CSDA 1020 - Big Data Analytics Tools
Project 4: ELK (Elasticsearch, Logstash, Kibana)

Prepared by: Rani Lottey

1.0 Business Problem

The City of New York has an open data set for 311 service requests which is updated daily. In an effort to gain valuable insights from this dataset an ELK (Elasticsearch, Logstash and Kibana) datastack and a Kibana dashboard will be set-up for some initial insights and analysis purposes.

2.0 Dataset

The open dataset for 311 service requests in New York City has 41 columns and to date 26.4 million records where each record is a 311 service request. The dataset was made public October 18, 2011 and the dataset owner is NYC OpenData. The dataset can be found at https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9.

3.0 ELK Datastack Set-up

As shown in Project 3, Elasticsearch, Kibana and Logstash were downloaded, configured and launched in a Hadoop single-node cluster using the GCP platform. Appendix A provides the commands used for this set-up. The Logstash configuration code shown in Figure 1 was used to bring in the dataset to the Logstash pipeline, apply some filters and then output to be used by Kibana.

The index that is created by the Logstash is called “new_nyc311”; this is the same index that will be created and used within Kibana once Logstash is initialized.

```
input {
   file {
     path => "/home/krlottey/311_service.csv"
     start_position => "beginning"
     sincedb_path => "/dev/null"
    }
 }
filter {
       csv {separator =>","
           columns => ["Unique Key","Created Date","Closed Date","Agency","Agency Name","Complaint
Type","Descriptor","Location Type","Incident Zip","Incident Address","Street Name","Cross Street 1","Cross
Street 2","Intersection Street 1","Intersection Street 2","Address Type","City","Landmark","Facility
Type","Status","Due Date","Resolution Description","Resolution Action Updated Date","Community
Board","BBL","Borough","X Coordinate (State Plane)","Y Coordinate (State Plane)","Open Data Channel
Type","Park Facility Name","Park Borough","Vehicle Type","Taxi Company Borough","Taxi Pick Up
Location","Bridge Highway Name","Bridge Highway Direction","Road Ramp","Bridge Highway
Segment","Latitude","Longitude","Location"]
                }
            date{match => ["Created Date", "MM/dd/yyyy hh:mm:ss a"]
            target => "Created Date"}
            date{match => ["Closed Date", "MM/dd/yyyy hh:mm:ss a"]
            target => "Closed Date"}
            date{match => ["Due Date", "MM/dd/yyyy hh:mm:ss a"]
            target => "Due Date"}
            date{match => ["Resolution Action Updated Date", "MM/dd/yyyy hh:mm:ss a"]
            target => "Resoultion Action Updated Date"
}

    mutate {convert => ["Incident Zip","integer"]}
    mutate {convert => ["BBL","integer"]}
    mutate {convert => ["X Coordinate (State Plane)","integer"]}
    mutate {convert => ["Y Coordinate (State Plane)","integer"]}
    mutate {convert => ["Latitude","float"]}
    mutate {convert => ["Longitude","float"]}
    mutate {copy =>
                        { "Longitude" => "[location][lon]"
                          "Latitude" => "[location][lat]" }
               }
    mutate {replace => { "Location" => "%{Longitude},%{Latitude}" }}
    }

output {
 elasticsearch {
 hosts => "localhost"
 index => "new_nyc311"}
stdout {codec => dots}
}
```

Figure 1: Logstash Configuration Code

4.0 Kibana Configuration

Prior to the index new_nyc311 being created and used in Kibana, this index will need to be processed further so that the location data, which was filtered in the Logstash configuration, can be used for mapping in Kibana. Using the Dev Tools console in Kibana the location type is changed to a geopoint which is recognized by Kibana mapping. Figure 2 shows the Dev Tools console and the commands used for this operation. Figure 3 shows the index new_nyc311 created in Kibana which has loaded properly.

