# MTA Traffic Analysis and Dynamic Pricing

Tanay Varshney
*New York University*
*tanay@nyu.edu*

Utkarsh Prakash
*New York University*
*utkarsh.prakash@nyu.edu*

Ashim Aggarwal
*New York University*
*aa6911@nyu.edu*

Jay Shah
*New York University*
*jjs816@nyu.edu*

***Abstract*—Transportation System of a city forms the arteries of any city. New York City's Metropolitan Transportation Authority (MTA) is one of the largest such network. The general perception of logistical in public domain ranges between indifferent to apathetic, and improvements to the same cost funds which are hard to secure due to frequent budget cuts. This work is aimed at securing investment capital from non conventional sources.**

## 1. Introduction

Metropolitan Transportation Authority (MTA) is a public benefit corporation responsible for the public transportation of 12 counties of New York and 2 counties of Connecticut. This study focuses on the transportation network of the 5 boroughs of New York City. The main theme of this study is to secure private/corporate sources of income to supplement/subsidize the ride costs. A lower ride cost will lead to a boost in ridership across various lines, reducing the increasing stress on the road network, as well as providing a supplementary source of income. This can be implemented by allowing targeted advertisements on specific tracks sponsored by private individuals or corporations. While there are existing advertisements, their nature, at best is passive. An insight in the traffic flow and better targeted efficient advertisements are the major focus this project.

## 2. Problem Statement and Formulation

### 2.1. Problem Statement

The formal problem statement being tackled is, Leveraging Public Transportation Ridership Patterns to attract private investments and/or revenue from corporate sources. The main goal is for corporations to adopt certain sections of the existing tracks in exchange for running blanket ads. They can also instead focus on already existing high traffic areas to get the best return on investment.

One of such examples is like that of Indian cities like Delhi, Mumbai and Gurugram which gained around 10% of their operational cost from such efforts.

### 2.2. Formulation

This process can be performed by aid of external data or corporate investors who can chose to "Promote Certain Regions". An example would be if a coffee chain plans to expand into Queens, the company can subsidize fares of certain sections of lines for certain time segments in exchange of running advertisements for that section.

## 3. Datasets and Tools

### 3.1. Datasets

We are primarily looking at the MTA data but have possible plans for other datasets. Following is a complete list:-

- MTA Turnstyle data
- MTA line data

### 3.2. Tools

Following are the major tools that have been used are:

- Kakfa
- Spark
- MongoDb
- NodeJs

## 4. Architecture and Proposed Approach

This work leverages kafka, spark and mongodb to guide the data flow to achieve the goals. Figure 1 shows the layout of the architecture being developed.

### 4.1. Data Cleaning and Streaming

The data we get is a dump from MTA. Figure 2 shows the raw data. We have 96GB of data which is un-ordered, un-grouped, and full of issues. Figure 3 shows the data post cleaning.

We performed the following operations:

- Broke the data apart into separate lines
- Grouping data for each station

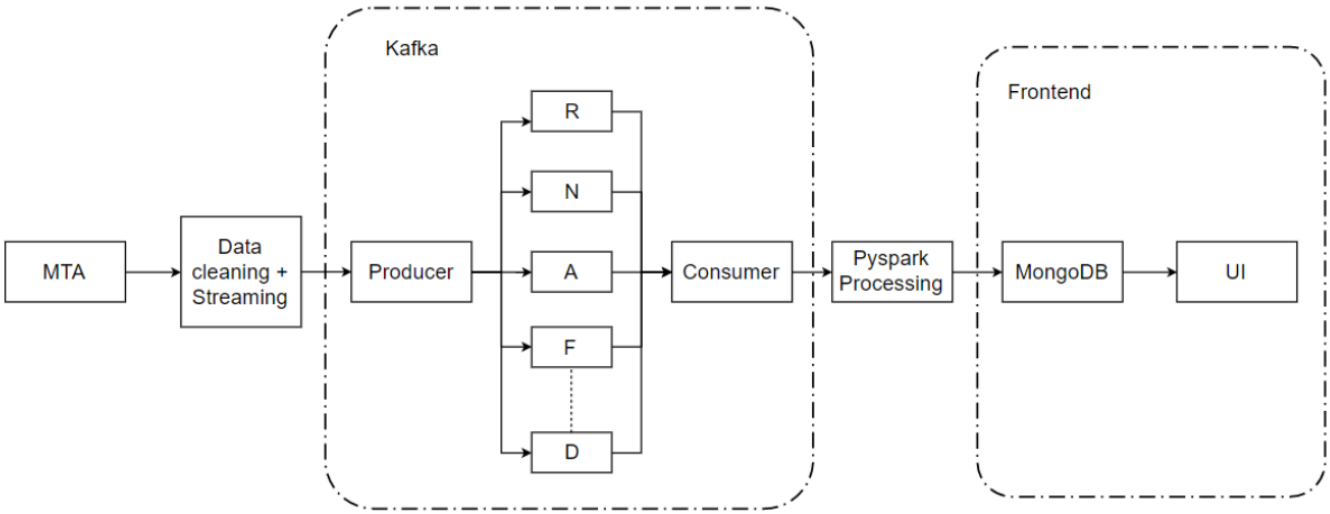


Figure 1. Architecture of project

```
C/A,UNIT,SCP,STATION,LINENAME,DIVISION,DATE,TIME,DESC,ENTRIES,EXITS
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/13/2019,00:00:00,REGULAR,0007018152,0002379496
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/13/2019,04:00:00,REGULAR,0007018180,0002379501
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/13/2019,08:00:00,REGULAR,0007018200,0002379549
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/13/2019,12:00:00,REGULAR,0007018283,0002379622
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/13/2019,16:00:00,REGULAR,0007018522,0002379696
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/13/2019,20:00:00,REGULAR,0007018849,0002379762
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/14/2019,00:00:00,REGULAR,0007019002,0002379787
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/14/2019,04:00:00,REGULAR,0007019026,0002379792
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/14/2019,08:00:00,REGULAR,0007019045,0002379811
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/14/2019,12:00:00,REGULAR,0007019127,0002379865
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/14/2019,16:00:00,REGULAR,0007019332,0002379919
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/14/2019,20:00:00,REGULAR,0007019562,0002379967
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/15/2019,00:00:00,REGULAR,0007019671,0002379985
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/15/2019,04:00:00,REGULAR,0007019678,0002379988
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/15/2019,08:00:00,REGULAR,0007019721,0002380065
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/15/2019,12:00:00,REGULAR,0007019852,0002380224
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/15/2019,16:00:00,REGULAR,0007020110,0002380277
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/15/2019,20:00:00,REGULAR,0007020841,0002380336
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/16/2019,00:00:00,REGULAR,0007021033,0002380362
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/16/2019,04:00:00,REGULAR,0007021054,0002380369
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/16/2019,08:00:00,REGULAR,0007021102,0002380468
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/16/2019,12:00:00,REGULAR,0007021261,0002380731
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/16/2019,16:00:00,REGULAR,0007021564,0002380794
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/16/2019,20:00:00,REGULAR,0007022365,0002380878
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/17/2019,00:00:00,REGULAR,0007022575,0002380908
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/17/2019,04:00:00,REGULAR,0007022585,0002380913
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/17/2019,08:00:00,REGULAR,0007022624,0002381023
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/17/2019,12:00:00,REGULAR,0007022796,0002381260
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/17/2019,16:00:00,REGULAR,0007023139,0002381326
A002,R051,02-00-00,59 ST,NQR456W,BMT,04/17/2019,20:00:00,REGULAR,0007023878,0002381403
```

Figure 2. Raw data

| STATION | DATE | TIME | ENTRIES | EXITS |
|---|---|---|---|---|
| JAY ST-METROTEC | 04/13/2019 | 4 | 514.0 | 473.0 |
| DEKALB AV | 04/13/2019 | 4 | 4082.0 | 1850.0 |
| ATL AV-BARCLAY | 04/13/2019 | 4 | 7139.0 | 8115.0 |

Figure 3. Clean data

- Calculating number of entries and exits from raw count
- Arranging the stations in the correct spatial order using external data
- Arranging the the correct temporal order

## 4.2. Producer Consumer Model

The producer consumer model relies on a Producer, a Consumer and Multiple topics. Currently there is a single producer and a single consumer but this will scale to a Producer per station and a topic for each line.

The producer is currently streaming the MTA data on to topics which then are read by the consumer, converted to a dataframe chunk and passed on to a analysis engine.

## 4.3. Analysis

The main ideas behind the analysis engine are as follows:

**4.3.1. Types of advertisement under consideration.** The analysis engine is designed to assign weights and compute the minimum bidding amounts for each station and the two different types of advertisements being offer, They are:

- In-coach Advertisement
- Bill board Advertisement

4.3.1.1. In-coach Advertisement. This type of advertisements refer to the in coach announcements and posters. The are more active form of advertisements as the riders can't avoid them. The computation for their bidding depends only on the number of entries at the station.

