Renters Beware

Pooja Hiranandani
New York University
New York, USA
ph1130@nyu.edu

Kun Jiang
New York University
New York, USA
kj850@nyu.edu

Fatima Mushtaq
New York University
New York, USA
fm1529@nyu.edu

***Abstract*—**

We have created a web analytics application for renters to access relevant but hard-to-acquire information on the neighborhoods and buildings that they are interested in renting in. The application provides users with a heat map view of New York City dwellings and neighborhoods based on statistics derived from our datasets. It also provides a query form for the user to lookup a specific building's data regarding the crimes committed around it, housing violations, 311 service requests, a neighborhood score, a safety score, an investment/gentrification score and a quality of life score for the address provided. It is hoped that, upon using the application, a potential renter has an extensive picture of the housing option under consideration before making the final decision to rent.

Keywords—analytics, rental, housing, violation, 311, crime, New York, city, big data, map.

I. Introduction

Renters are, in general, a vulnerable population, subject as they are to the vagaries of the real estate market, lackadaisical landlords, crumbling apartments and alarming neighborhood surprises. In service of this group, our application makes use of publicly available datasets such as 311 service requests, NYPD crime complaints, Department of Housing Preservation and Development registrations and violations and Department of Buildings violations and permit issuances to provide renters in New York City with pertinent but hard-to-access information about the neighborhoods and buildings that they are interested in. The application presents people with two modes of viewing the data - a query mode and a map mode. If the query mode is chosen, users are presented with a screen asking them to enter the address of a building in which they are considering renting an apartment. Given a building's address, the application fetches pertinent information about the building, specifically the number and types of Department of Buildings and Department of Housing Preservation and Development violations, if any, relevant 311 complaints made by former/current residents of that building, if any, the prevalence and types of crime in the area and neighborhood information and scores. If the map mode is chosen, users are able to view a heatmap with three layers - 311 calls for service, crime and housing violations - and an information map that will present pertinent information about all New York City neighborhoods. Neighborhood scores are compiled based on the statistics gathered and similar neighborhoods for each neighborhood will be presented to the user. K Means clustering is used to present the user with information on similar neighborhoods. Additionally, a correlation analysis has been done on the counts of all the data collated to provide insights into relationships between the variables under considerations and possibilities for future research.

II. Motivation

Renters generally have no more than a cursory insight into the apartment and area in which they plan to live in, provided to them by real estate websites that have no incentive to reveal the less than savory aspects of the rental units on their sites, as both, renters and real estate agencies and property owners are their clients. The motivation for this application is to provide a renter with information that will enable them to live in an area and apartment that matches their expectations and aspirations.

III. Related Work

(Glaeser et al, 2015) provides an overview of the uses of big data in improving urban planning, research, productivity, policymaking and satisfaction. Big data can be used to answer questions about urban development and the economy using longitudinal business database data and urban business sales data from credit card companies to evaluate the determinants of local productivity. Secondly, it can be used to measure the physical city and to measure outcomes that may be influenced by urban space. Social network maps can be derived from Facebook and LinkedIn. In principle, the GPS components of smartphones enable urban mobility to be tracked on a truly fine geographic scale. Thirdly, big data can be used to determine how much people value urban amenities through data like reviews on Yelp. Social patterns are also more observable from Twitter locales as well, and this lets researchers estimate which areas are popular on, say, weekend

