Assignment 2

Data Visualization with Kibana

Introduction

In this project we will be working with NYC Open Data published by the city of New York pertaining to 311 service requests collected since 2010 with over 21 million rows with 41 columns.

https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-toPresent/erm2-nwe9.

Problem Background

We have been hired as a data analyst by the city of New York to gain valuable insights from their huge data set for 311 service requests. Our task is to use the ELK stack you successfully installed and configured in your GCP platform. Successful completion of this task includes creating a Logstash configuration file (a sample is given to you) as well as a geo-point template (for maps), creating a GCP instance and firing Logstash to ingest the NYC 311 service request data into Elasticsearch and using Kibana to analyze and visualize the results as per the questions given.

Objectives

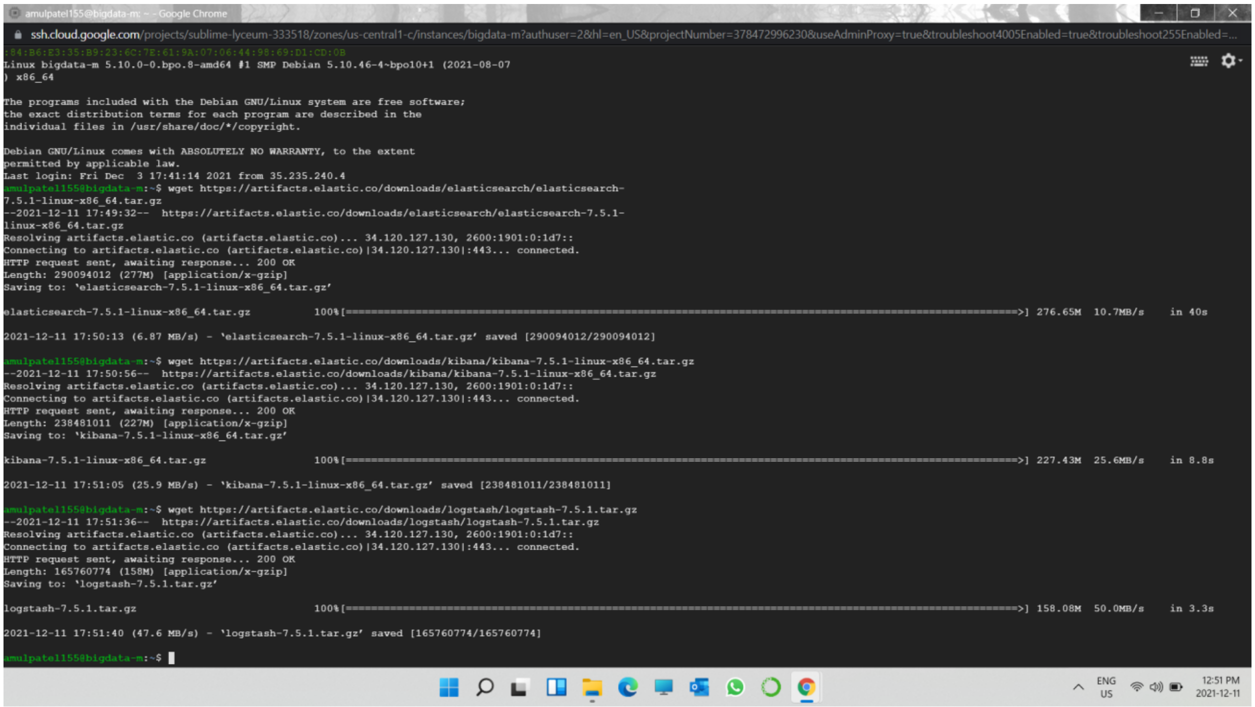
1. To further expose we are using ELK stack as an analytic tool to analyze streaming realistic big data.

2. To give us experience working with opened end problems, that are similar to problems that we will face in our career as a big data professional.

3. At the end of this project we should:

- Gain sufficient confidence in creating Logstash configuration files and creating Elasticsearch indices, advanced queries, charts, maps and dashboards using Kibana i.e. fully using ELK stack in real big data scenarios.
- Gain an appetite for working with large streaming datasets.
- Be aware of the potential and benefits of analyzing large streaming data using big data tools.

