Sentimental Analysis on the performance of domestic flights in an Airline Company

[1]Spardha Jhudeley, [2]Shivani Sharma, [3]Vyanjana Kishaniya
[4]Dr.Sachin Goyal, [5]Dr. Ratish Agrawal
Department of Information Technology,UIT RGPV Bhopal

ABSTRACT

The amount of data produced by mankind is growing rapidly every year. Analyzing this data and extracting information using is very important for any organization to use it to identify new opportunities. Sentiments essentially relates to feelings, attitude, emotions and opinions. Sentimental analysis is in one of the most popular research areas.it refers to identify and extract subjective information from a piece of text, to summarize overall sentiment about a particular product is a crucial task.

In this paper, Airline Company's data is analysed and the performance of domestic flights is checked. Summary concludes on the number of on-time, delayed, cancelled, and diverted flights in the report. The problem statement is to find out the most visited places by using airline services in a particular country. Sentimental analysis has been done by using Hadoop and Pig.

INTRODUCTION

The amount of data produced by mankind is growing rapidly every year. Rate at which data is increasing is so large that an entire cricket stadium can be filled. The rate is growing so enormously that analysing this data is a very complicated task but eventually helps in conclusions and providing the hypothesis .The conclusion of this analysis will be the overall review for an airline company. Later, these reviews can be used in decision making. The software used are Big data, Hadoop and Pig.

Big data[1] refers to data that is so huge and complicated that it becomes burdensome to process using conventional data processing software. Capture, storage, search, sharing, transfer, analysis and visualization are some of the major tasks in big data.[9]

Map reduce is an applications that helps to analyze large amount of data in concurrently and on huge blocks. Map-reduce includes two activities, mapping of data and reducing data.

CHALLENGES OF BIG DATA

- Heterogeneity
- Incompleteness
- Scale
- Data aggregation
- Privacy
- Human collaboration

- Timeliness

Heterogeneity

Big Data deals with large amount of data on a wider scale. The data on which analysis has to be performed may or may not be similar.

Incompleteness

Sometimes adequate information has not been provided. The data set remains incomplete and running of queries becomes a big task.

Example: In universities when analysis is done, the data records are designed such as they have fields for birth date, guardian/parent name, blood type, pursued course for each student. Sometimes because of the lack of the information the student is still placed in the data set but with the attribute null in the data field. Such errors are unavoidable and need to be tackled.

Scale

Since it deals with large amount of data of different sizes and types , scaling of data has to done for proper utilization of the resources. Data is growing rapidly and managing such a vast amount of data is not easy. Thus scaling of resources is required, so that without much manual interruption data can be managed automatically.

Data aggregation

Designing of data set and arranging them properly under different data fields under one big umbrella is a laborious task.

Privacy

The privacy[9] of individual might get offended as it requires large organizations to collect data sometimes with consent and sometime without the consent of the people.

Example: Student data at universities or patients' Dara in the hospital gets automatically dumped but the institutions at the time of their admission without taking the consent of the individual in particular.

Human collaboration

It requires human collaboration such as

- Collection of data
- Designing of data fields
- Designing of queries and running of queries

Timelines

The process of collection of data, data analysis[6] is time and power consuming. Human nature can consume heterogeneity but machines are not designed as such. So various commands and queries have to be designed to solve this, increasing the mental and physical work load. Thus consumig time and power.

BENEFITS

- Analyses of large amount of data can be done easily and efficiently
- Heterogeneous and homogeneous data can be analyzed
- Proper use of resources
- No wastage of resources as it allows proper distribution of resources according to the requirement.
- Helps a large organizations to develop strategies according to the current market needs and trends as the dumped data can be compared
- Avoids data replication and redundancy
- Decreases manual involvement
- Allows access to large amount of data at one single place without much browsing.

APACHE HADOOP

Hadoop[11] is an open source software framework for distributed storage and distributed data which is processed on large data sets. Doug Cutting developed Hadoop as a collection of open source project which provides a programming environment that is applied in a distributed system. Hadoop is composed of HBase, Pig, Hive, Zookeeper, HDFS and Map reduce.[4]. Hadoop[8] uses the HDFS file system, the function of HDFS is to divide, replicate and store the data sets in multiple nodes [3].

Hadoop[10] provides data parallelism and data and task replication schemes enable fault tolerance, but what is often criticized about it is the time required to load data into HDFS and the lack of reuse of data produced by mappers.

Advantages of Hadoop includes data parallelism and fault tolerance. Major disadvantage of Hadoop is the time required to load data into HDFS and also, lacking in reusing of data produced by mapping[7].

HDFS

Hadoop distributed file system[12] provides better data throughput and easy retrieval of data, high fault tolerance. HDFS stores files in blocks, the default block size is 64mb. In HDFS, blocks are made up by splitting data files and is replicated and stored into multiple nodes across the network.

