**ITCS 6122**

**SOFTWARE SYSTEM DESIGN AND IMPLEMENTATION**

**A PROJECT REPORT**

**ON**

**AIR TRAFFIC MANAGEMENT ANALYSIS USING BIG DATA**

**TEAM MEMBERS**

Divya Ravi (800902718)

Raghunath Nallapeta Viswanath (800909795)

Revati Dhananjayan Lalitha (800915306)

Roopa Shankargouda Patil (800910606)

Sushma Venkatesh Reddy (800908852)

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **S.NO** | **TOPIC** |
| 1 | Abstract |
| 2 | Introduction |
| 3 | Software & Hardware Requirements |
| 4 | Input file |
| 5 | Methods Used |
| 6 | UML Diagrams |
| 7 | Output |
| 8 | Result Analysis |
| 9 | Conclusion |
| 10 | Future Enhancements |
| 11 | References |

**ABSTRACT**

The goal of airline traffic management analysis is to maintain data of different airline companies and analyze it based on various factors. The long term objective is to assist airport authority in order to achieve reduction in cancellations of flights, accommodating new flight arrivals and departures based on traffic density, ways to reduce the flight delays and planning right paths for diversions. Our analysis strategy can be used as a basis for decision in order to add new flights based on the result of data mining process. Which means that our main evaluation criteria will be the operational results of applying the knowledge discovered by our data mining process and not the predictive accuracy of the models developed during this process.

The list of services provided by the application area as follows:

* **Analyzing flights delay**: To analyze data that finds various reasons for delays and their respective percentages in it.
* **Analyzing flights cancellations**: To analyze data that finds various reasons for cancellations and their respective percentages in it.
* **To study flight diversions:** A built in functionality will be used for displaying list of flights that were diverted.
* **Traffic Density**: This will help the airport authorities to study the traffic rates of the flights at different time periods which would help for future flight schedules.

**INTRODUCTION**

Every airlines and airport store the information about the flight related to their organization. The information includes year, month, departure time, flight number, source destination, distance and many more. We have collected such types of sample data from URL given below.

<http://stat-computing.org/dataexpo/2009/the-data.html>

The air traffic is analyzed by considering the inflow and outflow of the flights for a particular place. Flight delay, flight cancellation, flight diversions and traffic density are analyzed. Since this project deals with lots of data and particular details or information need to be queried, Hive programming language is going to be used. Hive Provides an SQL dialect called Hive Query Language for analyzing data stored and hence it is appropriate for the project. In addition, the data is analyzed using Pig scripting language, MapReduce and Spark.

We are analyzing on what factors the delay and frequency of the flights help the airline carriers improve their service to their customers and also the airport authorities to schedule new incoming or outgoing flights based on the analysis. Multiples factors such as weather delay, security delay, NAS delay etc. have considered for delay analysis. And also analysis of cancellation of the flights has been done to show the different reasons of cancellations over a period of time.

**WHY Big Data?**

Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch.

The importance of big data doesn’t revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

When you combine big data with high-powered analytics, you can accomplish business-related tasks such as:

* Determining root causes of failures, issues and defects in near-real time.
* Generating coupons at the point of sale based on the customer’s buying habits.
* Recalculating entire risk portfolios in minutes.
* Detecting fraudulent behavior before it affects your organization.

**SOFTWARE & HARDWARE REQUIREMENTS**

**Software Requirements:**

1. Oracle VM Virtualbox

2. HortonWorks

3. Tableau

4. Linux Operating system

**Hardware Requirements:**

1. RAM: 16 GB

2. Processor: i3 or i5 (Min)

3. 512 GB Hard Disk

**Steps Followed:**

The analysis process starts from the data sets. First the data sets of the airline carriers have been collected and grouped into single database.

In this project we have taken data from Statistical Computing Statistical Graphics (URL:<http://stat-computing.org/dataexpo/2009/the-data.html>). For sample analysis we have taken the data of ten years.

Once the data has been combined it is loaded into HDFS which stands for Hadoop Distributed File System. Once the data is available in HDFS Hive and Pig queries are run on the data to fetch the results and it is graphically visualized in Tableau.