Installation of Elasticsearch, Kibana, Logstash

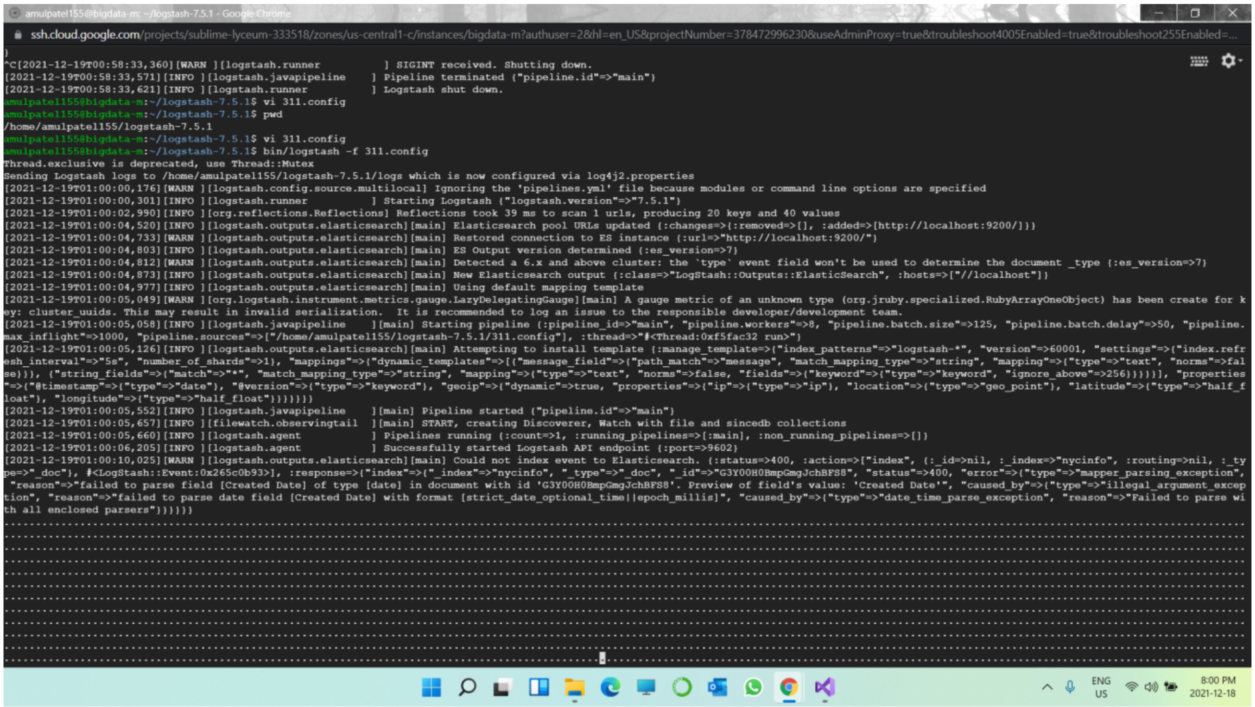

amulpatel155@bigdata-m: ~ - Google Chrome
ssh.cloud.google.com/projects/sublime-lyceum-333518/zones/us-central1-c/instances/bigdata-m?authuser=2&hl=en_US&projectNumber=378472996230&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=...
ENG
US
12:51 PM
2021-12-11

```
:84:B6:E3:35:B9:23:6C:7E:61:9A:07:06:44:98:69:D1:CD:0B
Linux bigdata-m 5.10.0-0.bpo.8-amd64 #1 SMP Debian 5.10.46-4~bpo10+1 (2021-08-07
) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Dec  3 17:41:14 2021 from 35.235.240.4
amulpatel155@bigdata-m:~$ wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-
7.5.1-linux-x86_64.tar.gz
--2021-12-11 17:49:32--  https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.5.1-
linux-x86_64.tar.gz
Resolving artifacts.elastic.co (artifacts.elastic.co)... 34.120.127.130, 2600:1901:0:1d7::
Connecting to artifacts.elastic.co (artifacts.elastic.co)|34.120.127.130|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 290094012 (277M) [application/x-gzip]
Saving to: ‘elasticsearch-7.5.1-linux-x86_64.tar.gz’

elasticsearch-7.5.1-linux-x86_64.tar.gz         100%[=========================================================================================================>] 276.65M  10.7MB/s    in 40s

2021-12-11 17:50:13 (6.87 MB/s) - ‘elasticsearch-7.5.1-linux-x86_64.tar.gz’ saved [290094012/290094012]

amulpatel155@bigdata-m:~$ wget https://artifacts.elastic.co/downloads/kibana/kibana-7.5.1-linux-x86_64.tar.gz
--2021-12-11 17:50:56--  https://artifacts.elastic.co/downloads/kibana/kibana-7.5.1-linux-x86_64.tar.gz
Resolving artifacts.elastic.co (artifacts.elastic.co)... 34.120.127.130, 2600:1901:0:1d7::
Connecting to artifacts.elastic.co (artifacts.elastic.co)|34.120.127.130|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 238481011 (227M) [application/x-gzip]
Saving to: ‘kibana-7.5.1-linux-x86_64.tar.gz’

kibana-7.5.1-linux-x86_64.tar.gz                100%[=========================================================================================================>] 227.43M  25.6MB/s    in 8.8s

2021-12-11 17:51:05 (25.9 MB/s) - ‘kibana-7.5.1-linux-x86_64.tar.gz’ saved [238481011/238481011]

amulpatel155@bigdata-m:~$ wget https://artifacts.elastic.co/downloads/logstash/logstash-7.5.1.tar.gz
--2021-12-11 17:51:36--  https://artifacts.elastic.co/downloads/logstash/logstash-7.5.1.tar.gz
Resolving artifacts.elastic.co (artifacts.elastic.co)... 34.120.127.130, 2600:1901:0:1d7::
Connecting to artifacts.elastic.co (artifacts.elastic.co)|34.120.127.130|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 165760774 (158M) [application/x-gzip]
Saving to: ‘logstash-7.5.1.tar.gz’

logstash-7.5.1.tar.gz                           100%[=========================================================================================================>] 158.08M  50.0MB/s    in 3.3s

2021-12-11 17:51:40 (47.6 MB/s) - ‘logstash-7.5.1.tar.gz’ saved [165760774/165760774]

amulpatel155@bigdata-m:~$ 
```