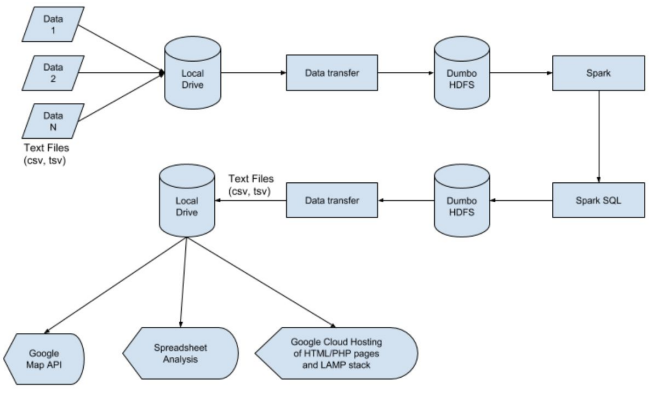
nights. Lastly big data can answer the question of how public policy can improve the quality of physical space. Perhaps the most natural use of big data in city governance is for management. The New York City Police Department, for example, credits the highly local measurement of crime as being a major part of making New York safe. Knowing exact locations of crimes enabled police to target resources and find criminals. Data made it easier to hold precinct chiefs accountable.The data-driven approach to policing works for other public services, as well. The StreetBump mobile app, for example, streams data about potholes from people's cars to Boston's street repairers; the data comes from jolts to the citizens' smartphones. Finally, the paper looks at the connection between big data and government service provision. There are many areas in which public services improve with information, including ensuring sanitary conditions in restaurants and hotels, targeting repairs of potholes, identifying struggling students, or deciding sentences for convicted criminals; in many of these contexts, machine learning can be used to make urban resource allocation more efficient. Complaints on Tripadvisor or Yelp, for example, can be used to guide public inspections via predictive algorithms. Big data thus provides a means of improving city services—often without explicitly requiring causal inference. [1]

(Olascoaga et. al. 2016) include in their analysis of the best neighborhoods of New York City not only datasets of residential building units, their complaints and 311 data but also socially produced datasets of the residents of those neighborhoods. In order to rank the 195 neighborhoods across New York City, the project acquired data from social media and live datasets through web APIs. It used NYC's Socrata Open Data that includes housing, transportation and municipal data of every neighborhood. It also used the 311 calls for service dataset of all the complaints reported from each neighborhood in the city. It used both supervised and unsupervised learning algorithms to make sense of disaggregated datasets. In the data exploration phase, Principal Component Analysis, K-Nearest Neighbors, K-Means and other hierarchical clustering algorithms are used to categorise neighborhoods where users most likely prefer to live. The algorithm suggested that the optimal number of clusters for the 195 New York City neighborhoods is between 3 and 7. The project used sentiment analysis on data retrieved from social media though web APIs to generalize the structure of the text, creating relationships among large amount of social data. This was done through Naive Bayes' Classification algorithm and Latent Dirichlet Allocation. To deal with textual inconsistencies in the data from tweets and Facebook groups, the media posts were first cleaned and parsed. Only English words were used for the sake of consistency. If only traditional measures: social capital, affordability and urban density are considered, the result of the authors align with Zillow and Trulia. However, if people choice is added as a factor, the results are more specific to neighborhood's eccentricities that are responsible for their quality. [3]

(Bansal et al, 2016) presents a good guideline of how a big data application can be developed, from data ingestion, research methodology, experiments all the way to the final presentation. The datasets used in the research paper is very similar with what we are using – housing data, crime data, 311 data, etc. in the city. Segregation of different urban areas was done and issues critical in a region were determined. Primarily, two phase clustering was performed. In the first phase, a dynamic grid based clustering was done on the basis of spatial attribute to analyze complaints that may have strong inter-dependency. In the second phase, the location based clusters are further clustered based on complaint category which helps in determining regions of city imitating similar complaint behavior. The analysis was done on real world data acquired for two cities New York (USA) and Bangalore (India). Experimental results were visualized to enable better interpretation. The results help in planning strategies to improve inhabitants' satisfaction rate and consequently improve their quality of life. The final presentation also aligns with what we are trying to achieve – showing results based on grids and aggregations on a map, and tables/charts. Grid based clustering would be an interesting future direction for our application. [2]

The application that we have developed is an attempt to harness publicly available city government datasets in service of the residents of the city to enable them to make enlightened choices about where they want to make a home and a life. The papers above informed our data sourcing and application design and provided insights into further research in this area.