HDFS[2] is required when the amount of data is too much for a single machine to handle. It is too complicated that other file systems cannot handle the complexities and uncertainty of networks. It contains following: Name node, secondary name node, data node, job tracker and task tracker. HDFS is a file system written in JAVA. It provides redundancy in storage for enormous amount of data

PIG

Pig is a scripting language and allows analyzers using apache Hadoop to concentrate more on analyzing data rather than to spend time in writing map -reduce algorithms.

Pig is one of the component of Hadoop system, initially developed at Yahoo! Pig is an apache open source project, perform for every kind of data. It works on HDFS cluster.[5] Storing and fetching of data is done by HDFS . The language for this platform is called pig Latin. Pig can execute its Hadoop jobs in Map reduce, Apache Tez or Apache spark.it supports various basic data types and also advanced data types like atom, tuple, bag and map.

Features

- Schemas are optional
- Pig is a dataflow language rather than declarative
- Defining data pipeline is easy
- It has Step by step query style
- Supports UDF's (user define function)
- Supports various datatype

PROBLEM STATEMENT

This analysis will help in analyzing the Airline data using Apache Pig. We have two datasets, i.e., late_Flights.csv and Airstation.csv.

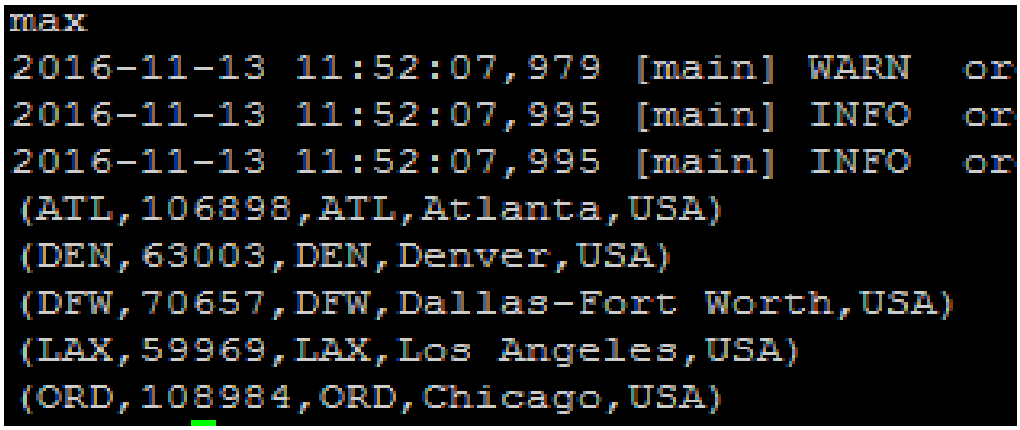
The problem statement is to find out the top 5 visited places by using airline services in a particular country

STEPS OF WORKING

- We need to register a jar file. The name of the file is *piggybank .it is* to make use of class CSVExcelStorage .
- Load the dataset in relation A using CSVExcelStorage
- In order to process the query further, columns are made in relation B.
- Filter "null" values from "dest" columns in relation C.
- Group C by "dest" in relation D.
- Count each column and group them.
- Result is deduced and top 5 destination are filtered.

  By above commands we were able to deduce top five destinations .by following commands the city and country will also be displayed as an output.
- Another table will be created which will include city and country name by using relation A1.
- Destination,city as well as country will be displayed in relation A2.
- The result and "dest" are joined by using joined table relation
- Finally the result is displayed by dumping.

OUTPUT



CONCLUSION

In this paper we have analyze the dumped data of airlines thereby giving us top 5 destination ,cities and countries which are most visited by passengers. The project has helped us to analyze larger data sets on a wider scale which was impossible to do manually, thus increasing the performance and efficiency.

REFERENCE

1. Bo Li, Survey of recent research progress and issues in big data. 20th December 2013.

2. Bakshi, Kapil "considerations for big datta. architecture and approach" Aerospace Conference, 2012 IEEE.

3. Zaslavsky, Arkady, Charith Perera, Dimitrios Geograkopoulus. "Sensing as a service and big data."

4. Mohammed Al-Zobbi, Seyed Shahrestani and Chun Ruan, Improving Map reduce privacy

5. Dokerglu T, Ozal S,Bayair MA,Cinar MS, Cosar A,. Improving the performance of Hadoop Hive by sharing scan and computation tasks. J Clooud Comptu.2014;3(1):1

6. Govindraju V. Big data analytics. Oxford: Elseveir Science;2015

7. Big data: survey, technologies, opportunities and challenges. 17th July 2014. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldein Mahmoud Ali, Muhammad Alam, Muhammad Shiraz,Abdullah Gani

8. Professional hadoop by Benoy Antony and Cazen Lee, Wiley publication. 2016

9. Big data computing and clouds: trends and future directions. 18th August 2014. Mrcos D. Assuncao, Rodrigo N. Calherios, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya

10. J.Polonsteky and O. Tene, “Privacy and big data: making ends meet”, Stanford Law Review online, vol66,25,2013

11. Hadoop in action by Chuck Lam, Dreamtech publication. 2016

12. S.Madden , “from databases to big data”,IEEE Internet Computing vol.16,no.3,pp.4,-6,2012