Project 1: Open Parking and Camera Violations

For this project, I loaded the Open Parking and Camera Violations dataset containing more than 50 million NYC parking violations and uploaded it to Elasticsearch. Then I connected to my Elasticsearch index in Kibana and created a dashboard.

Firstly, I created the Dockerfile and requirements.txt which includes requests, sodapy and elasticsearch.

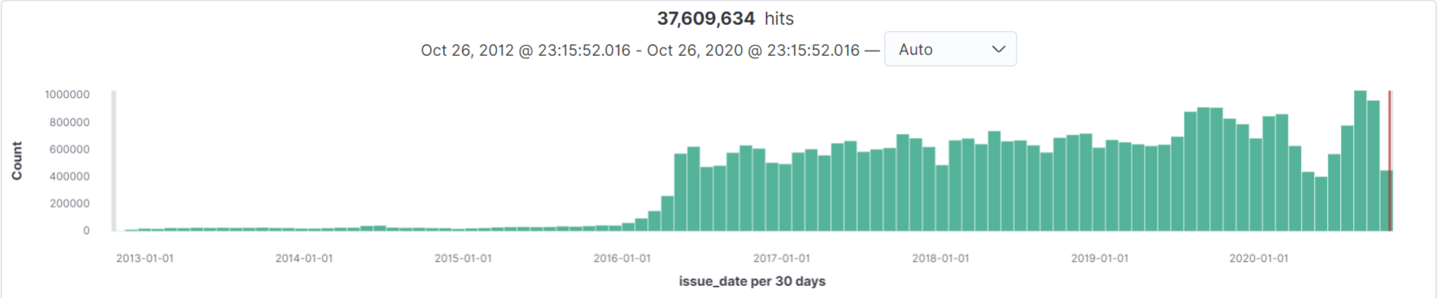
Secondly, I created the file main.py.

- This python file required 5 environment variables: DATASET_ID, APP_TOKEN, ES_HOST, ES_USERNAME, ES_PASSWORD.
- End users can input 2 command line arguments, page_size and num_pages. “page_size” is required which means how many records to request from the API per call. “num_pages” is optional. If it’s provided, continue querying for data num_pages times. Else, my script will continue requesting data until the entirety of the content has been exhausted.
- I created an Elasticsearch index and named it as “project01-opcv”. I skipped columns ‘violation_status’, ‘judgment_entry_date’, ’summons_image’ because there are too many NAs in these columns.
- Got the Elasticsearch instance.
- I used a while loop to load data from API and upload it to Elasticsearch. In a loop, I had a for loop to converted the data format, put them into an iterable “ACTIONS” and after the for loop, I used “helpers.bulk” to upload all rows into Elasticsearch.

Thirdly, I built and run the docker image and named it “project1”. Here is my command line.

```
docker build -t project1:1.0 project01/

docker run -d\

  -v ${PWD}:/app \

  --network="host" \

  -e DATASET_ID="nc67-uf89" \

  -e APP_TOKEN="zu1BRyhSkQPw4QilEeoS5QWgD" \

  -e ES_HOST="https://search-sta9760f2020xinlihou-gt4x4qah6zk72t7uzmawm4nxdq.us-east-2.es.amazonaws.com" \

  -e ES_USERNAME="xinli" \

  -e ES_PASSWORD="sta9760F2020@project1" \

  project1:1.0 --page_size=10000
```

Figure 1: Kibana Discover page. It took me more than 24 hours to upload 37.6 million records.