Logstash Config File

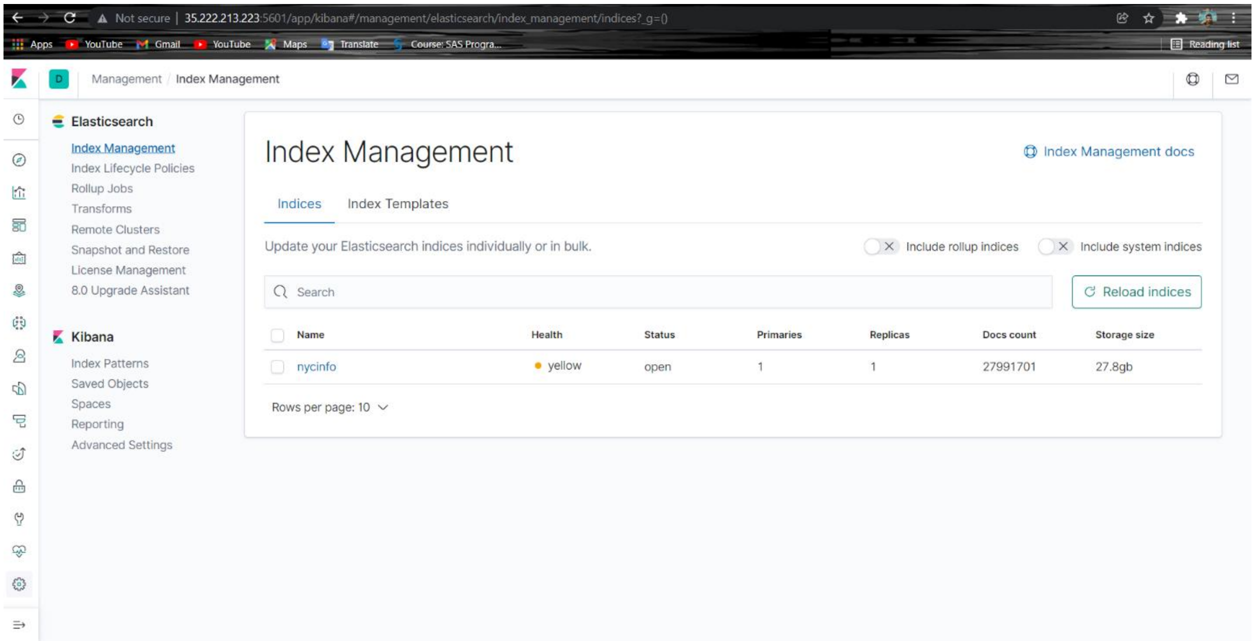
```
input {
    file {
       path => " /home/amulpatel155/logstash-7.5.1/311_service.csv"
       start_position => "beginning"
      sincedb_path => "/dev/null"
      }
  }

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 => "nycinfo"

  }
stdout {codec => dots}
```

Mapping


amulpatel155@bigdata-m: ~/logstash-7.5.1 - Google Chrome
ssh.cloud.google.com/projects/sublime-lyceum-333518/zones/us-central1-c/instances/bigdata-m?authuser=2&hl=en_US&projectNumber=378472996230&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=...
ENG
US
8:00 PM
2021-12-18

```
}
^C[2021-12-19T00:58:33,360][WARN ][logstash.runner          ] SIGINT received. Shutting down.
[2021-12-19T00:58:33,571][INFO ][logstash.javapipeline    ] Pipeline terminated {"pipeline.id"=>"main"}
[2021-12-19T00:58:33,621][INFO ][logstash.runner          ] Logstash shut down.
amulpatel155@bigdata-m:~/logstash-7.5.1$ vi 311.config
amulpatel155@bigdata-m:~/logstash-7.5.1$ pwd
/home/amulpatel155/logstash-7.5.1
amulpatel155@bigdata-m:~/logstash-7.5.1$ vi 311.config
amulpatel155@bigdata-m:~/logstash-7.5.1$ bin/logstash -f 311.config
Thread.exclusive is deprecated, use Thread::Mutex
Sending Logstash logs to /home/amulpatel155/logstash-7.5.1/logs which is now configured via log4j2.properties
[2021-12-19T01:00:00,176][WARN ][logstash.config.source.multilocal] Ignoring the 'pipelines.yml' file because modules or command line options are specified
[2021-12-19T01:00:00,301][INFO ][logstash.runner          ] Starting Logstash {"logstash.version"=>"7.5.1"}
[2021-12-19T01:00:02,990][INFO ][org.reflections.Reflections] Reflections took 39 ms to scan 1 urls, producing 20 keys and 40 values
[2021-12-19T01:00:04,520][INFO ][logstash.outputs.elasticsearch][main] Elasticsearch pool URLs updated {:changes=>{:removed=>[], :added=>[http://localhost:9200/]}}
[2021-12-19T01:00:04,733][WARN ][logstash.outputs.elasticsearch][main] Restored connection to ES instance {:url=>"http://localhost:9200/"}
[2021-12-19T01:00:04,803][INFO ][logstash.outputs.elasticsearch][main] ES Output version determined {:es_version=>7}
[2021-12-19T01:00:04,812][WARN ][logstash.outputs.elasticsearch][main] Detected a 6.x and above cluster: the `type` event field won't be used to determine the document _type {:es_version=>7}
[2021-12-19T01:00:04,873][INFO ][logstash.outputs.elasticsearch][main] New Elasticsearch output {:class=>"LogStash::Outputs::ElasticSearch", :hosts=>["//localhost"]}
[2021-12-19T01:00:04,977][INFO ][logstash.outputs.elasticsearch][main] Using default mapping template
[2021-12-19T01:00:05,049][WARN ][org.logstash.instrument.metrics.gauge.LazyDelegatingGauge][main] A gauge metric of an unknown type (org.jruby.specialized.RubyArrayOneObject) has been create for k
ey: cluster_uuids. This may result in invalid serialization.  It is recommended to log an issue to the responsible developer/development team.
[2021-12-19T01:00:05,058][INFO ][logstash.javapipeline    ][main] Starting pipeline {:pipeline_id=>"main", "pipeline.workers"=>8, "pipeline.batch.size"=>125, "pipeline.batch.delay"=>50, "pipeline.
max_inflight"=>1000, "pipeline.sources"=>["/home/amulpatel155/logstash-7.5.1/311.config"], :thread=>"#<Thread:0xf5fac32 run>"}
[2021-12-19T01:00:05,126][INFO ][logstash.outputs.elasticsearch][main] Attempting to install template {:manage_template=>{"index_patterns"=>"logstash-*", "version"=>60001, "settings"=>{"index.refr
esh_interval"=>"5s", "number_of_shards"=>1}, "mappings"=>{"dynamic_templates"=>[{"message_field"=>{"path_match"=>"message", "match_mapping_type"=>"string", "mapping"=>{"type"=>"text", "norms"=>fal
se}}}, {"string_fields"=>{"match"=>"*", "match_mapping_type"=>"string", "mapping"=>{"type"=>"text", "norms"=>false, "fields"=>{"keyword"=>{"type"=>"keyword", "ignore_above"=>256}}}}}], "properties
"=>{"@timestamp"=>{"type"=>"date"}, "@version"=>{"type"=>"keyword"}, "geoip"=>{"dynamic"=>true, "properties"=>{"ip"=>{"type"=>"ip"}, "location"=>{"type"=>"geo_point"}, "latitude"=>{"type"=>"half_f
loat"}, "longitude"=>{"type"=>"half_float"}}}}}}}
[2021-12-19T01:00:05,552][INFO ][logstash.javapipeline    ][main] Pipeline started {"pipeline.id"=>"main"}
[2021-12-19T01:00:05,657][INFO ][filewatch.observingtail  ][main] START, creating Discoverer, Watch with file and sincedb collections
[2021-12-19T01:00:05,660][INFO ][logstash.agent           ] Pipelines running {:count=>1, :running_pipelines=>[:main], :non_running_pipelines=>[]}
[2021-12-19T01:00:06,205][INFO ][logstash.agent           ] Successfully started Logstash API endpoint {:port=>9602}
[2021-12-19T01:00:10,025][WARN ][logstash.outputs.elasticsearch][main] Could not index event to Elasticsearch. {:status=>400, :action=>["index", {:_id=>nil, :_index=>"nycinfo", :routing=>nil, :_ty
pe=>"_doc"}, #<LogStash::Event:0x265c0b93>], :response=>{"index"=>{"_index"=>"nycinfo", "_type"=>"_doc", "_id"=>"G3Y00H0BmpGmgJchBFS8", "status"=>400, "error"=>{"type"=>"mapper_parsing_exception",
 "reason"=>"failed to parse field [Created Date] of type [date] in document with id 'G3Y00H0BmpGmgJchBFS8'. Preview of field's value: 'Created Date'", "caused_by"=>{"type"=>"illegal_argument_excep
tion", "reason"=>"failed to parse date field [Created Date] with format [strict_date_optional_time||epoch_millis]", "caused_by"=>{"type"=>"date_time_parse_exception", "reason"=>"Failed to parse wi
th all enclosed parsers"}}}}}}
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
..................................................................................................................................................................................................
```