Both Hive and Pig queries run based on map reduce approach which has been explained below:

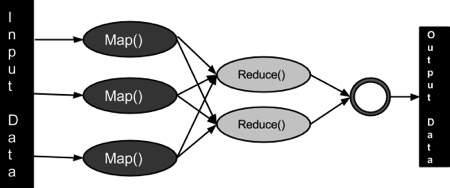
**The Algorithm used in MapReduce:**

Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

**Map stage:** The map or mapper’s job is to process the input data. Generally, the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

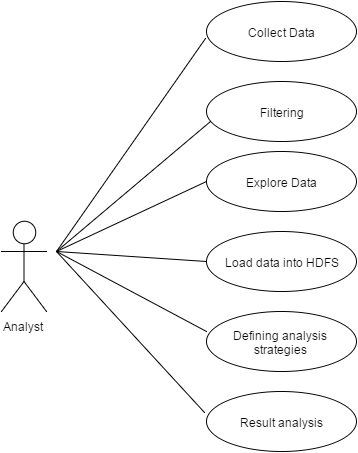
**Reduce stage:** This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS. During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.

Most of the computing takes place on nodes with data on local disks that reduces the network traffic. After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

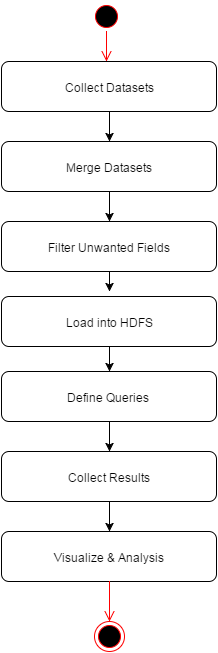


**UML DIAGRAMS**

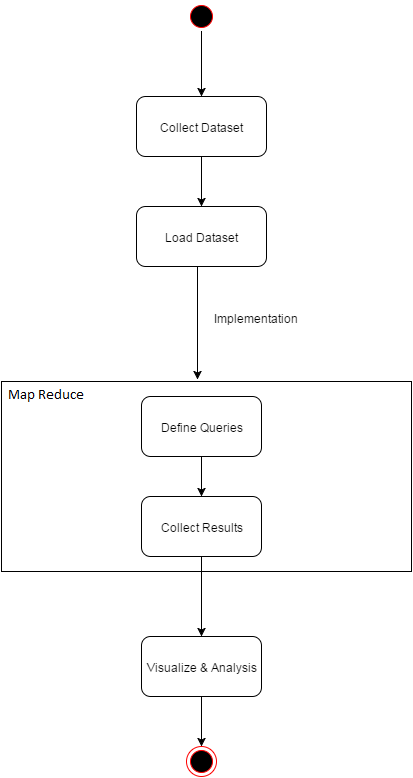
**1. Use Case Diagram**

****

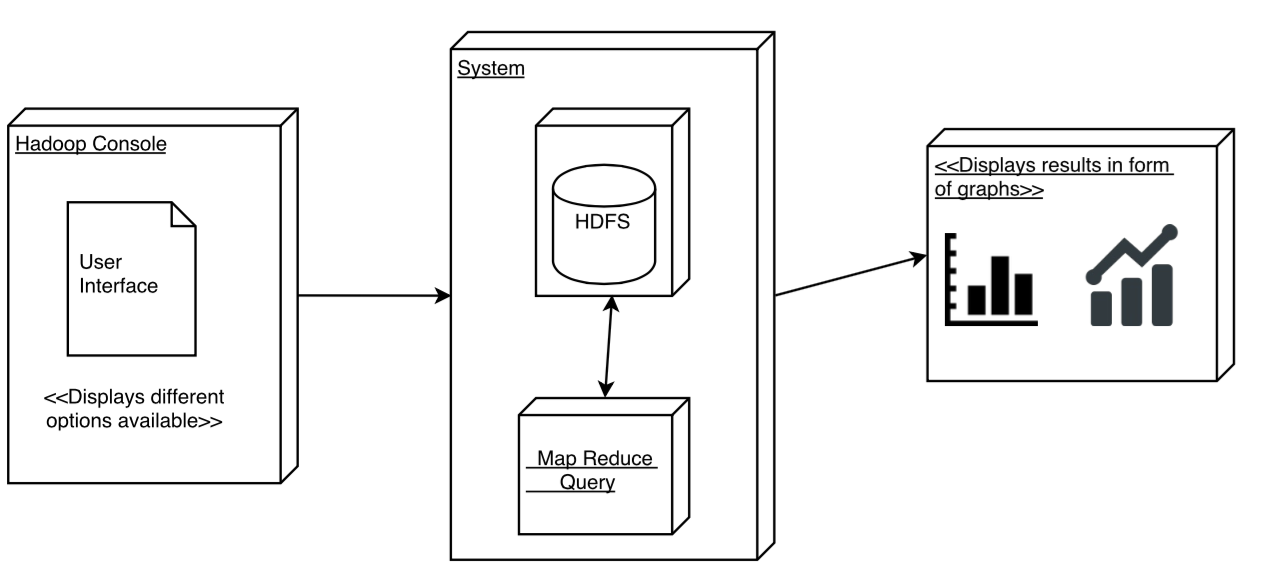
**2. Activity Diagram**

****

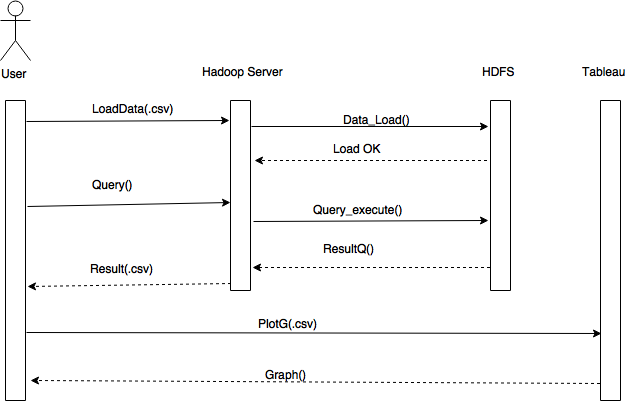
**3. State Chart Diagram**

****

**4. Deployment Diagram**

****

**5. Sequence Diagram**

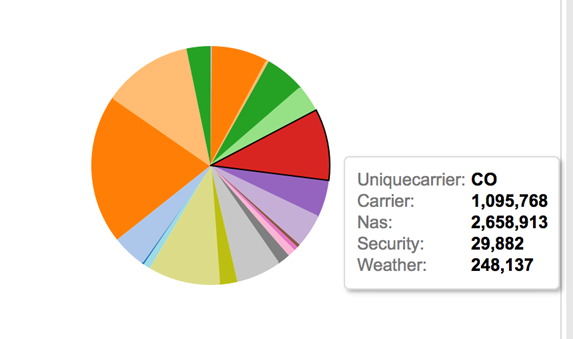
****

**RESULT ANALYSIS**

Below are the graphs which have been analyzed from querying the data in HDFS.

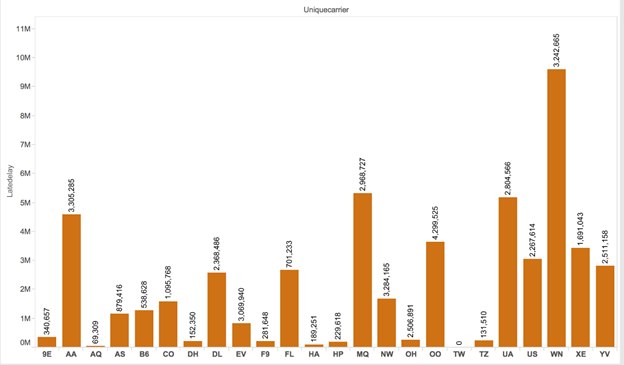
1) Below is the Tableau graph which represents count of each delay for a carrier and over all percentage of the delay of a carrier among all the carriers.

