CSCI 5408 ASSIGNMENT 5

By: Bhargav Dalal (B00785773) & Hardik Galiawala (B00777450)

DALHOUSIE UNIVERSITY

JULY 16, 2018

DALHOUSIE UNIVERSITY

Table of Contents

1. Objective: 2
2. Tool Selection: 2
3. Data Loading: 2
4. Data Cleaning: 4
5. Dashboard: 4
6. Output: 7
7. Code Submission: 8
Bibliography 8

1. Objective:

The main purpose of this assignment is to practice data analytics using MS Azure Data Lake and data analytics dashboard tools. The purpose of using Azure Data Lake is to store the raw and obtain the desired data in any kind of format using U-SQL queries. While, the purpose of using tools is to create data visualizations from the data, to display the intelligent information from the data.

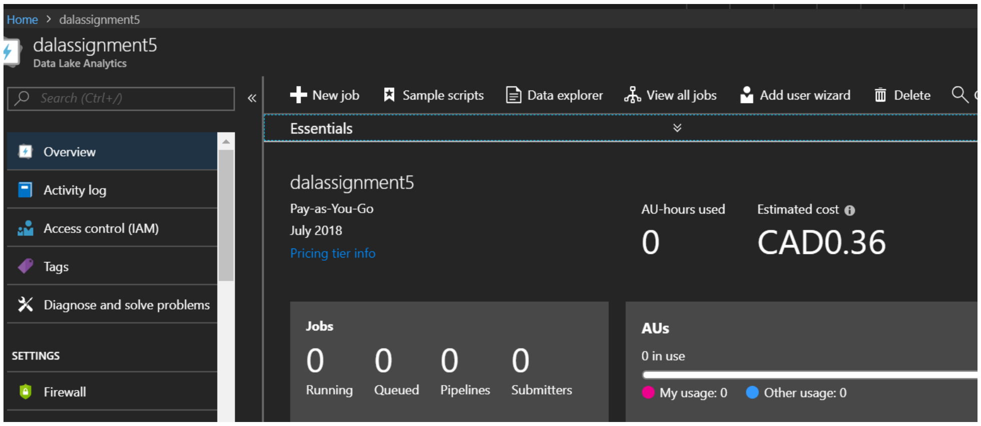
2. Tool Selection:

The tool selected for creating the dashboard was MS Azure Data Lake for storing as well as cleaning the data and Microsoft PowerBI for visualization. MS Azure Data Lake is very similar to Data Warehouse. But in MS Azure Data Lake, data can be stored in the raw format while the data requires to be normalized for storing it in the Data Warehouse. Thus, Azure Data Lake can provide a great platform to save the big data in the raw format and then it can give the capability to straight away to use it for the data analytics purpose.

PowerBI is a data visualization tool which creates the interactive dashboard. It is Microsoft's Business Intelligence tool in the market. PowerBI is the best tool for data visualization. The data visualization is quite simple, elegant and efficient in the tool. We need to just select the type of graph, fields on the x-axis and y-axis, your visualization is ready. It also provides the facility to summarize the data (count, sum, average, etc.) over the axis. It provides the filtering option too in the current visualization. The biggest advantage of using PowerBI is if you create multiple plots in the dashboard, it will automatically create relationships from the data. The tool also provides the option to change the relationship.

3. Data Loading:

The data once downloaded from the source[1], was than converted to csv format. After creating data lake instance on MS Azure platform, open it and choose the "Data Explorer" option from the menu bar.



