CISC 5950 Bigdata Programming Project 1

Suwen Gu, Qike Yan, Jiayuan Tong

Question 1: NY Parking Violations

Introduction:

Background:

The NYC department of Finance collects data on every parking ticket issued in NYC, almost 10M per year. For analyzing such big data, we need to set up a cluster based on three nodes. Store the data in Hadoop Distribute File System and do some data computations based on MapReduce framework along with the scheduler for resource management. Our final purpose is to find the date, type and color of cars and location of tickets most likely to be issued based on such big data stored in HDFS with Google Cloud serving.

Data:

There are 43 columns and 7.8 M rows in total for this dataset. It could be separated by years from 2014 to 2018. Most data type is plain text for describe information like types, colors of cars, issued locations and violation descriptions etc. Other data type are numbers and Date. For the following questions we need to figure out, we will focus on serval related columns such as issue_date, vehicle_color, vehicle_body_type, vehicle_year and combination of street_name and street_code as the mapping target.

Design:

Each file contains millions of data records. To handle such a large data set, we employed Hadoop's MapReduce model.

In this part of project, we designed one MapReduce procedure for all 4 problems. The followings are the design of our MapReduce procedure:

1. Set up the MapReduce framework along with yarn scheduler for resource management.
2. Mapper works on mapping the features we needed and adding 1 as the count. Reducer is responsible for aggregating the data from mapper and output the total occurrence of each item of interest.
3. Format the output and save all results into a new file.

Results:

Question 1: When are tickets most likely to be issued?

year range	**Date** of tickets most likely issued	**Number** of tickets
07/2013 – 07/2014	11/29/2013	46023
07/2014 – 07/2015	01/15/2015	83112
07/2015 – 07/2016	09/22/2015	49425
07/2016 – 07/2017	09/16/2016	46860
07/2017 – 07/2018	10/03/2017	52030
07/2018 – 07/2019	10/25/2018	50979

Here is the result from each separated year range. We could also get the top result based on the whole six years, showing as 01/15/2015 with 83112 tickets issued.

Question 2: What are the most common years and types of cars to be ticketed?

year range	**Common years**	**Type** of cars	Ticketed **Number**
07/2013 – 07/2014	2013	SUBN	265153
07/2014 – 07/2015	2014	SUBN	375463
07/2015 – 07/2016	09/22/2015		49425
07/2016 – 07/2017	09/16/2016		46860
07/2017 – 07/2018	10/03/2017		52030
07/2018 – 07/2019	10/25/2018		50979

Question 3: Where are tickets most commonly issued?

year range	**Location** of tickets most likely issued	**Number** of tickets
07/2013 – 07/2014	Broadway	201980
07/2014 – 07/2015	Broadway	240650
07/2015 – 07/2016	Broadway	210683
07/2016 – 07/2017	Broadway	206157
07/2017 – 07/2018	Broadway	249089
07/2018 – 07/2019	Broadway	165795

Question 4: Which color of the vehicle is most likely to get a ticket?

year range	**Color** of vehicle for tickets most likely issued	**Number** of tickets
07/2013 – 07/2014	WHITE	1349181
07/2014 – 07/2015	GY	1717522
07/2015 – 07/2016	GY	1603929

07/2016 – 07/2017	GY	1744312
07/2017 – 07/2018	WH	2149522
07/2018 – 07/2019	WH	1489224