AVIATION DATA ANALYSIS

Mohammed Khalil Bangdiwal, Khan Mohammed Aasif, Chafekar Mohd Umair, Rangwala Shahid
Information Technology Department, Mumbai University

Abstract - Big data is an act of gathering and storing large amount of information for analysis. Big data consists of massive volume of both structured and unstructured data. The analysis of airline data Is performed using hadoop. Hive statements have been used for querying the airline 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 airline data shows some patterns that exist between flight diversions and flight distance, flight cancellation and flight distance and so forth.

General terms - *Big data, Hadoop, Distributed file system, data analytics*

Keywords - *Airline data set, Hive tools*

I. INTRODUCTION

The term "Big Data" refers to all the data that is being generated across the globe at an unprecedented rate. It can also be referred to as "Data which was ignored in the last decades due to limitations of Database "This data could be either structured or unstructured. Data drives the modern organizations of the world and hence making sense of this data and unraveling the various patterns and revealing unseen connections within the vast sea of data becomes critical and a hugely rewarding endeavour indeed. Better data leads to better decision making and an improved way to strategize for organizations regardless of their size, geography, market share, customer segmentation and such other categorizations. Hadoop is the platform of choice for working with extremely large volumes of data. The most successful enterprises of tomorrow will be the ones that can make sense of all that data at extremely high volumes and speeds in order to capture newer markets and customer base. There are amazing benefits to real-time big data analytics. First it allows businesses to detect errors and fraud quickly. This significantly mitigates against losses. Second, it provides major advantages from a competitive standpoint. Real-time analysis allows businesses to develop more effective strategies towards competitors in less time, offering deep insight into consumer trends and sales.

II. AIRLINE DATA, HIVE, HDFS

This section briefly describes the characteristics of the Airline Dataset, Introduces HIVE and HDFS.

Airline Data

Legacy aircrafts used to capture 125+ flight parameters, but Boeing 787 captures more than 1000 flight parameters with some reports claiming half a terabyte of data per flight. This explains the big data explosion in aviation. Apart from these flight data, a large amount of data get generated in repair shops, inventory systems and by various regulatory organizations as well. Analyzing such big data can help in improving flight safety, reducing operational delay, better inventory management of spares, predictive maintenance of various equipments on board, improving fuel economy of the fleet etc.

Characteristics

Total number of files: 16
File type: csv (comma separated values)
Total file size: 100MB.
Total number of records: 10,000

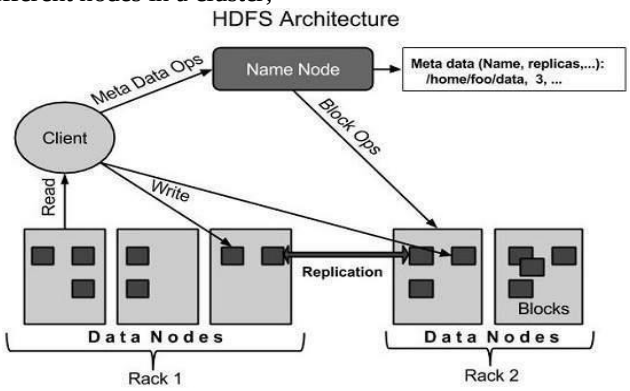
Hive:

Hive, allows SQL developers to write Hive Query Language(HQL) statements that are similar to standard SQL statements. HQL statements are broken down into Map Reduce jobs and execute data 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 Map Reduce 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].

HDFS:

HDFS is built to support applications with large data sets, including individual files that reach in to the terabytes. It uses a master/slave architecture,with each cluster consisting of a single Name Node that manages file system operations and supporting Data Nodes that manage data storage on individual compute nodes.When HDFS takes in data,it breaks the in formation down into separate pieces and distributes them to different nodes in a cluster,



allowing for parallel processing.The file systems copies each piece of data multiple times and distributes the copies to individual nodes, placing at least one copy on a different server rack than the others. As a result,the data on nodes that crash can be found elsewhere within a cluster ,which allows processing to continue while the failure is resolved.