Dataset is uploaded in Kibana



Not secure | 35.222.213.223:5601/app/kibana#/management/elasticsearch/index_management/indices?_g=()

Management / Index Management

Elasticsearch

- Index Management
- Index Lifecycle Policies
- Rollup Jobs
- Transforms
- Remote Clusters
- Snapshot and Restore
- License Management
- 8.0 Upgrade Assistant

Kibana

- Index Patterns
- Saved Objects
- Spaces
- Reporting
- Advanced Settings

Index Management

Index Management docs

Indices | Index Templates

Update your Elasticsearch indices individually or in bulk.

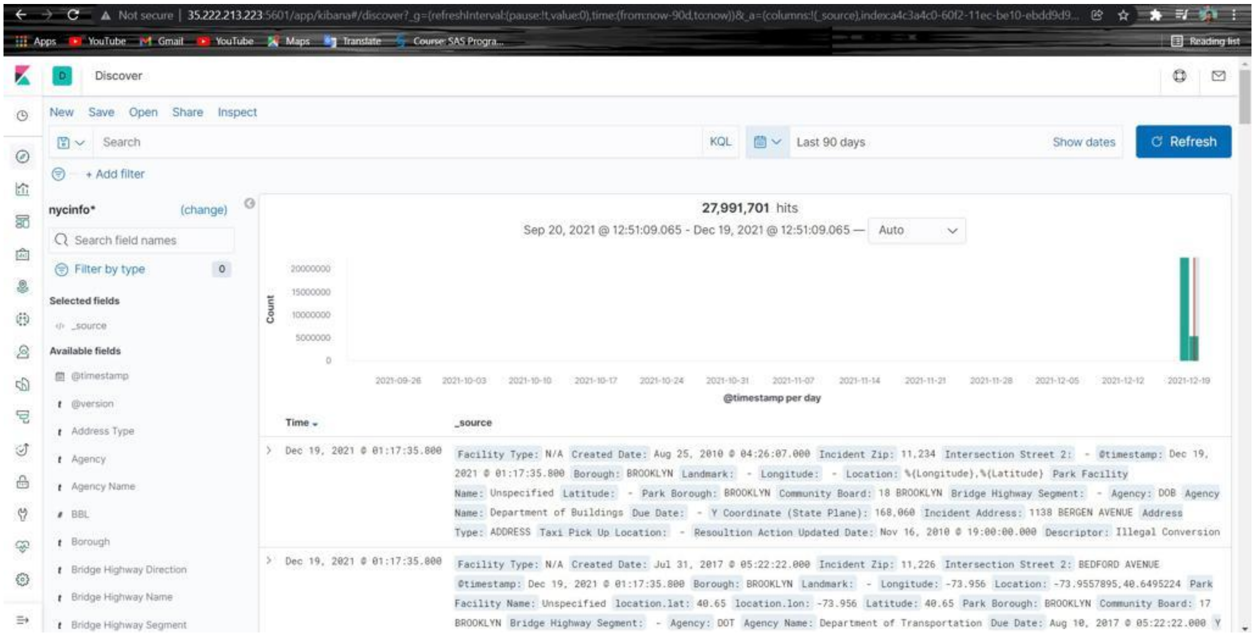
Include rollup indices   Include system indices

Search   Reload indices

☐	Name	Health	Status	Primaries	Replicas	Docs count	Storage size
☐	nycinfo	● yellow	open	1	1	27991701	27.8gb

Rows per page: 10

Dataset



Not secure | 35.222.213.223:5601/app/kibana#/discover?_g=(refreshInterval:(pause:!t,value:0),time:(from:now-90d,to:now))&_a=(columns:!(_source),index:a4c3a4c0-60f2-11ec-be10-ebdd9d9...

