Statistical Analaysis on Airline Data using Hadoop MapReduce and R

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Computer Science and Engineering

by

Suman Saurabh
Y11UC231
ss.sumansaurabh92@gmail.com

Under Guidance of
Dr. Ravi Prakash Gorthi

Department of Computer Science and Engineering
The LNM Institute of Information Technology, Jaipur

January 2015

Statistical Analaysis on Airline Data using Hadoop MapReduce and R

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Computer Science and Engineering

by

Suman Saurabh
Y11UC231
ss.sumansaurabh92@gmail.com

Under Guidance of
Dr. Ravi Prakash Gorthi

Department of Computer Science and Engineering
The LNM Institute of Information Technology, Jaipur

January 2015

The LNM Institute of Information Technology
Jaipur, India

CERTIFICATE

This is to certify that the project entitled Title submitted by Name (Roll NO Y15) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by him/her at the Department of Electronics and Electrical communication Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2014-2015 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this thesis is of standard required for the award of the degree of Bachelor of Technology (B. Tech).

Date

Adviser: Prof. NAME

To Common People

Acknowledgments

Acknowledgements goes here ... [?][?][?][?][?][?][?][?]

Abstract

Have you ever been stuck in a traffic and wondered you could have predicted the traffic pattern if you'd had more data? Respectfully digital India is evolving and getting enough traffic data of either Road, Railway or Airline would take years.

So to present our analysis on traffic we utilized Airline Data presented at ASA Data Expo 2009. This dataset is constructed from information made available by the Bureau of Transportation Statistics, USA. It consists of more than 120 million records corresponding to each commercial airline flight in the United States between 1987 and 2008.

As datasets get larger, real-time visualization becomes more difficult. Supposedly a dataset with a billion entries. If we compute a summary of the dataset and visualize it we will either need non-trivial parallel rendering algorithms or significant time to produce a drawing. This solutions would not scale well. To perform analysis we need to mine relevant data using MapReduce programming.

Airline dataset, because of it's large size we used Hadoop MapReduce to mine the relevant data. Graphical summary were than built in R and analysis were performed on it. It revealed the changing patterns in the flight traffic, cancellation, delays with a spot light on days after 9/11.

Before delving into analytics of airline data, we had a literary overview of how Data Analytics is transforming the digital world using the tools like MapReduce, Amazon EMC. The ongoing importance of consuming and extrapolating Total Data for new business processes and analytics approach has brought us tools like Hadoop, Spark that provides pragmatic, cost-effective, scalable infrastructure for building analytic solution on Big Data.

Big Data has not only changed the tools one can use for predictive analytics, it also changed the entire way of thinking about knowledge extraction and interpretation. Traditionally, data science has always been dominated by trial-and-error analysis, an approach that becomes impossible when datasets are large and heterogeneous.

Contents

List of Figures

List of Tables

Table Page

Chapter *1*

Introduction

1.1 The Area of Work

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.2 Problem Addressed

Aviation and air travel has established itself as a key economic and social resource in modern times. As the world population increases and becomes ever more interconnected, the demand for air travel will only increase. Currently there are over 100,000 commercial aviation flights and over 200,000 general aviation flights within the national airspace system (NAS) every day. This does not include military sorties or other, special purpose, flights within the NAS. The number of passengers flying to or from the U.S. is expected to grow an average of 4.5% annually, with cargo amounts showing a similar increase, while general aviation is expected to grow 1% annually. In addition, there is increasing interest, from both government and commercial sectors, in integrating unmanned aerial vehicles (UAVs) into the NAS. Though full UAV integration poses its own unique set of complications, nevertheless it is only a matter of time before they contribute to the air traffic over the NAS. This constant increase in air traffic within the increasingly congested NAS will require new methods and techniques to efficiently accommodate new traffic.

