**Air Traffic scaling, delay- analysis and prediction using PySpark on Hadoop Cluster**

Scaling, Simulation, delay analysis and delay prediction of 2019 US domestic air flights

Mohammed Zoher Guniem Megh Raj Upreti

University of Stavanger, Norway University of Stavanger, Norway

m.guniem@stud.uis.no mr.upreti@stud.uis.no

# ABSTRACT

As more commercial airplanes take off to the sky, data generated from air traffic is rapidly increasing in both size and complexity. According to the International Air Transport Association (IATA), Air travel industry is anticipated to annually grow by 3.5% in the next two decades. This growth raises big questions on how to gather, store, analyze and use the data generated from the those flying cities in the sky. In this research we start by using flight route information to produce a scalable and realistic simulation model, and in the second phase we analyze the delay distribution and then put some machine learning algorithms under test so we can have a better scientific foundation on what machine learning algorithm can produce better and more accurate prediction on flight delays.

# KEYWORDS

Hadoop, PySpark, Power BI, Python, Aviation, Scaling, delay, Prediction, Air traffic, Machine learning, Random forest, Logistic Regression, Decision Trees.

# 1 INTRODUCTION

Over the last few decades, air transport is increasing in popularity because of its speed and comfort which eventually increase the traffic in the airspace. With the great increase in air traffic comes a large increase in the demand for airport and airspace capacity. However, airspace and airport capacity cannot keep increasing at a rate necessary to match the rising demand.

During peak hours, the demand for resources in both airport and airspace is at is highest. Some of the most important resources are:

* Good trained human resources.
* Take-off and landing slots.
* Spacious airspace.
* Available Gates at airports.
* Available taxiways and runways.
* Many other factors.

Airspace congestion and flight delays are two of the most important bottlenecks factors that limits the available resources and causes multiple unhealthy side effects on both the operation of air industry and thus the growth of the economy.

It is therefore crucial to have a good trained human resource to manage the airspace and ground operations to avoid leak of capacity in air industry. In this paper we shall investigate the possibilities of using big data technologies to solve such challenges and limitations. Some of these cutting-edge technologies are:

* Apache Hadoop Clustering, which uses the MapReduce programming model and a network distribution to help solving problems related to big data and its demanding computation.
* Apache Spark framework is a cluster-computing framework that offers an interface to program entire clusters with implicit data parallelism and fault-tolerant. In this project we use the powerful python language to work with spark by its popular library PySpark.
* We will also be using some other python libraries and software’s like “basemap”, “matplotlib” and Power BI to further help and assist our research.

# 2 BACKGROUND AND MOTIVATION

# 2.1 Air Traffic Scaling and Simulation

Today’s modern aircrafts are equipped with numerous sensors which measures the performance and different states of each part and system of the aircraft. From the very basic of flying parameters like speed, altitude and location to more detailed data like temperature and pressure from the airplane’s engines and cabin. Data from each airplane is often sent to datacenters on the ground for use in maintenance and troubleshooting. Most airlines also store this same data about each aircraft for later use in analysis and simulation.

Through the age of aviation, Simulation has had great benefits in training staff to handle the routines, procedures and challenges of the aviation industry. All trainee pilots and airspace controllers on multiple levels must spend a big amount of their training in simulation environments. Such simulations enable them to be prepared for daily handling of air traffic that never stops around the clock. Unfortunately, making a realistic simulation model is a challenge for all software engineers. Simulation software’s can either generate air traffic randomly or by using data from realistic flights. And because the simulation must be as realistic as possible, the first option is not to consider. This makes any simulation software very dependent on data from real life.

The simulation algorithms must have a normal and not complicated flow and low running time, which means that data must be easily scalable with low overhead. In this research, we take advantage of the known capabilities of Hadoop and Spark to design a scaling algorithm that uses data from earlier US domestic flights to provide realistic data for general use in air traffic simulation and analysis.

# 2.2 Delay analysis

Another problem the aviation industry faces is in the delays caused on arrival and departure, these delays cost both the airlines, and passengers millions of dollars every year. A report about the effects of flight delays on economy was given by the United States Joint Economic Committee (JEC) and it has estimated that delayed domestic passenger flights cost the U.S. economy close to $41 billion in 2007 alone. The table in figure 1 below shows the distribution of delay costs between airlines, passenger and other related businesses. Understanding and predicting such delays should help all involved parts to be prepared and thus minimizing the effect of such delays on the economy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Airline Operating Cost** | **Value of Passenger Time** | **Spillover Costs to the Economy** | **Total** |
| $19.1 Billion | $12.0 Billion | $9.6 Billion | $40.7 Billion |

Different analysis software’s can be used to draw a picture on how and what might cause the delays. Since Microsoft launched its first public version of Power BI in 2015, the interests have grown around it because of its user-friendly GUI and flexibility in handling data. Although Power BI has a relatively big collections of different graphs, maps and other visualization tools, loading a big dataset into Power BI will produce a very heavy visualization model which slows the GUI speed and therefore degrade the quality of the user experience.