Discover

New Save Open Share Inspect

Search KQL Last 90 days Show dates Refresh

+ Add filter

nycinfo* (change)

Search field names

Filter by type 0

Selected fields

- _source

Available fields

- @timestamp
- @version
- Address Type
- Agency
- Agency Name
- BBL
- Borough
- Bridge Highway Direction
- Bridge Highway Name
- Bridge Highway Segment

27,991,701 hits

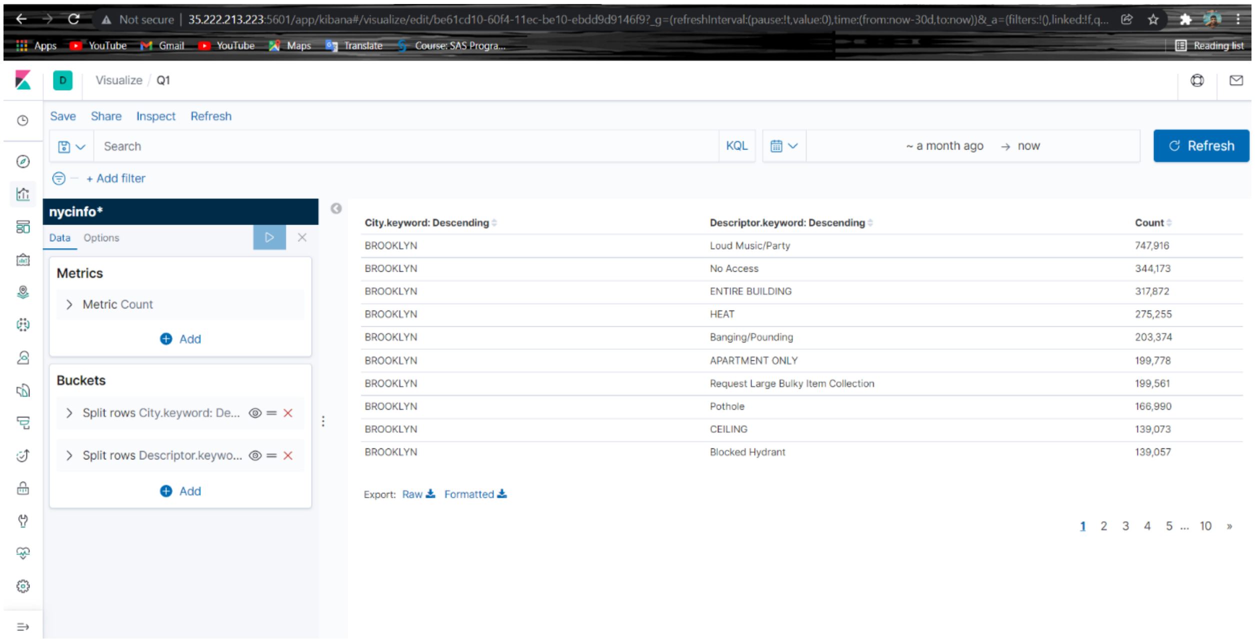
Sep 20, 2021 @ 12:51:09.065 - Dec 19, 2021 @ 12:51:09.065 — Auto

Time	_source
Dec 19, 2021 @ 01:17:35.800	Facility Type: N/A Created Date: Aug 25, 2010 @ 04:26:07.000 Incident Zip: 11,234 Intersection Street 2: - @timestamp: Dec 19, 2021 @ 01:17:35.800 Borough: BROOKLYN Landmark: - Longitude: - Location: %{Longitude},%{Latitude} Park Facility Name: Unspecified Latitude: - Park Borough: BROOKLYN Community Board: 18 BROOKLYN Bridge Highway Segment: - Agency: DOB Agency Name: Department of Buildings Due Date: - Y Coordinate (State Plane): 168,060 Incident Address: 1138 BERGEN AVENUE Address Type: ADDRESS Taxi Pick Up Location: - Resoultion Action Updated Date: Nov 16, 2010 @ 19:00:00.000 Descriptor: Illegal Conversion
Dec 19, 2021 @ 01:17:35.800	Facility Type: N/A Created Date: Jul 31, 2017 @ 05:22:22.000 Incident Zip: 11,226 Intersection Street 2: BEDFORD AVENUE @timestamp: Dec 19, 2021 @ 01:17:35.800 Borough: BROOKLYN Landmark: - Longitude: -73.956 Location: -73.9557895,40.6495224 Park Facility Name: Unspecified location.lat: 40.65 location.lon: -73.956 Latitude: 40.65 Park Borough: BROOKLYN Community Board: 17 BROOKLYN Bridge Highway Segment: - Agency: DOT Agency Name: Department of Transportation Due Date: Aug 10, 2017 @ 05:22:22.000 Y

Analysis

Create a table showing the top 10 cities with the highest calls alongside the count of top 10 complaint calls (by Descriptor) in each city.

