NYC Parking Ticket Data Analysis

Laksheen Mendis
Menuka Warushavithana
Philip Kirner
CS 455: Introduction to Distributed Systems, In Affiliation with:
Computer Science Department, Colorado State University

Background Information

- Parking tickets exist to ensure order and support public tranquility
- Parking tickets are an important part of city revenue
- While parking violations themselves aren't the most riveting of topics, they can provide insights into parking itself, and its impacts on cost of living and even the environment
- Because more parking spaces expand cities and increase total driving, they have impacts on "affordable housing, climate change, economic development, public transportation, traffic congestion, and urban design."[1]

Problem characterization

- By examining how,when, and to whom these tickets are given we may be able to attain insights into whether they are serving the purpose they exist to serve
- Obtaining all the desired data, without leaving out important correlations, and collating it in such a way that we can access real insights into the functioning of our public systems and their application towards the public good will be difficult
- If parking tickets are not acting as effective deterrents it may be possible to find the true causes of increases in violations
- No demographic data is included so some conclusions may be hard to make
- Correlations could be found using adjacent info like vehicle make/model and year, but these will lack much validity

Methodology

General outline of each analytic task (performed with Apache Spark)

1. Load data - Read data as a text file from HDFS and create a Dataframe/RDD
2. Pre-process data - Perform data cleaning on required fields of data
3. Analyze - Carry out computations (transformations and actions from Spark) to analyze preprocessed data
4. Save - Write the result RDD as text files in HDFS

Augmented analysis with U.S. Census data

1. Extract county level data for chosen census data aspects (population, median household income, poverty) from the American Community Survey [2]
2. Filter needed data (from the dataset) for the 5 counties that make up New York City

Methodology

- Analyzing violation time of day for each county
  - filter() operation to remove the data header row, each row in the dataset was then treated as a row in the RDD
  - Use VIOLATION_TIME as the the column of interest
  - map() to extract TIME and COUNTY, split items on commas, checked for null values, correctly map county codes to counties and remove rows with empty strings
  - Create a pair RDD to count total entries in each hour for each county using reduceBy()
- Analyzing types of violation occurring most each year
  - Select VIOLATION_CODE and ISSUE_DATE as columns of interest
  - Transform ISSUE_DATE column using a spark column transformation and substring() to extract the year, selecting the new year column and omitting the old date column
  - filter() to remove erroneous years and violation codes not found in our problem-scope
  - Finally perform an aggregated count grouped by year and code to retrieve the desired output

Methodology

- Analyzing the month with most violations per county
  - Similar to *analyzing the time of day when most violations were recorded*, but using ISSUE_DATA instead of VIOLATION_TIME
- Analyzing top 5 vehicle makes which contributed to most parking violations in each county
  - Similar to *analyzing the time of day when most violations were recorded*, but using the column VEHICLE_MAKE instead of VIOLATION_TIME
  - Preprocessing step generates a pair RDD (Key: COUNTY_CODE:VEHICLE_MAKE, Value: 1)
  - filter() out entries with 0 as key
  - Perform reduceByKey() to aggregate entries that have the same key
  - For each county, sort the entries using the value and select the top 5
  - Output (COUNTY_CODE:VEHICLE_MAKE, count) to file

Performance Benchmarks

- Used a cluster of 17 computers from CS 120 Linux lab to run the Spark jobs
  - Hadoop Distributed File System (HDFS)
  - YARN as the resource manager
  - Apache Spark (v3.1.2)
  - Replication factor of 3 for HDFS
  - 36 workers and 72 GB of memory
- For performance comparison, we ran some of the jobs after reducing the number of worker nodes to 8.
- Additionally, performance was compared for some jobs with a standalone (single-node) Spark setup

Performance Benchmarks

- Record the time it took to complete each analytic task (from YARN UI)
- Run some of the tasks in a cluster with a fewer number of nodes and compare the times

	Time Taken to Complete the Job	
Job	**17-node cluster**	**8-node cluster**
No. of Violations over each month for each county	21 min	15 min
Most recorded vehicle makes for each county	17 min	13 min
Month of year when the highest number of violations were recorded for each county	12 min	10 min

Performance Benchmarks

- Cluster with fewer nodes has better throughput
- A standalone (single node) performance was even greater
- Because our dataset is about 8GB in size, it can be loaded into the memory all at once perform the calculation. This explains the disparities in the performance as Spark computations are done entirely in-memory
- When the number of nodes increases, the communication overheads add up and jobs takes longer to run to completion. If the dataset had been larger, disk accesses would have slowed computation for the single node

Insights and Conclusions

- Most parking violations occur in morning hours
- Some abnormal variations were present in total number of violations in particular months
    - Warrants further investigation
- Certain vehicle makes are frequently associated with more parking violations. This could be due to high popularity of such vehicle makes and their affordability for middle-class population
- With the advancement of technology, more and more "parking" violations are issued automatically. With speed cameras playing an increasing role in law enforcement
- Some similarities can be observed in the change of number of violations and population in New York City
- There was no discernible association with the number of violations and median household income

References

[1] D.Shoup,"Parking reform will save the city," Sept 2019.[Online]. Available:
https://www.citylab.com/perspective/2019/09/parking-lot-urban-planning-transit-street-traffic-congestion/598504/

[2] U. C. Bureau, "American community survey (acs)," Apr 2020. [Online]. Available: https://www.census.gov/programs-surveys/acs