Naive baye's classification algorithm in prediction of Flight delays using MR

Ujjwala Urkude[1] , Pratibha Richariya[2]
[1]*M. Tech Scholar, CSE, Maxim institute of technology Affiliated to RGPV Bhopal*
[2]*Asst. Prof. CSE, Maxim institute of technology Affiliated to RGPV Bhopal*

***Abstract*- In proposed technique which is used Naive baye's Algorithm a set of data into a set of predefined classes or groups . This paper provides Flight dataset related queries. System is learning that 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.Big Data concept comes into existence .This is a tedious works for user to identify accurate data from huge unstructured data .As the years go on the web is overloaded with lots of information. While predictive schemes also use historical data, historical analytics bases assumptions and conclusions on carriers past performance with existing business, in existing markets. Predictive analytics, on the other hand, relies on insights about a group or market as a whole in order to help a carrier deliver the best possible experience.**

***Index Terms*- Big Data, Map Reduce, Hadoop, Classification Technique, Naive Baye's Algorithm, Flight Dataset.**

I. INTRODUCTION

In this project we analyse the flight big data collected from different airline companies and compute various parameters that affect the performance of flight journey. We compute the delay caused between various routes and the causes of delay in various flight journeys.

The main aim of our project is to compute the various types of delays associated with flight journeys. The Apache Hadoop Framework is a platform which helps us in extracting features from big data and divides the big data into blocks which are executed in parallel in clusters. The information about the clusters is saved in Job Configuration.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

Today, the problem is very different. Data sources are unpredictable, multi-structured (emanating from organized systems) and massive. Many are external to the enterprise. The techniques for mining data from these sources, and even the platforms most appropriate for doing so, are now somewhat in question. With the entry of Hadoop to the market – an approach that is separate from the relational data warehouse – the issue facing decision makers today is where and when to deploy these technologies for performing useful analytics. But first, a little clarity on what analytics means.

II. HADOOP FRAMWORK

we have used Apache Hadoop Platform and Map Reduce Framework to efficiently design algorithms that analyses the big data and compute various queries that affect the performance of flight journey.

Hadoop consolidates large volumes of information, stores it efficiently, processes it powerfully and does all this inexpensively. This makes Hadoop an ideal tool for many applications, and data scientists have tapped this potential by illustrating how Hadoop can be used to predict which airline flights will be delayed or cancelled. Obviously, this is useful information for consumer apps, not to

mention travel agencies and other stakeholders. In a nutshell, data scientists compiled data on flights from 1987-2008[7], including the airport of origin, the destination airport, and 26 other variables to build a learning model that accurately predicts flight delays using historical flight data and weather information.

2.1 Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is now an Apache Hadoop subproject.

2.2 Hadoop MapReduce

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks.

The MapReduce framework consists of a single master Job Tracker and one slave Task Tracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.

III. NAIVE BAYE'S ALGORITHM

In simple terms, a naive baye's classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive baye's classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features. For some types of probability models, naive baye's classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive baye's models uses the method of maximum likelihood; in other words, one can work with the naive baye's model without accepting Bayesian probability or using any Bayesian methods. An advantage of naive bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. The Naive Baye's classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms [6].

IV. EXPERIMENT AND RESULTS

The data from the project may be used to build computer modeling software that could predict the outcome of an infinite number of hypothetical flight, helping airlines mark weather delays in advance. Therefore, this knowledge could enable airlines to adjust their schedules to account for weather patterns. It may also lead to new options for passengers. For instance, airlines can take several steps ahead to predict a future flight delay and then offer passengers a pre-emptive rebooking to avoid it.

An experimental analysis indicated that the speedup achieved by the tool increases with the amount of data being processed. Additionally, the results showed that increasing the number of nodes in the cluster does not necessarily provide a corresponding reduction of execution times. Thus, the proper cluster configuration depends not only on the operations to be executed but also on the amount of input data; there must be a balance between the amount of data to be processed and the number of nodes to be used to achieve the best performance [1].

4.1 Flight Dataset

Flight links related to flights using the same aircraft and/or crew and/or passengers are the skeleton through which delays are propagated. Following the tree of reactionary delays allows studying the impact of different local strategies into the delays propagation through the network. The modelling approach in TREE is an agent-based approach, with aircraft as basic units, and includes mechanisms for simulating slot reallocation and slot swapping strategies as alternatives prior to flight cancellation [2].

