Big Data Analysis of Airline Data Set using Hive

Nillohit Bhattacharya, Jongwook Woo

Grad Student, Prof., Department of Computer Information Systems, California State University Los Angeles

Keywords: Hadoop, HDInsight, Big Data, Hive, Data Visualization, Data Analysis

ABSTRACT

In this paper, the analysis of the airline data set is performed using Microsoft Azure HDInsight which runs Hadoop in the cloud. Hive and Hive QL statements have been used for querying the data. Data visualization has been done by extracting the output of the HIVE query in excel and plotting the data using line and scatter plot charts. The visualization of the data shows some patterns that exist between flight diversions and flight distance, flight cancellation and flight distance and so forth.

Section 5 is the conclusion of this paper.

1. INTRODUCTION

There is no doubt that a lot of excitement exists with the term Big Data. Big Data in simple words can be large-scale data which does not have a well-defined structure. The size of the data itself is so huge that it is not practically easy for a single computer to store and process all the data by itself. Traditional computing approached the problem in a different way, the focus was always to increase the processing speed and power of the computer. As the data grows exponentially, the processing power of the single computer becomes a bottleneck and thus a new approach was needed to address the issue at hand. A new way was developed where many non-expensive commodity computers all working together in harmony with each other, in order to store and process this big data in parallel that allows us to extract meaningful information from a large data set. Moreover, current technologies using the cloud infrastructure allows us to easily create clusters of computers by renting them for as much time as required and then releasing the computing resources when no longer needed. Thus with cloud technologies we get the computing power of the clusters of computers with minimal investment.

The airline data has been taken from the United States Department of Transportation, Bureau of Transportation Statistics [1]. The data consists of the arrival and departure records of all US domestic flights from the period 2012 to 2014.

Section 2 gives a brief introduction of the Airline Data set, Hive and HDInsight. Section 3 describes the mechanism by which the data set is analyzed. Section 4 describes the experimental observations from the data set. Section 5 is the conclusion of this paper.

2. AIRLINE DATA, HIVE AND HDINSIGHT

This section briefly describes the characteristics of the Airline Data set, introduces HIVE and HDINSIGHT.

2.1 Characteristics of the Airline Data

Total number of files: 36.

File type: csv (comma separated values)

Total file size: 3.65 GB.

Total number of records: 18.2 Million (18,286,090)

2.2 Hive

Hive, allows SQL developers to write Hive Query Language (HQL) statements that are similar to standard SQL statements. HQL statements are broken down into MapReduce jobs and executed across a Hadoop cluster. Even though, HQL statements are similar to SQL statements, there are several key differences because Hive is based on Hadoop and MapReduce Operations.

The first is that Hadoop is intended for long sequential scans, and because Hive is based on Hadoop, queries tend to have a very high latency (many minutes). This means that Hive would not be appropriate for applications that need very fast response times. Hive is read-based and therefore not appropriate for transaction processing that typically involves a high percentage of write operations [2].

2.3 HDInsight

Azure HDInsight deploys and provisions Apache Hadoop clusters in the cloud, providing a software framework designed to manage, analyze, and report on big data with high reliability and availability. HDInsight uses the Hortonworks Data Platform (HDP) Hadoop distribution. Hadoop often refers to the entire Hadoop ecosystem of components, which includes Storm and HBase clusters, as well as other technologies under the Hadoop umbrella.

Azure HD Insight deploys and provisions Hadoop clusters in the cloud, by using either Linux or Windows as the underlying Operating System [3].

3. ANALYSIS OF AIRLINE DATA USING HDINSIGHT

In order to analyze the Airline Data, the data needs to be first saved to the Azure Blob Storage [4], which is a cloud data storage service provided by Microsoft Azure. For transferring the Airline Data to the Azure Blob storage, a client utility program “CloudBerry Explorer for Azure Blob Storage” was used [5].
Azure Blob storage is a robust, general-purpose storage solution that integrates seamlessly with HDInsight. Through a Hadoop distributed file system (HDFS) interface, the full set of components in HDInsight can operate directly on structured or unstructured data in Blob storage. Storing data in Blob storage provides the ability to safely delete the HDInsight clusters that are used for computation without losing user data. Azure HDInsight provides a full-featured Hadoop distributed file system (HDFS) over Azure Blob storage. It enables the full set of components in the Hadoop ecosystem to operate directly on the data it manages. Azure Blob storage and HDFS are distinct file systems that are optimized for storage of data and computations on that data [6].