This is where we can take advantage of the computational power of Hadoop Cluster computation and PySpark methods to easily extract target information about delays from big datasets and use it in fast and flexible power BI models.

# 2.3 Delay Prediction

To take the research in flight delays to the next level, we shall later apply machine learning to predict and expect such delays before they even occur based on what we already know. Hadoop PySpark has some great built-in libraries that helps runs the most popular and efficient machine learning algorithms on our datasets, and based on related work we should then test the accuracy of some popular machine learning algorithms like random forest, decision trees and logistic regression on our dataset.

An accurate machine learning algorithm can be very helpful when working with big datasets and trying to understand the hidden relations between attributes. But as we already know a machine learning algorithm can be highly accurate with one dataset but very much less accurate with another. Therefore, we should run tests where we compare different machine learning algorithms based on earlier related work and calculate their accuracy before moving on and selecting one machine algorithm to work with in the future.

We should also keep in mind that even though a prediction algorithm works with high accuracy today, this does not mean the same algorithm will necessary give high accuracy prediction in the future, this is why it is important to keep a continuous evaluation of the performance of any available machine learning algorithm on our working dataset.

# 2.4 Dataset

For this research, we had the need of a dataset that is realistic, includes detailed information about flights and convers the geographical size of nearly a continent. This why the choice has fallen on the United States of America. Since the airspace over the mainland USA is known to be heavily loaded with commercial airplanes around the clock, most of them are on domestic routes.

We are using an open source dataset which includes all the domestic flights on both mainland USA and overseas territories from the year of 2019. The dataset is downloaded from the website of the United States Bureau of Transportation Statistics. It is contained in 12 CSV files, one file per month. Its size can vary according to research need, because it is possible to only select the needed data parameters on each flight. In our case the initial size of the dataset is about 1.88 GB.

The most important advantages of this dataset are:

* Provides detailed information about the flight route from taxing into the runway until final arrival at the terminal, like actual and scheduled clock of taxi out, wheels off, wheels on, and taxi in.
* Reasons for flight delays are given in 5 clear and reasonable categories. Which gives us a clear picture on the actual causes of each delayed flight.

But the dataset is not perfect and has some issues that was discovered and dealt with under this research. Some of these issues are:

* Clock time like wheels off and wheels on is given in local time, which makes it essential to convert these into UTC times.
* Delay categories are numeric and needs to be converted into suitable interval categories for prediction.
* No airport coordinates included in the main dataset. Which makes it necessary to use some support dataset to provide this information.

The support data set comes from the same bureau and enables us to gain more information about the geographic coordinates of the departure and arrival airports, along with UTC time variation. This support dataset is of size 1.17 MB for our needs, but it comes in a separate CSV file and includes data from almost all airports inside and outside the USA, which is very useful in case of a global extension of our research.

# 2.5 Tools & Setup

The following highlights some of the tools and technologies used in this research:

* Apache Hadoop, version 3.1.1

Using a distributed cluster consisting of one master node and 3 slave nodes.

* Apache Spark, version 2.4.5

Installed on top of the Hadoop cluster and utilizes its resources.

* PySpark

A powerful library that enables us to use its effective functions to write a consistent and god structured code using python. The PySpark library takes care of translating the python code into Scala which is the native programming language of Spark.

* Power BI

A popular visualization tool introduced to the IT industry by Microsoft. And has the main advantage of analyzing data in a user-friendly, and reactive way. It has also multiple visualization tools and models that makes the analysis model more flexible to changes in the future.

* mpl\_toolkits.basemap

An open-source python library that enables us to retrieve the flight path using the great circle calculations. The library also helps with the actual simulation by plotting the results in a geographic map.

* matplotlib

A popular open source library that is helpful for making simulation and statistical plots.

* Another helpful open-source libraries

Like datetime, argparse, sys, imageio, json, time, os, shutil, cv2 and os.