4.3.1.2. Bill Board Advertisement. This type of advertisements refer to the bill boards and posters on the station premise. The are more passive form of advertisements as the riders can avoid them. The computation of their bidding depends on the number of entries and number of exits on the particular station.

**4.3.2. Sliding Windows and weights.**

4.3.2.1. Sliding Window. The concept of sliding windows is used to normalize the traffic in the local vicinity of every station. The window length is currently five. The reason why this idea is being used is because the analysis of traffic at one location not affected by the traffic at another location. If traffic is normalized globally the model will lose "Local Hubs". Figure 4 gives a visual overview of what the semantics look like visually.

4.3.2.2. Computing the weights. There are two different sets of weights; weights for in coach advertisements and weights for billboard advertisements

Following is the equation to compute the weights for in coach advertisements:

$$Weight_i = \sum(Numberofentries_{ij})/TotalEntries_i \quad (1)$$

where $i$ in the window and $j$ is a station in window

Following is the equation to compute the weights for in bill board advertisements:

$$Traffic_{ij} = \sum(Numberofentries_{ij} + Numberofexits_{ij}) \quad (2)$$

$$WindowTraffic_i = TotalEntries + TotalExits \quad (3)$$

$$Weight_i = Traffic_{ij}/WindowTraffic_i \quad (4)$$

where $i$ in the window and $j$ is a station in window

These weights will always be in the range of 0 to 1, which will be used as the dollar amount for the ticket rebate to be handed out.

The minimum bidding amount for rebate which is used is expressed as

$$RebateBid = Weight_i * Numberofentries_i \quad (5)$$

### 4.4. MongoDB and User Interface

To connect the pyspark analysis engine with the User interface we needed to use a NoSQL based storage platform. We are using a mongodb connector to bridge the analysis engine and UI. Figure 5 shows a basic User interface. It shows the stations, current amount to bid, and option to bid on a station.

## 5. Achieved Goals

The main goals that were achieved are:

- Develop an analytics engine to gain insights from the data
- Build a scalable platform to handle high velocity, veracity and volume data.
- Build an end to end platform to and provide a platform for bidding on

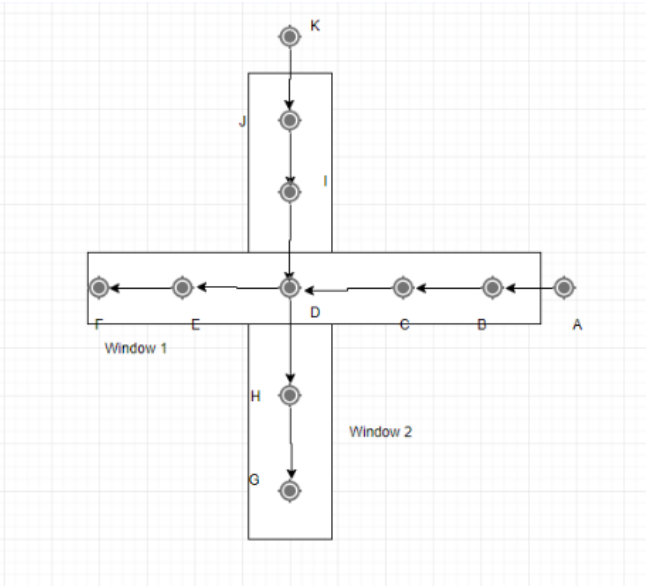


Figure 4. Sliding Window

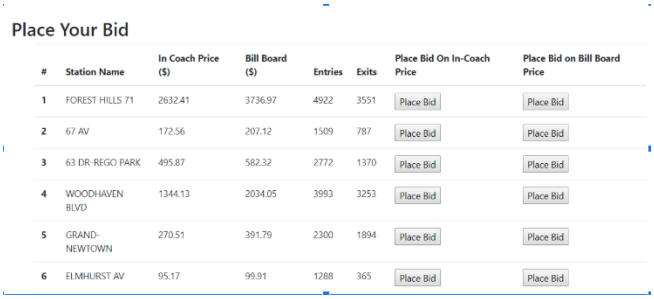


Figure 5. User Interface

## 6. Conclusion

An end to end product has been developed to act as a proof of concept. The key idea was to develop a platform which analyzes the user traffic and provide insights and take into consideration the volumetric and spatial nature of data. The key idea of sliding window mimics the page rank algorithm and seems to scale well.

## 7. Future Work

Following are the directions that are slatted to be explored.

- Stress testing the subsidy metric
- Building modular and appealing front-end for various user personas
- Providing cheaper options for local ads