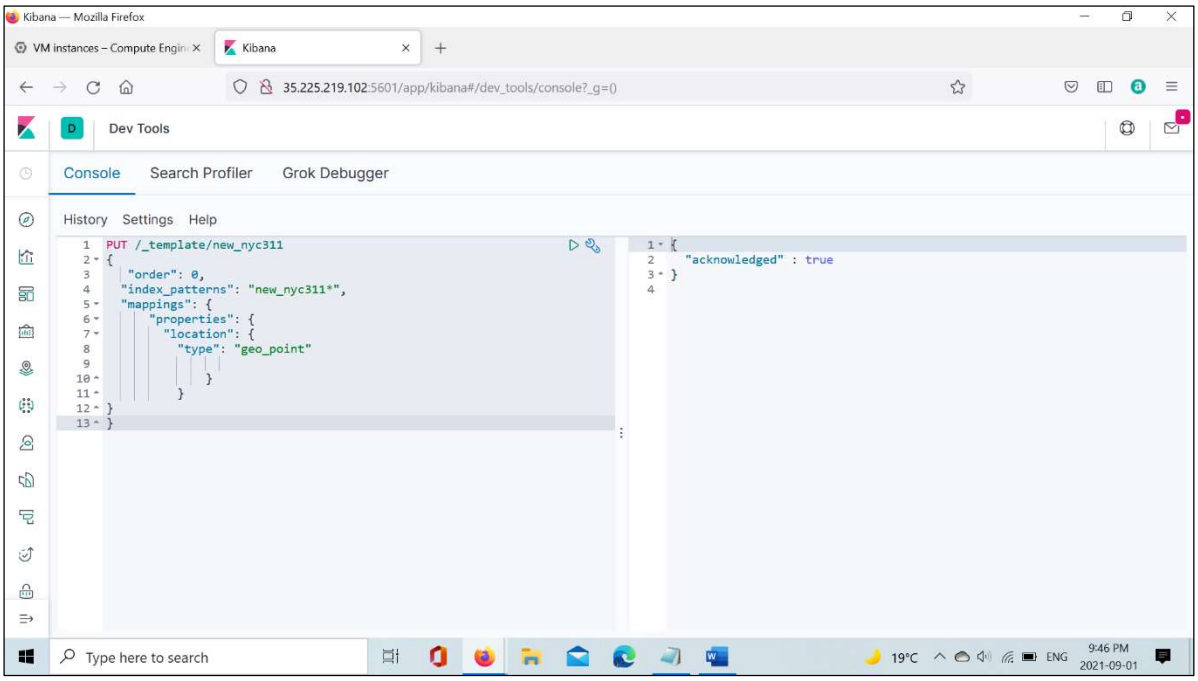

Figure 2: Kibana Dev Tools Console Used for Geopoints Set-up

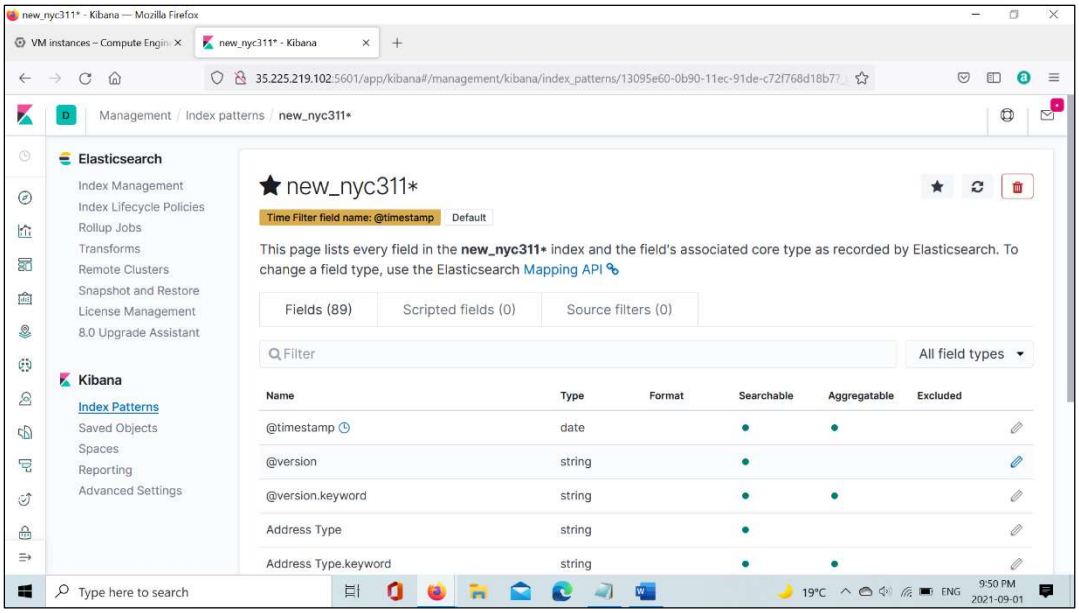

Figure 3: Index new_nyc311 Created in Kibana

5.0 Analysis of the Dataset

Figures 4, 5, 6 and 7 show examples of the data displayed