# 2.6 Related Work

There are several works in the literature that focus on air-traffic and airport delays. Airline operations are highly complicated processes that are intended to regulate expensive, tightly constrained, and interdependent resources, such as the crew, aircraft, airports, and maintenance facilities. Myriads of research have been carried out on aviation planning problems such as delays, but only a few have been performed on the characteristics of flight scaling, airspace simulation, airline delays and the use of machine learning to predict flight delays. Following are some useful researches that we used to build get started on our research:

* *Agent-based Modelling and Simulation of Air Transport Technology* in 2013 by Grether and Nagel.

The research used scaling and simulation software for flight data. They normalized the data, performed pre-processing and simulated the results between inbound flights, outgoing flights. They also changed local times to UTC (Coordinated Universal Time) for consistency.

* Predicting Flight Delays in 2012 by Dieterich Lawson and William Castillo.

This research used a dataset of flights including several years, which resulted in 135 million flights. However, work on weather data only obtaining 40% recall. They used algorithms like support vector machines, Naïve Bayes, and Random forest.

* Application of Machine Learning Algorithms to predict flight arrival delays in 2017 by Nathalie Kuhn and Navaneeth Jamadagni.

The contributors to this project used machine Learning algorithms like decision trees, Neural Network and Logistic regression algorithms and concluded that the departure delays are the main factor in aircraft delay. They also performed some statistical analysis like average delays of flights and basic summary statistics.

* Predicting flight delays and cancellations using weather as a feature in 2016 by Movva, N. and Menon

The report of this research has compared algorithms like Random forest, XGboost (Extreme Gradient), Linear regression, and SVM and finally concluded with AUC (Area Under Curve) 0.81 highest among all algorithms.

# 3 AIR TRAFFIC SIMULATION

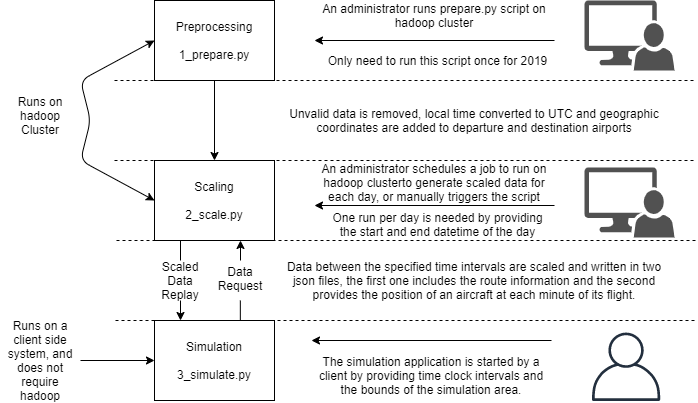
# 3.1 Design

The main goal of this part of the project is to provide simulation applications with easily scalable data that contains the following:

* The geographic position of an aircraft at each minute in a route.
* The route information of each aircraft.

The data should also be easily scalable between a time interval in the day, and according to which geographic area the simulation is based on.

Based on the requirements above we decided to split the solution into 3 main steps which are explained in figure 2.2



# 3.2 Analysis

The divided design in 3.1 should improve the running time and flow of the simulation application, this is because much of the preparing and scaling steps are performed on the distributed cluster of Hadoop which is expected to be much powerful than the client computer.

For instance, the preprocessing step takes care of converting to UTC time and adding the needed coordinates for scaling. The scaling on its side filters out those rows that does not fit into the time frame of a day and generates the full path of a flight in geographic points. This leaves the client with the only need to check if a flight is inside or outside the bounds of the simulation area.

On further examination, we found out that the preparing script takes about 13 minutes to finish for the entire dataset, and the scaling script takes only 5 minutes for one day. These running times durations can off course change to be shorter or higher based on multiple factors like background processes in the operating system and other Hadoop jobs executed on the same cluster, but we could assume having a dedicated cluster to service the preparing and scaling scripts. However, the running time of the most critical script which is “scale.py” is not expected to exceed 30 minutes per day. Which is considered a good result when considering the massive size of our initial dataset.

# 3.3 Optimization

To optimize our solution and make it run faster, we decided to isolate the preprocessing in the “1\_prepare.py” script, this relaxes the scaling script from needing a local time conversion to UTC, adding coordinates and other data validation steps.

The Scaling script “2\_script.py” can then focus on finding out which flight belong to the given time frame and calculating the great circle route path. In addition, this script is made more robust by enabling the administrator to choose any different time interval according to his need and can be run as a scheduled job for each day or week, etc. in the future.

Replacing the scaled data into json files, makes it easier for the client application to retrieve and process data which in turn improves the running time of the client application.

For further details on the final experimental phase on the application level see section 6.2 on Air Traffic Simulation, but first let’s dive into the details of out implementation.