IV. Application Design

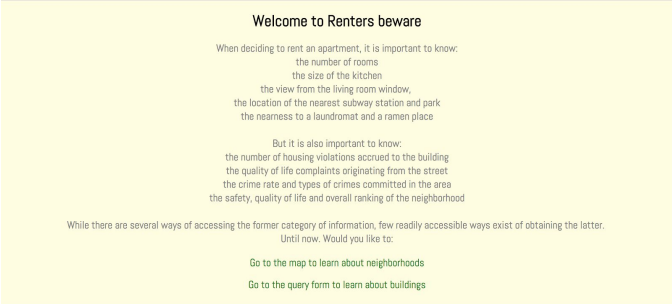


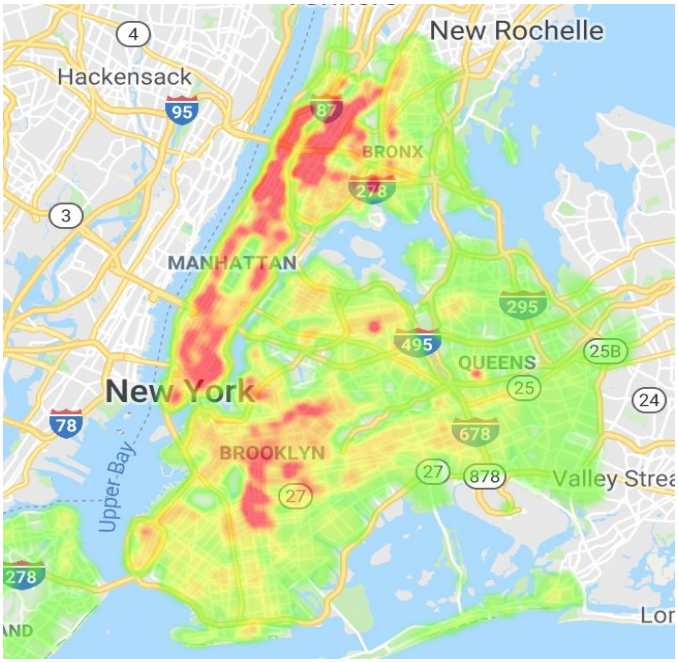
We downloaded the data from our data sources and transferred them to the HPC Dumbo Hadoop cluster at NYU. Then we used Spark and Spark SQL to profile, parse, clean, and create output datasets.[4,5,6] Next, we transferred these output datasets to a local drive and used them as data sources for our frontend which is a web application that consists of a map

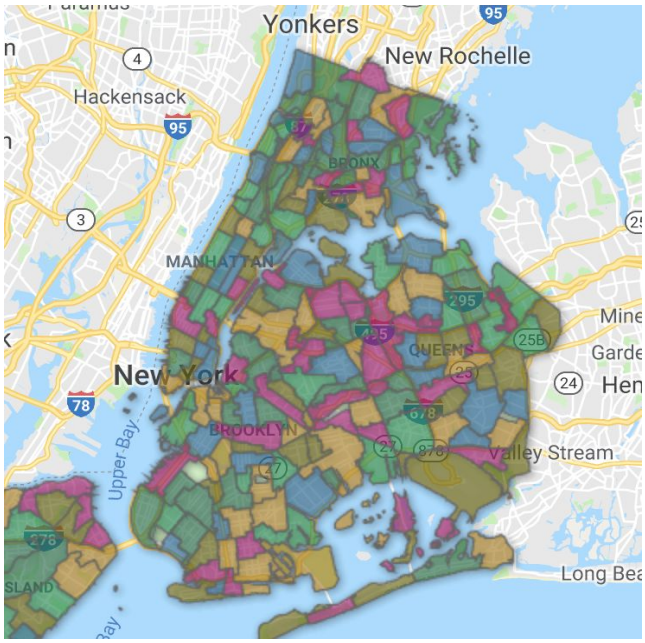
interface and a query form. The map interface is hosted in a bucket on Google Cloud Platform storage and the data is rendered using Google Maps Javascript API. [7,8] The query form is served from a MySQL database hosted on a LAMP stack in Google Cloud Platform Compute Engine.[9] Additionally, the neighborhood dataset was analysed to compile neighborhood scores using a spreadsheet application called Numbers. [10]