in a table, pie chart, tag cloud and map. Figure 8 shows an example of a dashboard created with the table, pie chart, tag cloud and map.

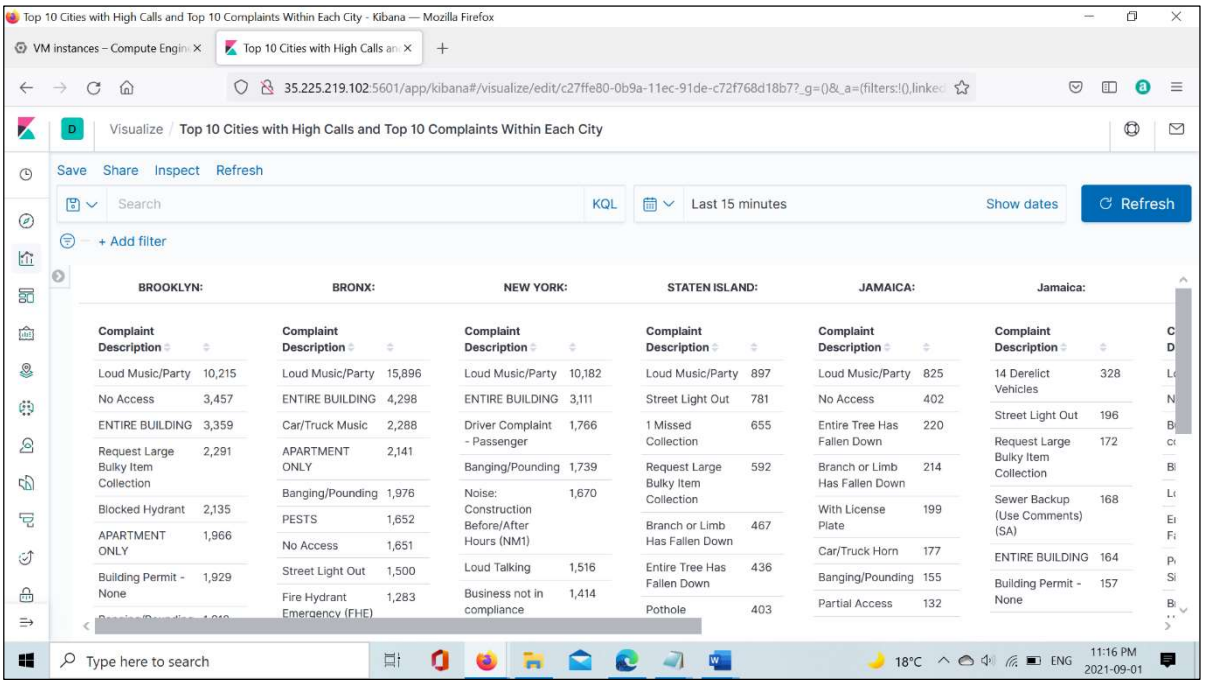

Figure 4: Top 10 Cities with High Calls and Top 10 Complaint Calls (by Descriptor) within Each City

