A Study of Performance Enhancement of Map-Reduce Algorithm for Flight Data Analysis using Hadoop and Utility Tools

Ankita Kadre[1], Prof. S. R. Yadav[2]
PG SCHOLAR, MITS (Bhopal), HEAD PG, CSE, MITS (Bhopal)

Abstract - Big Data analysis is an emerging field of research where the data gleaned from traditional and non-traditional sources is structured and then analyzed to derive important conclusions. The main aim of this paper is to use a distributed architecture for the big data analysis of flight data. Apache Hadoop Platform is a technology that helps in providing a Distributed File System (HDFS) where flight data can be stored and queries can be executed in parallel. An efficient algorithm using the Map-Reduce Framework is proposed to compute various queries that affect the performance of flight journey. A list of 30 flight and airline carrier related features were identified and applied to the algorithm. It can be seen that the analysis of flight data using distributed framework lead to statistical observations that can help in improving the flight services. Hadoop improved speed of querying through parallelism.

Keywords - Map-Reduce, Big Data, Hadoop, HDFS, Flight Data

1. INTRODUCTION

Air journey is ever best way of transport around the world, and decrease the hours or days in the trip. Other convenience plan rather than airlines result a factor of months even numerous rural zones can't be reached without it. A normal flight has a lot of information such as flight name, travel time, capacity, distance covered and average delay etc. There is huge database for a flight data that always manages by the owner.

The flights data actually deal with Big Data, terabytes of data generated by each flight that requires real-time Specification optimize flight operations, maintain safety. The ability to process terabytes of data in real-time, a 'big fast data' is a vast competitive help for an airline as it leads to improved service with lower operational costs. It also reduces the cost on storage hardware as not all data is crucial and by filtering it immediately, only the relevant data can be stored. It would help for the travelling easier if it is known how to be better prepared for flight delays when flight tickets are booked.

2. SYSTEM MODEL

Flight data is a huge data which contains several information about the flight for all the airports. This data contains Flight id, airport id, arrival time, departure time and many more. So to analyze such type of data to rate flight and provide reviews to the users is a challenging task. As this data is of GBs and quick results are needed to provide reviews in real time environment so a faster solution to solve this big data problem is needed.

3. PREVIOUS WORK

Javier Conjero Et.Al in [1], used the COSMOS platform supporting sentiment and tension analysis on Twitter data. They demonstrated how this platform could be scaled using the Open Nebula Cloud environment with the Map/Reduce based analysis using Hadoop. Twitter data can be labeled as "Big data". Cardiff Online Social Media Observatory (COSMOS) platform aims to provide mechanisms to capture analyze and visualize data harvested from online repositories and feeds in particular interactive and openly accessible social networking sites such as Twitter. Understanding people's opinion and public moods is called as sentiment analysis. Senti-Strength is a benchmarking tool used for sentiment analysis and is gathering momentum in the research community.

In [2], Park Et.Al.built a predicting model that could resolve issues related traffic accidents. They say that traffic accident related prediction is a data mining task achieved by classification analysis, a data mining procedure requiring enough data to build a learning model. In order to resolve the imbalance of large data, the Hadoop map reduce system was used to pre-process traffic big data and analyze it to create data for the learning system.

In [3], Jyoti Nandimath Et.Al reviewed distributed architecture paradigm to solve the data processing problems in Apache Hadoop. Using Hadoop's various components such as data clusters, map-reduce algorithms and distributed components, they concluded that it can be used to resolve various location-based complex data problems and provide the relevant information back into the system, thereby increasing the user experience. Hadoop framework is used by many big companies like Google, Yahoo, and IBM for applications such as the search engine, advertising and information gathering and processing.

In [4], researchers have compared the existing MPI algorithms with Map-Reduce technique to for parallel and concurrent execute. For applications where the computations are rigorous, the MPI model works efficiently. However, it only works well in the theoretical world. In practical world, Map-Reduce are better used as it allows expressing distributed computations on massive amounts of data. It is used in highly clustered environment where each data-centre cluster having commonly used hardware performs parallel execution. Distributed computing is a paradigm that is ubiquitous whenever, large scale computing is termed.