dalassignment5
Pay-as-You-Go
July 2018
Pricing tier info

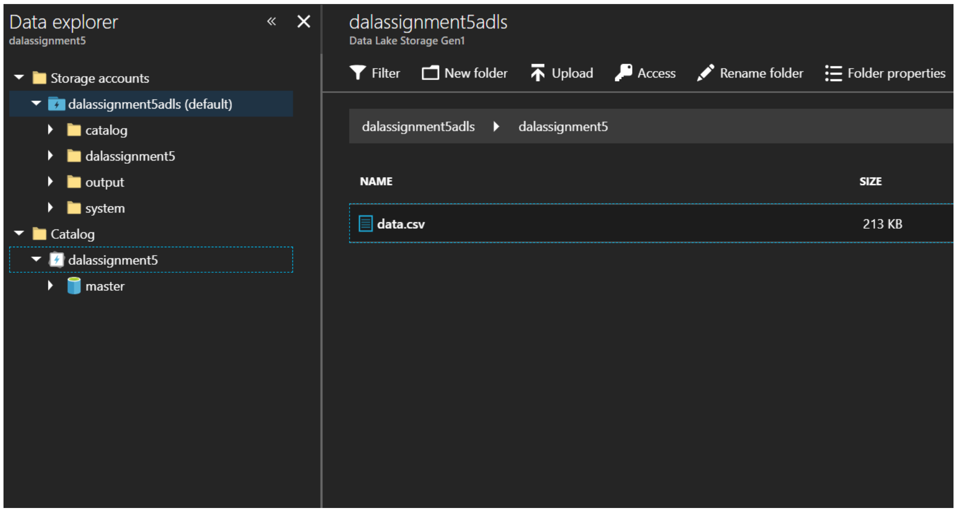
AU-hours used 0
Estimated cost CAD0.36

Jobs
0 Running
0 Queued
0 Pipelines
0 Submitters

AUs
0 in use
My usage: 0
Other usage: 0

Figure 3.1 – Data Lake Dashboard

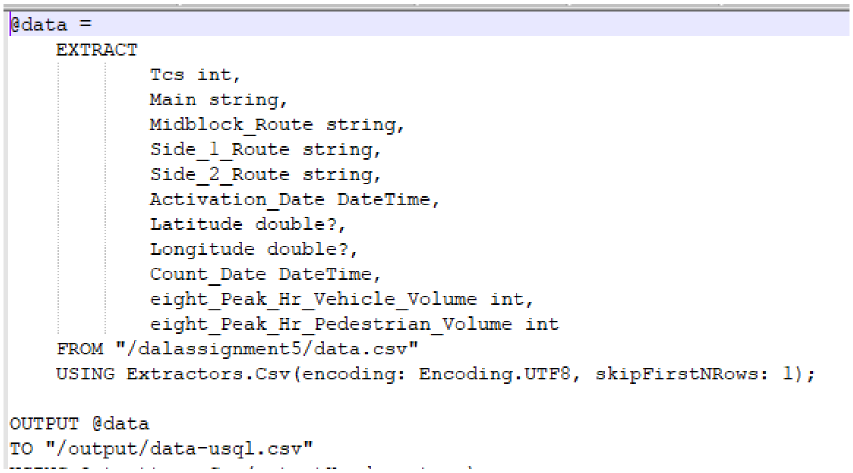
In that screen, click on storage accounts and create folder name same as you data lake name and upload the data (csv) file here.



NAME	SIZE
data.csv	213 KB

Figure 3.2 – Uploading Data on Azure Data Lake

Further when the data is uploaded successfully, go to the data lake dashboard. Click on the New Job and below scripts was used to load the data:



```
@data =
    EXTRACT
            Tcs int,
            Main string,
            Midblock_Route string,
            Side_1_Route string,
            Side_2_Route string,
            Activation_Date DateTime,
            Latitude double?,
            Longitude double?,
            Count_Date DateTime,
            eight_Peak_Hr_Vehicle_Volume int,
            eight_Peak_Hr_Pedestrian_Volume int
    FROM "/dalassignment5/data.csv"
    USING Extractors.Csv(encoding: Encoding.UTF8, skipFirstNRows: 1);

OUTPUT @data
TO "/output/data-usql.csv"
```

Figure 3.3 – Data Loading Script

The data will be than loaded and ready to use for data analysis purpose.

4. Data Cleaning:

During the extraction of data through usql, we have taken care of defining the data-type of each column explicitly with its respective data-type. Apart from that, we have not done any cleaning as the data was already appropriate. While working with the raw data we have changed column names where we added underscore (_) instead of spaces. Though this is not a data cleaning process, but we thought it is worth noting.