**All Delays according to carrier**:



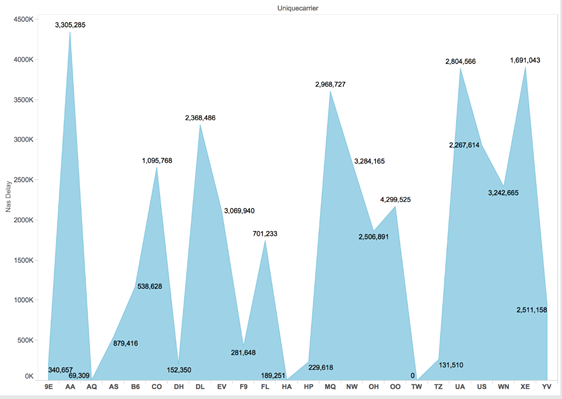
2) Below bar graph represents the count of each carrier for late delay. This analysis can be used to see the different carrier standings in time delays. It can easily be known that the carrier WN is having the highest delays among all the carriers.

**Late Delay:**

****

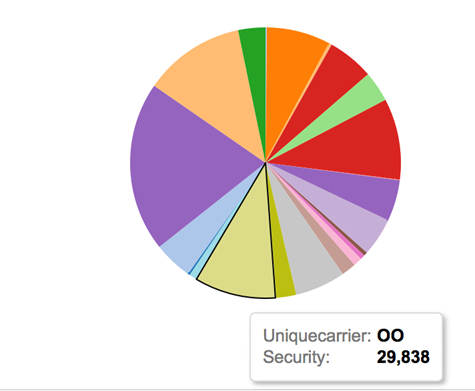
**3)** Below graph shows the NAS delay for each carrier over a period of time. NAS delay includes delay due to airport operations, air traffic control, heavy traffic volume etc.

**National Aviation System (NAS) Delay:**

****

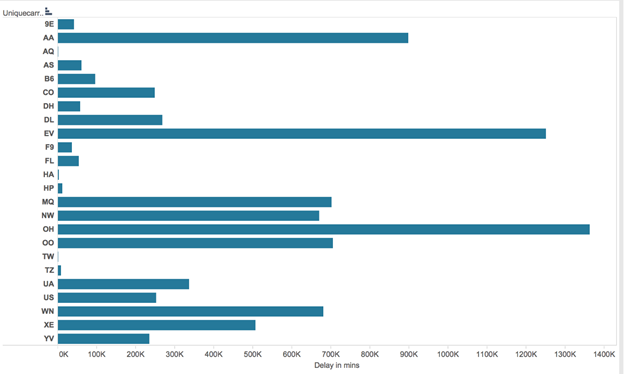
**4) Below pie chart shows the count of security delays of each carrier and percentage of each carrier.**

**Security Delay:**

****

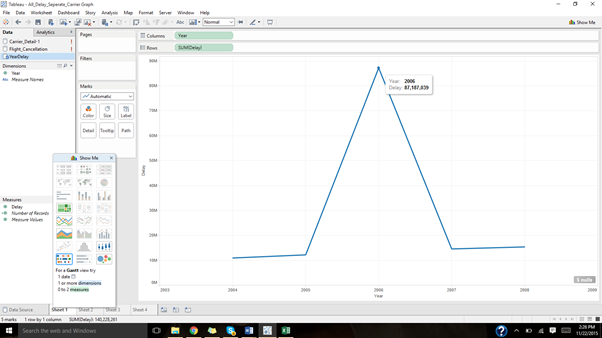
5) Below bar graph shows the weather delays of each carrier.

**Weather Delay:**

****

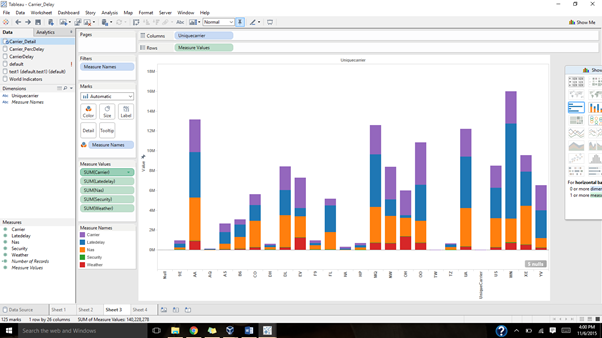
6) Below graph represents the growth of delays of airline carriers.

