**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 about how to gather, store, analyze and use the data generated from the those flying cities in the sky.

# KEYWORDS

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

# 1 INTRODUCTION

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 variables 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 analyzation 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 air traffic controllers must spend a big amount of their training in simulation environment. 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 big 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 normal 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.

Another problem the aviation industry faces is in the delays caused on the ground and in the air. These delays cost both the airlines, and passengers millions of dollars every year. Understanding and predicting such delays should help all involved parts to be prepared and thus minimizing the effect of such delays on the operations.

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 has 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 to extract target information about delays from big datasets and use it in fast and flexible power BI models.

But what really takes the delay research to the next level is to be able 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.