Visualization: -

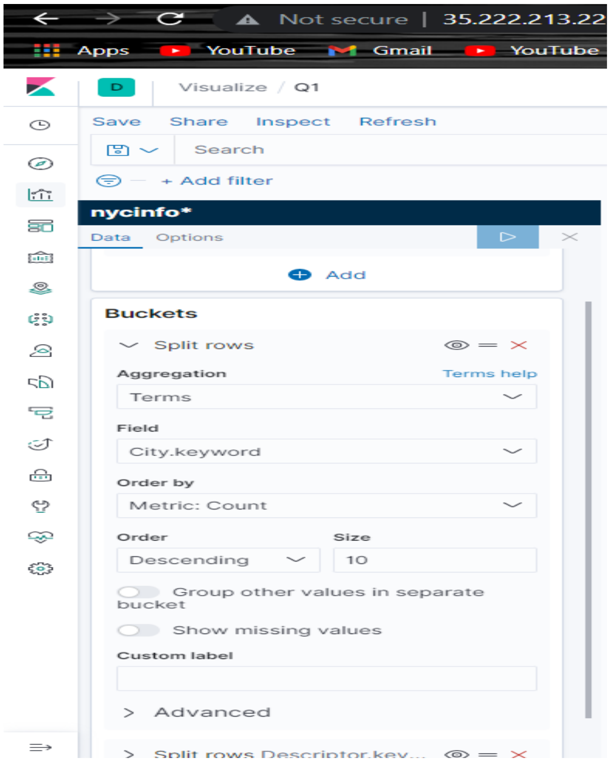


City.keyword: Descending	Descriptor.keyword: Descending	Count
BROOKLYN	Loud Music/Party	747,916
BROOKLYN	No Access	344,173
BROOKLYN	ENTIRE BUILDING	317,872
BROOKLYN	HEAT	275,255
BROOKLYN	Banging/Pounding	203,374
BROOKLYN	APARTMENT ONLY	199,778
BROOKLYN	Request Large Bulky Item Collection	199,561
BROOKLYN	Pothole	166,990
BROOKLYN	CEILING	139,073
BROOKLYN	Blocked Hydrant	139,057

Export: Raw Formatted

1 2 3 4 5 ... 10 »

