Big Data Analysis of Public Transportation Data

A picture containing drawing

Description automatically generated

Thesis - Rough Draft

Date:

2nd September 2020

Author:

Aausuman Deep

Supervisor:

Dr. Alejandro Arbelaez

Second Reader:

Dr. Gregory Provan

Table of Contents

[Abstract 3](#_Toc47050897)

[Keywords 3](#_Toc47050898)

[Introduction 4](#_Toc47050899)

# Abstract

The data generated in the public transportation sector has an enormous potential to be cleaned, treated and processed to extract relevant information. This can assist us in making key decisions to improve those services over which a vast majority of population depends for their travels and daily commutes. In the following project I have subjected a month of public bus network data of Dublin, Ireland and performed a few big data analysis algorithms over them to harness some important details that can help the concerned department in better understanding the service that they work with and how it is performing and whether there is any room for improvement. Since the data generated at each instant in this network is quite detailed, the level of data analysis that can be performed in this regard is also quite comprehensive. Therefore, usage of a proper big data technology was on the cards and which is why I have used Apache Spark on Hadoop as my main tool here. The primary concerns related to public bus transportation networks is undoubtedly the delays with which they run on their respective lines throughout the day and the major factors which directly affect these delays like congestions. An algorithm has been initially produced to identify the most affected lines by these delays, with stepwise delve into deeper and more detailed delay analysis. A prediction has also been made on the expected arrival times of the buses in the future dates from the data that we have as our principal repository. The algorithm is also capable of handling real time data by the help of Spark streaming, to analyze data from second to second and give us valuable insights on to how the network is performing.

# Keywords

Spark, Delay Analysis, Spark Streaming, Big Data.

# Introduction

Big Data has been around for a while now. The need for processing the vast amounts of data being generated every minute in every industry is ever increasing, but in my opinion, it is also quite important to derive methods to reduce the production of redundant data in this big data. Technologies are there at our disposal, but how we effectively use them is what makes an analytical algorithm into an analytical funnel for giving us relevant data. This can then be used to perform further tasks for predictions and decision making.

One such instance where we can apply these techniques is the transportation sector. The public transportation network is one which is used profoundly on a daily basis by countless people and hence it naturally produces a lot of data to this end. From the timings of the buses leaving and arriving at stations to the geological positions of these vehicles at different times of the day, the collection opportunities are endless. In the following project we are going to concern ourselves with the data specifically generated by the Dublin public bus network.