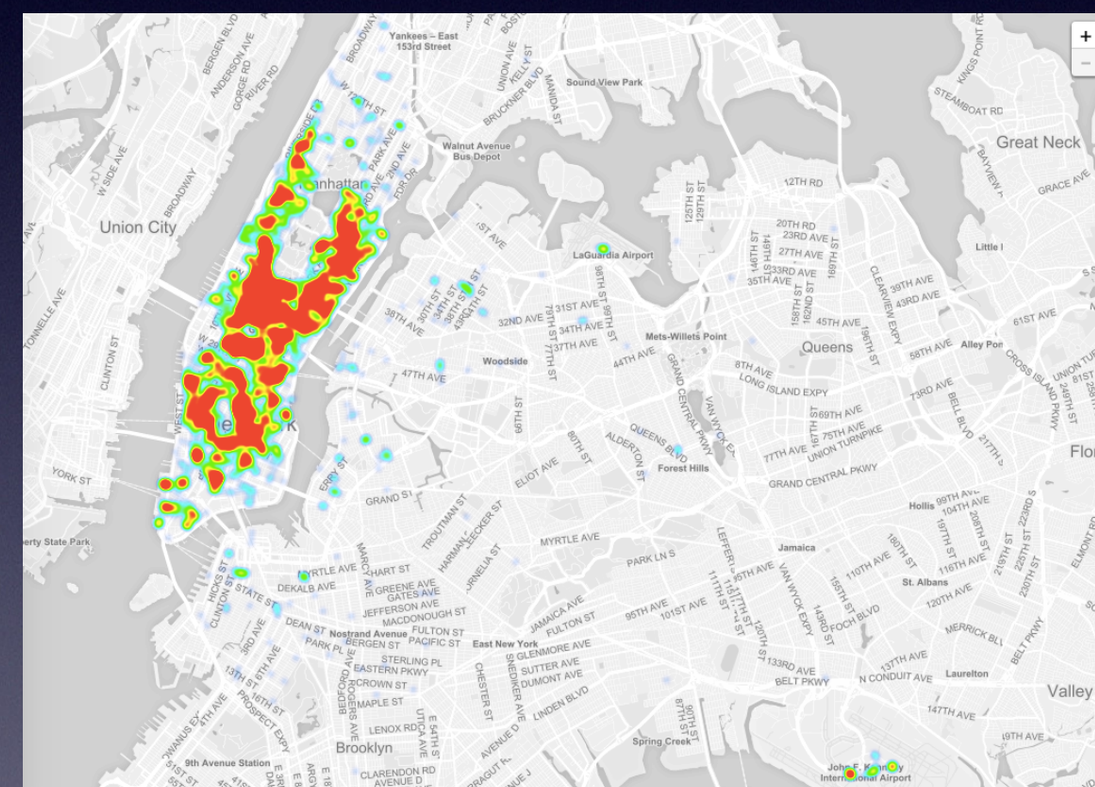


Drop

Pick Up / Drop Off Regions 20:00



Pick



Drop

So What ?

- Drivers — Where to pick up customers
- Customers — Where to take taxis
- Governors — How traffic flows change eg: vary buses numbers in different regions

Demo

Future Work

- Hardware based speeding for millions of data plotting on web browser
- Make the system friendly to other similar datasets (eg. bus, bicycle)

Acknowledge

- Professor James Abello
- Haiying Liu
  Zhen Sun
  Yuhua Ni
  Qi Dong
  Christin Jose
  Soundarajan Thiagarajan