ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# Big Data Analytics on Aviation data for the prediction of Airline Trends in Seasonal Delay

Dr.M.Sornam
Department of Computer Science
University of Madras
Chennai, TamilNadu, India

M.Meharunnisa
Department of Computer Science
University of Madras
Chennai, TamilNadu, India

Parthiban Nagendren
Department of Computer Science
University of Madras
Chennai, TamilNadu, India

*Abstract:* The direction of the worldwide carrier industry is basically similar to that of a flying machine. Now and again it takes off for the high skies and on occasion, it plunges to ground levels. In the middle of these highs and lows, lies the tale of the business – of its survival, of the new and rising patterns that fuel its development. The issue of carrier deferrals, as measured by the quantity generally landings as a percent of aggregate operations, has been of expanding significance as of late as a large portion of the populace picks air go as a favored method of transportation. This paper provides the result about the total flight delay for a specific period of time caused due to climate, security, carrier, NAS, arrival and departure based on total number of flights getting delayed over the past few years (2006, 2007 and 2008). The historic data which is to be analysed is stored on the databases such as MongoDB and Hive. The usage of time series analysis along with the integration of heterogeneous database helps to achieve the Airline Seasonal Delay which is implemented and visualized in R. The reports are generated by using time series modelling to provide the insights for the aviation industry to take future measures to avoid delays and manage them.

***Keywords:*** Flight Delay, heterogeneous database, Time Series Modeling.

## INTRODUCTION

### *A. Research Motivation*

The colossal increment in air movement comes an extensive increment in the interest for air terminal limit. Notwithstanding, airspace and airplane terminal limit can't continue expanding at a rate important to coordinate the rising interest. When an airport's capacity is reduced during “peak hours", the demand for an airport's resources exceeds the capacity that the airport can afford. This is known as a capacity-demand imbalance. Demand refers to the number of flights scheduled to arrive or depart in a given time period (rate of flight arrivals or departures). Capacity is the maximum number of flight arrivals or departures in a given time period. The direct result of the capacity-demand imbalance is the airport congestion and flight delay. Many major airports around the world have significant delay problems as a result of an imbalance between capacities and demand [1]. Flight delays are obviously frustrating to air travelers and costly to airlines. Airline companies are the most important customers of the airport [2].

Flight deferral is mind boggling to clarify, in light of the fact that a flight can be out of timetable because of issues at the airplane terminal of inception, at the goal air terminal, or amid the airborne. A blend of these components regularly happens. Postponements can now and then additionally be owing to aircrafts. A few flights are influenced by reactionary deferrals, because recently entry of past flight. These reactionary postponements can be exasperated by the timetable operation. Flight timetables are regularly subjected to abnormality. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions are taken. Even if complex, there exist some pattern of flight delay due to the schedule performance and airline itself [10].

### *B. Causes of the Delay*

The airlines report the causes of delay in broad categories that were created by the Air Carrier On-Time Reporting Advisory Committee. The categories are Air Carrier, National Aviation System, Weather, Late-Arriving Aircraft and Security. The causes of cancellation are the same, except there is no late-arriving aircraft category [3].

The categories are defined as [3]:

- 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.).
- Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- 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.
- Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.

- 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.

### *C. Types of Database and Tool incorporated*

***a)MongoDB and Hive:***

MongoDB (from humongous) is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database. MongoDB can run over multiple servers, balancing the load or duplicating data to keep the system up and running in case of hardware failure. MongoDB can be used as a file system with load balancing and data replication features over multiple machines for storing files. MongoDB supports fixed-size collections called capped collections. This type of collection maintains insertion order and, once the specified size has been reached, behaves like a circular queue.

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. It is a platform used to develop SQL type scripts to do MapReduce operations. Mapreduce is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware.

***b)Tool incorporated:***

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering …) and graphical techniques, and is highly extensible. One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

## II. TIME SERIES MODELING

Time series modeling is a dynamic research area which has attracted attentions of researcher’s community over last few decades. The main aim of time series modeling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series[4]. This model is then used to generate future values for the series, i.e. to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past [4]. Due to the indispensable importance of time series forecasting in numerous practical fields such as business, economics, finance, science and engineering, etc., [5,6,7] proper care should be taken to fit an adequate model to the underlying time series. It is obvious that a successful time series forecasting depends on an appropriate model fitting.

### *A.Definition of Time Series:*

A time series is a sequential set of data points, measured typically over successive times. It is mathematically defined as a set of vectors $x_t$, t = 0, 1, 2... Where t represents the time elapsed. The variable $x_t$ is treated as a random variable. The measurements taken during an event in a time series are arranged in a proper chronological order.

