CISC 5950 Big Data Programming Project 1

Weifeng Yan, Alma Marko, Jiarui Zhang

Part 1: NY Parking Violations

DATA:
In this part, we are working with data on parking violations issued in 2020. The data set contains 43 columns and 3.6M rows. For the purposes of testing our code, we have reduced the number of rows we will use to the first 65,535 rows of data. The missing values in the file will be represented with an empty column or as 0. Therefore in the mapper phase, we are only focusing on values with a length greater than 0 and greater than one if a value can be turned into an integer. In this project, we will be mainly exploring the information from the following columns: Violation Time, Vehicle Body Type, Vehicle Year, Violation Location, and Vehicle Color.

DESIGN:
In our mapper, we are making simple counting for each of the columns that we are interested in and in our reducer, we are summing up the counts for each of our key values. Due to our Data Volume, we added a combiner for each of our questions in order to reduce the shuffling cost of our machine.

QUESTIONS

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

For this question, we chose the feature of "VIOLATION TIME" to represent the time and the value of this feature like *"0515P", "0935A".* In order to solve the question in a better way, we translate the standard time into military time, and in an hour. For example, *0515P → 17:00, 0935A → 09:00*

Design:
Mapper:
For each row in our data, we are outputting(hour, 1)
Combiner:
For each output from the mapper, we are outputting (hour, sum_count))
Reducer:
For each output from the combiner, we are outputting the top three (hour, sum(sum_count)

Result:
Below are the top three most common hours to get a ticket with their counts.

Hour	Count
08:00	7232
09:00	6424
11:00	6043

```
8:00    7232
9:00    6424
11:00   6043
```

Question 2: What are the most common years and types of cars to be ticketed?
For this question, we used the data in the Vehicle Body Type column and the Vehicle Year column to find the most common years and types of cars to be ticketed. For data cleaning, we removed the rows that had an empty value for Vehicle Body Type or the value 0 for Vehicle Year.

Design:
Mapper:
For each row in our data, we are outputting ((year, body type), 1)
Combiner:
For each output from the mapper, we are outputting ((year, body type), sum_count)
Reducer:
For each output from the combiner, we are outputting the top three ((year, body type), sum(sum_count))

Result:
Below are the top three most common years and types of cars that were ticketed with their count.

Vehicle Year-Vehicle Body Type	Count
2018-SUBN	4402
2017-SUBN	3545
2019-SUBN	3345

```
2018-SUBN       4402
2017-SUBN       3545
2019-SUBN       3345
```

Question 3: Where are tickets most commonly issued?

For this question, we are using “Violation Location” to represent the location where the violations occurred. For data cleaning, we removed all “Violation Location” values with a length less than 0 and int(value) equals 0.

Design:

Mapper:

For each row in our data we are outputting (violation location, 1)

Combiner:

For each output from the mapper, we are outputting (violation location, sum_count)

Reducer:

For each output from the combiner, we are outputting the top three (violation location, sum(sum_count)).

Result:

Below are the top three most common location where the violation occurred with their count.

Violation Location	**Count**
79	1353
83	1088
72	961

```
79      1353
83      1088
72      961
```

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

For this question, we chose the data in the “VEHICLE COLOR” feature to find the most color of cars to be ticketed. For data cleaning, we removed the rows that had an empty value for VEHICLE COLOR or the value 0 for Vehicle.

Design:

Mapper:

For each row in our data, we are outputting(color, 1)

Combiner:

For each output from the mapper, we are outputting (color, sum_count)

Reducer:

For each output from the combiner, we are outputting the top three (color, sum(sum_count))

Result:
Below are the top three most common vehicle colors that were ticketed with their count.

Color	**Count**
GY	13369
BK	11141
WH	8750

```
GY      13369
BK      11141
WH      8750
```