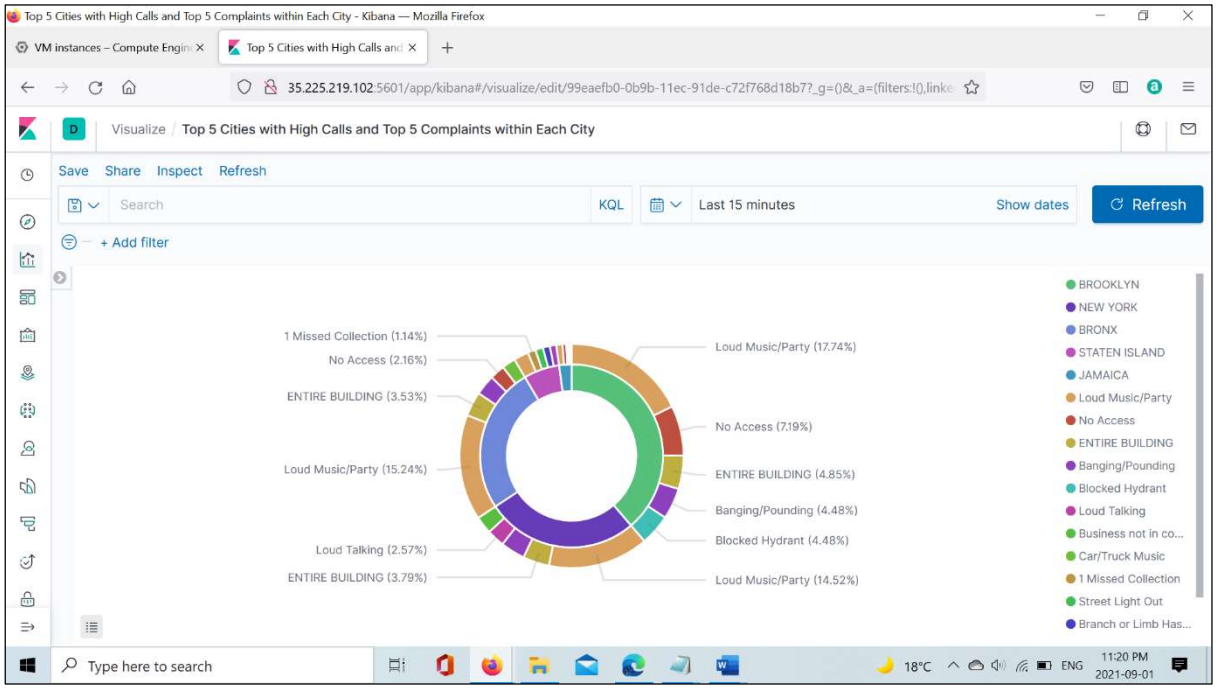

Figure 5: Top 5 Cities with High Calls and Top 5 Complaint Calls (by Descriptor) within Each City