5. Dashboard:

After cleaning and loading the data, different tasks were given to extract the intelligent information from the data. The data was extracted using U-SQL language. There were five different tasks and so according to the task five CSVs are generated and hence five plots for visualizing each CSV (or each task). The following dashboard consist of all the five plots:

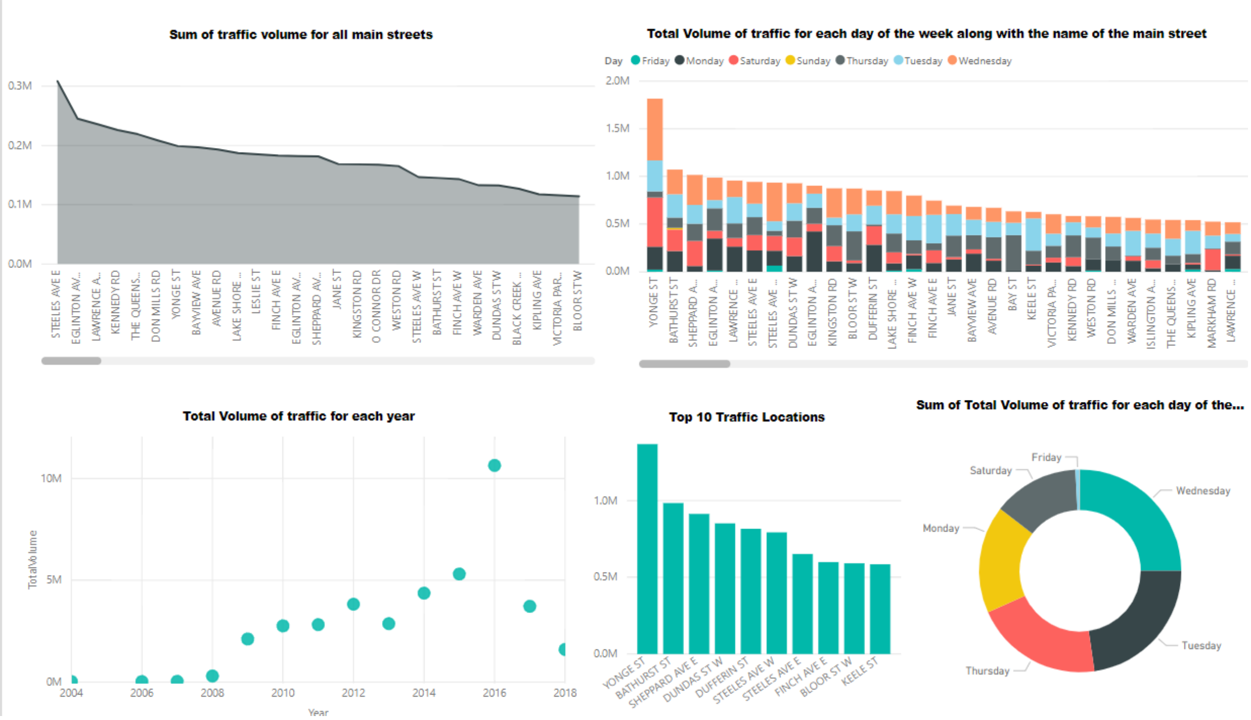


Fig 4.1 – Dashboard

Each plot in the dashboard is interactive. There were five different type of plots used for each requirement.

