Big Data Analysis by Classification Algorithm Using Flight Data Set

Ujjwala Urkude
Masters of Technology in Computer Science & Engineering
University of RGPV , Bhopal.

Abstract- **In this work, we proposed a technique with broad application which is used to classify each item in a set of data into a set of predefined classes or groups .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 .Where Classification is a process of generalizing the data according to different instances .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 learn 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.**

I. PROPOSED WORK

''Data classification by Naive Baye's Algorithm Using Flight Dataset'' Our proposed scheme is using Naive Baye's Classifier are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) . under a variety of names, including simple Bayes and independence Baye's. Naive Baye's is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm 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 any possible correlations between the colour, roundness and diameter features.

We are applying this concept in finite dataset; big data is a big part of the aviation industry. Each flight generates terabytes of data that requires real-time analysis to optimize flight operations, maintain safety and meet all compliance requirements. Huge competitive advantage for an airline as it leads to better service with lower operational costs. 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.

In order to evaluate the risk of missing a connection, we need like to know the probability of the incoming flight being too late to be able to catch the second flight, taking into account the incompressible time necessary to go from the arrival gate to the departure gate of the second flight (possibly including immigration control). Models already exist to estimate the gate- to-gate transfer time. The goal of this master thesis is to build a model for the prediction of flight arrival delays [3].

We are using Big data is unstructured data that exceeds the processing complexity of conventional database systems. Big Data applications and the difficulties raised by Big Data volumes, distributed data distribution and by complex and dynamic characteristics. The data is too big, moves too fast, or doesn't fit the rule restricting behaviour of our database architectures. This information comes from multiple, distinct, independent sources with complex and evolving relationships in a Big Data which is keep on growing day by day[2].

II. RESULT

In this work we take the Flight Dataset For experiments, in each experiment Data will generate the Dataset and then calculate the Average Delay, maximum delay Per Airline Company and Average Delay, Maximum Delay per source to Destination. The average performance is tabulated in Table1. From Table 1, we can say that proposed scheme are better than that obtained using the other considered. In the processed Data, Data will be operates on

various Nodes with respect to large amount of dataset .The data size is in giga byte. According to the proposed method, the Nodes can accept data set and provide the execution time.

Table1 comparative analysis of proposed Classification technique

Data set	Execution time by using Factor1	Proposed scheme
2GB	1	0.96
4GB	2.1	1.9
10GB	4.8	4.5
Overall Accuracy	72.26 %	80.75%

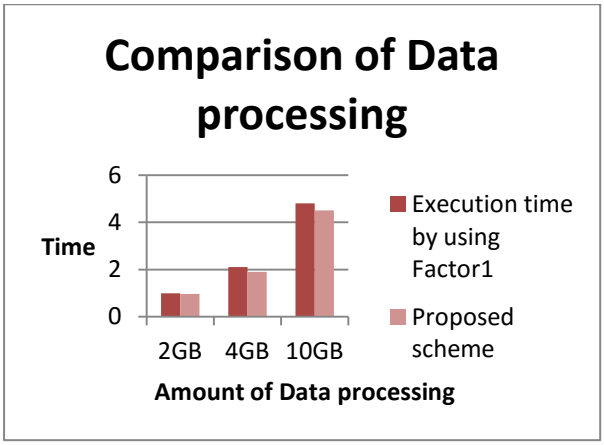

Fig1.1: shows difference between amount of data processing by method

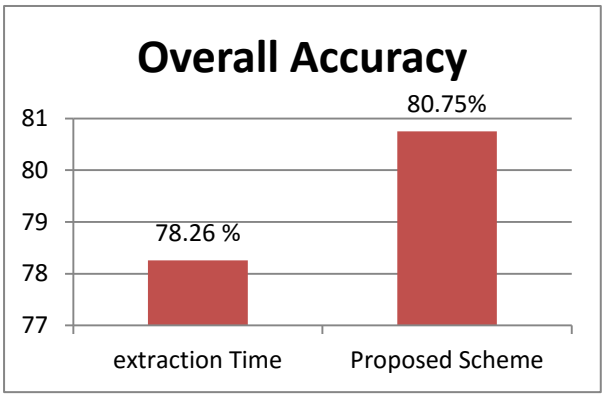

Fig1.2: Average Accuracy of Data between proposed scheme and extract Data

III. CONCLUSION

In this paper we use the classification algorithm like Naive Baye's Algorithm has been done in details .This paper presented aviation prediction, which can show flight prediction according to large dataset .The analysis also found that the dependence of individual link delays on state varied from link to link. Similar promising results were obtained for the prediction of departure delays relative to the scheduled flight time.

The research aim of this dissertation is to decrease the Flight data execution time. In this dissertation, we proposed a novel and secure method Based on Classification. In the first phase, Flight Dataset use for processing the dataset After that data will be stored in HDFS storage .training data hides the additional data into the Naive Baye's After this process jar file send and provide the output based on optimization.

REFERENCES

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

[3] Nagesh Nayak , "Estimation of the Impact of Single Airport and Multi-Airport System Delay on the National Airspace System using Multivariate Simultaneous Models" , January 2012.