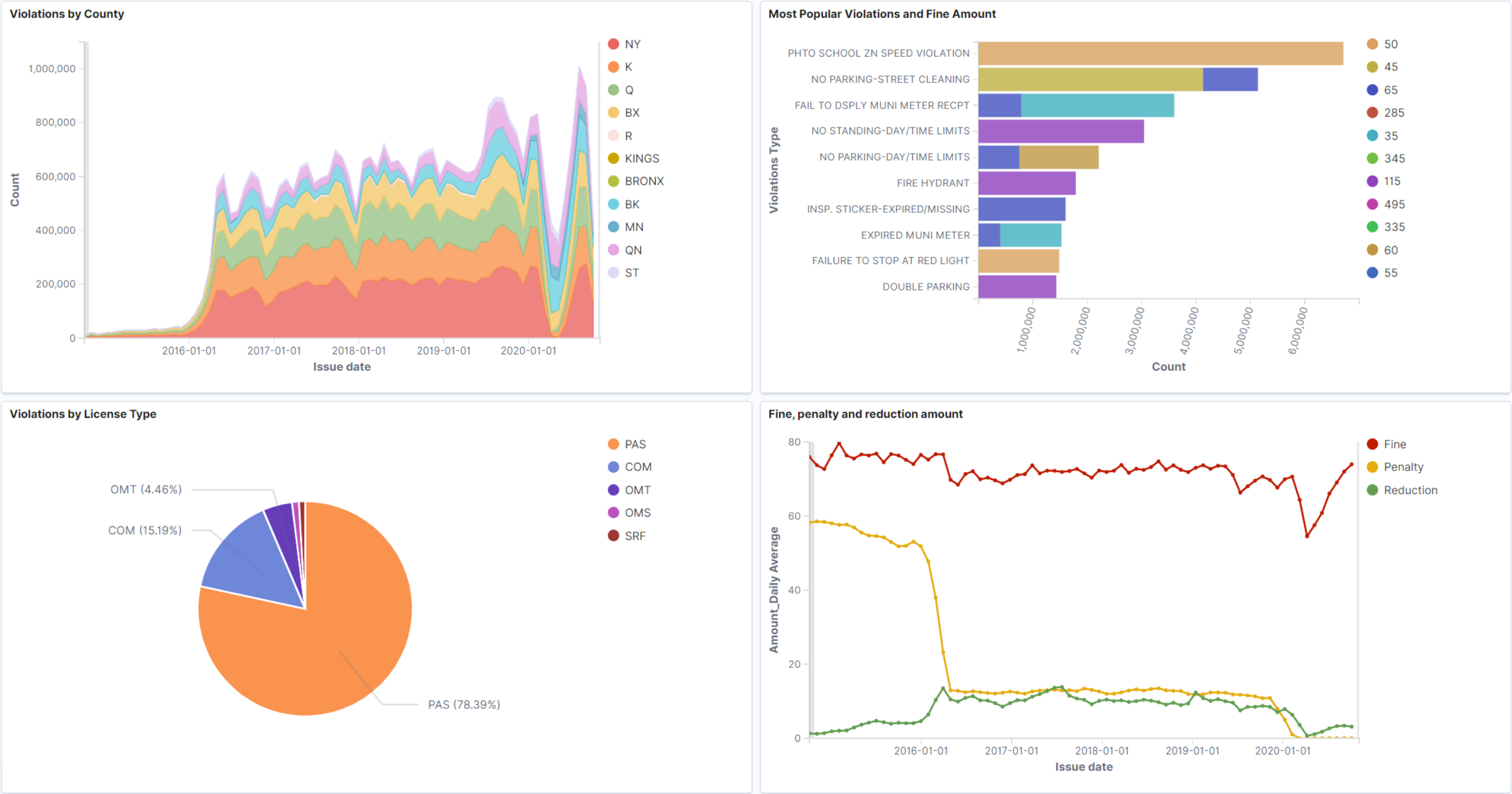


37,609,634 hits

Oct 26, 2012 @ 23:15:52.016 - Oct 26, 2020 @ 23:15:52.016 — Auto

Finally, I added the data into Kibana, created 4 meaningful visuals and made a dashboard. Here is my Kibana dashboard:

Figure 2: Kibana dashboard



- Visual 1: this is a stacked area chart which shows the daily number of violations by county. The horizontal axis is “Issue date” and the vertical axis is the count of violations. It suggests that NY, K BX and Q have much more violations than other counties.
- Visual 2: this is a horizontal stacked bar chart shows the top 10 most popular violations and corresponding fine amounts. It suggests that the most common violation is “PHTO SCHOOL ZN SPEED VIOLATION” and typically the fine amount is 50 dollars. The second one should be “NO PARKING-STREET CLEANING” and the fine amount could be 45 or 65 dollars.
- Visual 3: this is a pie chart which shows violations by license type. It suggests that the most common license type is PAS (Passenger Vehicles). The second one should be COM (Commercial Vehicles).
- Visual 4: this is a line chart and also a time plot which shows how fine amounts, penalty amounts and reduction amounts changes overtime. In general, payment amount = fine amount + penalty amount - reduction amount, so these 3 variables are critical. It shows that the fine amount has been relatively stable over the years and has fluctuated in 2020. Penalty amount used to be quite large and both the penalty and reduction amount have been stable since 2016 and have been decreasing since 2020.