Query: -



Split rows

Aggregation: Terms

Field: City.keyword

Order by: Metric: Count

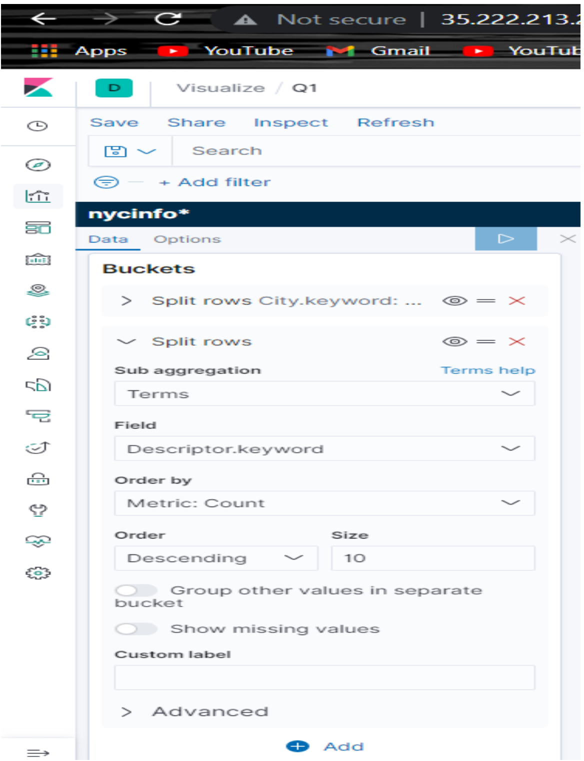
Order: Descending

Size: 10

Group other values in separate bucket

Show missing values

Custom label



Split rows

Sub aggregation: Terms

Field: Descriptor.keyword

Order by: Metric: Count

Order: Descending

Size: 10

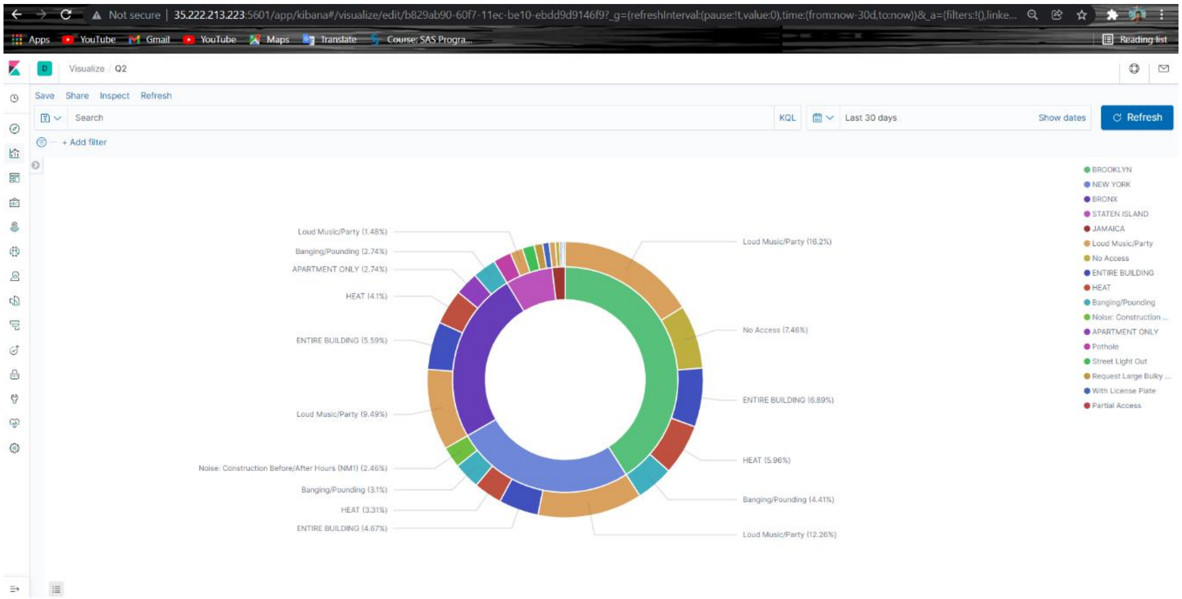
Group other values in separate bucket

Show missing values

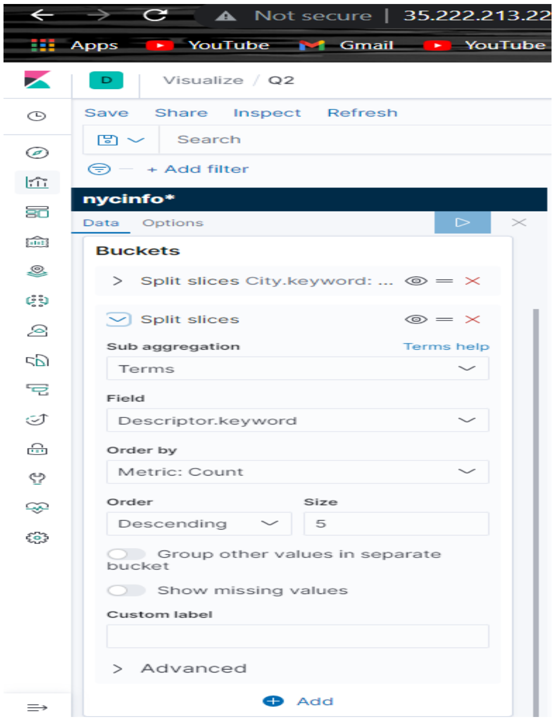
Custom label

Create a pie chart showing the top 5 cities with the highest calls alongside the top five calls (Descriptor) in each city.

Visualization: -



Query: -



nycinfo*

Buckets

> Split slices City.keyword: ...

Split slices

Sub aggregation: Terms

Field: Descriptor.keyword

Order by: Metric: Count

Order: Descending

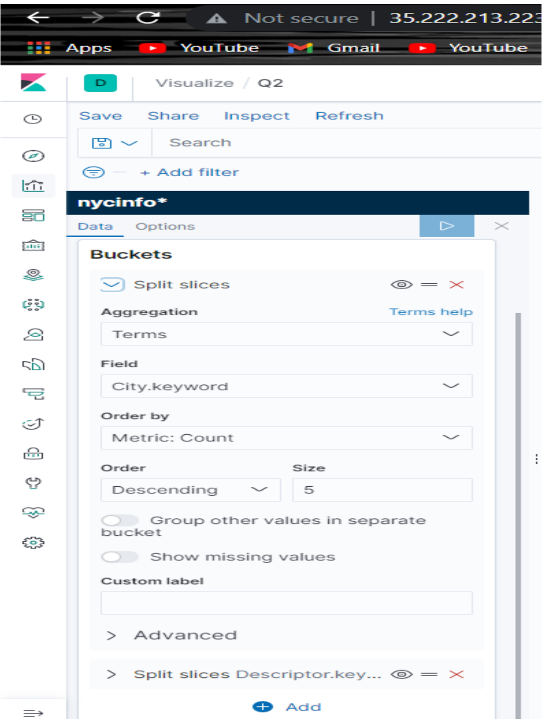
Size: 5

Group other values in separate bucket

Show missing values

Custom label

> Advanced



nycinfo*

Buckets

Split slices

Aggregation: Terms

Field: City.keyword

Order by: Metric: Count

Order: Descending

Size: 5

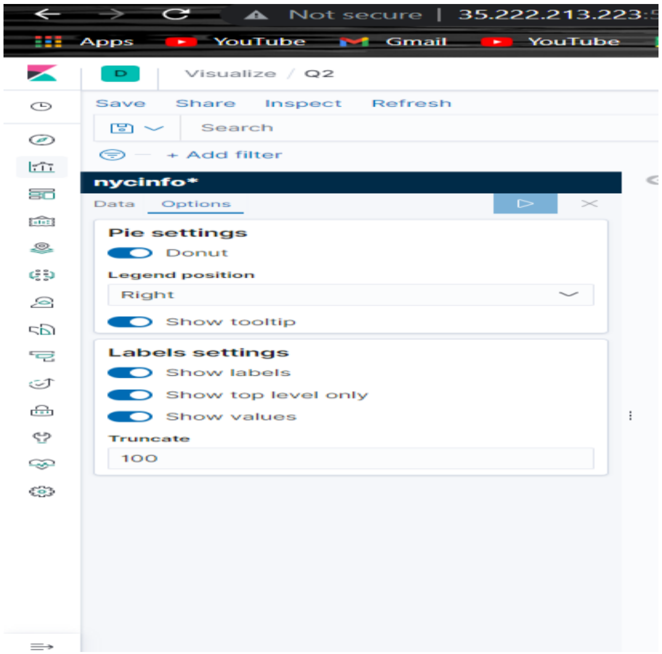
Group other values in separate bucket

Show missing values

Custom label

> Advanced

> Split slices Descriptor.key...