Manikandan and Siddharth Ravi have studied the Map-Reduce programming model of Hadoop to a good extent and implemented in their paper in [5]. A clear understanding of the Map-Reduce technique is necessary for proper implementation of Hadoop. Map-Reduce are a programming model for processing large scale datasets in computer clusters. The Map-Reduce programming model consists of two functions, map () and reduce (). Implementation of user's own processing logic can be specified and customized in their map () and reduce () function. The map () takes an input key/value pair and produces a list of intermediate key/values pairs.

4. PROPOSED METHODOLOGY

In this flight big data, following algorithms are used for analysis and provide review of different airlines. The queries are aimed to get following results:

- Average delay for every airline.
- No. of times flight was cancelled.
- Arrival delay for every airline.
- No. of times flight was diverted.
- Long route flights.
- Traffic at different airports

Algorithm 1. Calculation of average delay per Airline Company.

```
1.  class Mapper
2.      method Map(string s , integer i)
3.          Emit(string s, integer i)

1.  class Reducer
2.      method Reduce (string s , integer [i1, i2, i3, . . . . in])
3.          sum ←0
4.          count ←1
5.  forall values n ← integer[i1,i2, i3, . . . . . in ] do
6.              sum ←sum+i
7.              count ← count+1
8.  avg← sum/count
9.  end for
10.             Emit(string s, double avg)
```

Algorithm 2.Calculation of No. of cancelled flights per Airline Company.

```
1.  class Mapper
2.      method Map(string s, Boolean b)
3.  if b==1 then
4.              Emit(string s, 1)
5.  end if

1.  class Reducer
2.      method Reduce (string s , Boolean [i1, i2, i3, . . . . in])
3.          sum←0
4.  for all values n ← integer[i1,i2, i3, . . . . . in ] do
5.              sum ←sum+i
6.          end for
7.              Emit(string s, integer sum)
```

Algorithm 3. Calculation of Maximum arrival delay per Airline Company.

```
1.  class Mapper
2.  method Map(string s , integer i)
3.          Emit(string s, integer i)

1.  class Reducer
2.      method Map( string s , integer i1, i2, i3, . . . . . in)
```

```
3.  max←0
4.  for all values n ← integer [i1, i2, i3, . . . . . in] do
5.  if max <val then
6.                  max ←val
7.  end if
8.          end for
9.            Emit( string s, integer max)
```

Algorithm 4. Calculation of No. of diverted flights per Airline Company.

```
1.  class Mapper
2.      method Map(string s, Boolean b)
3.  if b==1 then
4.              Emit(string s, 1)
5.  end if

1.  class Reducer
2.  method Reduce (string s , Boolean [i1, i2, i3, . . . . . in])
3.          sum←0
4.  for all values n ← integer[i1,i2, i3, . . . . . in ] do
5.              sum ←sum+i
6.          end for
7.            Emit(string s, integer sum)
```

Algorithm 5. Calculation of Longest Distance travelled by Airline Company from source to destination.

```
1.  class Mapper
2.  method Map(string s , integer i)
3.          Emit(string s, integer i)

4.  class Reducer
5.      method Map( string s , integer i1, i2, i3, . . . . . in)
6.  max←0
7.  for all values n ← integer [i1, i2, i3, . . . . . in] do
8.  if max <val then
9.                  max ←val
10. end if
11.         end for
12.           Emit( string s, integer max)
```

Algorithm 6. Calculation of average traffic per month on airport.

```
1.  class Mapper
2.  method Map(string s , integer i)
3.          Emit(string s, integer i)

4.  class Reducer
5.      method Reduce (string s , integer [i1, i2, i3, . . . . in])
6.          sum ←0
7.  forall values n ← integer[i1,i2, i3, . . . . . in ] do
8.              sum ← sum+1
9.  end for
10.         Emit(string s, double)
```

5. SIMULATION/EXPERIMENTAL RESULTS

Following analysis is done on flight dataset to retrieve results using hadoop.

