NYC Parking tickets

Big Data Project

Team Members
Karim Mohamed Ibrahim
Mohamed Alaa Farghaly
Rohanda Hamed El-Sayed

1. Brief problem description

The NYC Department of Finance collects data on every parking ticket issued in NYC (~10M per year!). This data is made publicly available to aid in ticket resolution and to guide policymakers.

There are two files, covering 2015 and 2016. The files are roughly organized by fiscal year (July 1 - June 30) with the exception of the initial dataset.

The main problem is to answer these question:

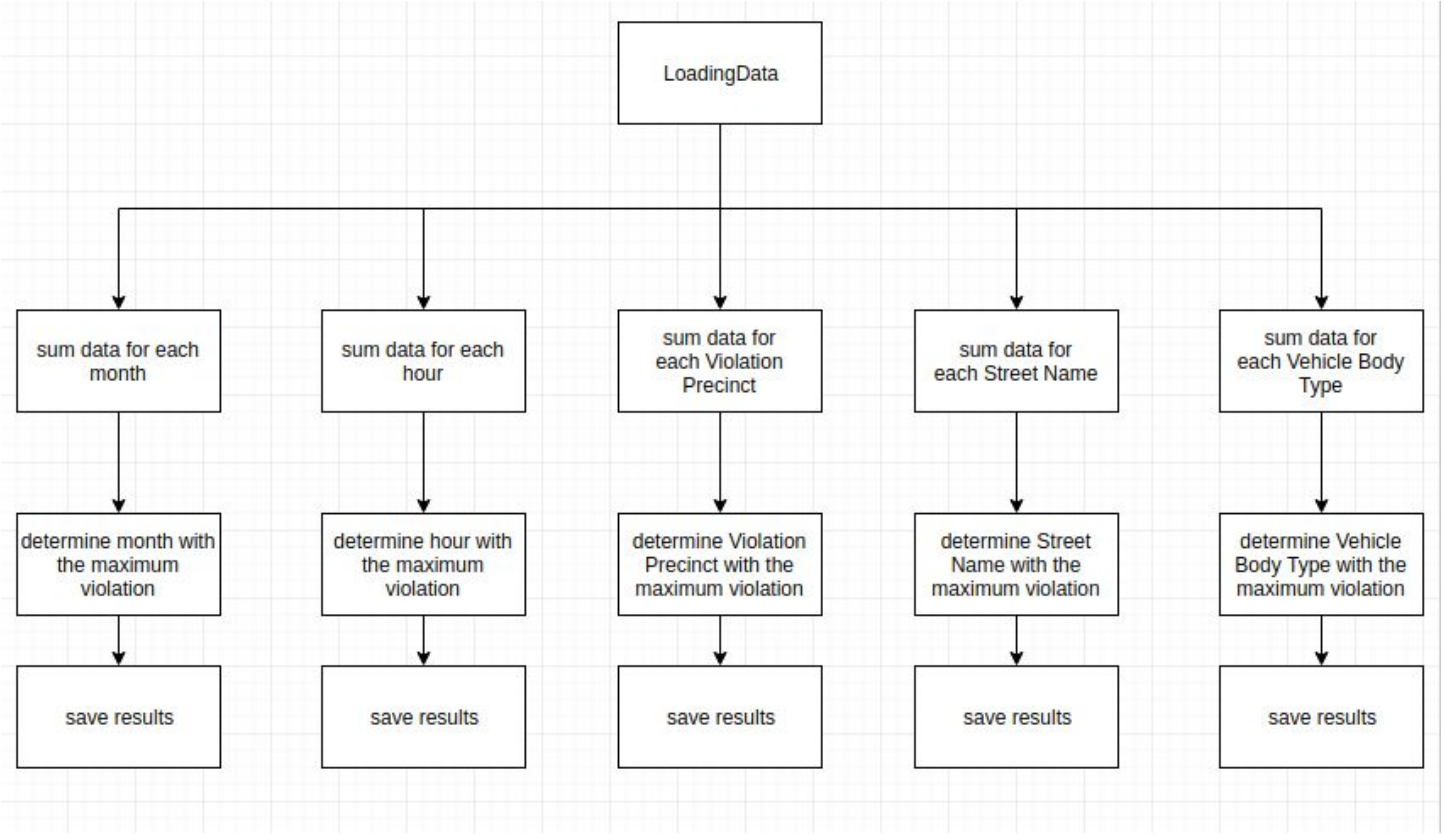
1. When are tickets most likely to be issued? Any seasonality?
2. Where are tickets most commonly issued?
3. What are the most common years and types of cars to be ticketed?