To address these issues, the US Congress approved plans for the development of the Next Generation Air Transportation System (NextGen). It is an overhaul of the current NAS with the goals of allowing more aircraft to safely fly closer together with more direct routes. It is scheduled for implementation in stages between 2012 and 2025 with 5 major elements: (i) Automatic dependent surveillance-broadcast (ADS-B) will replace radar systems with satellite based global positioning information for each air-

craft. This infor- mation will be broadcast in realtime to airports another aircraft within a 150 mile radius allowing them to fly closer without jeopardizing safety. (ii) Systemwide information management (SWIM) is a consolidation of multiple information systems into a single coherent system and will reduce redundancy and facilitate information sharing. (iii) NextGen data communication will add data links between aircraft and air traffic controllers to the current two-way voice communication. (iv) NextGen network enabled weather is an ambitious effort to fuse data from tens of thousands of ground, air, and space based sensors into a single national weather information system to provide realtime weather information. (v) NAS voice switch (NVS) will replace multiple existing voice switching systems into a single consolidated air/ground and ground/ground voice communication system.The NextGen system will provide the infrastructure to allow aircraft to safely fly closer together thereby making more efficient use of limited airspace. It will allow aircraft to use more direct routes instead of being constrained to predetermined sky highways thereby reducing congestion and reduce fuel costs. With pieces of the NextGen infrastructure coming into place, there is an opportunity to further their benefits by developing software tools that provide added value.

This paper focuses on visual analysis tools to study the changes on air traffic congestion in span of 21 years which would allow policy makers to see the effects of changing the aircraft separation volume on congestion. The same tool can also be used as a decision aid for processing requests for unmanned aerial vehicle operations. Specifically, this paper will discuss methods and tools used to calculate and render air traffic densities over areas of interest, as well as methods for aggregating such traffic densities over different time scales to extract fluctuations and periodic cycles in traffic patterns. We apply these tools to study the effects of possible modifications to the current en-route aircraft separation requirements. These modification, which are based on the characteristics of large fixed wing aircraft, has the potential of increasing the amount of available air space, allowing for future increases in overall air traffic numbers. In addition, we apply the same suite of tools to provide a quick visual inspection of planned UAV operation under different aircraft separation requirements. The studies conducted in this paper are based on a data set which is constructed from information made available by the Bureau of Transportation Statistics,

There are over 300,000 flightswithin the United States every day. In the future, daily air traffic number of all varieties are expected to continue rising. In addition, there is increasing interest in integrating unmanned aerial vehicles, for both government and commercial interests, into the national airspace system (NAS). This large growth in aviation operations will only increase traffic within the already limited NAS, leading to higher congestion and less free airspace. In this report, we present visual analysis tools to study the changes on air traffic congestion in span of 21 years. The tools support visualization of time-varying air traffic density over an area of interest using different time granularity. We use this visual analysis platform to investigate how changing the aircraft separation volume can reduce congestion while maintaining key safety requirements. The same tool can also be used as a decision aid for processing requests for unmanned aerial vehicle operations.

To present our analysis on traffic we utilized Airline Data presented at ASA Data Expo 2009. This dataset is constructed from information made available by the Bureau of Transportation Statistics, USA. It consists of more than 120 million records corresponding to each commercial airline flight in the United States between 1987 and 2008. As datasets gets larger, real-time visualization becomes more difficult. Supposedly a dataset with a billion entries. If we compute a summary of the dataset and visualize it we will either need non-trivial parallel rendering algorithms or significant time to produce a drawing. This solutions would not scale well. To perform analysis we need to mine relevant data using MapReduce programming.

1.2.1 Data

Twenty years of data (120 million observations) on commercial domestic flights in the United States.

