IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

Survey on Flight Data Analysis Using Hadoop Map-Reduce Algorithm

Musham Sravani[1], Myla Haripriya[2], Repala Sri Maheswari[3], R. Sirisha[4]

Professor & Head, Department of Computer Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India [4]

B. Tech Students, Department of Computer Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India[1,2,3]

ABSTRACT: The aviation industry generates vast amounts of data on a daily basis, encompassing diverse parameters such as flight schedules, passenger information, aircraft performance, and operational metrics. Analyzing this data is critical for optimizing flight operations, enhancing safety measures, and improving overall efficiency. Traditional data processing techniques often struggle to handle the scale and complexity of this data. In this study, we propose a novel approach leveraging the MapReduce algorithm, a parallel processing framework, to efficiently analyze flight data at scale. Our methodology involves preprocessing raw flight data to extract relevant features, followed by the implementation of MapReduce jobs to perform distributed computation tasks across a cluster of computing nodes.

Specifically, we focus on tasks such as flight delay prediction, route optimization, and anomaly detection. By harnessing the parallel processing capabilities of MapReduce, we are able to overcome the challenges associated with large-scale flight data analysis, including processing time and resource utilization. We demonstrate the effectiveness of our approach through a series of experiments using real-world flight datasets. Our results indicate significant improvements in computational efficiency and scalability compared to traditional data processing methods. Overall, our study contributes to advancing the field of flight data analysis by offering a scalable and efficient solution for extracting valuable insights from massive datasets.

KEYWORDS: Flight data analysis, MapReduce algorithm, Big Data, Parallel processing, Aviation industry.

I. 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 flight data actually deal with Big Data, terabytes of data generated by each flight that requires real-time specification to optimize flight operations, maintain safety and rally all compliance desires. 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 we know how to be better prepared for flight delays when we book the flight tickets.

Every day, there are many no. of people travel by airplane to get to their destinations Unfortunately many of these flights do not leave on-time this is problem for airline travelers. The Flight delays are almost always the top complaints for the airline companies, and every year such delays can cause huge economic lost, flight delays are a major problem within the airline industry. The reasons of flight delays vary, depending on waivers, airports, airlines, and others. The airlines also have to face with weather delays before they could happen, It's like a proactive cause, for doing a much better job of communicating with passengers and optimizing resources.

II. LITERATURE REVIEW

Javier Conejero Et. Al (2013) 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 labelled as "Big data". Cardiff Online Social Media Observatory (COSMOS) platform aims to provide mechanisms to capture analyze and visualize data harvested from

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

online repositories and feeds in particular interactive and openly accessible social networking sites such as Twitter. SentiStrength is a benchmarking tool used for sentiment analysis and is gathering momentum in the research community. They conducted their experiments on a single desktop computer with 4GB of RAM, 2.83 GHz Intel Core Quad 2 CPU. And used the Apache Hadoop Map/Reduce framework to analysis 1TB of tweets in plain text format.
After surveying this paper, it was concluded that the it is feasible to use Hadoop and Map/Reduce paradigm for distributed data computing. Instead of using sequential version of the application, usage of multiple "virtual" worker nodes improves the performance significantly. Better the number of workers deployed, the better the performance achieved, limited by the amount of VMs that can run parallel without affecting each other.

Park Et.Al.built a predicting model that could resolve issues related traffic accidents in their paper in [2]. 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.

The actual size of the data related to traffic analysis was around 300 GB. The overall system was based on Hadoop with implementing data pre-process, learning data creation, over swamping by Hive. The cluster and classification analysis is operated by Mahout. They had used Hadoop version 1.0.4 with 1 Name Node 1 2ndNameNode and 30 data nodes on a CPU with 3.10 GHz quad-core and 16 GB RAM with Ubuntu server 12.04 LTS (64 bit). The number of created accident and non-accident data is 1,023,120 with 1,161 mapper created and 300 reducers. The entire process got finished in 768 seconds. The accident data accounted for 0.14% of the total data processed.

It was found that Hadoop was efficient to process the big data and to create the learning data in the future. The total and target precision was found to be 80.56% and 42.71%. With regards to future works, it was found that real time data learning and prediction cannot be applied with Hadoop, since it is impossible to process real time data.