III. RELATED WORK

1. Big Data Analysis Of Airline

Data Set Using Hive

Authors: P. Swathi, J. Kumari

Year: 2017

- In this paper, the analysis of the airline data set is performed using Microsoft Azure, HD Insight which runs Hadoop in the cloud. Hive and Hive QL statements have been used for querying the data.
- Data visualization is done by obtaining the output of the HIVE query in excel and plotting the data using the techniques available.
- The data visualization shows some patterns that exist between flight diversions and flight performance, flight cancellation and flight distance and so forth

2. A Review on Flight Delay Prediction

Authors: Alice Sternberg, Jorge Soares, Diego Carvalho, Eduardo Ogasawara

Year: 2017

- Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became cumbersome due to the complexity of air transportation system, the number of methods for prediction, and the deluge of flight data. In this context, this paper presents a thorough literature review of approaches used to build flight delay prediction models from the Data Science perspective. It proposes a taxonomy and summarizes the initiatives used to address the flight delay prediction problem, according to scope, data, and computational methods, giving particular attention to an increased usage of machine learning methods. Besides, it also presents a timeline of significant works that depicts relationships between flight delay prediction problems and research trends to address them.

3. Analysis of Airport Data using Hadoop-Hive: A Case Study

Authors: S. K. Pushpa, Manjunath T. N., Srividhya.

Year: 2016

- It is increasingly difficult to handle huge amount of data that gets generated no matter what's the business is like, range of fields from social media to finance, flight data, environment and health.
- Big Data is also used to monitor things as diverse as wave movements, flight data, traffic data, financial transactions, health and crime. The challenge of Big Data is how to use it to create something that is value to the user.
- In this paper Big Data is depicted in a form of case study for Airline data based on hive tools.
- Here data is stored in HDFS and processed using hive tools.
- This paper proposes a method to analyze few aspects which are related to airline data such as

a) list of airports operating in the country India,

b) list of airlines having zero stops

c) list of airlines operating with code share

d) list highest airports in each country

e) list of active airlines in United State

4. Analysis of Operational Flight Data in Hadoop using MapReduce and the MATLAB Distributed Computing Server (MDCS)

Authors: Lukas Höhndorf, Javensius Sembiring, Robin Karpstein, and Florian Holzapfel.

Year: 2016

- It focuses on the various accidents occurred in the past to make quantitative statements about the future state based on previous experience and knowledge.

General concept: To quantify accidental probabilities in aviation, there are a lot of computations and interactions with huge volume and uses MATLAB distributed computing server to perform operatons.

- MDCS allows to run computationally intense MATLAB programs and Simulink models on clusters , clouds, and grids enabling to speed up computations and solve larger problems.
- It focuses on detecting departure and arrival airports and runways using mapreduce and the MCDS.

5. Big Data Analysis by Classification Algorithm Using Flight Data Set

Author: Ujjwala Urkude
Year: 2016

- In this work, a technique is proposed with broad application which is used to classify each item in a set of data into a set of predefined classes or groups
- It is a tedious works for user to identify accurate data from huge unstructured data.
- Several major kinds of classification algorithm C4.5 , Decision Tree ,J48, ID3, includes Naive Baye's algorithm This paper provides Flight dataset related queries.
- System is capable of predicting the number of aircraft in certain region of the airspace at a given time with greater accuracy than similar Model.
- The Naive Baye's Classifier on the data set on different size for different cluster configuration provides the potential data as well as aspects that affect its performance.

6. Big Data Analysis of Airline Data Set using Hive

Authors: Nillohit Bhattacharya, Jongwook Woo
Year: 2015

- 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.
- 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.
- Here 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.
- 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 .
- Once the HD Insight cluster is up and running, Hive queries can be executed on the cluster from the Hive query console window.

7. Big Data Infrastructure for Aviation Data Analytics

Authors:- Anandavel Murugan Chandramohan , Dinkar Mylaraswamy , Brian Xu , Paul Dietrich.
Published in: Cloud Computing in Emerging Markets (CCEM), 2014 IEEE International Conference.
Date of Conference: 15-17 Oct. 2014
Conference Location: Bangalore, India

- This paper describes the approach towards developing and using a big data infrastructure for analyzing aviation data.
- In this paper, we briefly introduce our data sources, nature of data collected, cluster
- Design, data loading and storage strategy and library of choice for analytics and visualization.
- It represents some of the analytics we have implemented for health monitoring of auxiliary
- Power units (APUs) using our big data infrastructure. Best practices to implement big data analytics and pitfalls are discussed and substantiated with our experiences.