### *B.Time Series Analysis:*

In practice a suitable model is fitted to a given time series and the corresponding parameters are estimated using the known data values. The procedure of fitting a time series to a proper model is termed as Time Series Analysis

### *Time Series and Stochastic Process:*

A time series is non-deterministic in nature, i.e. we cannot predict with certainty what will occur in future. Generally a time series $\{x_t, t = 0,1, 2,...\}$ is assumed to follow certain probability model [8] which describes the joint distribution of the random variable $x_t$. The mathematical expression describing the probability structure of a time series is termed as a stochastic process [9]. Thus the sequence of observations of the series is actually a sample realization of the stochastic process that produced it.

A usual assumption is that the time series variables $x_t$ are independent and identically distributed, following the normal distribution [8]. For example if the temperature today of a particular city is extremely high, then it can be reasonably presumed that tomorrow’s temperature will also likely to be high. This is the reason why time series forecasting using a proper technique, yields result close to the actual value.

## III. EXISTING MODEL

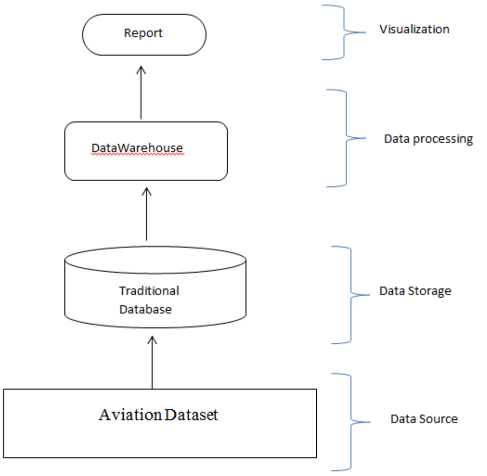


*Figure I*

## IV. PROPOSED MODEL

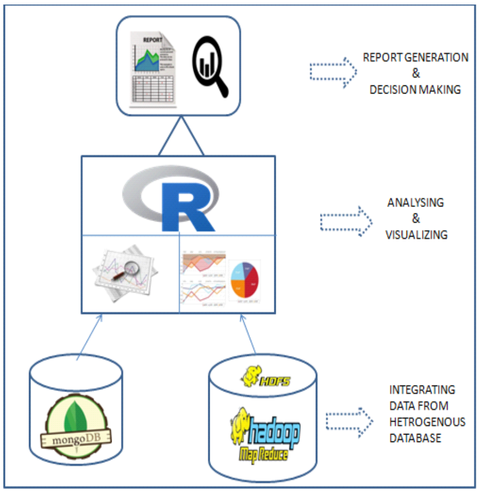

REPORT
REPORT GENERATION & DECISION MAKING
R
ANALYSING & VISUALIZING
mongoDB
HDFS
hadoop Map Reduce
INTEGRATING DATA FROM HETROGENOUS DATABASE

*Figure II*

## V.WORK FLOW

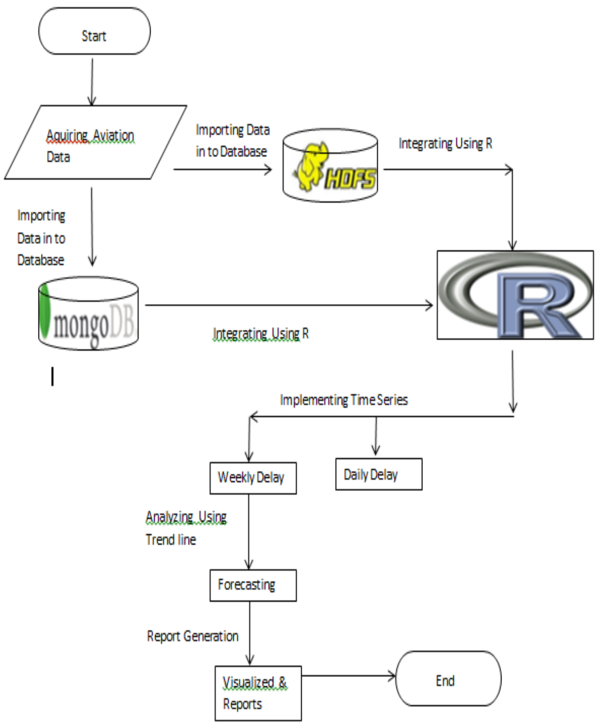
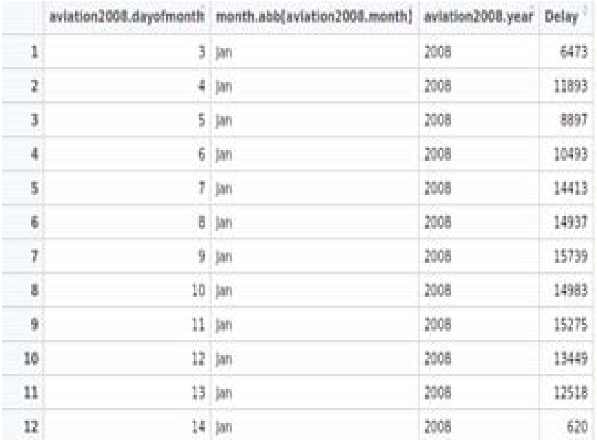