# 3.4 Implementation

To match the design that was mentioned in 3.1, this implementation has been divided into 3 steps, these steps are explained below:

* Preparing the dataset

As mentioned earlier it is essential to deal with the weaknesses and deficiencies in the dataset before moving on with designing a scaling algorithm. To help us overcome this we design a preprocessing script that runs on the Apache Hadoop cluster using PySpark. The script does the following on the entire dataset:

- Some minority number of the flights in the dataset does not have any airtime, wheels off and/or wheels on time due to cancelation or some other reasons. It is then essential to remove these rows before proceeding on.

- Because flights can often connect two regions located in different time zones, It is difficult to use local time when considering the position of an aircraft in respect to time, its is then better to convert all times from local time to Coordinated Universal Time (UTC).

- Our original dataset does contain the sequential ID of the departure and arrival airports, but it does not contain their geographic coordinates that are needed to calculate the flight path later in the scaling step. This is solved by joining our dataset with a support dataset that provide geographic coordinates of each airport.

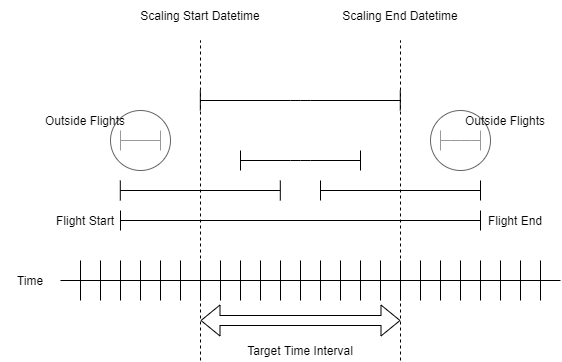
All the above steps are performed using PySpark functions and the script is meant to run only once on the entire dataset with no command line input arguments. The script is to be found under the following path and name in the project code folder. “.../1\_scaling\_and\_simulation/1\_prepare.py”.

* Scaling the dataset

The output of the preparing algorithm is written on the Hadoop file system in csv format and will be used by the next step which scales the dataset to produce a useful route information and location at each time in json format which can be used by a simulation application. The following steps explains the general structure of the scaling algorithm:

- A start and end datetime, along with a folder name where the output should be stored is provided by the user using command line arguments.

- After reading the prepared data, PySpark functions are used to filter the dataset according to the provided start and end datetimes. By keeping the rows that have a wheels off and wheels on time before, during or after the provided datetime bounds, we ensure that we do not miss flights that started and/or ended before the start and end time bounds. Figure 2.3 explains this stage.



- To be able to provide a location of the aircraft at each minute of the flight we rely on great circle calculations. The spherical shape of the earth makes it looks like pilots want to take a circle shape route from source to destination which seems longer on a traditional 2-dimensional map, but it is indeed the shortest path between source and destination. In addition to this pilots has to take into account the wind directions along the path and they do also fly according to established non-physical traffic checkpoints on the map, but often these checkpoints are placed in a path that is very much similar to a great circle.

To calculate the great circle of each flight, we take advantage of the python open-source library “mpl\_toolkits.basemap” which takes in the coordinates of the source and destination as an argument along with the number of points to be generated along the path, by letting the airtime in minute represents the number of points to be generated by the method “gcpoints” we can make sure that the average speed is taking into consideration. Average speed lays very much near the cruising speed of a flight. Which is the speed that the flight is having most of the time during its duration.

All of the scaling steps are performed using PySpark function, except the great circle calculation which we experienced to be faster if done by iterating over the remaining scaled routes using the “rdd.collect()” PySpark function.

The script is designed to produce 2 json files, one that contains the information of each different flight and the other that provide the geographic position of each flight between the given time interval.

Although it is possible to change the time intervals, it is recommended to run the script for one day at a time, either in advance or a week before. The output can then be served to the client simulation application by using ftp or by giving access to a NoSQL database like mongoDB. The json structure makes it easy to establish a NoSQL database for this purpose, but for our project we feel it is enough to try to save some time and use the scp command to be able to use and test the scaled data on the client side.