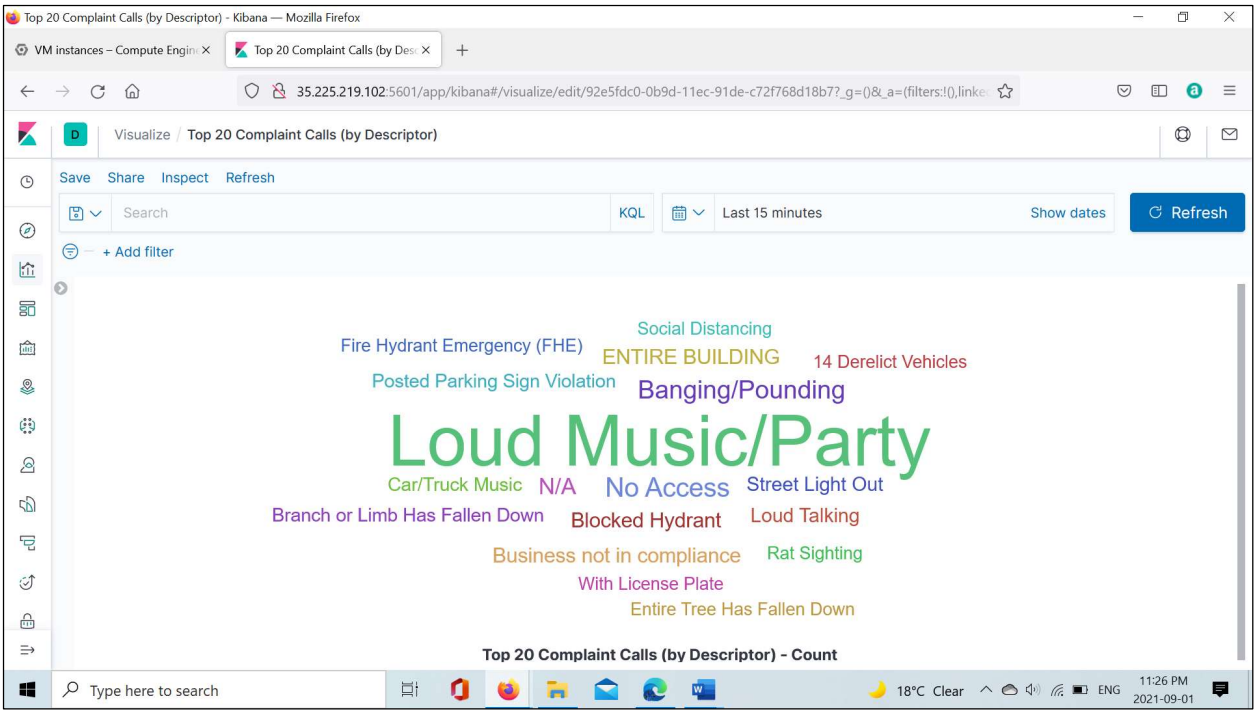

Figure 6: Tag Cloud with the Top 20 Complaint Calls (by Descriptor)