Pie settings

Donut

Legend position

Right

Show tooltip

Labels settings

Show labels

Show top level only

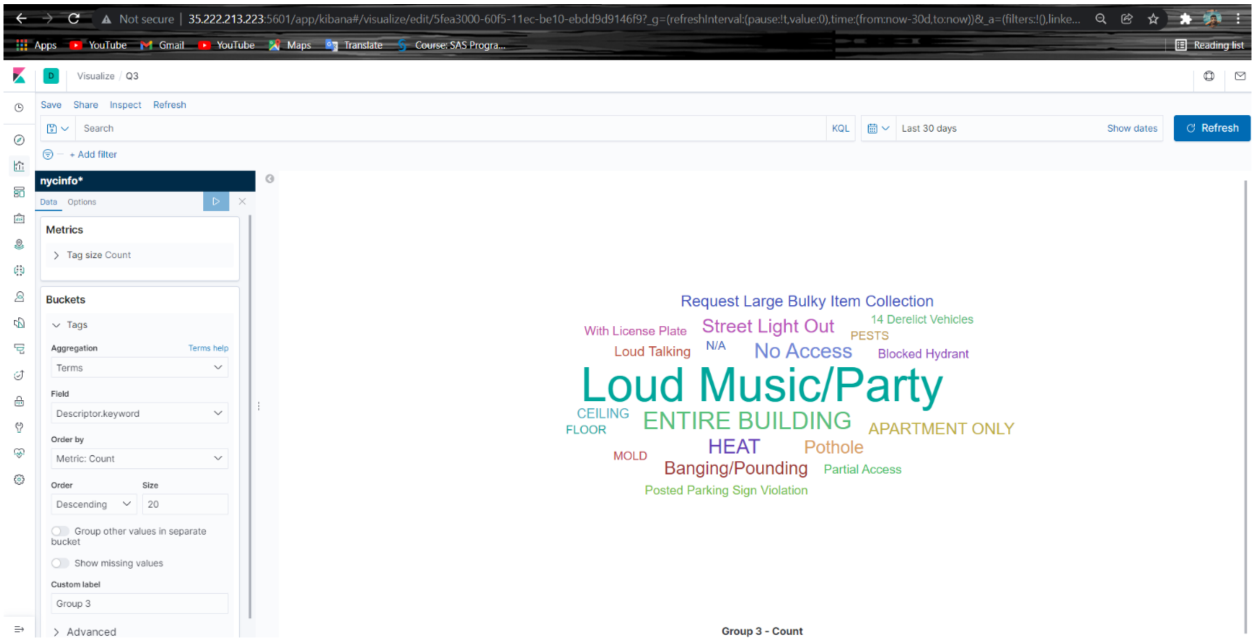
Show values

Truncate

100

Create a tag cloud representing the top 20 call descriptors.

Visualization & Query: -



Not secure | 35.222.213.223:5601/app/kibana#/visualize/edit/5fea3000-60f5-11ec-be10-ebdd9d9146f9?_g=(refreshInterval:(pause:!t,value:0),time:(from:now-30d,to:now))&_a=(filters:!(),linke...

Visualize / Q3

nycinfo*

Metrics

Tag size Count

Buckets

Tags

Aggregation: Terms

Field: Descriptor.keyword

Order by: Metric: Count

Order: Descending

Size: 20

Group other values in separate bucket

Show missing values

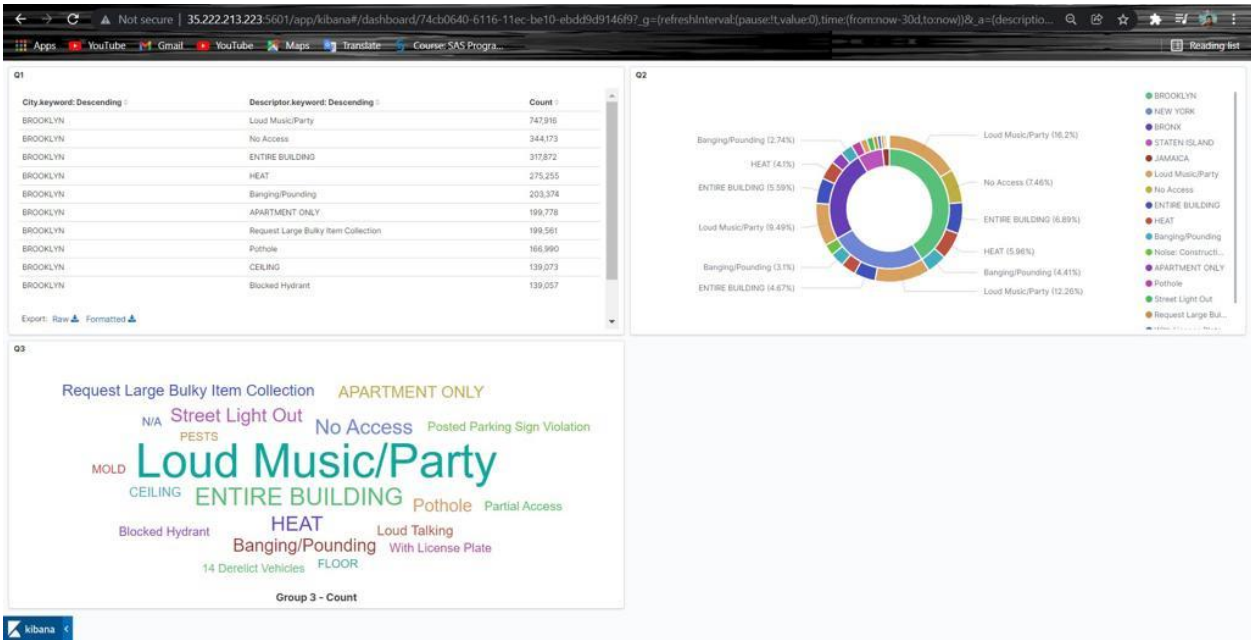
Custom label: Group 3

Last 30 days

Create a coordinated map of all the major call descriptors in each city

Create a dashboard for all visualizations of 1to 4 above.

Visualization: -



Q1

City.keyword: Descending	Descriptor.keyword: Descending	Count
BROOKLYN	Loud Music/Party	747,916
BROOKLYN	No Access	344,173
BROOKLYN	ENTIRE BUILDING	317,872
BROOKLYN	HEAT	275,255
BROOKLYN	Banging/Pounding	203,374
BROOKLYN	APARTMENT ONLY	199,778
BROOKLYN	Request Large Bulky Item Collection	199,561
BROOKLYN	Pothole	166,990
BROOKLYN	CEILING	139,073
BROOKLYN	Blocked Hydrant	139,057

Teamwork: -

As being team member, all the members have equally contribution in this project. Karun has installed the ELK, create firewall for Kibana and setup the Kibana for further steps. Main task of Amish was uploaded data set into Kibana. Amish has insert data into Logstash, he did some changes in Logstash file, create config file and uploaded dataset into Kibana. Amul has perform all the research question after competition of Karun and Amish task. As well as Amul has created this document file.

Thank You