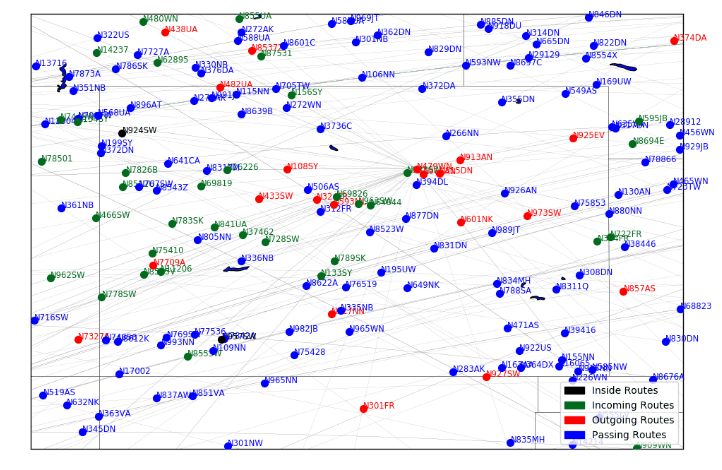
The scaling script is to be found under the following path and name in the project code folder. “.../1\_scaling\_and\_simulation/2\_scale.py”.

* Simulation

In this step, we decided to develop a small simulation application that uses the scaled data from the scaling step above, the results of this step should determine the reliability and efficiency of our implementation. Further details on the experimental setup and result is described in the section of experimental evaluation for air traffic simulation 3.5

The simulation application is to be found under the following path and name in the project code folder “.../1\_scaling\_and\_simulation/3\_simulate.py”.

And following is an example of its output images



# 3.5 Experimental evaluation of air traffic scaling and simulation

1. **Experimental setup**

We have developed a simulation application for general use purposes that uses the ouput of the Hadoop scaling step and takes in the following command line argument from a user:

* North-, south-, east- and west bounds.
* Tag of the aircraft whether it is the tail number of the aircraft or route between airports, cities or states.
* Start and end clock of the simulation.
* The name of the folder where the output should be stored.
* Whether to keep snapchat images or not
* The gif duration of each minute interval in the output.

We randomly select a day, run the scaling script on Hadoop for this day and then use the scaled results to test this simulation application. And measure its running time by dividing the day into 8 intervals each interval containing 3 distinct hours of that day. Also, for this task we select the entire airspace of the state of Colorado as a simulation area because of its strategic location in the middle of the United State, and its busy airspace. We will also be running the client simulation on a normal PC with RAM equals to 32 Gb. Which is much lower than must machines that are dedicated to simulation can offer in modern days. The simulations are also executed concurrently at the same on the same machine and operating system.

1. **Results**

The running time results of our experimental setup is giving in the table below

|  |  |
| --- | --- |
| ***Time interval in the 10th of April*** | ***Simulation running time to produce output, hh:mm:ss*** |
| 00:00 – 03:00 | 0:06:55.823885 |
| 03:01 – 06:00 | 0:06:19.043982 |
| 06:01 – 09:00 | 0:04:18.139255 |
| 09:01 – 12:00 | 0:04:18.947238 |
| 12:01 – 15:00 | 0:05:17.992771 |
| 15:01 – 18:00 | 0:06:41.680611 |
| 18:01 – 21:00 | 0:06:21.062542 |
| 21:01 – 23:59 | 0:05:57.069641 |
| Total Running time | 0:46:6 |
| Average per simulated minute in running seconds |  |

The sum of running times in the table above is around 46 minutes and 6 seconds for all 8 simulations. If we divide this sum on the number of seconds in a day, we get the amount of running time spent on each minute of simulation.

(46\*60)+6 / 24\*60\*60 ~ 0,08 running time per minute of simulation.

As we see from the calculations above, it takes a client machine around 0,08 seconds to retrieve simulation data from one minute of airspace traffic above the State of Colorado.

# 4 FLIGHT DELAY ANALYSIS

# 4.1 Design

# 4.2 Analysis

# 4.3 Optimization

# 4.4 Implementation

# 5 FLIGHT DELAY PREDICTION

# 5.1 Design

How is code executed? Where? Which input parameters are used? What is the output?...

# 5.2 Analysis

Highlight some advantages and disadvantages of you design above!

Evaluate your design and argue about it!

What can we do better? And how?

Also highlight some advantages of your design.

# 5.3 Optimization

What did you do to make your script go faster, better and why it does that? This is based on your statement and observation above

# 5.4 Implementation

What are the major and critical steps in the scripts? Explain any important calculations, or lines of code?

Write only pseudo code, or in descriptive brief words.

# 6 EXPERIMENTAL EVALUATION

# Delay Analysis

1. **Experimental setup**
2. **Results**

# Delay Prediction

1. **Experimental setup**

Explain how you plan on proving the capabilities of your design? What measurements did you use and How will you plot the result?

1. **Results**

Now explain the result with nice plots? And comment them? What can we say about the result? Are we satisfied and what could be done better in the future?

# CONCLUSION

# 8 FURTHER WORK

# REFERENCES