International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

BIG DATA ANALYSIS OF AIRLINE DATASET USING HIVE

Dr. C K GOMATHY[1], Miss. B JAHNAVI[2], Mr. C H RAKESH[3], Mr. S ARAVIND[4]

Abstract - In this article, airline database analysis is performed using Microsoft Azure HDInsight manages Hadoop in the cloud. Hive and HiveQL statements are used for the following purposes:

The data show flight deviations and distances, some patterns between flights, Flight cancellations, distances, etc. Please refer to the data. Data visualization was performed by extracting the results from HIVE. Run the query in Excel and draw the data using line and scatter plots. Visualization

Key Words: Big Data, Dataset, Microsoft Azure Database, Hive QL, Data visualization, etc.

1.INTRODUCTION

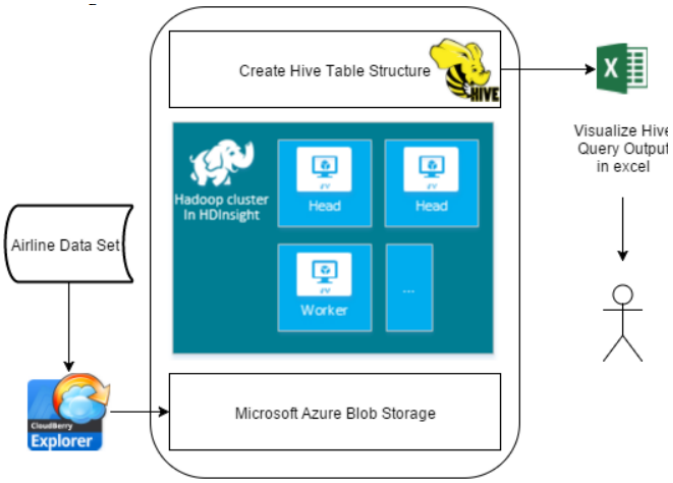
There is no doubt that there is a lot of excitement in the word big data. Simple big data Words can be large amounts of data that do not have a well-defined structure. The data itself is so large that it is virtually impossible to store and process it on a single computer. All the data itself. Traditional computers have tackled the problem differently. The focus has always been on increasing the processing speed and power of computers. Like the data. Exponential growth, the processing power of a single computer becomes a bottleneck. Therefore, a new approach was needed to address this issue. A new method has been developed. Where many cheap basic computers work in harmony with each other. The other enables extraction by storing and processing this big data in parallel. Meaningful information from large datasets. In addition, current technology using the cloud. The infrastructure makes it easy to create groups of computers for rent. Time as needed and free up computer resources when you no longer need them. Therefore, use cloud technology to gain the computing power of a group of computers.

1.1 Apache HIVE

Apache Hive is an information warehouse software program constructed for presenting information question and evaluation on pinnacle of Apache Hadoop. Traditional SQL queries need to be carried out withinside the MapReduce Java API to execute SQL programs and queries over allotted information. Hive indicates an SQL-like interface to asked information saved those mixes with Hadoop in diverse databases and report systems. Hive affords the important SQL abstraction to combine SQL-like queries into the Java without the want to put into effect queries withinside the low-degree Java API. Since nearly all information warehousing programs paintings with SQL-primarily based totally querying languages, Hive enables portability of SQL primarily based totally programs to Hadoop.

2. METHODOLOGY

To analyze airline data, you must first store the data in Azure Blob Storage is a cloud data storage service provided by Microsoft Azure. For Transfer airline data to Azure Blob storage, customer service uses " Cloud Berry ". Azure Blob Storage Explorer. Azure Blob Storage is a robust general purpose storage solution that integrates seamlessly with HDInsight. Via Hadoop Distributed File System (HDFS) interface, full set of components can work with HDInsight directly to structured or unstructured data in Blob storage. Save data to Blob storage, it provides the ability to safely remove HDInsight clusters used on your computer without losing user data. Azure HDInsight provides a complete Hadoop distributed archive system on Azure Blob Storage. Activate the complete set of Hadoop components ecosystem that directly processes the data it manages. Clear file system optimized for data storage and data computation. Once the Azure Blob storage account is created and the data is transferred, the HDInsight cluster can be launched from Microsoft Azure Portal.