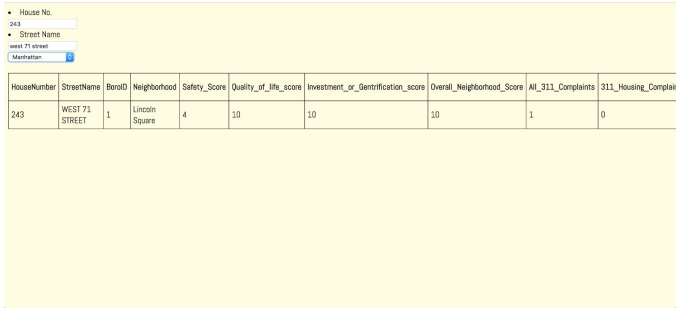
User interface

Link to Web Application:
https://storage.googleapis.com/renters-beware/welcome.html









V. Datasets

1. Multiple Dwelling Registrations – Pursuant to New York City's Housing Maintenance Code, the Department of Housing Preservation and Development (HPD) collects registration information from owners of residential rental units. Owners are required to register if they own residential buildings with three or more units or if they own one or two-family homes and neither they nor members of their immediate family live there. Registrations are required upon taking ownership of a qualifying building, and once a year thereafter. Our application will be concerned only with the addresses found in this dataset. [11]
   This dataset was around ~ 17 MB, collected in July 2018 via a one time download.

2. Department of Housing Preservation and Development violations – Pursuant to New York City's Housing Maintenance Code, the Department of Housing Preservation and Development (HPD) issues violations against conditions, in rental dwelling units and buildings, that have been verified to violate the New York City Housing Maintenance Code (HMC) or

the New York State Multiple Dwelling Law (MDL). This is a dataset of such violations, both open and closed, recorded since October 1, 2012. It contains information such as date, location and type of violation, current status of violation etc. [12]
This dataset was ~ 2 GB, collected in July 2018 via a one time download.

3. Neighborhood Tabulation Areas - Shapefile of New York City neighborhoods for use in the information map and a dataset that maps census tracts to neighborhoods in order to map each dwelling, 311 call, building permit, crime and housing violation to a neighborhood. [13]
   The Shapefile and dataset was ~ 2 MB, collected in July 2018 via a one time download.

4. Department of Buildings Violations - This dataset contains information of violations that have been issued by the Department of Buildings. It contains information such as date, location and type of violation, current status of violation etc. [14]
   This dataset was ~ 400MB, collected in July 2018 via a one time download.

5. Department of Buildings Permit issuance - A list of Department of Buildings (DOB) permits issued for a particular day and associated data. It contains information such as building location, building owner, type of job for which permit has been issued, etc. This dataset was used to extract building construction, renovation and maintenance history. [15]
   This dataset was about 17 MB, collected in July 2018 via a one time download.

6. NYPD Complaint Data - This dataset includes all valid felony, misdemeanour, and violation crimes reported to the New York City Police Department (NYPD) from 2016 onwards to the most recent completed quarter. It contains information such as date, time, location, type of alleged crime etc. [16]
   This dataset was about 34.7 MB, collected in July 2018 via a one time download.

7. 311 service requests - 311 is a telephone number that provides access to non-emergency municipal services. Examples of service requests include abandoned vehicles in roadway, graffiti removal, illegal burning, debris in roadway, code and housing violation etc. The dataset contains 311 Service Requests from 2010 till the present day. It includes details such as the building from which the complaint was made, type of complaint, action taken, date of resolution, etc. [17]
   This dataset was ~10GB in July 2018 and was collected through a one time download.