8. Predictive analytics with aviation big data.

Authors:- Samet Ayhan , Johnathan Pesce , Paul Comitz , David Sweet, Steve Bliesner, Gary Gerberick.
Published in: Integrated Communications, Navigation and Surveillance Conference (ICNS), 2013
Date of Conference: 22-25 April 2013 Conference Location: Herndon, VA, USA.
Conference Location: Herndon, VA, USA.

- This paper describes a novel analytics system that enables query processing and
- Predictive analytics over streams of big aviation data. As part of an Internal Research and Development project, Boeing Research and Technology (BR&T) Advanced Air Traffic Management (AATM) built a system that makes predictions based upon descriptive patterns of massive aviation data. Boeing AATM has been receiving live Aircraft Situation Display to Industry (ASDI) data and archiving it for over two years

- At the present time, there is not an easy mechanism to perform analytics on the data. The incoming ASDI data is large, compressed, and requires correlation with other flight data before it can be analyzed. The service exposes this data once it has been uncompressed, correlated, and stored in a data warehouse for further analysis using a variety of descriptive, predictive, and possibly prescriptive analytics tools.
- The service is being built partially in response to requests from Boeing Commercial
- Aviation (BCA) for analysis of capacity and flow in the US National Airspace System (NAS).
- The service utilizes a custom tool developed by Embry Riddle Aeronautical University.
- (ERAU) that correlates the raw ASDI feed, IBM Warehouse with DB2 for data management, WebSphere Message Broker for real-time message brokering, SPSS Modeler for statistical analysis, and Cognos BI for front-end business intelligence (BI) visualization tools. This paper describes a scalable service architecture, implementation and value it adds to the aviation domain.

9. Data Mining and Data Warehousing in the Airline Industry

Published in : Academy of Business Research Journal, Vol. III

Month of conference : September 2013

- Organizations are constantly looking to enhance their decision-making activities in order to improve business processes and build a competitive advantage.
- Data mining, which is the automated extraction of predictive information from large databases, helps connect large volumes of this heterogeneous data and allows organizations to analyze it from multiple perspectives.
- Designed for query and analysis rather than transaction processing, a data warehouse is a relational database that centralizes data coming from multiple sources. It translates the information into common models, names, and definitions while also providing a mean to make information available for decision making. Although data mining and data warehousing are powerful tools for organizations they can present several challenges.
- Studying the successes and failures of the industry to conduct data mining and data warehousing activities as airlines struggle in an increasingly competitive environment can be beneficial to other economic sectors as well

10. ANALYSIS OF AIRCRAFT ARRIVAL DELAY AND AIRPORT ON-TIME PERFORMANCE

Author: YUQIONG BAI

Year:2001

- In this research, statistical models of airport delay and single flight arrival delay were developed. The models use the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the Surface Airways Weather Data from the National Climatic Data Center (NCDC).
- Multivariate regression, ANOVA, neural networks and logistic regression were used to detect the pattern of airport delay, aircraft arrival delay and schedule performance.
- These models are then integrated in the form of a system for aircraft delay analysis and airport delay assessment. The assessment of an airport's schedule performance is discussed.
- The results of the research show that the daily average arrival delay at Orlando International Airport (MCO) is highly related to the departure delay at other airports.
- This research also investigated the delays at the flight level, including the flights with delay ≥0 minute and the flights with delay ≥15min, which provide the delay pattern of single arrival flights.

11. Characterization Of Delay Propagation In The US Air Transportation Network.

Authors: Pablo Fleurquin, José J.Ramasco, Víctor M. Eguíluz

- In this work, the focus is on the properties of flight delays in the US air transportation network.
- Flight performance data in 2010 is analyzed and the topological structure of the network as well as the aircraft rotation is studied. The properties of flight delays, including the distribution of total delays, the dependence on the day of the week and the hour-by-hour evolution within each day, are characterized paying special attention to flights accumulating delays longer than 12 hours.
- It was discovered that the distributions are robust to changes in takeoff or landing operations, different moments of the year or even different airports in the contiguous states. However, airports in remote areas (Hawaii, Alaska, Puerto Rico)
- Can show peculiar distributions biased toward long delays.

12. How data can help reduce aviation accidents.

Authors: Carmen Serpa, Sunilkumar Kakade.

- This project is based on the data from the National Transportation Safety Board (NTSB). It involves classifying a set of aircraft accident/incident data covering the years 2000 to 2015 in United States.