1. **Aggregate results based on "Main Street Name" and calculate average volume of vehicles for all available years provided in the dataset. You need to calculate one average value for each individual "Main Street Name".**

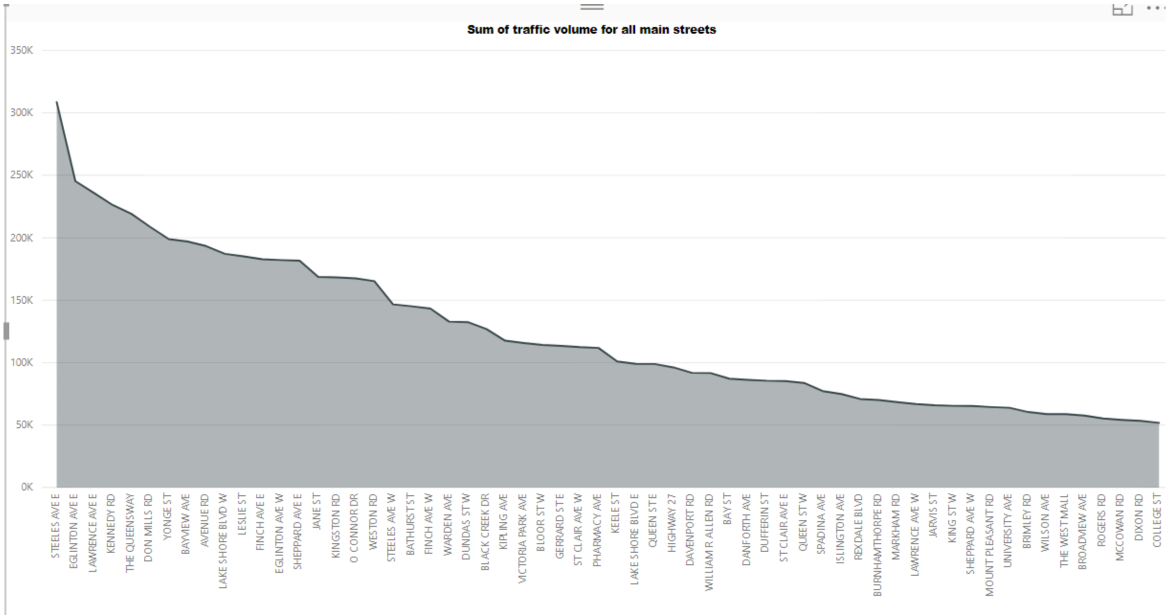


Figure 4.2 – Sum of traffic volume for all Main Streets

2. **Based on past 5 years of data, identify which 10 traffic locations are busiest during peak hours (consider both vehicle traffic and pedestrian traffic).**

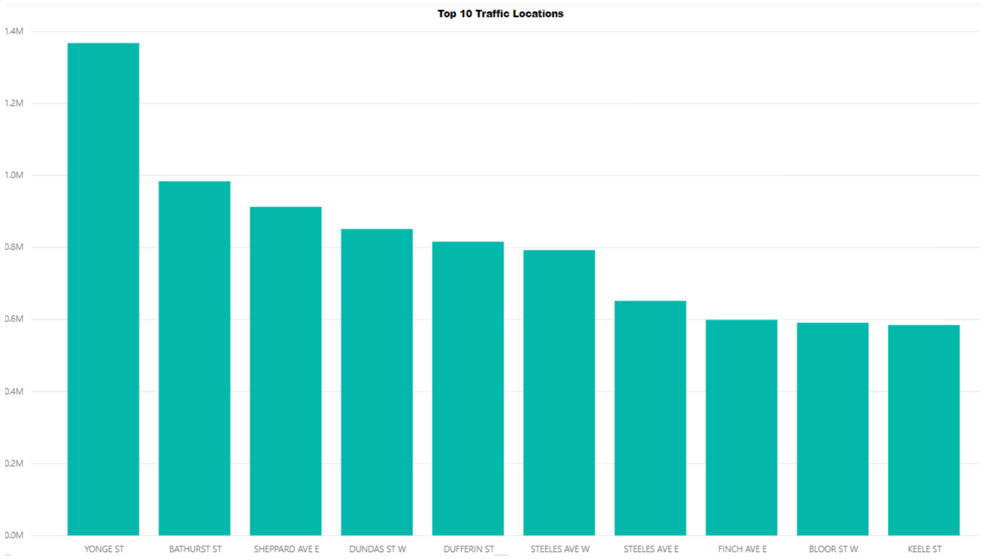


Fig 4.3 – Top 10 Traffic Location for last 5 years

3. **Aggregate results based on individual year (2017, 2016, 2015, etc.) and calculate sum of vehicles and pedestrians traffic count for all available locations. You need to do sum on all available locations and group them based on individual years.**