8. PLUTO - The Primary Land Use Tax Lot Output (PLUTO™) data file was developed by the New York City Department of City Planning's Information Technology Division (ITD)/Database and Application Development Section. It contains extensive land use and geographic data at the tax lot level. This dataset was used to map geographical coordinate information to the multiple dwellings dataset which was missing this data. [18]
   This dataset was ~250 MB in July 2018 and was collected through a one time download.

The schemas of the datasets that were used in our application are as follows:

Information Map schema

Neighborhood Name : *Text*	Count of 311 Streets and Sidewalks Complaints : *Number*
Neighborhood shape in the form of KML text	Count of all 311 complaints : *Number*
Count of 311 housing complaints : *Number*	Count of 311 Graffiti complaints: *Number*
Count of 311 noise complaints : *Number*	Count of all building violations : *Number*
Count of low pressure boiler violations: *Number*	Count of elevator violations : *Number*
Count of construction violations : *Number*	Count of no boiler inspection violations : *Number*
Count of extremely hazardous violations : *Number*	Count of all crimes : *Number*
Count of rapes and other sex crimes : *Number*	Count of grand larcenies: *Number*
Count of burglaries : *Number*	Count of frauds and thefts : *Number*
Count of harassment : *Number*	Count of murders and manslaughters : *Number*
Count of arson : *Number*	Count of building permits issued : *Number*
Overall Neighborhood Score : *Number*	Neighborhood safety score : *Number*
Neighborhood quality of life score : *Number*	Neighborhood Investment/gentrification score : *Number*

Heatmap schema

Latitude: *Float*
Longitude : *Float*
Count of 311 calls: *Number*
Count of crimes: *Number*
Count of rental housing violations : *Number*

Query form schema

House Number : *Text*	Street Name : Text
Borough Code : *Number*	Count of all 311 complaints : *Number*
Count of 311 Streets and Sidewalks Complaints : *Number*	Count of 311 Graffiti complaints: *Number*
Count of 311 housing complaints : *Number*	Count of 311 rodent complaints *: Number*
Count of 311 Air and Water Quality complaints : *Number*	Count of 311 Public Drinking complaints: *Number*
Count of 311 garbage complaints : *Number*	Count of 311 noise complaints : *Number*
Count of all rental housing violations : *Number*	Count of elevator violations : *Number*
Count of low pressure boiler violations: *Number*	Count of no boiler inspection violations : *Number*
Count of construction violations : *Number*	Count of extremely hazardous violations : *Number*
Count of Housing Preservation and Development Department Violations : *Number*	Count of all crimes : *Number*
Count of rapes and other sex crimes : *Number*	Count of grand larcenies: *Number*
Count of burglaries : *Number*	Count of frauds and thefts : *Number*
Count of harassment : *Number*	Count of murders and manslaughters : *Number*
Count of arson : *Number*	Overall Neighborhood score : *Number*
Neighborhood safety score : *Number*	Neighborhood quality of life score : *Number*
Neighborhood Investment/gentrification score : *Number*	

VI. Remediation

After viewing information on the neighborhoods and buildings under consideration, the user is provided with a link to the http://streeteasy.com profile of the building they are interested in to fix an appointment for a viewing (if there are apartments available).

VII. Experiments

The key features of our application are:

1. **Neighborhood Rankings**

   Neighborhood rankings were compiled in the following way:

   Crime per capita and 311 calls for service per capita and housing violations per capita were calculated for every neighborhood using the counts collated from our datasets and the Census 2010 population data. These per capita columns were sorted and the neighborhoods were divided into ten equally sized groups. Each group was given the same score.

   For the safety score, the crime per capita column was sorted in ascending order and the scoring started at 10, all the way down to 1.

   For the quality of life score, the 311 calls for service per capita and housing violations per capita were sorted in ascending order, one after the other, and the scoring started at 10, all the way down to 1.