*Figure III*

## VI. TIME SERIES ANALYSIS OF AIRLINE DELAY

In this paper, we made an exploratory data analysis on trends and seasonality for delay on 2006, 2007 and 2008 datasets. The data is taken from the bureau of transportation statistics –United States department of transportation [3]. 2008 dataset has been stored in HIVE and 2006, 2007 datasets are stored in MongoDB.

### A. *Table I : Daily Delay in 2008*

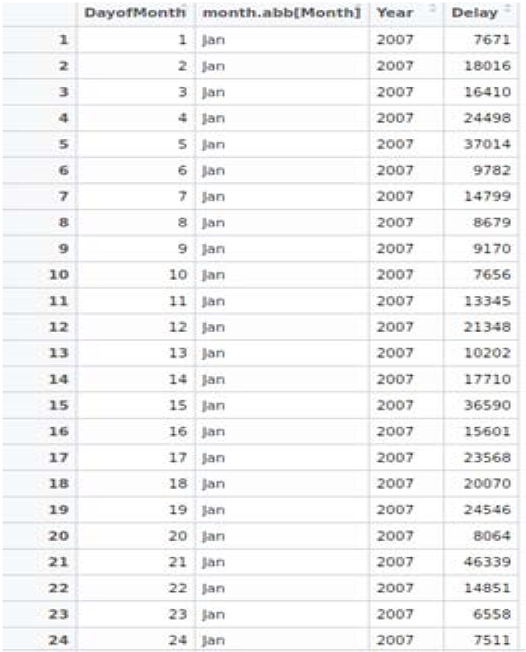


| | aviation2008.dayofmonth | month.abb[aviation2008.month] | aviation2008.year | Delay |
|---|---|---|---|---|
| 1 | 3 | Jan | 2008 | 6473 |
| 2 | 4 | Jan | 2008 | 11893 |
| 3 | 5 | Jan | 2008 | 8897 |
| 4 | 6 | Jan | 2008 | 10493 |
| 5 | 7 | Jan | 2008 | 14413 |
| 6 | 8 | Jan | 2008 | 14937 |
| 7 | 9 | Jan | 2008 | 15739 |
| 8 | 10 | Jan | 2008 | 14983 |
| 9 | 11 | Jan | 2008 | 15275 |
| 10 | 12 | Jan | 2008 | 13449 |
| 11 | 13 | Jan | 2008 | 12518 |
| 12 | 14 | Jan | 2008 | 620 |

### *B.Weekly Delay in 2008:*

"WEEK: 1 Date: 7 Total Delay: 56908"
"WEEK:2 Date: 14 Total Delay: 109225"
"WEEK: 3 Date: 21 Total Delay: 100032"
"WEEK: 4 Date: 28 Total Delay: 94533"