**Year wise Delay:**

****

7) Below graph shows all types of delays of airline carriers and their percentages.

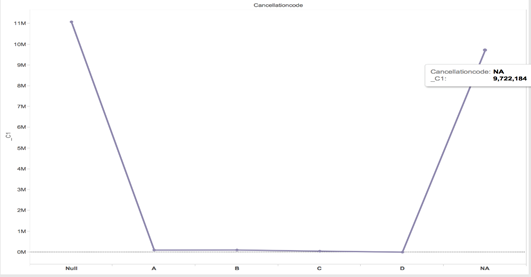
**Carrier Percentage Delays:**



8) Below chart is the analysis of cancellation of flights based on different reasons.

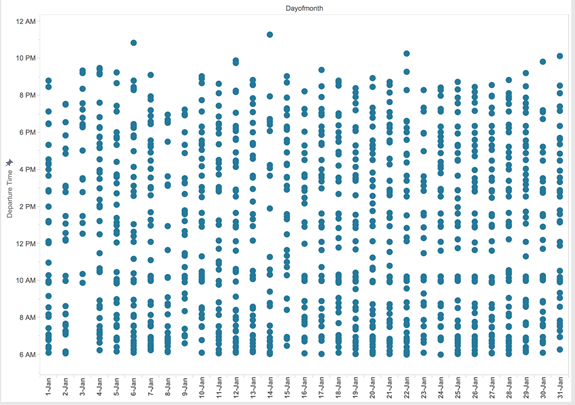
|  |  |
| --- | --- |
| **Cancellation Code** | **Description** |
| A | Carrier |
| B | Weather |
| C | National Air System |
| D | Security |
| NA | Not Known |

**Cancellation code Analysis:**

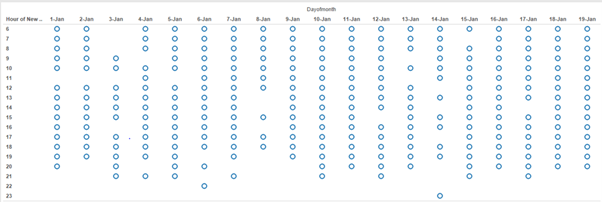


**9) Traffic Density:** Here we have analyzed by selecting data of a particular month say January in our case for a particular place ‘ROC’.

**Departure traffic density:**

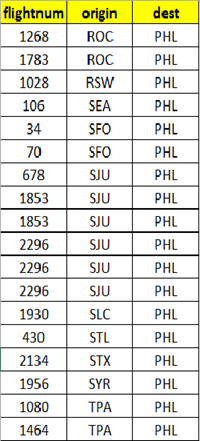


**Arrival Traffic density:**

****

**10)** Below is the snapshot of all the flights which were diverted with their flight numbers.

**List of Diverted flights:**

****

**CONCLUSION**

The following are the conclusions of our work on analysis of airline management

* Successfully analyzed different reasons for the delay.
* Successfully analyzed the flight cancellation reasons of different airline carriers.
* Successfully analyzed the traffic density which helps to add new flights for an airport or a location.

**FUTURE ENHANCEMENTS**

* A website can be created to load and analyze the data.
* More accurate results can be generated using machine learning concepts.
* Enabling ticket booking functionality to the application.
* Expanding this to other means of transports also such as bus, taxi etc.

**REFERENCES**

1. [**http://stat-computing.org/dataexpo/2009/the-data.html**](http://stat-computing.org/dataexpo/2009/the-data.html)
2. [**https://pig.apache.org/docs/r0.11.0/basic.html#Data+Types+and+More**](https://pig.apache.org/docs/r0.11.0/basic.html#Data+Types+and+More)
3. [**https://cwiki.apache.org/confluence/display/Hive/LanguageManual**](https://cwiki.apache.org/confluence/display/Hive/LanguageManual)
4. **http://download.bigbata.com/ebook/oreilly/books/Hadoop.The.Definitive.Guide.3rd.Edition.May.2012.pdf**