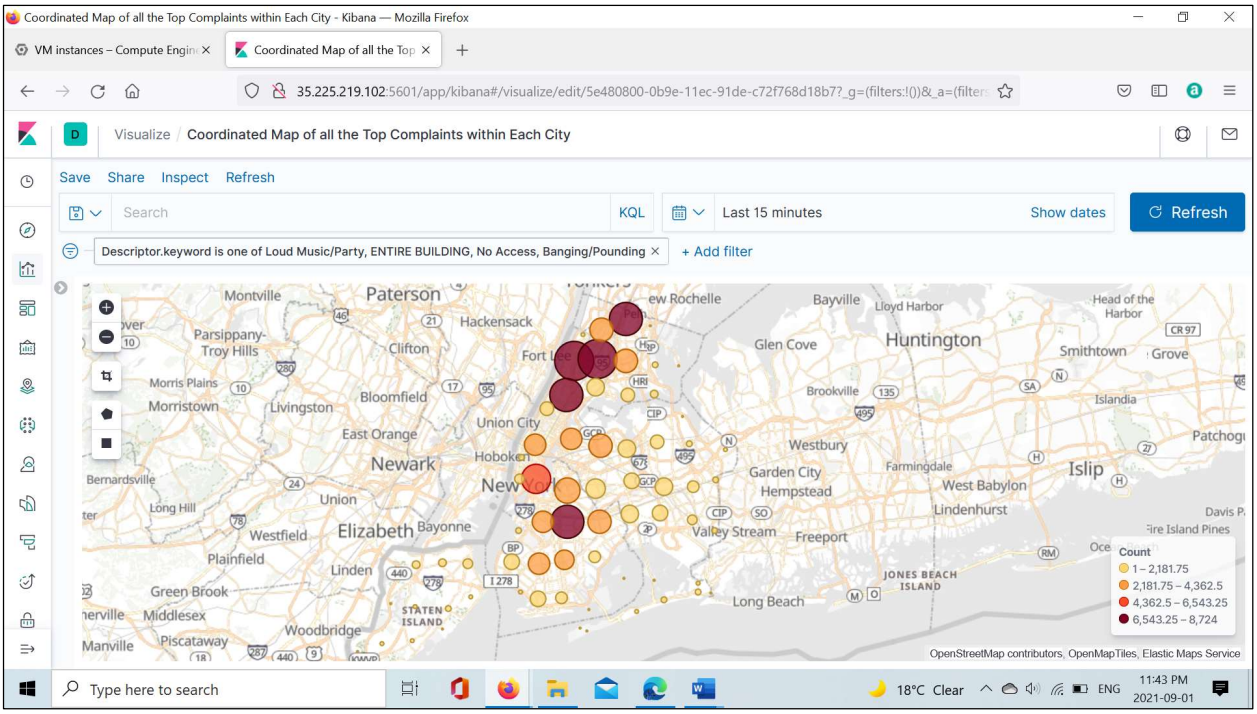

Figure 7: Coordinated Map of All the Top Complaints (by Descriptor) in Each City

In Figure 7, the top complaints were filtered as loud music/party, entire building, no access and banging/pounding using Figure 4 as the reference.