### C. *Table II :Daily Delay in 2007*

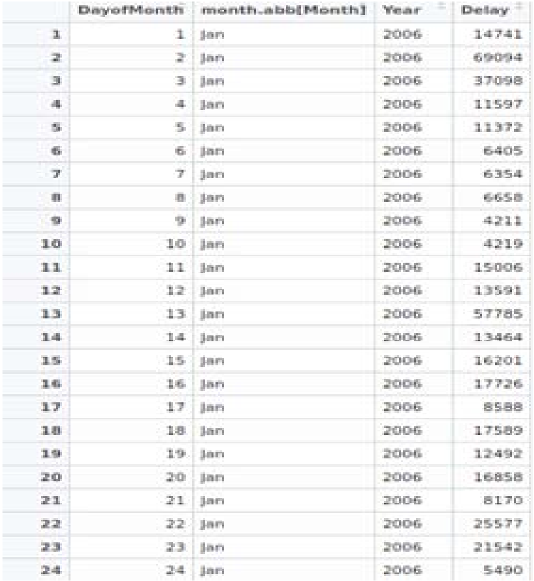


| | DayofMonth | month.abb[Month] | Year | Delay |
|---|---|---|---|---|
| 1 | 1 | Jan | 2007 | 7671 |
| 2 | 2 | Jan | 2007 | 18016 |
| 3 | 3 | Jan | 2007 | 16410 |
| 4 | 4 | Jan | 2007 | 24498 |
| 5 | 5 | Jan | 2007 | 37014 |
| 6 | 6 | Jan | 2007 | 9782 |
| 7 | 7 | Jan | 2007 | 14799 |
| 8 | 8 | Jan | 2007 | 8679 |
| 9 | 9 | Jan | 2007 | 9170 |
| 10 | 10 | Jan | 2007 | 7656 |
| 11 | 11 | Jan | 2007 | 13345 |
| 12 | 12 | Jan | 2007 | 21348 |
| 13 | 13 | Jan | 2007 | 10202 |
| 14 | 14 | Jan | 2007 | 17710 |
| 15 | 15 | Jan | 2007 | 36590 |
| 16 | 16 | Jan | 2007 | 15601 |
| 17 | 17 | Jan | 2007 | 23568 |
| 18 | 18 | Jan | 2007 | 20070 |
| 19 | 19 | Jan | 2007 | 24546 |
| 20 | 20 | Jan | 2007 | 8064 |
| 21 | 21 | Jan | 2007 | 46339 |
| 22 | 22 | Jan | 2007 | 14851 |
| 23 | 23 | Jan | 2007 | 6558 |
| 24 | 24 | Jan | 2007 | 7511 |

### *D. Weekly Delay in 2007*

"WEEK: 1 Date: 7 Total Delay: 156661"
"WEEK: 2 Date: 14 Total Delay: 114934"
"WEEK: 3 Date: 21 Total Delay: 97624"
"WEEK: 4 Date: 28 Total Delay: 69285"

### *E. Table III :Daily Delay in 2006*



| | DayofMonth | month.abb[Month] | Year | Delay |
|---|---|---|---|---|
| 1 | 1 | Jan | 2006 | 14741 |
| 2 | 2 | Jan | 2006 | 69094 |
| 3 | 3 | Jan | 2006 | 37098 |
| 4 | 4 | Jan | 2006 | 11597 |
| 5 | 5 | Jan | 2006 | 11372 |
| 6 | 6 | Jan | 2006 | 6405 |
| 7 | 7 | Jan | 2006 | 6354 |
| 8 | 8 | Jan | 2006 | 6658 |
| 9 | 9 | Jan | 2006 | 4211 |
| 10 | 10 | Jan | 2006 | 4219 |
| 11 | 11 | Jan | 2006 | 15006 |
| 12 | 12 | Jan | 2006 | 13591 |
| 13 | 13 | Jan | 2006 | 57785 |
| 14 | 14 | Jan | 2006 | 13464 |
| 15 | 15 | Jan | 2006 | 16201 |
| 16 | 16 | Jan | 2006 | 17726 |
| 17 | 17 | Jan | 2006 | 8588 |
| 18 | 18 | Jan | 2006 | 17589 |
| 19 | 19 | Jan | 2006 | 12492 |
| 20 | 20 | Jan | 2006 | 16858 |
| 21 | 21 | Jan | 2006 | 8170 |
| 22 | 22 | Jan | 2006 | 25577 |
| 23 | 23 | Jan | 2006 | 21542 |
| 24 | 24 | Jan | 2006 | 5490 |

### *F.Merging the Data from Heterogeneous Database:*

[1] WEEK: 1 Date: 7 Total Delay: 166174
WEEK: 2 Date: 14 Total Delay: 124447
WEEK: 3 Date: 21 Total Delay: 107137
WEEK: 4 Date: 28 Total Delay: 78798
[2] WEEK: 1 Date: 7 Total Delay: 128190
WEEK: 2 Date: 14 Total Delay: 88110
WEEK: 3 Date: 21 Total Delay: 174778
WEEK: 4 Date: 28 Total Delay: 77531

[3]WEEK: 1 Date: 7 Total Delay: 56908
WEEK: 2 Date: 14 Total Delay: 109225
WEEK: 3 Date: 21 Total Delay: 100032
WEEK: 4 Date :28 Total Delay: 94533