Dates : day of week, date, month, year
Arrival and Departure times : actual and scheduled
Flight times : actual and scheduled
Origin and destination : airport code, latitude, longitude
Carrier : American, Aloha Air, United, US Air,e.t.c

1.2.2 Goal

This is intentionally vague in order to allow different entries to focus on different aspects of the data, but here are a few ideas to that we focussed on :

- Summarize data by time periods, airport, and carrier
- Temporal effects
    - Are some time periods more prone to delays than others?
    - Relationships between delays and *Seasonal factors*: winter, summer, holidays *Weather factors*: Blizzards and severe weather *Daily factors*: Time of day, day of week
- Spatial effects
    - Are some airports more prone to delays than others?
- Carrier effects
    - Are some carriers more prone to delays than others?
- Analysis of traffic on New York and Chicago a densly populated metropolitan cities in USA?

1.2.3 Tools Used

1.2.4 (

end)

1.3 Existing System

The main thrust of this paper is on visual analysis of air traffic data. Hence, this section focuses on work related to visualizing air traffic data. One of the most popular technique for visualizing air traffic data is to represent the trajectory of each aircraft as an animated particle. Many such visualizations are available on the web via sites such as youtube. A version that was designed by Aaron Koblin demonstrates several techniques and embellishments for presenting the flight trajectories. More recently, the discrete nature of the flight tracks were smoothed out to obtain a continuous estimate of air traffic density using a view dependent kernel density estimator. Representing air traffic data as a density plot is not new. Kellner [8] also used density plots of the arrival and departure rates of aircraft at different airports to assess their capacity. This paper will use similar techniques in visualizing the air traffic data. More importantly, our work examines the impact of varying minimum aircraft separation policy on air traffic density, and also examines if a flight plan, e.g. of a UAV operation request, will endanger existing flight patterns.

There are many factors affecting air traffic congestion and airport capacity. One of those that is controllable and fall under policy decisions is the specification of minimum separation between aircraft. Currently, this is set to 5 nautical miles horizontally, and 1,000 feet vertically [4] when the aircraft is en-route. This limit is adjusted as the aircraft approaches an airport and can drop to 3 miles horizontally on landing approaches to airports. The relative weight class of the leading and following aircraft are also taken into con- sideration in such situations in order to reduce risks due to wake turbulence [3]. The en-route limit accounts for aircraft speed (typical passenger jets fly at average speed of 500 miles per hour or just over 8 miles per minutes), weather impact on visibility, and wake turbulence from leading aircraft, among other factors. With the touted capabilities of ADS-B, the NextGen enabled weather system, and integrated information system, one can theoretically safely reduce the minimum separation requirements between aircraft. This paper provides visual analysis tools to examine the effects of different shapes and parameters describing the minimum separation volume between aircraft.

With regards to UAV operation, they are more generally referred to as Unmanned Aircraft Systems (UAS)[7, 2]. Over the past few years, interest in UAS has rapidly increased. This is because of the possibilities they offer to both government and commercial interests. They would enable a broad range of satellite-like abilities, but at a much lower cost. Aerial photography, communications, environmental monitoring, and security are some of the abilities that UAS deployment could make possible on a large scale. Currently, UAS are predominantly used by the Department of Defense and the Department of Homeland Security, and often outside of national air space (NAS). A handful of UAS are allowed to operate inside our NAS, though almost exclusively for national security or research purposes. However, each UAS operation must be pre-approved by the FAA on a case by case basis. This process is very tedious and does not scale well to large numbers of flights. There are a few studies on risk managment of operating UAS. A recent study uses a site-specific non- uniform probabilistic background air traffic to study the risks [11]. Using the visual analysis tools presented in this paper, checking whether the flight plan for a UAS will allow for a safe operation within the NAS can be accomplished expeditiously.

1.3.1 The Oracle Airline Data Model

The Oracle Airline Data Model is a powerful logical and physical data model that will help airlines effectively store, manage, and analyze airline data that currently resides in passenger service systems (includes reservation systems and departure control systems), global distribution system (GDS), loyalty management systems, and customer data warehouses. It provides a single scalable repository for transactional and historical data that can be used to provide real-time business intelligence and strategic insights youre your airline. Using sophisticated trending and data mining capabilities based on Oracles OLAP and data mining technology, airline personnel will now have the data analysis capabilities to develop Airline -specific insights that are relevant, actionable, and can improve both top-line and bottom-line results.