Once the Azure Blob storage account is created and the data is transferred, the HDInsight cluster can be launched from Microsoft Azure Portal. The overall structure of the system is as shown in the Fig 1 below.

![Figure 1: System architecture for analyzing the airline data set](image)

### 3.1 Hive Query for Creating an External Table in Hive

Once the HD Insight cluster is up and running, Hive queries can be executed on the cluster from the Hive query console window. The following HQL statement is executed to create an external table “Airline”. Note that the following HQL statements are very similar to SQL statements.

```
DROP TABLE IF EXISTS Airline;
CREATE EXTERNAL TABLE Airline
(YEAR BIGINT, MONTH BIGINT, FL_DATE STRING, CARRIER STRING, TAIL_NUM STRING, ORIGIN STRING, DEST STRING, CRS_DEP_TIME STRING, DEP_TIME STRING, DEP_DELAY BIGINT, DEP_DELAY_NEW BIGINT, DEP_DELAY_GROUP BIGINT, DEP_TIME_BLK STRING, TAXI_OUT BIGINT, WHEELS_OFF STRING, WHEELS_ON STRING, TAXI_IN BIGINT, CRS_ARR_TIME STRING, ARR_TIME STRING, ARR_DELAY BIGINT, ARR_DELAY_NEW BIGINT, ARR_DEL15 BIGINT, ARR_DELAY_GROUP BIGINT, ARR_TIME_BLK STRING, CANCELLED BIGINT, CANCELLATION_CODE STRING, DIVERTED BIGINT, CRS_ELAPSED_TIME STRING, ACTUAL_ELAPSED_TIME STRING, AIR_TIME BIGINT, FLIGHTS BIGINT, DISTANCE BIGINT, DISTANCE_GROUP BIGINT, CARRIER_DELAY BIGINT, WEATHER_DELAY BIGINT, NAS_DELAY BIGINT, SECURITY_DELAY BIGINT, LATE_AIRCRAFT_DELAY BIGINT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION
'wasb://hdgradstudy@gradstudy.blob.core.windows.net/Airline_Data';
```

The total time that was required to complete the above job is 4.971 seconds.

### 4. EXPERIMENTAL RESULTS

For analyzing the Airline Data, a cluster of 4 data nodes (4 compute machines) running Microsoft Windows Server 2012 R2 Datacenter operating system was launched. Hive runs on the cluster by default when the cluster is up and running. Using the Hive query console, the data is analyzed as follows:

#### 4.1 Total Number of Flights Cancelled Each Month for 2012-2014

The Hive QL for querying the total number of flights that were cancelled every month from 2012 to 2014 is:

```
SELECT YEAR, MONTH, COUNT(CANCELLED) AS TOTAL_CANCELLED
FROM Airline
WHERE CANCELLED = 1
GROUP BY YEAR, MONTH
ORDER BY YEAR, MONTH
LIMIT 50;
```

The total time that was required to complete the above job is 122.496 seconds.

#### 4.2 Total Number of Flights Diverted Each Month for 2012-2014

The Hive QL for querying the total number of flights that were diverted every month from 2012 to 2014 is:

```
SELECT YEAR, MONTH, COUNT(DIVERTED) AS TOTAL_DIVERTED
FROM Airline
WHERE DIVERTED = 1
GROUP BY YEAR, MONTH
ORDER BY YEAR, MONTH
LIMIT 50;
```
The total time that was required to complete the above job is 128.767 seconds.

The figure Fig (2) below shows the trend of cancelled and diverted flights by month from January to December for the period 2012 to 2014.