Researchers have compared the existing MPI algorithms with Map-Reduce technique to for parallel and concurrent execution. Each of these algorithms yield different performance and usability characteristics. 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 is 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. Map-Reduce was first researched by Google in 2005 and is used by the big 5 companies i.e Google, Microsoft, Yahoo, Amazon and eBay. It provides fault tolerance and scalability when used in a distributed environment. Working of the Map-Reduce is in batch mode. Google has used their Percolator or Yahoo S4 Distributed Steam Computer Platform when implemented in a real time manner. Google implemented their file system effectively in this context. Analysis of the Google File System (GFS) lead to the conclusion that the key infrastructure in Google's Map-Reduce was the underlying distributed file system to ensure data locality and availability.

When we try to conjoin the Hadoop's architecture with and efficient distributed file system, distributed computing becomes easy and data parallelism over thousands of computing nodes is possible reducing the processing time by a considerably amount. The working of traditional DBMS is such that it is not suitable for solving extremely largescale data processing tasks. According to [5], Hadoop framework won the 1st prize in sorting (Gray Sort) benchmark tests for 100 TB sorting with over 3800 nodes working in parallel.

After surveying several papers, it was found that Map-Reduce has various advantages over traditional DBMS. They are listed as follows:

1. Simple and easy to use- A programmer can easily define their own functions to effectively perform any task in hand.
2. Flexibility- There is no data dependency or underlying schema that defines Map-Reduce, hence can be changed perfectly.
3. Independent of the storage
4. Fault Tolerance-It can continue to work even if an average of 1.2 failures are recorded amongst the nodes
5. Scalability- Yahoo reported that there could be more than 4000 worker nodes doing their tasks simultaneously.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

2.1 ANALYSIS OF TECHNIQUES DISCUSSED IN LITERATURE SURVEY

Author	Datasets Used	Methodology	Algorithm Used	Results
Javier Conejero Et. Al	Twitter data collected from the COSMOS platform	Sentiment analysis using the SentiStrength tool and the Apache Hadoop Map/Reduce framework.	Processing tweets with SentiStrength at the Map stage in Hadoop to obtain sentiment scores (positive or negative).	The results demonstrate the feasibility and scalability of using Hadoop and the Map/Reduce paradigm for distributed data computing.
Park Et. Al.	Traffic data related to accidents	Pre-processing, Learning Data Creation, Hadoop MapReduce.	Hadoop MapReduce, Hive, Mahout.	Parallel Processing: Utilized 1,161 mappers and 300 reducers for efficient parallel processing. Data Processed: Processed a total of 1,023,120 instances of traffic data.
B. N. A. Syed, S. Huan, L. Kah, and K. Sung	Diabetes dataset, German dataset, Heart dataset, Ionosphere dataset.	Incremental learning with support vector machines	SV-incremental algorithm	Image classification, Text classification.
S. Guha	The dataset consists of daily counts of chief complaints from emergency departments (EDs) of the Indiana Public Health Emergency Surveillance System (PHESS).	The methodology involves using the RHIPE (R and Hadoop Integrated Programming Environment) for the statistical analysis of large and complex data.	RHIPE integrates the R programming language with Hadoop's distributed computing framework.	The result of using RHIPE is the potential for improved scalability, efficiency, and flexibility in analysing massive datasets within the R programming environment.
F. Berman	Data preservation in the information age and comparing Pig and its associated language Pig Latin with other data processing languages and systems.	The methodology involves comparing Pig and Pig Latin with other data processing languages and systems.	Filtering, grouping, joining, and aggregating data, among others.	The strengths and weaknesses of Pig and Pig Latin in comparison to alternative approaches.
Brad Brown, Michael Chui, and James Manyika	Sales data from each store Customer, transaction data, Supplier data, Location data,Data from corporate wikis.	Collecting, integrating, and analysing the diverse datasets and make data-driven decisions.	Predictive modelling algorithms, Clustering algorithms, Association rule mining algorithms.	Increased market share across profitable segments, Improved agility and responsiveness to market dynamics.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

T. White	Historical data	Analysing historical data on storage capacities and transfer speeds of hard drives to understand their evolution over time	MapReduce programming model and Hadoop	Handling lots of data from remote sensing sources, doing it faster and using resources smarter than before.
L. Wang, M. Kunze, J. Tao, and G. von Laszewski	Satellite imagery, aerial photographs, LiDAR data, and other geospatial datasets	Design and implementation of a cloud-based remote sensing production system aimed at efficiently processing large volumes of remote sensing data.	Data scheduling and optimization to minimize data transfer among distributed data.	Development and implementation of a cloud-based remote sensing production system that addresses inefficiencies in existing systems and improves productivity and performance in data-intensive computing tasks.
S. Humbetov	Climate sensors, social media posts, digital media, transaction records, GPS signals, etc.	The utilization of big data technologies like Hadoop for managing and analyzing the vast amounts of data generated. Specifically, the focus is on utilizing R Pig, which is a scalable framework for machine learning and advanced statistical functionalities.	various algorithms- Regression algorithms, Classification algorithms, Clustering algorithmstc.	The rapid growth in the volume of data generated, particularly in recent years.