The Oracle Airline Data Model provides detail transaction storage and advanced analysis into a full range of airline subject areas, including reservations, sales, operations, loyalty, and finance. Using reservation data, the data model can provide detailed insight into passenger bookings by time period, fare class, and flight. It provides insights into channel performance looking at bookings, cancellations, and revenues through travel agency, OTA, ticket counter, call center, and web channels. It allows you to analyze passenger revenues by geography, time period, and flight. Finally it provides insights into loyalty program member activity through a variety of reports. The data model fits the needs of large network carriers and low-cost carriers.

1.4 Creation of bibliography

Use sampleBib.bib file to save your bib format citations. Use the command [**?**] for referring to a particular article.

Chapter *2*

Literature survey

2.1 Introduction

Air traffic data usually consists of a collection of flight trajectories of different aircraft. Each flight trajectory usually contains information about the type of aircraft, origin and destination airports, followed by a series of entries that records the time, location, and altitude of the aircraft. The flight tracks are usually recorded in 10 second intervals. Other information such as date, heading, velocity, etc. are generally recorded as well, but were not available in the data set used in our study. The data set used to test and demonstrate our visual analysis tool has an area of interest that is New York an Chicago. Two of the largest metropolitan cities in United States of America. It includes all flight path information from flights that took place from the begining of Jan 1987 to the end of Dec 2008. There are 349,992 unique flight path records in this particular data set. This data set is comprised of uniquely identified flight paths, each containing latitude, longitude, and altitude information at 10 second intervals for the duration of the flight within the area of interest. The time of day and month in which the flights took place are specified. However, the specific date the flight took place is not included.

2.1.1 Flight Data Monitoring

Many airlines collect and analyze flight data of routine flights. The process is generally referred as flight data monitoring, which involves data acquisition, transmission, storage and analysis, which are described in detail in this section. By reviewing a number of software tools for flight data analysis, a benchmark of current flight data analysis methods was established. Improvement opportunities were identified from the literature review, which motivated this research.

2.1.2 Anomaly Detection

The approach for anomally detection is as described in[1]. The approach that detects abnormal flights from routine airline operations using FDR data and asks domain experts to interpret the results and operational implications. Thus, anomaly detection algorithms will be developed to detect anomalies from FDR data. Anomaly detection refers to the problem of detecting an observation (or patterns of obser-

vations) that is inconsistent with the majority members of the dataset. It is also referred to as novelty detection, anomaly detection, fault detection, deviation detection, or exception mining in different application domains. A significant number of anomaly detection techniques have been developed. While some of the techniques are generic and can be applied to different application problems, many of them are focused on solving particular types of problems in an application domain.

2.1.3 General Anomaly Detection Techniques

Many anomaly detection techniques have been developed to address anomaly detection problems in many application domains. Three main approaches have been taken: statistical approach, classification approach, and clustering approach. The categories are not mutually exclusive as some of the techniques adopt concepts from more than one basic approach. For eg. there is anomaly in the flight traffic pattern at the end of year 2001 related to the 9/11 terrorist attack.

Chapter *3*

Proposed Work

In this thesis, we investigate and visualize data for domestic flights for US airline data in focus with flight orignated at New York and Chicago, the two most densly populated cities in United States of America.The data set contains more than 120 million flight records from October 1987 to December 2008. The thesis reflects the process followed by analysts working with big data: sampling is used to generate hypotheses that are then tested against the complete dataset.

