Analysis of Airport Authority Using Hadoop

Akshay Phadatare, Tushar Gadage, Vinod Pawar, Amit Raut,

Roshankumar Bauskar

Abstract—

Nowadays an airport has a huge amount of data like number of flights, time of arrival and dispatch, flight routes, number of airports operating in each country, list of active airlines in each country etc. So far, for storing and analyzing the database of airport DBMS is used. SQL is used for applying the queries related to all the entries of an airport database management system. My SQL uses B-tree or distributed hash tables for analyzing the database. It is not capable of analyzing a huge amount of data. It uses a parallel database for storing data. The drawback of this system is that it is not capable of the huge amount of data increase to a very large extent. So for overcoming this problem, Hadoop database file system is used. Hadoop stores the huge amount of data using HDFS and Map-reduce the computational model for providing the data in parallel. We attempted to explore detailed analysis on airline data sets such as listing airports operating in India, list of airlines having zero stops, list of airlines operating with codeshare which country has highest airports and list of active airlines in united state. Here we focused on the processing the big data sets using hive component of Hadoop ecosystem in a distributed environment. This work will benefit the developers and business analysts in accessing and processing their user queries.

General Terms

Big data, Hive Tools, Data Analytics, Hadoop, Distributed File System

Keywords

Airline data set, Hive Tools.

I. INTRODUCTION

The main motivations behind Hadoop are Big Data, the problems which exists with the Traditional Large Scale Computing system, and what requirements and Alternative Approach should have. Big Data is a relative term. If Big

Data is referred by volume of transactions and transaction history, and then hundreds of terabytes (10^{12} bytes) may be Big Data. Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two year along. This data comes from everywhere. The magnitude of Big Data is colossal but Big

Data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business agiler, and to answer questions that were previously considered beyond your reach. Until now, there was no practical way to harvest this opportunity. Today Hadoop a platform for big data is the state of art technology to open the door to a world of possibilities. The existing systems are mostly based on a traditional database which is difficult to process such a huge data. Mostly this system processed only structured data. By using Big data technology we can be structured, semi-structured and unstructured data.

II. RELATED WORK

As far as data storage model considered by B-trees or distributed hash tables using key-value pair is too limited to handle large datasets. Many projects have attempted to provide solutions for distributed storage at higher-level services over wide area networks, often at Internet scale. This incorporates take a shot at

disseminated hash tables that started with ventures, for example, CAN, Chord, Tapestry, and Pastry. These frameworks address worries that don't emerge from a Big table, for example, profoundly variable data transfer capacity, untrusted members, decentralized control and Byzantine adaptation to internal failure are not Big table objectives.

Several database developers have created parallel databases that can store huge volumes of information. Oracle's Real Application Cluster database utilizes shared disks to store information (Big table uses GFS) and an appropriated lock director (Big table uses Chubby). IBM's DB2 Parallel Edition depends on a shared-nothing design like a big table. Each DB2 server is accountable for a subset of the columns in a table which it stores in a relational database. Both databases afford a complete relational model with transactions. The limitation is that it is not scalable for a huge amount of data as data increases to a very larger extent. Hence apache hive supports for a huge amount of data

In this paper, Apache Hive is considered for analyzing large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 file system. It provides a SQL-like language called HiveQL with the schema for reading and transparently converts queries to Map Reduce, Apache Tez, and Spark jobs. All three execution engines can run on Hadoop YARN. To accelerate queries, it provides indexes, including bitmap indexes.

III. CHALLENGES IN BIG DATA

The uses of Big Data in various fields of knowledge are immense in the sense its potentiality of micro and macro levels of analysis of the data. For instance, the tools in Big Data help the Institutions to study the quantitative and qualitative learning abilities of students from different strata of the society. Even the behavioral learning and the psychological attitudes of the student may also be estimated through the tools of Big Data. Big Data can also be used in analyzing the cognitive abilities and the impact of health on acquiring the knowledge of health condition of the students usually affects the learning process.

Further, the scope of big data is so vast that it has been used in globalized urban societies in planning the locality, intelligence transportation, air ambulance monitoring system, road mapping, environment and natural disaster prediction.

Big Data is supported by the range of technologies such as Hadoop [4]. Traditional relational database skill is still in high demand but increasingly, so are the skills needed to work with the generation of non-relational databases known as NoSQL. These databases which are often open source are built to handle the processing of large volumes of data and use different design strategies, architectures and query languages. One of the biggest challenges in Big Data is Big Data analytics, where analyze examining and interpret Big Data.

In this paper first tables were created for the below mentioned Data Set [6]. The Dataset was loaded into the created tables on an HDFS system. The Hive queries were applied and the results were analyzed.

IV. PROPOSED SYSTEM

As our project is based Airlines Data Analysis using Hadoop which shows the sorting of a large amount of airlines datasets to get outputs

- Predicting flight delay
- Busiest Routes
- Month delay of flights.
- Yearly delays of flights

Cloudera open-source Apache Hadoop distribution, CDH (Cloudera Distribution Including Apache Hadoop), targets enterprise-class deployments of that technology. Cloudera says that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects (Apache Hive, Apache Avro, Apache HBase, and so on) that combine to form the Hadoop platform. Cloudera is also a sponsor of the Apache Software Foundation.