The main aim of this study is to enable airlines to anticipate and deal with delays before they happen. As we all know, most airlines compensate for delays by adding slack to the system. They usually re-schedule the flight time during the winter on the same day or keep additional staff members on call. It is a fact that, all these days, they used to ignore the large-scale weather patterns when they were building flight schedules, thinking that, not much can be done with emerging weather patterns which are limited due to the ire of mother nature.

The research wants to alert the airlines with weather delays before they could happen. It's more like a proactive cause, for doing a much better job of communicating with passengers and optimizing resources. While predictive schemes also use historical data, historical analytics bases assumptions and conclusions on carriers past performance with existing business, in existing markets.

4.2 Experimented Results

Execute the Flight Data set we are using Ubuntu operating system, Shown in below

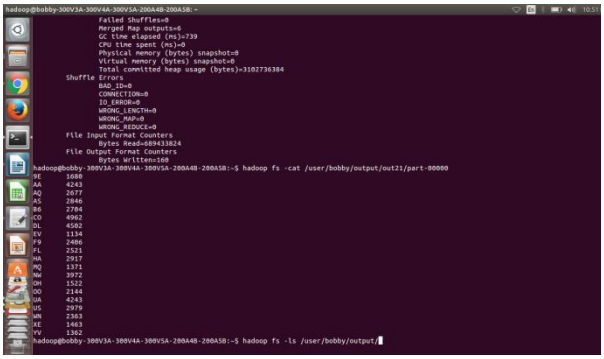

Fig 4.1 show the output in ubuntu

1 Average Delay:

Classification technique provides prediction according to dataset.The average delay per Airline Company by number of arrivals, where the actual arrival counts. The amount of data set depends on number of node, if number of node increase the execution time will be decrease.

Table 4.1 : Quality parameter of Average delay

Data Set	Execution Time (second)		
	1 Node	4 Node	7 Node
1 GB	360	170	136
1.2 GB	386	182	147
1.4 GB	403	193	158

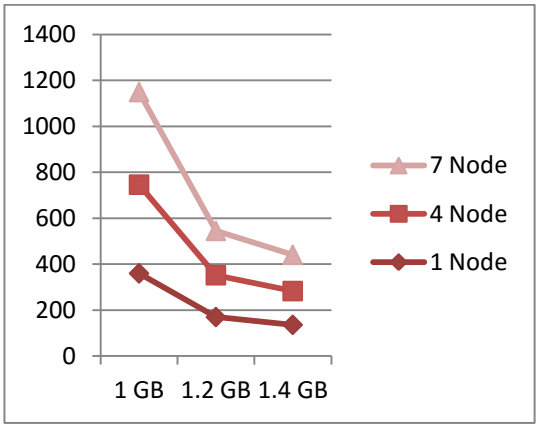

Fig 4.2: The Average Delay per Airline company

2. Maximum Delay:

The amount of data set depends on total arrival delay and total departure delay. The Maximum delay per airline company calculates by total number of delay , where the actual delay counts

Table 4.2:Quality parameter of Maximum Delay

Data Set	Execution Time (second)		
	1 Node	4 Node	7 Node
1 GB	300	151	123
1.2 GB	323	173	143
1.4 GB	341	186	153

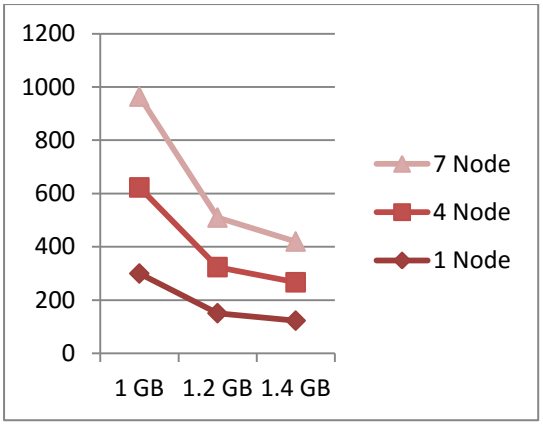

Fig 4.3: The Maximum Delays per Airline company

3. Average Delay Source to Destination

Flight dataset where set key map as source to destination then set the total delay after that calling total delay of each airline company .