The computation for the comparison of their informal ”rule” and analyses of the distribution of the population values requires coding MapReduce programming modal. R to be used as a tool as a major component for data mining and visulization as R is one of the most powerful tool for statistical data analysis. Many have argued that statistics students need additional facility to express statistical computations. By introducing students to commonplace tools for data management, visualization, and reproducible analysis in data science and applying these to real-world scenarios, prepares them to think statistically. The statistical data analysis cycle involves the formulation of questions, collection of data, analysis, and interpretation of results. Data preparation and manipulation is not just a first step, but a key component of this cycle. When working with data, analysts must first determine what is needed, describe this solution in terms that a computer can understand, and execute the code.

For analysing airline data we first looked at the variables provided by the data. Since the of dataset is very large, real-time visualization is not scalable. If we try to compute a summary of the dataset and visualize it we will either need non-trivial parallel rendering algorithms or significant time to produce a drawing.

For analyzing such big dataset MapReduce programming modal has been proposed to mine the relevant information from dataset and perform statisitical modelling. For e.g. To plot the air traffic pattern originating at NewYork, it is scalable to perform mining operation on the data and extract the flight details orignated at New York. We can than use this data to build a graphical models on Cancellation Rate, Delay Rate monthly as well as weekly. This modal can be plotted graphically to understand the anomally present in 20 decades New York air traffic history.

The proposed work in thesis builds visulization modal described in four sections. First section mines the data from 120MM airline traffic using Hadoop MapReduce and R. Details of the package is provided above in Introduction section. Section 2 brings the modal of traffic pattern for top 20 busiest air traffic

destinations. Chicago the city with highest airline traffic is palced at the top succeeding Atlanta and Dallas Fort-Worth. Section 3 provides the detail of overall airline traffic orignated at Chicago. This section also visualizes the pattern for flight cancellation and delay in Chicago in respect with its two airports O'Hare International and Chicago Midway. Section 4 modals the similar graph as of Chicago for the airport in New York: LaGuardia and John F Kennedy.

3.0.4 Mining Data Using Hadoop and R

Hadoop MapReduce programming modal is built in Java but hadoop provides utility to allow MapReduce programs to run from most of the famous languages. Hadoop streaming is a utility that comes with the Hadoop distribution. The utility allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer in any programming language. Both the mapper and the reducer are executables that read the input from stdin (line by line) and emit the output to stdout. The utility will create a Map/Reduce job, submit the job to an appropriate cluster, and monitor the progress of the job until it completes. resently the supported languages are Python, C, R, Scala,Ruby.

Many pacakages in R are present which allows to build mapreduce programs but due to good doucmentation of Rhipe, it has been used by author to develop mapreduce code to mine data. RHIPE is the R and Hadoop Integrated Programming Environment.It means "in a moment" in Greek. RHIPE is a merger of R and Hadoop. R is the widely used, highly acclaimed interactive language and environment for data analysis.Hadoop consists of the Hadoop Distributed File System (HDFS) and the MapReduce distributed compute engine. RHIPE allows an analyst to carry out D&R analysis of complex big data wholly from within R. RHIPE communicates with Hadoop to carry out the big, parallel computations.

```
map <- expression({
  # For each input record, parse out required fields and output new record:
  extractTop20 = function(line) {
    fields <- unlist(strsplit(line, "\\,"))
    # Skip header lines and bad records:
    if (!(identical(fields[[1]], "Year")) & length(fields) == 29) {
      origin <- fields[[17]]
      #print(origin)
      destination <- fields[[18]]
    # Skip records where departure dalay is "NA":
     if (!(identical(origin, "NA"))) {
       # field[9] is carrier, field[1] is year, field[2] is month:
       rhcollect(paste(fields[[1]], "\t", origin, sep=""), 1)
     }
     if (!(identical(destination, "NA"))) {
       # field[9] is carrier, field[1] is year, field[2] is month:
       rhcollect(paste(fields[[1]], "\t", destination, sep=""),1)
       }
    }
  }
  # Process each record in map input:
  lapply(map.values, extractTop20)
})
```

Figure 3.1 Code Logo

Chapter *4*

Results

Chapter *5*

Conclusions and Future Work

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5.1 Scope of further work

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.