Flight/Airline Data Analysis Program (FDAP):

A pro-active non-punitive program for gathering and analyzing data recorded during routine flights to improve flight crew performance, operating procedures, flight training, air traffic control procedures, air navigation services, or aircraft maintenance and design.

This Civil Aviation Advisory Publication (CAAP) was written to provide background information and guidance material for any operator of an aircraft that intends to develop and establish a Flight Data Analysis Program (FDAP) and for Civil Aviation Safety Authority (CASA) in the assessment of those programs

IATA Forecasts that by 2017 total passenger numbers are expected to rise to 3.91 billion an increase of 930 million passengers over the 2.98 billion carried in 2012.

All these Passengers generate a large number of data such as reservation, consumer, financial, geolocation, social information etc.

By capturing and processing these data customer trends, preferences, identification, profiling, segmentation is possible.

Aircraft generate 2 terabytes of data per flight from their sensors.

Real or near real-time decision support processing requires capturing data sources generating data at high speeds.

Sources such as Video Analytics and CCTV data streams require the capturing and processing of fast data streams to identify passenger patterns in real time such as security screening queue management, passenger emergencies, security incidents.

Hadoop Distributed File System:

The **Hadoop Distributed File System** (**HDFS**) is designed to store very large data sets reliably and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks.

Map Reduce:

To take the advantage of parallel processing of Hadoop, the query must be in Map Reduce form. The Map-Reduce is a paradigm which has two phases, the mapper phase, and the reducer phase. In the Mapper, the input is given in the form of key-value pair. The output of the mapper is fed to the reducer as input. The reducer runs only after the mapper is over. The reducer too takes input in key-value format and the output of reducer is final output.

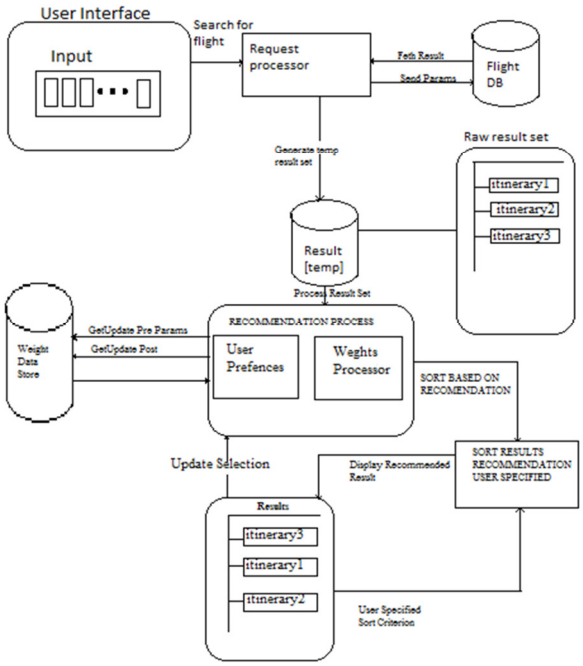
Map Reduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduced and scatter operations.

V. METHODOLOGY

In this paper, the tools used for the proposed method is Hadoop and Hive which is mainly used for structured data. Assuming all the Hadoop tools have been installed and having semi-structured information on airport data. The methodology used is as follows:

- Our project is focused on the extraction of data and analysis of Airlines data.
- Predicting flight delay
- Busiest Routes
- Month delay of flights.
- Yearly delays of flights
- Find the list of active airlines in the country
- The big issue to manage or to handle such an anonymous or huge data.

Fig: Architecture Dig.



VI. RESULTS AND DISCUSSION

This paper emphasizes on data analysis on airline data set. The paper address the usage of modern analytical tool Hive on Big Dataset which focuses on common requirements of an airport. It shows the create table and load data commands for HDFS system. It also gives the number of Map and Reduces that are internally taken care of the underlying tools of Hadoop System. It is found that Hive is effective in terms of processing huge datasets when compared to traditional databases with respect to time and data volume.

VII. CONCLUSION

This paper addresses the related work of distributed databases that were found in literature, challenges ahead with big data, and a case study on airline data analysis using Hive. The author attempted to explore detailed analysis on airline data sets such as listing airports operating in India, list of airlines having zero stops, list of airlines operating with codeshare which country has highest airports and list of active airlines in united state. Here author focused on the processing the big data sets using hive component of Hadoop ecosystem in the distributed environment. This work will benefit the developers and business analysts in accessing and processing their user queries.

REFERENCES

[1] S. K. Pushpa VTU, Manjunath T. N. VTU, Srividhya VTU, Bengaluru Dept. of ISE, BMSIT & amp. 2016: Analysis of Airport Data using Hadoop-Hive.

[2] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. May 2011 Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[3] Manjunath T N et.al, Automated Data Validation for Data Migration Security, International Journal of Computer Applications (0975 – 8887), Volume 30– No.6, September 2011.

[4] Ratnasamy S., Francis P., Handley M., Karp R., and Shenker S. A scalable content-addressable network. In Proc. of SIGCOMM (Aug. 2001)

[5] Rostrum A, and Druschel P.Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In Proc. of Middleware 2001 (Nov.2001)

[6] Zhao B. Y., Kubiatowicz J., and Joseph A. D. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Tech. Rep. UCB/CSD-01- 1141, CS Division, UC Berkeley, Apr. 2001.