- Average delay per airline company
- Number of cancelled flights
- Max arrival delay per airline company
- Number of diverted fights per airline company
- Longest distance travelled per airline company
- Average traffic per month in different airports

1. Average delay per Airline Company

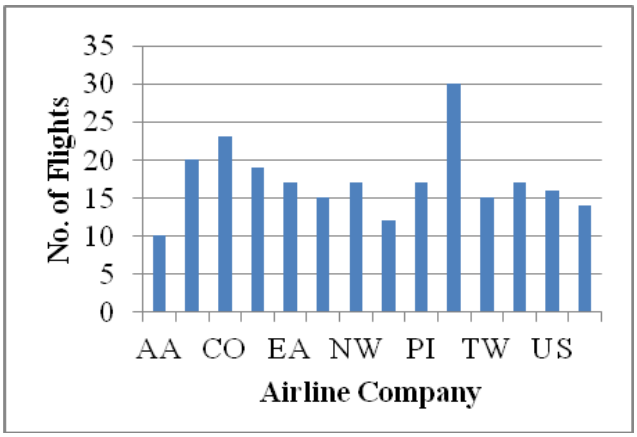

Table 1: Average delay per Airline Company

2. Number of Cancelled Flight per Airline Company

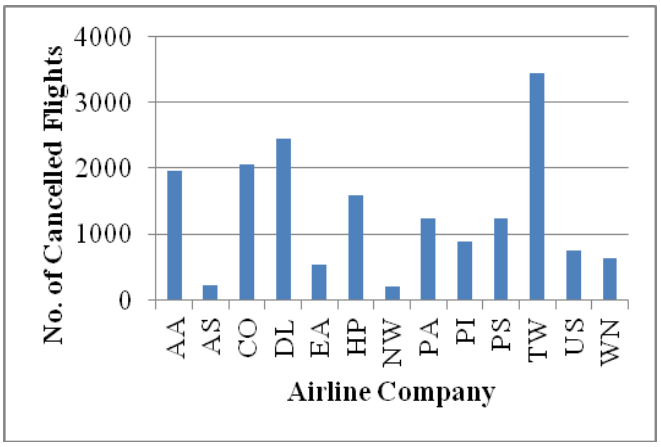

Table 2: No. of Cancelled Flight per Airline Company

3. Number of Cancelled Flights per month Airline Company

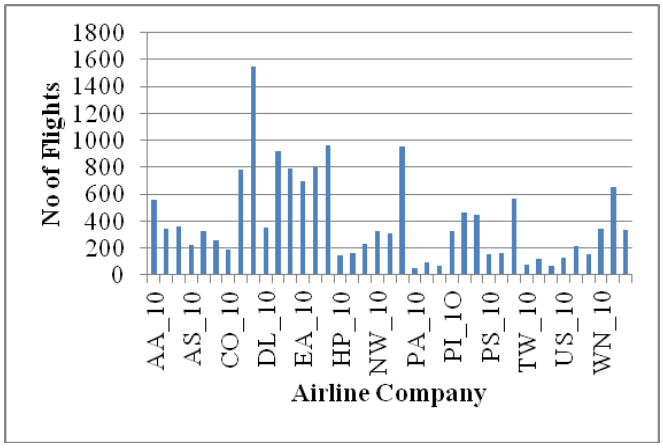

Table 3: No. Cancelled Flights per month Airline Company

4. Number of diverted Flights per Airline Company

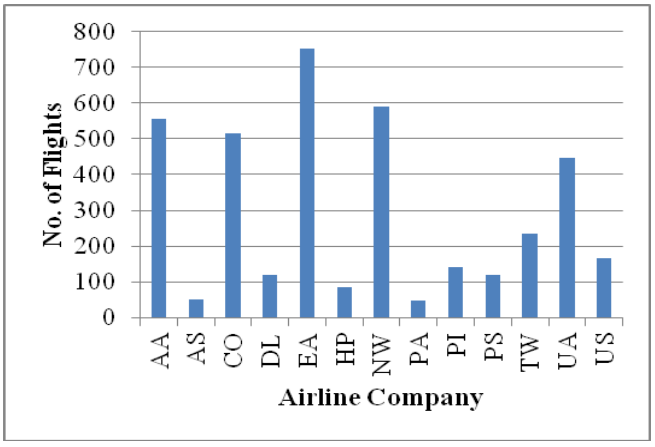

Table 4: diverted Flights per Airline Company

5. Longest distance travelled by the Airline Company.

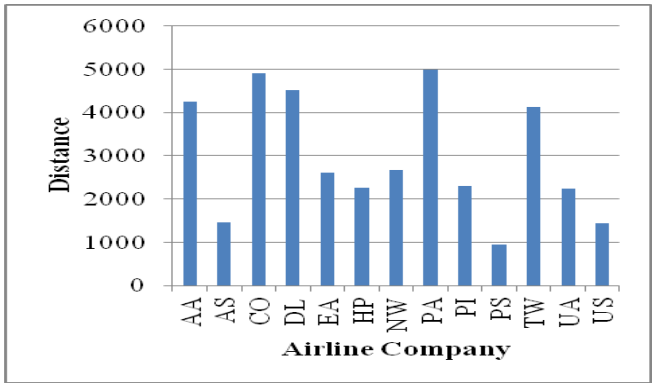

Table 5: Longest distance travelled by the Airline Company

6. Average traffic per month on Airport

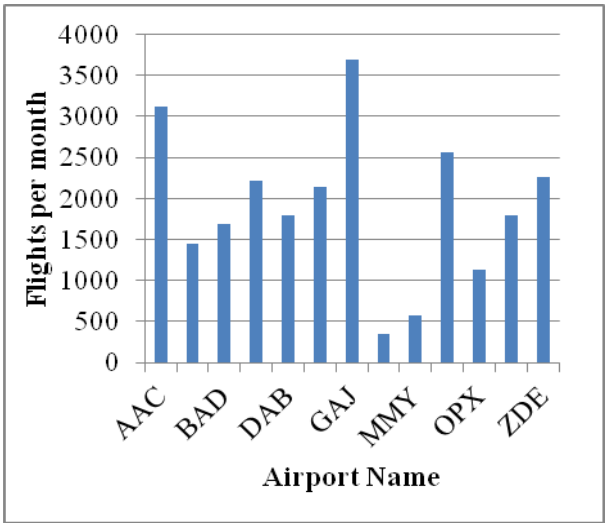

Table 6: Average traffic per month on Airport

6. CONCLUSION

In this paper, enormous volumes of flight data collected from different airline companies for reducing flights delays and improving the flight journey has been analyzed so that one of best probable solution is selected in order to selecting flights. This defines the efficient algorithms that, improves the flights services and it will help the user to find the routes that would minimize the delay. Analyzing the result, the ability is being improved to deliver comprehensive knowledge in order to selecting flights.

7. FUTURE SCOPES

In this paper flight big data analysis is done using parallel computing framework for big data called Hadoop. The results obtained helps to provide review of different airlines. However such type of analysis can be combined with machine learning algorithms to automatically analyze the data using Apache Mahout with Hadoop.