2. How you made your analysis

We decided to answer the questions using HIVE as it is optimized and integrated with tools that can answer our question. the task is divided into parts:

1. The data is not combined in one file and combining them using copy and paste is impossible as each file is about 2.5 GB. As a result, we need a way to combine them to import them as a one file into a table inside HIVE.
2. To answer the first question "When are tickets most likely to be issued? Any seasonality?", we need to define which column is responsible for reclaiming time and date. Then, we sum up among the data for each month and for each hour and take the maximum number for each of them.
3. To answer the first question "Where are tickets most commonly issued?", we need to define which column is responsible for reclaiming street names and precinct. Then, we sum up among the data for each street names and for each precinct and take the maximum number for each of them.
4. To answer the first question "What are the most common years and types of cars to be ticketed?", we need to define which column is responsible for reclaiming types of cars. Then, we sum up among the data for each types of cars and take the maximum number.
5. Compare between our results and the results of the other activities on kaggle such as https://www.kaggle.com/argha48/preliminary-data-visualization

3. The final pipeline of your solution and its diagram



- number of maps = 19 and number of reducers= 20
- To answer the first question, we needed to extract the month from a complete date and extract the hours from a complete hour-minutes-PMorAM data. we managed to do so using

4. The trials you made and not included in the final solution

To answer the first question, when we selected the data that contain the date, we used group by function by mistake which ended in a small results.

in case of month:

10,271
07,263
11,258
04,245
06,242
03,236
12,235
08,232
09,230
02,216
05,215
01,211

in case of hours:

09,135
11,135
08,133
10,132
01,131
02,131
04,130
07,130
06,128
12,127
05,127
03,127
00,121
18,10
16,8
17,6
13,5
15,5
19,5
20,4
21,4
22,4

while the number of rows is above 500,000 rows.

5. Results and evaluation

Question 1:

our results for hours:

hour	AM/PM	Frequency	Percentage
9	A	2379019	10.64%
11	A	2374771	10.62%
1	P	2182811	9.76%
8	A	2056538	9.20%
12	P	1988880	8.89%
10	A	1974533	8.83%
2	P	1860299	8.32%
3	P	1346890	6.02%
4	P	1222738	5.47%
7	A	1138574	5.09%
5	P	879562	3.93%
6	P	522017	2.33%
6	A	462016	2.07%
9	P	271328	1.21%
8	P	259618	1.16%
1	A	205699	0.92%
10	P	200767	0.90%
5	A	187340	0.84%
2	A	173186	0.77%
11	P	158252	0.71%
7	P	155843	0.70%
3	A	136759	0.61%
0	A	128720	0.58%
12	A	98079	0.44%
	sum	22364239	

Our results for month:

Month	Frequency	Percentage
1	2208882	9.85%
2	1572801	7.01%
3	1979592	8.82%
4	1852958	8.26%
5	1915332	8.54%
6	1926001	8.58%
7	1855329	8.27%
8	1814882	8.09%
9	1969587	8.78%
10	2063916	9.20%
11	1734802	7.73%
12	1541918	6.87%
sum	22436000	

Question 2:

Our results for Street:

Street	Frequency	percentage
Broadway	451333	14.83%
3rd Ave	342034	11.24%
5th Ave	225690	7.42%
Madison Ave	203026	6.67%
Lexington Ave	175618	5.77%
2nd Ave	167913	5.52%
1st Ave	152423	5.01%
7th Ave	137818	4.53%
Queens Blvd	127833	4.20%
Amsterdam Ave	125027	4.11%
8th Ave	121661	4.00%
6th Ave	113993	3.75%
Jamaica Ave	102424	3.37%
EB HORACE HARDING EX	99350	3.26%
Columbus Ave	92034	3.02%
Park Ave	85180	2.80%
37th Ave	84836	2.79%
Coney Island Ave	82001	2.69%
Roosevelt Ave	78534	2.58%
White Plains Rd	74926	2.46%
Sum	3043654	

Our results for precinct:

Precinet ID	Frequency	percentage
0	3667824	28.25%
19	1152801	8.88%
18	759209	5.85%
14	733528	5.65%
1	632848	4.88%
114	612299	4.72%
13	593614	4.57%
109	484522	3.73%
17	459270	3.54%
20	427420	3.29%
84	418319	3.22%
70	383691	2.96%
115	364255	2.81%
61	347314	2.68%
112	345245	2.66%
103	325878	2.51%
6	325191	2.51%
10	320244	2.47%
108	314996	2.43%
66	312762	2.41%
Sum	12981230	

Question 3:

Our results for Car type:

Vehicle type	Frequency	percentage
SUBN	7195351	32.96%
4DSD	6332087	29.00%
VAN	3227378	14.78%
DELV	1648039	7.55%
SDN	948639	4.35%
2DSD	595500	2.73%
PICK	561830	2.57%
REFG	173661	0.80%
UTIL	162589	0.74%
TRAC	150592	0.69%
TAXI	133439	0.61%
BUS	109316	0.50%
4 DR	107235	0.49%
CONV	100146	0.46%
TRLR	72496	0.33%
TK	70330	0.32%
WAGO	66136	0.30%
MCY	63709	0.29%
4D	61840	0.28%
P-U	52410	0.24%
Sum	21832723	

6. Any Enhancement and future work

1. We may search for data that say if the car is with some aspects it may or may not make a violation.
2. Then, we use these data to predict for any features if any car is suspected to make a violation.