Table4.3: Quality Parameter of Average Delay per source to Destination

Data Set	Execution Time (second)		
	1 Node	4 Node	7 Node
1 GB	361	172	135
1.2 GB	386	180	148
1.4 GB	401	191	157

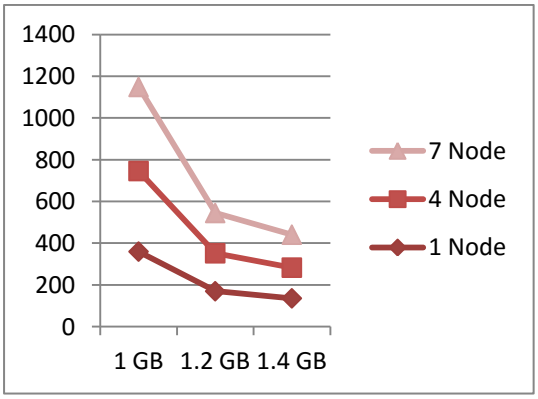

Fig4.4: Average Delay per source to Destination

4 Maximum Delay Source to Destination

The maximum delay calculates from source to destination by the flight dataset using arrival delay and departure delay.

Table4.4: Quality Parameter

Data set	Execution Time (second)		
	1 Node	4 Node	7 Node
1 GB	420	228	195
1.2GB	446	235	201
1.4GB	463	253	219

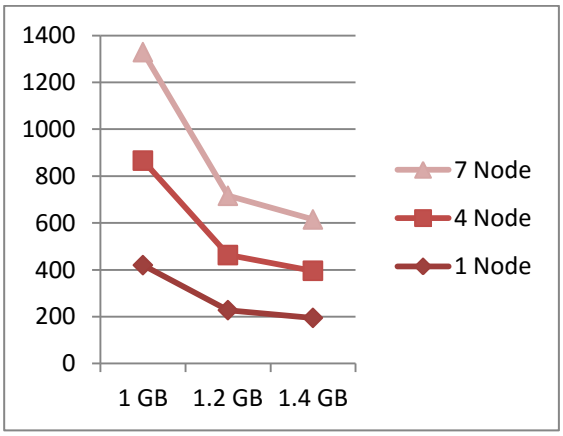

Fig4.5: Maximum Delay per source to Destination

V. CONCLUSIONS

Big data which provides high velocity, high variety, high volume of data . We are applying big data classification in flight dataset which provide high dimensions of decision making process. Big data classification using Naïve Baye's Algorithm is an appropriate tool of classification algorithm. Developing technique provides flight prediction according to weather.

The Flight data execution time, proposed a novel and secure method Based on Classification. Flight Dataset use for processing the dataset After that data will be stores ,the training data hides the additional data into the Naive Baye's After this process jar file send after provide the optimized output

REFERENCE

[1] V.A.Ayma , R.S. Ferreira ,P. Happ, D.oliveira , G. Costa , A.Plaza , P . Gamba, " Classification Algorithm For Big Data Analysis ,A MapReduce Approach" joint ISPRS conference 2015 .

[2] C. Ciruelos , A . Arranz , I . Etxebarria , S . Peces , B. Campanelli , P . Fleurquin , V.M. Eguiluz , J.J.Ramasco, " Modelling Delay Propagation Trees for Scheduled Flights" 2015.

[3] Bharti Thakur, Manish Mann , "Data Mining for Big Data: A Review" , May 2014.

[4] SONGTAO ZHENG, "NAÏVE BAYES CLASSIFIER: A MAPREDUCE APPROACH" , October 2014.

[5] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane, Rahul Pitale , "COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES USING DIFFERENT DATASETS" May 2013

[6] Tina R. Patil, Mrs. S. S. Sherekar , "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification" , Vol. 6, No.2, Apr 2013 ISSN: 0974-1011 (Open Access).

[7] Banavar Sridhar, Yao Wang, Alexander Klein, Richard Jehlen, "Modeling Flight Delays and Cancellations at the National, Regional and Airport Levels in the United States" , 2009.

[8] Praful Koturwar , Sheetal Girase , Debajyoti Mukhopadhyay " A survey of Classification Techniques in the Area of Big Data" .

[9] Balaji Lakshminarayanan , Denial M . Roy , Yee Whye Teh , " Mondrian Forests for Large Scale Regression when Uncertainty Matters" 15 October 2015 .

[10] Devapriya Singaravelu " Airline Analytics" Management Progress Report 1- 3, 2013-2014.

[11] Juan Jose Rebollo and Hamsa Balakrishnan "Characterization and Prediction of Air Traffic Delays" , March 2014.

[12] Michelle Normoyle, "Analysis of US Domestic Flights" , 2013/2014 .

[13] CHEN HUAIXIN , "An improved Naive Bayes Classifier for Large Scale Text" .