   For the investment/gentrification score, the building permits count column was sorted in descending order and the scoring started at 10, all the way down to 1.

   The overall neighborhood score is an aggregation of these three scores, with most weightage given to the quality of life score, second most weightage given to the safety score and the least weightage given to the investment/gentrification score.

2. Google Map View and Query Form

A heatmap with three layers representing crime, 311 calls for service and rental housing violations was developed. A neighborhood level information map with all statistics related to neighborhoods was developed. A query form was also developed to enable users to acquire information about a specific building of interest. The schemas described above detail the fields displayed in the various views. For the maps, data from the last 1.5 - 2 years was considered. For the query form, the time range was longer, about 5 years, for the 311 calls and housing violations as anything shorter would mean a very sparsely populated dataset as counts of 311 calls/violations per building are being sought. Crime data however is more recent as these were compiled for entire areas/streets. We took crime data of the year 2017 upto the first quarter of this year, 2018.

3. K-Means Clustering

K-Means clustering was performed on the neighborhoods to cluster together similar neighborhoods based on the variables we had gathered data on. The clusters are presented to the user in the neighborhood information map and query form under the label 'similar neighborhoods'. 30 clusters were formed because experimentation revealed that any fewer and the clusters got too big and if we went beyond 30, there were many clusters with just one or two neighborhoods.

4. Correlation Analysis

We performed a correlation analysis on all the variables we had gathered data on - to learn of significant relationships between them that could point to possibilities for future application development and research.

It was discovered; there is a strong positive correlation between counts of permits and housing violations (0.82) and between counts of permits and crime (0.76) and between crime and housing violations (0.77).

Some insights from the analysis:

- Often the investment/gentrification score is very high for neighborhoods widely known as currently gentrifying, but other statistics such as crime and housing violations take time to improve and catch up with the increasing levels of investment in the area.
- Areas that have high levels of crimes also have high levels of housing violations signalling a lack of investment in areas of high crime.
- Permits and violations go hand in hand which is a curious discovery that deserves more investigation.

We faced a few challenges along the way. Among the most significant were:

- **Multiplicity of datasets:** We had multiplicity of datasets of varying sizes and formats, including a shapefile. Cleaning, understanding and joining them sensibly to meet our objectives proved to be more time-consuming than we had initially imagined.
- **Mapping geographical coordinates to textual addresses:** Our base dataset of rental housing did not have any geographical coordinate information. The set had to be mapped against the PLUTO dataset to obtain NY State Plane coordinates which then had to be converted to latitude-longitudes for display on the map page. This took some time to figure out and involved the use of an external library called GeoSpark.[19]
- **Mapping textual addresses to Crime:** The crime dataset did not have any textual address information that could be mapped to the rental housing data organically. It only had geographical coordinate information. Therefore, to map crime to buildings, we initially used another dataset Center Street Line from NYC Open Data, that has street information with geographical coordinates for each block belonging to that street. If we mapped crime to only the block of buildings where it was committed, it would exclude all the nearby buildings that are equally affected by the crime.

  Haversine distance is used to measure distance between two geographical coordinates. Therefore, we opted for a circular mile radius approach using haversine distance. That is, if a building is within 0.15 miles radius of crime location, it is mapped to that crime. 0.15 mile buffer was chosen as it covers slightly more than a block distance on a street.

  Haversine Formula:

  $$a = sin^2(\Delta\phi/2) + cos(\phi_1).cos(\phi_2).sin^2(\Delta\lambda/2)$$

  $$c = 2.atan2(\sqrt{a}, \sqrt{(1-a)})$$

  $$d = R.c$$

  where ϕ is latitude, λ is longitude, R is the earth's mean distance in miles (3959.8 miles).

  The angles (two lat, long difference : ϕ, λ) is converted to radians before passing them to the functions.

