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.

Index Terms— Big Data, Classification Algorithm, Delays Prediction, Flight Dataset, Hadoop MAPREDUCE, Data analysis.

I. INTRODUCTION

“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 color, 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].

II. METHODOLOGY

The simulation of project is to calculate delays that affect the performance of flight journey. The airlines report the causes of delays in five broad categories:

A. Air Carrier

The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).

B. Extreme Weather

Significant meteorological conditions (actual or
forecasted) that, the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.

C. National Aviation System (NAS)
Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.

D. Late-arriving aircraft
A previous flight with same aircraft arrived late, causing the present flight to depart late.

E. Security
Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

III. CLASSIFICATION TECHNIQUE
Classification can consists of predicting according to certain outcome based on a given input. In prediction outcome, the algorithm processes in a training set that containing attributes and the outcome, usually called goal or prediction attribute. Advance technology has great growth in volume of data which is available on the internet. The huge number of database and information available the algorithm discovers relationships between the attributes that would make is possible to predict the outcome. Aim of the classification is based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects.

We are using Big data is structured 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 behavior 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].

IV. HADOOP MAPREDUCE
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. 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.

V. RESULT
In this work 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:

<table>
<thead>
<tr>
<th>Data set</th>
<th>Execution time by using Factor1</th>
<th>Proposed scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>2GB</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>4GB</td>
<td>2.1</td>
<td>1.9</td>
</tr>
<tr>
<td>10GB</td>
<td>4.8</td>
<td>4.5</td>
</tr>
<tr>
<td>Overall</td>
<td>72.26 %</td>
<td>80.75%</td>
</tr>
</tbody>
</table>
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.

VII. FUTURE WORK

In the future, the research can be extended in the following direction:

(i). Perform a scalability of data analysis to view the relationship between running time and size of the cluster.
(ii). The proposed scheme still need to record extra information for restoring the dataset
(iii). In future this algorithm can also implementation of algorithm.

REFERENCES