**Figure 2:** Cancelled and Diverted Flights by Month

### 4.3 Total Number of Flights Cancelled Each Year for 2012-2014

The Hive QL for querying the total number of flights that were cancelled every year from 2012 to 2014 is:

```
SELECT YEAR, COUNT (CANCELLED) AS TOTAL_CANCELLED FROM Airline WHERE CANCELLED = 1 GROUP BY YEAR ORDER BY YEAR LIMIT 50;
```

The total time that was required to complete the above job is 136.229 seconds.

### 4.4 Total Number of Flights Diverted Each Year for 2012-2014

The Hive QL for querying the total number of flights that were diverted every year from 2012 to 2014 is:

```
SELECT YEAR, COUNT (DIVERTED) AS TOTAL_DIVERTED FROM Airline WHERE DIVERTED = 1 GROUP BY YEAR ORDER BY YEAR LIMIT 50;
```

The total time that was required to complete the above job is 123.71 seconds.

The figure Fig (3) below shows the scatter plot of the total number of flight diversions with respect to flight distance (in miles) for the period 2012 to 2014.

**Figure 3:** Number of flight diversions vs flight distance

### 4.5 Effect of Flight Distance on Diversions for 2012-2014

The Hive QL for querying the total number of diverted flights as a function of flight distance from 2012 to 2014 is:

```
SELECT DISTANCE, COUNT (DIVERTED) AS DIVERTED_COUNT FROM Airline WHERE DIVERTED = 1 GROUP BY DISTANCE ORDER BY DIVERTED_COUNT DESC LIMIT 1500;
```

The total time that was required to complete the above job is 137.1 seconds.

The figure Fig (4) below shows the scatter plot of the total number of flight diversions with respect to flight distance (in miles) for the period 2012 to 2014.

**Figure 4:** Number of flight diversions vs flight distance

### 4.6 Effect of Flight Distance on Cancellations for 2012-2014

The Hive QL for querying the total number of cancelled flights as a function of flight distance from 2012 to 2014 is:

```
SELECT DISTANCE, COUNT (CANCELLED) AS CANCELLED_COUNT FROM Airline WHERE CANCELLED = 1 GROUP BY DISTANCE ORDER BY CANCELLED_COUNT DESC
```
The total time that was required to complete the above job is 102.096 seconds.

The figure Fig (5) below shows the scatter plot of the total number of flight cancellations with respect to flight distance (in miles) for the period 2012 to 2014.

**Figure 5:** Number of flight cancellations vs flight distance

### 4.7 Effect of Flight Distance on Average Departure Delay for 2012-2014

The Hive QL for querying the average flight departure delay as a function of flight distance from 2012 to 2014 is:

```
SELECT DISTANCE, AVG (DEP_DELAY) AS AVG_DELAY
FROM Airline
GROUP BY DISTANCE
ORDER BY AVG.Delay DESC
LIMIT 1000;
```

The total time that was required to complete the above job is 111.169 seconds.

The figure Fig (6) below shows the scatter plot of the average flight departure delay (in minutes) with respect to flight distance for the period 2012 to 2014.

**Figure 6:** Average departure delay vs flight distance

### 4.8 Monthly Average Departure Delay for 2012-2014

The Hive QL for querying the monthly average flight departure delay for 2012 to 2014 is:

```
SELECT MONTH, AVG (DEP_DELAY) AS AVG_DELAY
FROM Airline
GROUP BY MONTH
ORDER BY MONTH;
```

The total time that was required to complete the above job is 105.263 seconds.

The figure Fig (7) below shows the trend of the average flight departure delay as a function of month from January to December for the period 2012 to 2014.

**Figure 7:** Average departure delay vs month

### 4.9 Yearly Average Departure Delay for 2012-2014

The Hive QL for querying the yearly average flight departure delay for 2012 to 2014 is:

```
SELECT YEAR, AVG (DEP_DELAY) AS AVG_DELAY
FROM Airline
GROUP BY YEAR;
```

The total time that was required to complete the above job is 76.179 seconds.

The figure Fig (8) below shows the trend of the average flight departure delay as a function of year for the period 2012 to 2014.