### *G. Comparison of Weekly Delay of 2006,2007, 2008*

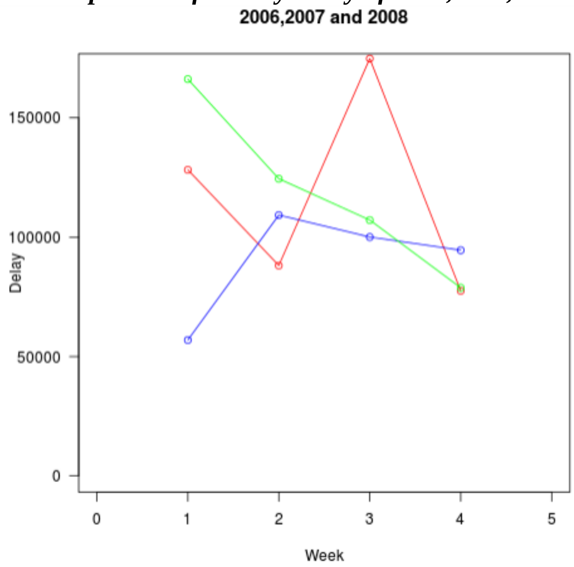


***Figure - IV***

## VII. RESULTS

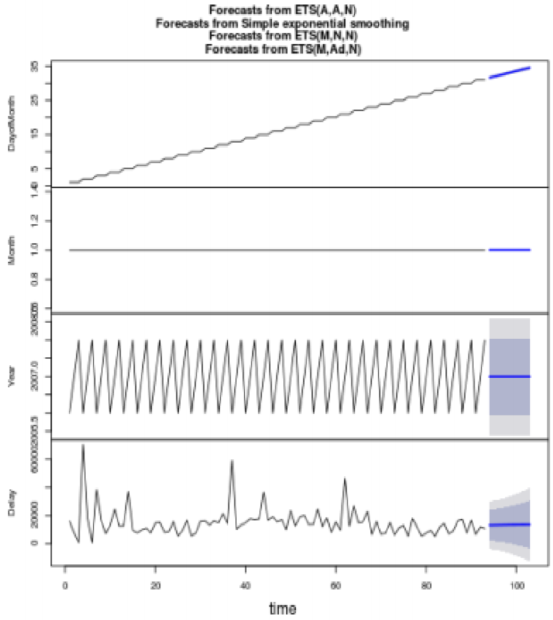


***Figure V***

The trend line provides us the possibility of total delay in the upcoming years.

## VIII. CONCLUSION

With the advent of the big data era, dealing with large amounts of data is challenging. Big data can help operations for airline companies and airports to reduce redundant variability.

This Project provides the result about the total flight delay for a specific period of time caused due to climate, security, carrier, NAS, Arrival & Departure based on total number of flights getting delayed over the past few years. This result helps us to setup a trend line to take necessary measures to avoid future delays. The data is integrated from MongoDB and Hive, which is used to provide the insights for the aviation industry to take future measures to avoid delays and manage them.

## IX. REFERENCES

[1] Aisling, R., and J.B. Kenneth, "An assessment of the capacity and congestion levels at European airports", ERSA conference papers ersa 1999, page no 241, European Regional Science Association.

[2] Ashford and Wright, "Airport Engineering",1992, John Wiley & Sons, Inc,

[3] Bureau of Transportation Statistics, Airline On-Time Statistic. U.S. Department of Transportation. Washington, D.C. http://www.bts.gov/programs/airline_information

[4] T. Raicharoen, C. Lursinsap, P. Sanguanbhoki, "Application of critical support vector machine to time series prediction", Circuits and Systems, 2003. ISCAS'03.Proceedings of the 2003 International Symposium on Volume 5, 25-28 May, 2003, pages: V-741-V-744.

[5] G.P. Zhang, "A neural network ensemble method with jittered training data for time series forecasting", Information Sciences 177 (2007), pages: 5329–5346.

[6] G.P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing 50 (2003), pages: 159–175.

[7] H. Tong, "Threshold Models in Non-Linear Time Series Analysis", Springer-Verlag, New York, 1983.

[8] John H. Cochrane, "Time Series for Macroeconomics and Finance", Graduate School of Business, University of Chicago, spring 1997.

[9] K.W. Hipel, A.I. McLeod, "Time Series Modelling of Water Resources and Environmental Systems", Amsterdam, Elsevier 1994.

[10] Yuqiong Bai, "Analysis of aircraft arrival delay and airport on-time performance ",M.S. Tongji University, China, 2004