Project Report – Analyzing flight delays

W205. Rohitha Bhushan

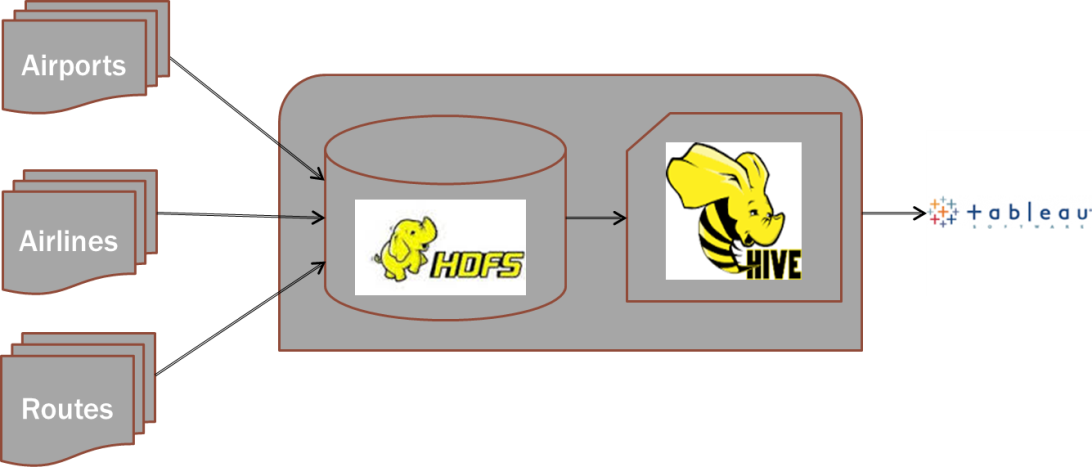
Introduction:

The Airline Industry suffers cancellations and delays routinely resulting in significant costs to both airlines and travelers. In this project, we aim to analyze the airline dataset from the United States Department of Transportation, Bureau of Transportation Statistics. The data consists of all arrival and departure records of all US domestic flights in the year 2013. The file is of type csv and about 1GB with 6 million records per year.

Architecture:

Hive is based on Hadoop and MapReduce operations. Hive allows SQL developers to write Hive Queries. These queries are broken down into MapReduce operations and executed across a Hadoop cluster. Amazon AWS deploys and provisions Apache Hadoop clusters in the cloud, providing a software framework.

The datasets are first saved to Amazon S3. Through a Hadoop distributed file system interface, the full set of components in S3 can operate directly on structured or unstructured data in EBS volume. Storing data in EBS Volume provides the ability to safely delete the clusters used for computation without losing user data.



Experimental results:

Once HDFS is running, Hive queries can be executed on the cluster from the Hive CLI. Start off by creating three tables for my analysis – airports, airlines and routes. I also created a climate table but owing to the complexity of the dataset, didn’t use all the data available.

Using the Hive query CLI, the data is analyzed by running the following queries. The total number of flights cancelled was 14,747 for the month of December alone. Diverted flights stood at 1156. Two MapReduce jobs were launched for this query. Queries took about 55s.

Analyzing departure/arrival delays showed us that ATL airport (Atlanta, GA) had the maximum number of delays at 16073 and 15158 respectively. The Western Samoa airport at Pago Pago had the least delays. Weather delays bogged down Chicago’s O’Hare International the most in December.

I performed some statistical data analysis by first calculating average departure delay and the standard deviation of departure delay. When I calculated the correlation between departure delays and arrival delays, the co-efficient was closer to 1 which indicated a strong likelihood. For the month of December, it appears that there are no flights that can compensate for the departure delay and arrive early to their destination. I also calculated the Pearson’s co-efficient ‘r’ to analyze relationship between distance and departure delay.

Finally, for a visual display I connected to Tableau using Cloudera ODBC driver and deployed my datasets. Friday consistently saw delays in all areas - weather , security and late aircraft. Arrival delays for a month had an interesting trend of spikes with very little consistency. Other than the fact that the least delays happened in the first week, nothing useful came up there. Delays due to weather were often on weekends.

Improvements:

1. In the next version of this project, I’d build a REST API to serve the data.
2. Streaming ATC for live data to ensure real-time analysis is something to explore for the future.
3. Machine learning models could be attempted with additional datasets to predict delays efficiently.

Variations:

1. Usage of Weather dataset: I modeled the architecture initially using Climate dataset but did not retrieve any useful analysis from it, more than what I could using Hive queries.
2. Cascading delays: Utilized complex Hive queries to perform statistical analysis for this in place of Python script that I wrote.