REFERENCES

[1] J. Conejero, P. Burnap, O. Rana, and J. Morgan, "Scaling Archived Social Media Data Analysis using a Hadoop Cloud," 2013.

[2] S. H. Park and Y. G. Ha, "Large Imbalance Data Classification Based on Map-Reduce for Traffic Accident Prediction," 2014 Eighth Int. Conf. Innov. Mob. Internet Serv. Ubiquitous Comput. pp. 45–49, Jul. 2014.

[3]J. Nandimath, "Big Data Analysis Using Apache Hadoop," pp. 700–703, 2013.

[4] J. Shafer, S. Rixner, and A. L. Cox, "The Hadoop Distributed File system : Balancing Portability and Performance."

[5] S. G. Manikandan and S. Ravi, "Big Data Analysis Using Apache Hadoop," 2014 Int. Conf. IT Converg.Secur., pp. 1–4, Oct. 2014.

[6] S. Maitrey and C. K. Jha, "Handling Big Data Efficiently by Using Map Reduce Technique," 2015 IEEE Int. Conf. Compute. Intell.Commun. Technol., pp. 703–708, Feb. 2015.

AUTHOR'S PROFILE

Ankita kadre has received his Bachelor of Engineering degree in Computer Science & Engineering from PCST, Bhopal in the year 2012. At present she is pursuing M.Tech. with the specialization of Computer Science & Engineering in Millennium Institute of Science & Technology, Bhopal. His areas of interests are database, networking, big data, clustering.

Prof. S. R. Yadav has received his Bachelor of Engineering in Computer Science and Engineering from G.I.E.T. Gunupur under B.U. Orissa in the year 2006. M.Tech. in Computer Science and Engineering from P.G. Department of Computer Science Engineering under B.U. Berhampur, Orissa in the year 2009. M.B.A. in HR from Academy of Management Bhopal under B.U. Bhopal, M.P. in the year 2014.He is a Ph.D. Scholar of Computer science and engineering PAHER Univ. Udaipur, Rajasthan, India. At present he is working as an Associate Professor at Millennium Institute of Technology and Science, Bhopal (India). His areas of interests are Data Mining, Intrusion Detection System using Data Mining and Neural Networks.