3. RESULTS

To analyze airline data, run a group of 4 data nodes (4 computers) using the Microsoft Windows Server 2012 R2 Datacenter operating system has been released. Hive keeps running the default for a running cluster. With the Hive query console, the data is analyzed as follows:

3.1 No. Of Flights Cancelled in Each Month from 2012 to 2014

Hive QL checks the total number of flights canceled each month since then from 2012 to 2014 are:

SELECT YEAR, MONTH, COUNT (CANCELLED) AS TOTAL_CANCELLED

FROM Airline

WHERE CANCELLED = 1

GROUP BY YEAR, MONTH

ORDERED BY YEAR, MONTH

LIMIT 50;

3.2 No. Of Flights Detoured in Each Month from 2012 to 2014

HiveQL to see the total number of flights detoured each month from 2012 to 2014 are

SELECT YEAR, MONTH, COUNT (DIVERTED) AS TOTAL_DIVERTED

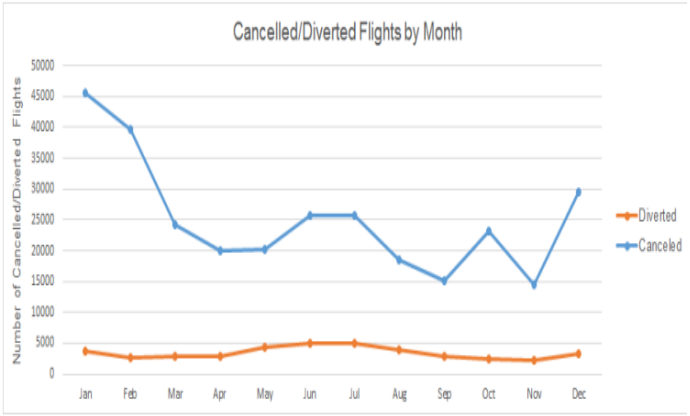
FROM Airline

WHERE DIVERTED = 1

GROUP BY YEAR, MONTH

ORDER BY YEAR, MONTH

LIMIT 50;



Cancelled and Detoured Flights by Month

3.3 No. Of Flights Cancelled in Each Year from 2012 to 2014

Hive QL checks the total number of flights that are cancelled each year since then from 2012 to 2014:

SELECT YEAR, COUNT (CANCELLED) AS TOTAL_CANCELLED

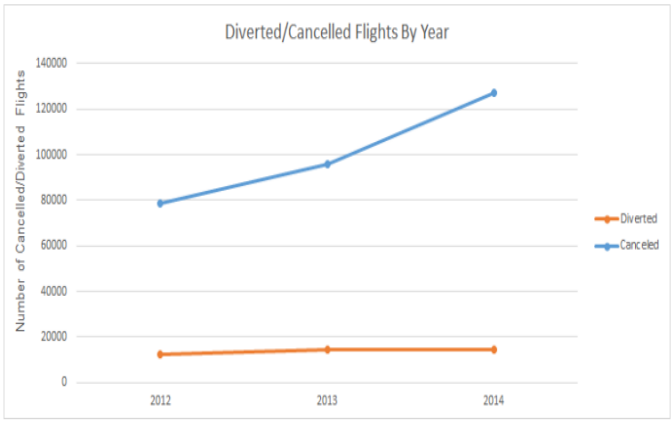
FROM Airline

WHERE CANCELLED = 1

GROUP BY YEAR

ORDER BY YEAR

LIMIT 50;



CONCLUSIONS

The experimental results above show this interesting set of trends and patterns. It exists in large datasets that help you better understand your data. Recently advances in cloud technology will help increase the power of parallel processing with the help of group of computers with little investment and little underlying maintenance of computer hardware.

From the experimental results we also see the following observations:

- The average delay in flight departures is highest between June and July each year so the average delay increases sharply from November to December.
- The average delay in flight departures is constantly increasing during the period 2012-2014.Despite the decrease in total flights from 2013 to 2014.
- The shortest average departure delay was observed on short flight distances 500 miles.
- Maximum number of cancelled flights has a flight distance of less than 1000 Miles.

- From 2012 to 2014, the number of cancelled flights tends to increase every year.
- The number of cancelled flights has increased sharply since the month of November to January of all years from 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.
[2] What is Hive? http://www-01.ibm.com/software/data/infosphere/hadoop/hive/.
[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.simpletalk.com/cloud/cloud-data/an-introduction-to-windows-azure-blob-storage-/
[5] Explorer for Microsoft Azure Storage:FreewareClient, 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/
[7] "Market Basket Analysis Algorithms with MapReduce", Jongwook Woo, DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28, 2013, Volume 3, Issue 6, pp445-452, ISSN 1942-479

BIOGRAPHIES

1. Dr. C.K. Gomathy is an Assistant Professor in Computer Science and Engineering at Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her area of interest is Software Engineering, Web Services, Knowledge Management and IOT.

2. Miss. Bandi Jahnavi, 11189A023, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

3. Mr. Chaganam Rakesh, 11189A037, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

4. Mr. Aravind Srinivasan, 11189A018, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.