**Phase 3: Proposal for**

**Analyzing Flight Data**

Fall 2016

10/23/2016

Yeshwant Sai Sandeep Guntupalli

1002318

Hassan Musafer

917307

Aruna Veeramreddy

1004293

Table of Contents

1.Introduction 3

2. Problem Definition 3

3. Data Set 4

4. Modeling MapReduce Programming 6

5. Evaluation Plan 6

6. Project Plan 6

7. Conclusions 7

# Introduction

Big Data is not only a complex term but it is also the latest approach to solve complex and huge amount of data. Many people working on big data have different approaches to solve the tasks. As a great Data Scientist once said, in olden days we used oxen for heavy pulling, and when it couldn’t pull we didn’t try to create a big ox instead we used more ox. In the same way we need not create big computers to solve the problems, we should use more computers. We should remember the 4 V’s that are the basic tools for the big data. Here we will explain each one separately. First is the Volume, as there are over 100,000 flights everyday carry people from one source to destinations, there flights produce huge amount of data. This data cannot be processed with the regular file systems. So here we use Big data. Second is the velocity, this is about how speed the data is generated. In our data set each flight generates heavy amount of data and calculate that for a year it is huge. Third is the variety, which has different varieties of data like the altitude, time and so on. Last it’s the Veracity which is the accuracy of the data that is being generated. The data produced by these flights are different compared to the data generated by the conventional data sources.

1. **Problem Definition**

As we all know that flying is the major source for travelling all over the world. Many

people will travel from one part to another part of the world for many reasons like travel,

to find a better job or study. They need to get there on time that they are promised.

There will be many delays caused by many circumstances such as weather, security etc.

If just one flight gets delayed all the other flights which are supposed to land at the same

airport gets delayed. This is not good for the airline industry. People gets indolence due

to the delay of the flights. Suppose in the United States the flights get delayed very

frequently. There may be many factors for this delay. Because of these delays the airline

industry gets affected. For example, the air route between Chicago and Knoxville get

delayed by whopping 44.1%, the same is between Chicago and Wichita where the

delay is 43.9%. these statistics are for the year 2014. If one airport gets delayed by just

one airplane it causes the whole airport to get stranded. The flights that are scheduled to

arrive after the delayed flights gets more delayed. So, airports and carriers try to

minimize the delay time as much as possible.

So in order to limit these delays, we use Big Data to analyze the data that we have. By analyzing the data, we can concentrate on which department the delays are happening and try to improve on those. We are trying to get the black box data of the flight so that we can track the flights even when they are in the air. This helps us to know at what altitude the flight is travelling and also helps us track them easily. By having the black box data we can easily identify whether the flight is travelling in correct path and also lets us to know whether the flight will arrive at the time mentioned at the start of the flight.

# 3. Data Set

# We use the dataset that I have downloaded from the following website.

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

# The data set consists of 29 attributes

|  |  |
| --- | --- |
| Name | Variable |
| Year | 2008 |
| Month | 1 - 12 |
| Dayofmonth | 1 - 31 |
| Dayofweek | 1(Monday) – 7 (Sunday) |
| DepTime | Actual Departure Time |
| CRSDepTime | Scheduled Departure Time |
| ArrTime | Actual Arrival Time |
| CRSArrTime | Scheduled Arrival Time |
| UniqueCarrier | Unique Carrier Code |
| FlightNum | Flight Number |
| TailNum | Plane Tail Number |
| ActualElapsedTime | In Minutes |
| CRSElapsedTime | In Minutes |
| AirTime | In Minutes |

|  |  |
| --- | --- |
| ArrDelay | Arrival Delay in minutes |
| DepDelay | Departure delay in minutes |
| Origin | Origin IATA Airport Code |
| Dest | Destination IATA Airport Code |
| Distance | In Miles |
| TaxiIn | Taxi in Time in Minutes |
| TaxiOut | Tax Out Time in Minutes |
| Cancelled | Was The Flight Cancelled |
| CancellationCode | Reason for Cancellation (A= Carrier, B = Weather, C= NAS, D = Security) |
| Diverted | 1 = Yes, 2 = No |
| CarrierDelay | In Minutes |
| WeatherDelay | In Minutes |
| NASDelay | In Minutes |
| SecurityDelay | In Minutes |
| LateAircraftDelay | In Minutes |

# There are the attributes of the data set that we have taken. By using this Dataset, we can know the main reasons for the delay and from this we can easily compute the large data.

# 4. Modeling MapReduce Programming

# MapReduce is a programming technique for analyzing the data that does not fit in the memory. MapReduce consists of two parts Map and Reduce. In the Map part, the Map algorithms are applied to individual chunks of the data file that we have. In the Reduce part, the results of all the applications of Map are combined in the application of a Reduce function. The framework takes care of scheduling tasks, monitoring them and re- execute the failed tasks. The framework sorts the outputs of the map and then assign them as inputs to the reduce function.

# As in the case of our project, the purpose is to find the best airport and the worst airport in terms of the total number of flight delayed. In the mapper, the key will be the name of the airport, and the value will be the comparison between the actual departure time and the scheduled departure time. Therefore, if the real departure time more than the scheduled, then diffidently there is a delay. In the Shuffle and Sort phase, all the intermediate key/ value pairs will be organized and feed to the Reducer. In the reducer phase, it will combine all the in-between key/value pairs and producing the result. The airport has the largest delay will be the worst, and the airport has the smallest delay will be the best.

# 5. Evaluation Plan

# As we know we use a large data set, we should use know how to collect data. As we use a CSV file we should transfer the huge file in to Hadoop using the get command. The main goal is to how we can compute the data and get the required output. As we use MapReduce, Hive, Spark we need to have three separate complex queries which we have to solve using 3 of them. We use an existing data set that we acquired from the statistical website. We analyze the data for the year 2008. If we want to increase the data set, we can use the previous year’s data as well. The website consists of data from 1987 to 2008. At present we are trying to analyze the single year’s data. We can use the following methods to analyze the data that we have acquired.

# Map Reduce: In the map-reduce programming, we are going to find what is the best airport and what is the worst airport.

# Hive and Pig: In Hive we are going to discover what is the optimal time for travelling? However, in Pig we will find what the best airline to book.

# Spark: In Spark, we are going to find what kinds of weather conditions cause flights get delayed.

# 6. Project Plan

# The project plan is that we have to use the data set and integrate it with Hadoop and solve the complex queries using MapReduce, Hive and Pig and Spark. In the next we try to solve some queries using the MapReduce function. The main obstacle will be to send data in to HDFS. Then by writing the MapReduce function to the data that we have we can solve queries.

# 7. Conclusions

# The airline data set is one of the most popular problems that are still interesting by researchers over the time so that they can optimize the flight industry and make it more reliable using distributed computing software and models.  In this project, we are going to analyze the flight data set and get knowledge on what is the best airport and what is the worst airport regarding to the number of delayed flights that have happened in a particular airport using big data tools of programming software such as Map-Reduce. Another important thing to mention is what the optimal time for traveling or in other words, what is the time that causes lower flights get delayed? And what is the best airline to book using Pig and Hive. Lastly, we are going to find what kinds of weather conditions that affected flights to become delayed and observe a relation using Spark since our data set are rich with so many details about the delayed flights.