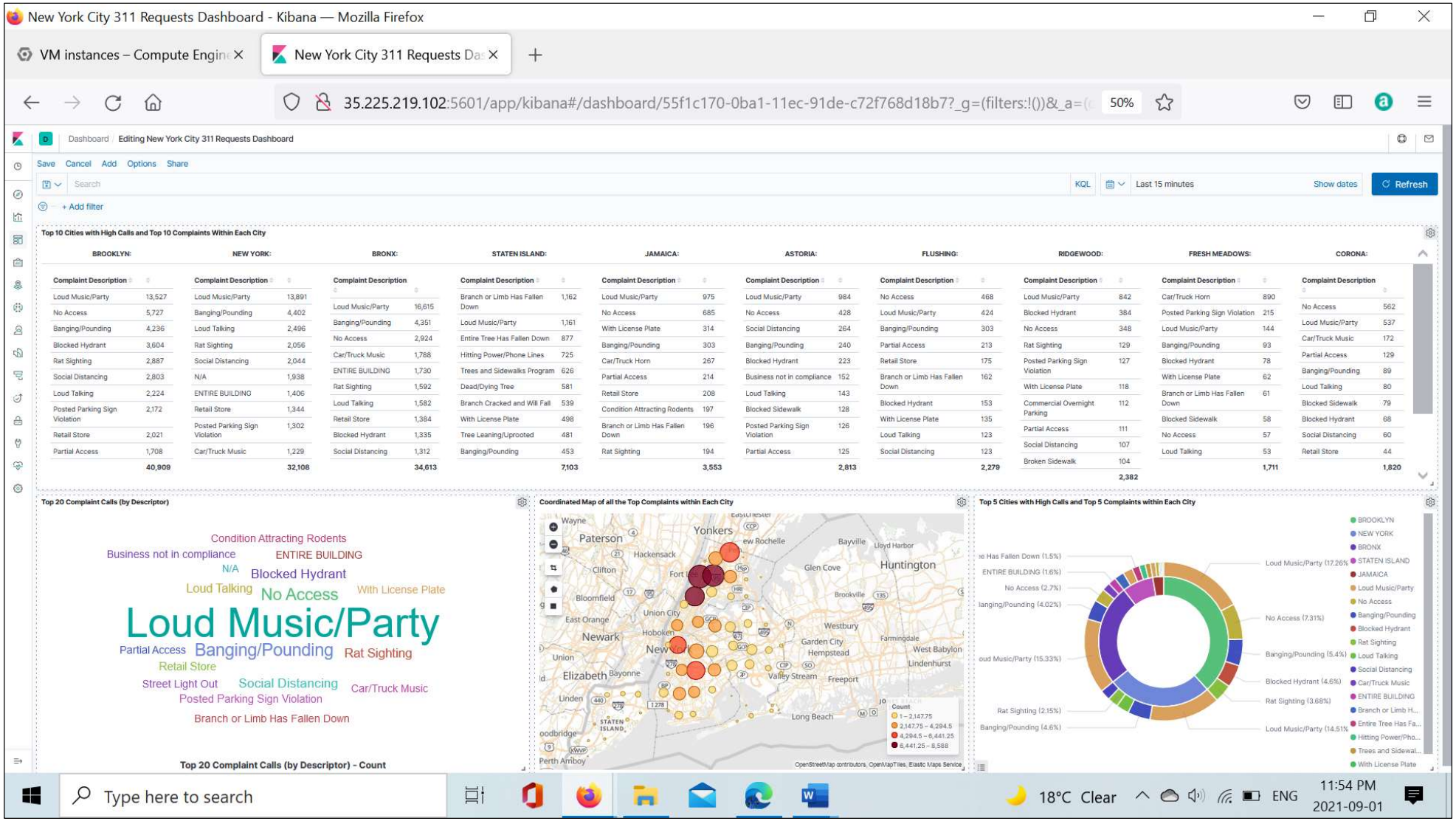

Figure 8: Dashboard using the Visualizations Created Previously

APPENDIX A

-- To download the components, use wget commands --

wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.5.1-linux-x86_64.tar.gz
wget https://artifacts.elastic.co/downloads/kibana/kibana-7.5.1-linux-x86_64.tar.gz
wget https://artifacts.elastic.co/downloads/logstash/logstash-7.5.1.tar.gz
ls

-- Elasticsearch Configuration --

tar -xzf elasticsearch-7.5.1-linux-x86_64.tar.gz
ls
cd elasticsearch-7.5.1
cd config/
ls
vi elasticsearch.yml

-in vi editor-

i for insert - remove hashtag

change the following:
network.host: 0.0.0.0
discovery.seed_hosts: ["10.128.0.xx:9300"] - internal IP address of instance
cluster.initial_master_nodes: ["bigdata-m"]

esc to exit insert mode
:wq to save and quit
cd ..

ls

-Before starting elasticsearch in GCP-

sudo sysctl vm.max_map_count=262144

bin/elasticsearch (should get a bunch of info loading)

-- Kibana Configuration --

tar -xzf kibana-7.5.1-linux-x86_64.tar.gz
ls

cd kibana-7.5.1-linux-x86_64

cd config/
ls
vi kibana.yml

-in vi editor-

i for insert - remove hashtag

server.port: 5601
server.host: "0.0.0.0"

esc to exit insert mode
:wq to save and quit
-- Firewall Rules Set-up --

-- Start Kibana --

cd ..
bin/kibana

-- Kibana Website Check --

Connect to Kibana by using external IP address from instance followed by :5601

-- Logstash Configuration --

tar -xzf logstash-7.5.1.tar.gz
ls
-- Upload NYC 311 File --

wget https://www.dropbox.com/sh/smx7s2f32y4izkk/AADHeFrdwIGAOjHr-ZYMtY87a/311_Service_Requests_from_2010_to_Present.csv

wget https://www.dropbox.com/s/dzop4gsuvn3esby/311_service.csv

ls

-- Upload Logstash Config File and start Logstash --

vi logstash_311nyc.config (copy and paste into a new vi file created the config setup)
ls

cd logstash-7.5.1

bin/logstash -f /home/krlottey/logstash_311nyc.config

-- Kibana Interface --

Get extenal IP address from the VM instance and adding :5601 to the IP address and open in Google to use Kibana

https://www.dropbox.com/s/cmbncfyki7wyjnu/logstash_nyc311.config