

**School of Computer Science**

**University College Cork**

**CS6500: Dissertation Report**

**Big Data Analysis of Public Transportation in Ireland**

*(Rough Draft)*

To be submitted to:  
Dr. Alejandro Arbelaez, Department of Computer Science, UCC

To be submitted by:

Student Name: Aausuman Deep

Student Number: 119220605

Final submission date – 4th September 2020

Disclaimer

This document is a very rough draft of the final version of the dissertation. It consists of information that I may need to construct in the final phases of the project to cumulate a proper report.

Any information in the further pages is not part of the final reviewed document as of date and is not be entrusted or lied upon for further research or reference.

Table of Contents

[Introduction 3](#_Toc43828343)

[Big Data and its impact on 21st century 3](#_Toc43828344)

[Apache Spark – What it is 4](#_Toc43828345)

[Introduction to our dataset 5](#_Toc43828346)

[Environment and Control flow setup 8](#_Toc43828347)

[Core Analysis 10](#_Toc43828348)

# Introduction

## Big Data and its impact on 21st century

Placeholder - {One page of filler information about how big data technologies have been useful for data processing in recent times and how we can harness that power in the transportation sector with few examples

}

## Apache Spark – What it is

Placeholder - {One page talking about Apache Spark and its capabilities

}

# Introduction to our dataset

The data that we will be conducting our research on is that of Dublin Bus Transportation system for the month of January 2013. It records all the relevant information that has been gathered by the department with a general glance at its structure as follows.

We have a csv file for each day of the month, i.e. for 1st January 2013, 2nd January 2013 and so on.

Each datapoint (row in the CSV file) has the following entries:

1. Timestamp micro since 1970 01 01 00:00:00 GMT'
2. Line ID
3. Direction
4. Journey Pattern ID
5. Time Frame (The start date of the production timetable - in Dublin the production timetable starts at 6am and ends at 3am)
6. Vehicle Journey ID (A given run on the journey pattern)
7. Operator (Bus operator, not the driver)
8. Congestion [0=no,1=yes]
9. Lon WGS84 (Longitude)
10. Lat WGS84 (Latitude)
11. Delay (seconds, negative if bus is ahead of schedule)
12. Block ID (a section ID of the journey pattern)
13. Vehicle ID
14. Stop ID
15. At Stop [0=no,1=yes]

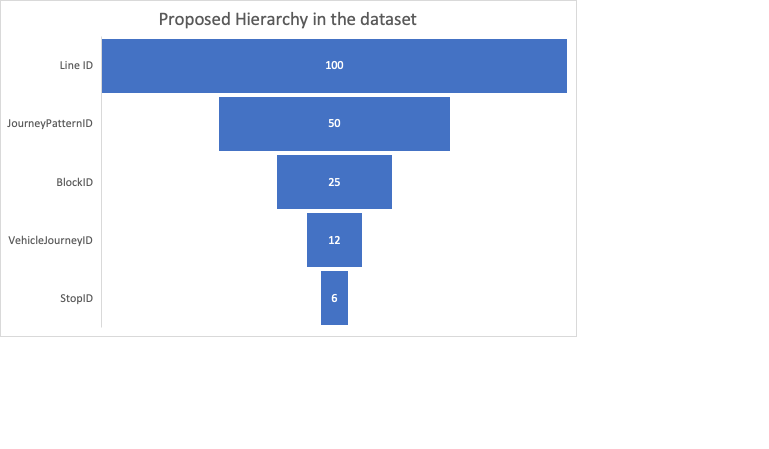
This above structure is the raw format and depicts the data gathering process conducted by them. A quick snapshot at the head of data of 1st January 2013 is as follows

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Timestamp | LineID | Direction | JourneyPatternID | TimeFrame | VehicleJourneyID | Operator | Congestion | Lon | Lat | Delay | BlockID | VehicleID | StopID | AtStop |
| 1.357E+15 | 747 | 0 | 7470001 | 31/12/12 | 3493 | SL | 0 | -6.236852 | 53.425327 | -709 | 747006 | 40040 | 7411 | 0 |
| 1.357E+15 | 27 | 0 | null | 31/12/12 | 3883 | RD | 0 | -6.233417 | 53.342232 | 0 | 27017 | 33521 | 395 | 0 |
| 1.357E+15 | 40 | 0 | null | 31/12/12 | 2226 | HN | 0 | -6.27825 | 53.416683 | 0 | 40206 | 33142 | 6071 | 0 |
| 1.357E+15 | 7 | 0 | 71003 | 31/12/12 | 6106 | D1 | 0 | -6.231633 | 53.317768 | 0 | 7019 | 43004 | 3222 | 1 |
| 1.357E+15 | 747 | 0 | 7471001 | 31/12/12 | 3531 | SL | 0 | -6.254617 | 53.355484 | -454 | 747007 | 40039 | 1445 | 0 |
| 1.357E+15 | 56 | 0 | 056A1001 | 31/12/12 | 1830 | RD | 0 | -6.233183 | 53.342201 | 0 | 56001 | 33488 | 2379 | 0 |
| 1.357E+15 | 25 | 0 | 025A0001 | 31/12/12 | 2866 | CD | 0 | -6.296867 | 53.3475 | 0 | 25007 | 33604 | 4604 | 0 |
| 1.357E+15 | 747 | 0 | 7470001 | 31/12/12 | 3493 | SL | 0 | -6.238668 | 53.425789 | -687 | 747006 | 40040 | 7411 | 0 |
| 1.357E+15 | 27 | 0 | null | 31/12/12 | 3883 | RD | 0 | -6.2334 | 53.342232 | 0 | 27017 | 33521 | 395 | 0 |
| 1.357E+15 | 4 | 0 | null | 31/12/12 | 4243 | HN | 0 | -6.279 | 53.416683 | 0 | 4001 | 43043 | 7226 | 0 |

As mentioned above there is definite hierarchy in the structured data. This information might be very useful to us when we perform complex analysis in the project as we move ahead.

1. Line ID is the definite segregation point in order to distinguish separate entities in this bus network.
2. The secondary grouping criteria within Line ID would be Journey Pattern ID
3. The third subgrouping criteria within Journey Pattern ID would have to be Block ID (i.e. sections of journey pattern)
4. The next subsequent reduction parameter will be Vehicle Journey ID as it is mentioned to be a given run within a Journey Pattern ID
5. Then finally come the stops at which the bus will stop on its run.

This may be subject to change as we identify in the future if there is any error in my assumptions, but as of now this is the hierarchy is what we will be following.



**The values mentioned above are not correct.** They will be done as part of our analysis. In the end they will depict exact number of line ID in the network, and then further values will all be the average number of journey patterns -> average number of blocks -> average number of vehicle journeys -> average number of stops

There will be a preliminary analysis script which will contain essential pieces of information, which as of now I have partially written in R (Analysis\_Script.R has been committed and pushed to Github), which will need to be converted to python.

Placeholder – {Any other piece of information that I feel relevant to our dataset can be mentioned later here

}

# Environment and Control flow setup

Placeholder – {Here we will mention in detail what all software requirements have been utilized for completion towards this project. This will a standard but shortened version of an industry standard SRS document of 1 to 2 pages also curtailing on the project and code repositories

}

# Core Analysis