- The NTSB provides one of the most comprehensive online aircraft accident and incident databases. It includes dates, places, aircraft and engine types, scheduled and non-scheduled certificated air carrier, and name of air carrier.
- The purpose of this study is to optimize analyses of the NTSB massive data with Hadoop and uncover hidden patterns that can give some ideas of how we can reduce the rate of aviation accidents and thus decrease risk and improve safety
- The focus is to see what we can learn from the past in order to identify appropriate measures to prevent repetitive errors in the future. Hadoop is used to extend the ability of human pattern recognition to uncover accidents' causes in multivariate data.

IV. FINDINGS

- The project focuses on finding patterns in ticket booking using data from the past in order to improve sales of tickets and to provide seasonal sales to improve businesss.
- The project also takes into consideration finding primary reasons for flight cancellations taking place through the past data available.
- The project aims to gather useful information about sales of tickets in bad weather conditions and its effect on the business.
- It tries to find out the reasons for delays occurring in the flight routine and if possible provide suggestions to improve the service for better customer experience.
- The project also tries to gather information about the regular customers using the service so as to to provide them with interesting schemes which in turn will attract more customers thus benefiting the business.

V. REFERENCES

[1]. M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak. Detecting periodic patterns of arrival delay. Journal of Air Transport Management, Nov. 2007.

[2]. YUQIONG BAI . ANALYSIS OF AIRCRAFT ARRIVAL DELAY AND AIRPORT ON-TIME PERFORMANCE,2001.

[3]. ANAC. Agencia Nacional de Aviacao Civil. Technical report, http://www.anac.gov.br/,2017.

[4]. E. Balaban, I. Roychoudhury, L. Spirkovska, S. Sankararaman, C. Kulkarni, and T. Arnon. Dynamic routing of aircraft in the presence of adverse weather using a POMDP framework. In 17th AIAA Aviation Technology, Integration, and Operations Conference, 2017.

[5]. H. Balakrishnan. Control and optimization algorithms for air transportation systems. Annual Reviews in Control, 2016.

[6]. S. B. Boswell and J. E. Evans. Analysis of downstream impacts of air tra_c delay. Lincoln Laboratory, Massachusetts Institute of Technology, 1997

[7]. J. Cox and M. Kochenderfer. Ground delay program planning using markov decision processes. Journal of Aerospace Information Systems, 2016.

[8]. T. Diana. Validating delay constructs: An application of confirmatory factor analysis. Journal of Air Transport Management, Mar.2014.

[9]. EUROCONTROL. CODA Digest - Delays to Air Transport in Europe. Technical report, https://www.eurocontrol.int/articles/coda-publications, 2017.

[10]. EUROCONTROL. European Organisation for the Safety of Air Navigation. Technical report, https://www.eurocontrol.int/, 2017.

[11]. FAA. Federal Aviation Administration. Technical report, http://www.faa.gov/, 2017.

[12]. M. Hansen. Micro-level analysis of airport delay externalities using deterministic queuing models: a case study. Journal of Air Transport Management, Mar. 2002.

[13]. L. Ionescu, C. Gwiggner, and N. Kliewer. Data Analysis of Delays in Airline Networks. Business and Information Systems Engineering, 2016.

[14]. C.-Y. Hsiao and M. Hansen. Air transportation network flows: Equilibrium model. Transportation Research Record, 2005.

[15]. T. Kotegawa, D. De Laurentis, K. Noonan, and J. Post. Impact of commercial airline network evolution on the U.S. air transportation system. In Proceedings of the 9th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2011.

[16]. T. Krsti´c Simi´c and O. Babi´c. Airport traffic complexity and environment efficiency metrics for evaluation of ATM measures. Journal of Air Transport Management, Jan. 2015.

[17]. S. A. Morrison and C. Winston. The effect of FAA expenditures on air travel delays. Journal of Urban Economics, Mar. 2008.

[18]. E. R. Mueller and G. B. Chatterji. Analysis of aircraft arrival and departure delay characteristics. In AIAA aircraft technology, integration and operations (ATIO) conference, 2002.

[19]. V. Pai. On the factors that a_ect airline flight frequency and aircraft size. Journal of Air Transport Managemen, July 2010.

[20]. T. Pejovic, R. B. Noland, V.Williams, and R. Toumi. A tentative analysis of the impacts of an airport closure. Journal of Air Transport Management, Sept. 2009.