III. METHODOLOGY

3.1 WORK FLOW

Here, we propose a Map-Reduce technique in Hadoop for data intensive use. Map-Reduce programming consists of writing two functions, Map and Reduce function. Map function takes a key; value pair after some process gives list of intermediate values with corresponding the key. Map function is written in such a way that many map functions can be executed once, so that it will part of the program that divides up tasks. Reduce function will takes the output of the map functions, and do some process on them, usually combining values, to generate the appropriate result in an output file. Figure representing the execution of a Map-Reduce job.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

Figure.1. Workflow of Map-Reduce

We propose a technique which will take suitable features from dataset and perform the queries which are given below. At the initial stage, selected features are mapped as <key, value> pair with the help of record reader. Here key represents the group no. and the value represents the data related to the task. Each <key, value> pair is give as input to the mapper and a new <key, value> pair is generated for the further process. These data is shuffle and sorted in appropriate order and assign to the reducer. Reducer will take the intermediate key pairs and value set which will relevant to the keys. Reducer will merge these values to the small set of values. Here input for the reducer is the output of the mapper which is grouped according to their key because different mappers output may give same key. At last final output is generated from different reducer and combined the final output.

3.1.1. Proposed Queries for Flight data analysis:

1. Calculate average delay per Airline Company.
2. Calculate No. of cancelled flights per Airline Company.
3. Calculate Maximum arrival delay per Airline Company.
4. Calculate No. of diverted flights per Airline Company.

3.1.2. ***Pseudo code for Proposed Queries on Flight Data Analysis*****:**

1. Pseudo code for the average delay per Airline Company is given below. Here we are selecting the feature, delay corresponding to every flight of each airline company. We can consider delay like extreme weather delay, late-arriving aircraft delay, security delay, arrival delay, departure delay, carrier delay, national aviation system delay etc. But we are considering the arrival and departure delay. Because this delay caused by most of the above given delay.

3.1.3. Algorithm 1: Calculate average delay per Airline Company.

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 [i_1, i_2, i_3, ... i_n])
6. sum =0
7. count =1
8. **forall** values n = integer[i_1,i_2, i_3, ... i_n] **do**
9. sum =sum+i
10. count = count+1
11. avg= sum/count
12. **end for**
13. Emit(string s, double avg)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

1. Pseudo code for the cancelled flight per Airline Company is given below. Here we are selecting the feature, cancelled corresponding to every flight of each airline company. There are many reasons for cancelling the flight it may cancelled because of weather condition, late-aircraft delay, security problems etc.

3.1.4. Algorithm 2: Calculate 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**
6. **class** Reducer
7. **method** Reduce (string s , boolean [i_1, i_2, i_3, ... i_n])
8. sum=0
9. **for all** values n = integer[i_1,i_2, i_3, ... i_n] **do**
10. sum =sum+i
11. **end for**
12. Emit(string s, integer sum)

2. Pseudo code for the average delay per Airline Company is given below. Here we are selecting the feature, delay corresponding to every flight of each airline company. We can consider delay like extreme weather delay, late-arriving aircraft delay, security delay, arrival delay, departure delay, carrier delay, national aviation system delay etc. But we are considering the arrival and departure delay. Because this delay caused by most of the above given delay. Here the max arrival delay of the airline company is given.

3.1.5. Algorithm 3. Calculate Maximum arrival delay per Airline Company.

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 i_1, i_2, i_3, ... i_n)
6. max=0
7. **for all** values n = integer [i_1, i_2, i_3, ... i_n] **do**
8. **if** max <val**then**
9. max = val
10. **end if**
11. **end for**
12. Emit(string s, integer max)

3. Pseudo code for the diverted flight per Airline Company is given below. Here we are selecting the feature, diverted corresponding to every flight of each airline company. There are many reasons for diverting the flight it may divert because of weather condition, security problems etc.

