

CISC 5950 — Project 1

In CISC 5950, we have learned the following topics,

1. Set up a 3-node cluster with Hadoop Distributed File System and run examples.
2. On top of HDFS, set up the cluster with MapReduce programming framework.
3. Run examples of MapReduce programs.
4. Scheduling on the Cloud.

In this project, we are going to design our own Hadoop MapReduce-based program to analyze the data. The project consist of two parts.

NY Parking Violations

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 policy-makers.

You can find the data from the [Link of NYC Parking Data](#).

| | # Summ... | Plate ID | Registr... | Plate Ty... | Issue D... | # Violatio... | Vehicle ... | Vehicle ... | Issuing ... | Street ... |
|----|------------|----------|------------|-------------|------------|---------------|-------------|-------------|-------------|------------|
| 1 | 1283294138 | GBB9093 | NY | PAS | 08/04/2013 | 46 | SUBN | AUDI | P | 37250 |
| 2 | 1283294151 | 62416MB | NY | COM | 08/04/2013 | 46 | VAN | FORD | P | 37290 |
| 3 | 1283294163 | 78755JZ | NY | COM | 08/05/2013 | 46 | P-U | CHEVR | P | 37030 |
| 4 | 1283294175 | 63009MA | NY | COM | 08/05/2013 | 46 | VAN | FORD | P | 37270 |
| 5 | 1283294187 | 91648MC | NY | COM | 08/08/2013 | 41 | TRLR | GMC | P | 37240 |
| 6 | 1283294217 | T60DAR | NJ | PAS | 08/11/2013 | 14 | P-U | DODGE | P | 37250 |
| 7 | 1283294229 | GCR2838 | NY | PAS | 08/11/2013 | 14 | VAN | | P | 37250 |
| 8 | 1283983620 | XZ764G | NJ | PAS | 08/07/2013 | 24 | DELV | FORD | X | 63430 |
| 9 | 1283983631 | GBH9379 | NY | PAS | 08/07/2013 | 24 | SDN | TOYOT | X | 63430 |
| 10 | 1283983667 | MCL78B | NJ | PAS | 07/18/2013 | 24 | SDN | SUBAR | H | 0 |
| 11 | 1283983679 | M367CN | NY | PAS | 07/18/2013 | 24 | SDN | HYUND | H | 0 |
| 12 | 1283983734 | GAR6813 | NY | PAS | 07/18/2013 | 24 | SDN | TOYOT | H | 0 |

The above figure shows several records, where each row represents a parking ticket and the columns are the details of the tickets.

To start the project, you have to,

1. Start the 3-node cluster
2. Set up the HDFS

3. Store the data in HDFS
4. Set up the MapReduce framework along with the scheduler for resource management.

By analyzing the data, we need to answer the following,

- When are tickets most likely to be issued?
- What are the most common years and types of cars to be ticketed?
- Where are tickets most commonly issued?
- Which color of the vehicle is most likely to get a ticket?

NBA Shot Logs

<https://www.kaggle.com/dansbecker/nba-shot-logs>

This is the [DATA](https://www.kaggle.com/dansbecker/nba-shot-logs) (<https://www.kaggle.com/dansbecker/nba-shot-logs>) on shots taken during the 2014-2015 season, who took the shot, where on the floor was the shot taken from, who was the nearest defender, how far away was the nearest defender, time on the shot clock, and much more. The column titles are generally self-explanatory.

The above figure shows several records, where each row represents a shot and the columns are the details of the shot, e.g. the game ID, who is the defender, what is the distance between them.

| | # GAME_ID | A MATCH... | A LOCATI... | A W | # FINAL_... | # SHOT_... | # PERIOD | 📅 GAME_... | # SHOT_... | # DRIBB... |
|---|-----------|--------------------------|-------------|-----|-------------|------------|----------|------------|------------|------------|
| 1 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 1 | 1 | 1:09 | 10.8 | 2 |
| 2 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 2 | 1 | 0:14 | 3.4 | 0 |
| 3 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 3 | 1 | 0:00 | | 3 |
| 4 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 4 | 2 | 11:47 | 10.3 | 2 |
| 5 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 5 | 2 | 10:34 | 10.9 | 2 |
| 6 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 6 | 2 | 8:15 | 9.1 | 2 |
| 7 | 21400899 | MAR 04, 2015 - CHA @ BKN | A | W | 24 | 7 | 4 | 10:15 | 14.5 | 11 |

By analyzing the data, we need to answer the following,

- For each pair of the players (A, B), we define the **fear score** of A when facing B is the hit rate, such that B is closet defender when A is shooting. Based on the **fear score**, for each player, please find out who is his "most unwanted defender".
- For each player, we define the **comfortable zone** of shooting is a matrix of,

$\{\text{SHOT_DIST}, \text{CLOSE_DEF_DIST}, \text{SHOT_CLOCK}\}$

Please develop a MapReduce-based algorithm to classify each player's records into 4 comfortable zones. Considering the hit rate, which zone is the best for James Harden, Chris Paul, Stephen Curry and LeBron James.

Bonus Question

The biggest challenge when using K-Means is to decide on the number of clusters. Having more clusters creates some small classes with very few records, while having less clusters leads to classes that are too general.

Based on a K-Means algorithm above, try to answer the following question,

- Given a Black vehicle parking illegally at 34510, 10030, 34050 (street codes). What is the probability that it will get an ticket? (very rough prediction).
- At 10 am, I want to go to Lincoln Center and I just want to walk within 0.5 mile. Where should I park? (Divided into zones).

Grading Rubric

You should complete the lab in groups of 4 students.

(70%) P1: NY Parking Violations (17.5% * 4);

(20%) P2: NBA Shot Logs (10% * 2);

(10%) Two Reports the your design and experiments, please as detail as possible and must include your screenshots; In addition, you also need to write two README files for P1 and P2.

(10%) Bonus Question (5% + 5%);

Submission

You are expected to email me a zip(or tar) file by the deadline (Nov. 8th, 2020). The zip file should include two (or three) folders,

- Part1: your codes, report and README
- Part2: your codes, report and README

- Bonus: your codes, report and README

Useful Links

1. [Analysis of NYC Parking Tickets.](#)
2. [Preliminary Data Visualization.](#)
3. [Exploring 42.3M NYC Parking Tickets.](#)
4. [NY Parking Violations Issued .](#)
5. [Insights From Raw NBA Shot Log Data.](#)
6. [Investigating the hot hand phenomenon in the NBA \(CODE\).](#)
7. [Parallel K-Means Clustering Based on MapReduce.](#)
8. [NBA 16-17 regular season shot log.](#)
9. [The Fear Factor.](#)
10. [The Best And Worst Defenders.](#)
11. [NBA Classification.](#)
12. [Stephen Curry's Decision Tree.](#)
13. [Points per Match \(ATL vs WAS only\).](#)
14. [MapReduce-kmeans.](#)