**Figure 8:** Average departure delay vs year

### 4.10 Total Number of Flights Each Year for 2012-2014

The Hive QL for querying the total number of flights per year for 2012 to 2014 is:
SELECT COUNT (*) FROM Airline
WHERE YEAR = 2014;

The total time that was required to complete the
above job is 110.428 seconds.

The figure Fig (9) below shows the trend of the
total number of flights each year for the period 2012 to
2014.

![Total Flights By Year](image)

**Figure 9:** Total number of flights vs year

5. CONCLUSION

From the above experimental results, we can see
that interesting sets of trends and patterns exists in large
data sets which helps us to get a better understanding of
the data.

Recent advancement in cloud technologies helps
us to harness the power of parallel processing of a cluster
of computers with little investment and almost no
maintenance of the underlying computer hardware.

From the experimental results we also see the
following observations:

a. Average flight departure delay is at peak during
   the months of June and July every year and there
   is a sharp increase in the average delay from
   November to December.

b. Average flight departure delay is increasing
   continuously over the period 2012 to 2014 in
   spite of the fact that the total number of flights
   have decreased from 2013 to 2014.

c. Highest average departure delay for flights has
   been observed for flight distance of less than 500
   miles.

d. Highest numbers of flights which are cancelled
   have a flight distance of less than 1000 miles.

e. There is a trend of increasing flights which are
   cancelled every year from 2012 to 2014.

f. There is a sharp rise in the number of flights
   which are cancelled from the month of November to
   January every year for 2012 to
   2014.

REFERENCES

[1] Airline Data Set, United States Department of
   Transportation, Office of the Assistant Secretary
   for Research and Technology, Bureau of
   Transportation Statistics,
   http://www.transtats.bts.gov/DL_SelectFields.asp?
   Table_ID=236


[3] Introduction to Hadoop in HDInsight: Big-data
   analysis and processing in the cloud,
   https://azure.microsoft.com/en-
   us/documentation/articles/hdinsight-hadoop-
   introduction/

[4] An Introduction to Windows Azure BLOB Storage,
   https://www.simple-talk.com/cloud/cloud-data/an-
   introduction-to-windows-azure-blob-storage/-

   Client, http://www.cloudberrylab.com/free-
   microsoft-azure-explorer.aspx

[6] Upload data for Hadoop jobs in HDInsight,  
   https://azure.microsoft.com/en-
   us/documentation/articles/hdinsight-upload-data/

   MapReduce”, Jongwook Woo, DMKD-00150,
   Wiley Interdisciplinary Reviews Data Mining and
   Knowledge Discovery, Oct 28 2013, Volume 3,
   Issue 6, pp445-452, ISSN 1942-4795

AUTHOR PROFILE

Nillohit Bhattacharya is currently pursuing his Master’s
degree in Information Systems at California State
University Los Angeles. He is a member of the HiPIC
(High-Performance Information Computing Center) at the
Computer Information Systems Department of the
University. He completed his Bachelor degree in
Electronics from Pune University, India in 2005 and Post
Graduate Program in Management from Symbiosis
International University, India in 2011. His interests
include learning new and emerging technologies in Big
Data and Analyzing Data Sets to extract patterns and
information from them.

Jongwook Woo is currently a Full Professor at Computer
Information Systems at California State University Los
Angles. He is a director of the HiPIC (High-Performance
Information Computing Center) at the university.
He received his BS and the MS degree, both in Electronic
Engineering from Yonsei University in 1989 and 1991,
respectively. He obtained his second MS degree in
Computer Science and received the PhD degree in
Computer Engineering, both from University of Southern
California in 1998 and 2001, respectively. His research
interests are Information Retrieval /Integration /
Sharing on Big Data, Map/Reduce, In-Memory Processing, and
functional algorithm on Hadoop Parallel/Distributed/Cloud Computing, and n-Tier
Architecture application in e-Business, smartphone, social
networking and bioinformatics applications. He has
published more than 40 peer reviewed conference and
journal papers. He also has consulted many entertainment
companies in Hollywood.