3.1.6. Algorithm 4: Calculate 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**
6. **class** Reducer
7. **method** Reduce (string s , boolean [i_1, i_2, i_3, ... i_n])
8. sum = 0
9. **for all** values n = integer[i_1,i_2, i_3, ... i_n] **do**
10. sum = sum+i
11. **end for**
12. Emit(string s, integer sum)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

IV. RESULT ANALYSIS

In the present work, analysis of the flight data using Hadoop framework which gives the different results between different attribute values. Using these values, it can select better option and optimize the solution in order to select the flights. Some of the important attributes here.
Calculate average delay per Airline Company.
Calculate No. of cancelled flights per Airline Company.
Calculate Maximum arrival delay per Airline Company.
Calculate No. of diverted flights per Airline Company.

4.1. AVERAGE DELAY PER AIRLINE COMPANY

Here we analyze the relation between the Airline Company and their average delay. Delay may be happened from many reasons so with the help of this analysis travelers can select a suitable flight for their journey.

Airline Company	Delay (in min)
AA	10
AS	20
CO	23
DL	19
EA	17
HP	15
NW	17
PA	12
PI	17
PS	30
TW	15
UA	17
US	16

Table 1. Average delay per Airline company

4.2. NO. OF CANCELLED FLIGHT PER AIRLINE COMPANY

Many flights will suffer from weather problem, delay problems etc., so here it can calculating the cancelled flights per Airline Company which will help the travelers to select a suitable flight.

Airline Company	Delay (in min)
AA	10
AS	20
CO	23
DL	19
EA	17
HP	15
NW	17
PA	12
PI	17
PS	30
TW	15
UA	17
US	16

Table 2. No. of Cancelled Flight per Airline Company

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305062 |

4.3. NO. OF CANCELLED FLIGHTS PER MONTH AIRLINE COMPANY

From the dataset total no. of flights that were cancelled is analyzed with the help of this comparison travelers can plan which flight have the more probability to cancellation of the flights.

Airline Company	Delay (in min)
AA	10
AS	20
CO	23
DL	19
EA	17
HP	15
NW	17
PA	12
PI	17
PS	30
TW	15
UA	17
US	16

Table 3. No. of Cancelled Flights per month Airline Company

V. CONCLUSION

The thing that impressed us while starting this project was the simplicity with which one could program a distributed application using Map/Reduce or Hadoop. Without any considerable prior information about distributed file systems we were able to easily install and configure Hadoop on our machines.

Being an open-source produce and free to download, it is an inexpensive solution for someone looking to harnessing the power of a distributed system for solving problems. It is indeed an inexpensive and simple, yet powerful solution for exploiting the potential of parallel processing.

In this project, it analyzing the enormous volumes of flight data collected from different airline companies for reducing flights delays and improving the flight journey so that select one of best probable solution in order to selecting flights. Our thesis defines the efficient algorithms that, improves the flights services and it will be help the user to find the routes that would minimize the delay. Analyzing the result, that improve the ability to deliver comprehensive knowledge in order to selecting flights.

REFERENCES

[1] S. H. Park and Y. G. Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction," 2014 Eighth Int. Conf. Innov. Mob. Internet Serv. Ubiquitous Comput., pp. 45–49, Jul. 2014.
[2] B.N. A. Syed, S. Huan, L. Kah, and K. Sung, "Incremental learning with support vector machines," in InternationalKnowledge Discovery and Data Mining conference, 1999.
[3] S. Guha, "Computing environment for the statistical analysis of large and complex data," Ph.D., Purdue University, 2010.
[4] F. Berman, "Got data?: a guide to data preservation in the information age," Commun. ACM, vol. 51, pp. 50–56, December 2008. [Online]. Available: http://doi.acm.org/10.1145/1409360.1409376
[5] Brad Brown, Michael Chui, and James Manyika, Are you ready for the era of „big data_? ,McKinseyQuaterly,Mckinsey Global Institute, October 2011
[6] T. White, Hadoop: The Definitive Guide, 2nd ed. O Reilly Media, 2011, pages 7.
[7] L. Wang, M. Kunze, J. Tao, and G. von Laszewski, "Towards building a cloud for scientific applications," Advances in Engineering Software, vol. 42, no. 9, pp. 714–722, 2011.
[8] S. Humbetov, "Data-Intensive Computing with," 2012.
[9] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.
[10] Nillohit bhattacharya and JongwookWoo,"Big Data Analysis of Airline Data Set using Hive"2015 Access peer-reviewed scientific research Journal of engineering.pp 22-26,August 2015

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details