We had wanted to use spark streaming to update our application in realtime but were unable to, for lack of time. We had also wanted our application to contain information about landlords but we did not have the time to integrate that into the application. Additionally it would be nice to include

some predictions in the application such as areas most likely to experience a rise/fall in crime rate, housing violations, areas most likely to gentrify or decline in the next few years based on investment information. This would involve building some linear regression models. It might also be interesting to see how well a linear regression model involving the variables we have focussed on as features might perform at predicting rental or property prices.

VIII. Conclusion

All our analytics are deduced solely from numbers in datasets provided by the NYC Open Data website, which has datasets sourced directly from governmental departments. We compared our rankings with results from other online ranking websites for New York City neighborhoods like https://www.niche.com which compiles rankings based on crime, public schools, cost of living, job opportunities, and local amenities. The rankings are equivalent in most cases (top 50 in niche.com found are scored 7-10 in our system when the neighborhood names match exactly) and where there is a divergence, either the neighborhoods are conceived of differently or our investment/gentrification score (based on building permits issued) explains the situation or both.

For example, Park-Slope-Gowanus has a neighborhood score of 3 but a gentrification score of 10. Park Slope is in the top 10 neighborhoods in niche.com but because of the difference in neighborhood conceptions (Park Slope-Gowanus vs. Park Slope) and perhaps because other statistics have not yet caught up with the ongoing gentrification, it scores low in our system which gives a higher weightage to 311 calls for service, crime rate and housing violations.

We hope that this web application is filled with information useful to prospective renters and property investors alike. Sources of this information are all available in the public domain but lie dormant in a format that is not easily accessible. There is a need for more applications of this sort that weaponise publicly available datasets to empower and enlighten citizens and residents alike.

Acknowledgments

Google Cloud Platform https://cloud.google.com/ for storage and hosting

Google Maps API https://cloud.google.com/maps-platform/ for the maps interface

GeoSpark http://datasystemslab.github.io/GeoSpark/ for converting between coordinate systems

NYC Open data https://opendata.cityofnewyork.us/data/

NYU HPC department and Prof. Suzanne McIntosh.

References

1. Edward L. Glaeser, Scott Duke Kominers, Michael Luca, Nikhil Naik. Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. NBER, Boston, MA, December 2015.
2. P. Bansal and D. Toshniwal. Analyzing civic complaints for proactive maintenance in smart city. IEEE, Okayama, Japan, June 2016.
3. Sandoval Olascoaga, Carlos S Xu, Wenfei Flores, Hector. Crowd-Sourced Neighborhoods - User-Contextualized Neighborhood Ranking. Berkeley, CA, August 2016.
4. S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP ‘03, 2003.
5. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S. and Stoica, I. (2010) Spark: Cluster Computing with Working Sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud’10), USENIX Association, Berkeley, 10 p.
6. Armbrust, Michael, Xin, Reynold S, Lian, Cheng, Huai, Yin, Liu, Davies, Bradley, Joseph K, Meng, Xiangrui, Kaftan, Tomer, Franklin, Michael J, Ghodsi, Ali, et al. Spark SQL: Relational data processing in Spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1383–1394. ACM, 2015.
7. https://developers.google.com/maps/documentation/javascript/tutorial
8. https://cloud.google.com/storage/
9. https://cloud.google.com/compute/
10. https://www.apple.com/numbers/
11. https://data.cityofnewyork.us/Housing-Development/Multiple-Dwelling-Registrations/tesw-yqqr
12. https://data.cityofnewyork.us/Housing-Development/Housing-Maintenance-Code-Violations/wvxf-dwi5
13. https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page
14. https://data.cityofnewyork.us/Housing-Development/DOB-Violations/3h2n-5cm9
15. https://data.cityofnewyork.us/Housing-Development/DOB-Permit-Issuance/ipu4-2q9a
16. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-YTD/5uac-w243
17. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
18. https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page
19. http://datasystemslab.github.io/GeoSpark/