Chicago Traffic Crashes Analysis

By DATA CHAMPS

Team Members

Shivani Konyala

Yeshwanth Nellikanti

ASSIGNMENT – 3

1. What research question did you decide to program in Pig or Spark and why?

We have decided to run all our research questions in Spark. This is because it's fast and we are familiar with Python and Pandas. In spark, we ran one of our queries which has an aggregate function and it showed up the result in seconds which did not happen in Hive.

2. Are you running into any problems running your queries?

Initially, we started to execute our queries in Hive and for one of the queries it took a long time and the results are not as expected. So, we tried changing some of the advanced configurations in Hive and this showed no improvement in processing or reducing the time.

3. What percentage of your queries have you completed?

Currently, we have completed 60% of the queries in the spark.

4. What do you have remaining for the project?

As of now, we are done with executing one of our research questions successfully and the second and third queries are partially done.
Once we are done with successful execution of all the three queries, we will then work on visualization reports.