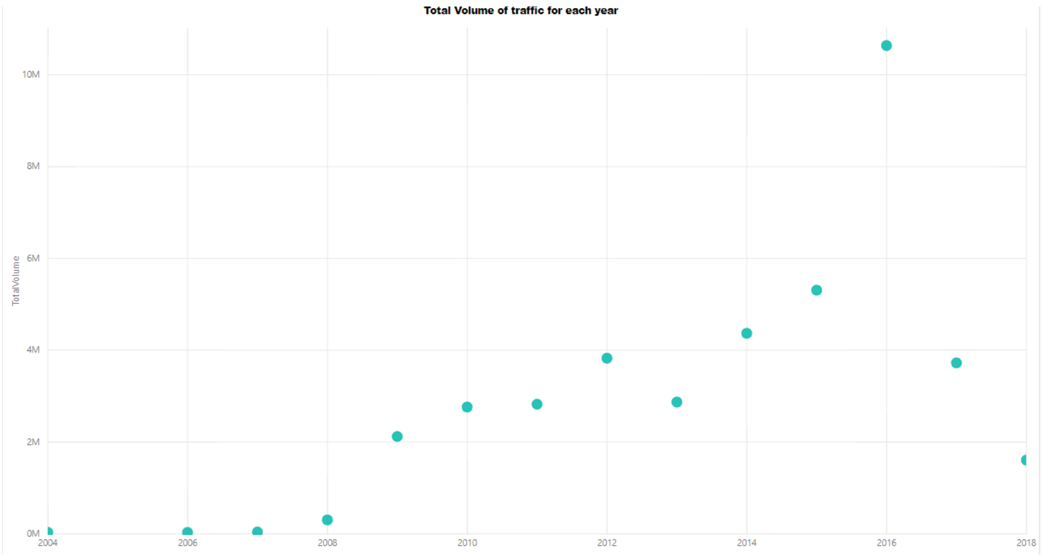


Fig 4.4 – Yearly based total volume of traffic

4. **Considering all historic years of data and all available locations, identify which day of the week (out of 7 days in a week) has been the busiest with vehicle and pedestrian traffic. Export sum of final counts for all 7 days of week for plotting.**

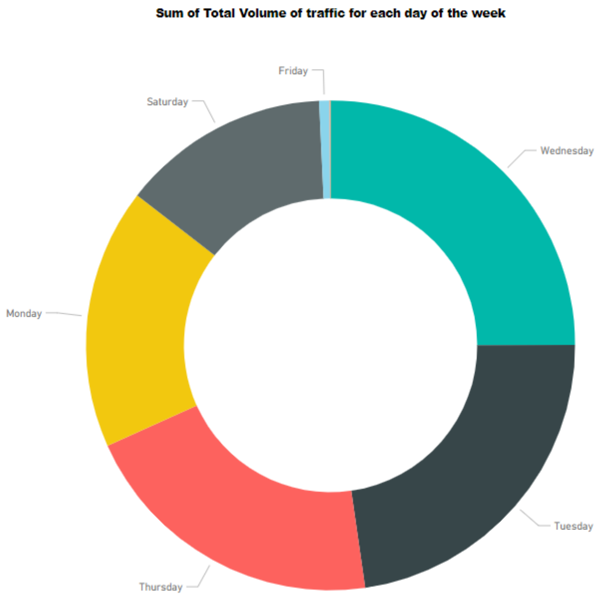


Fig. 4.5 – Sum of total volume of traffic based on the day of the week

5. **Aggregate results based on "Main Street Name", identify which day of the week (out of 7 days in a week) has been the busiest with vehicle and pedestrian traffic for each individual location. Include all historic data in observation. [HINT: GroupBy Main Street Name, Day of Week and calculate SUM (Vehicle Traffic + Pedestrian Traffic).**

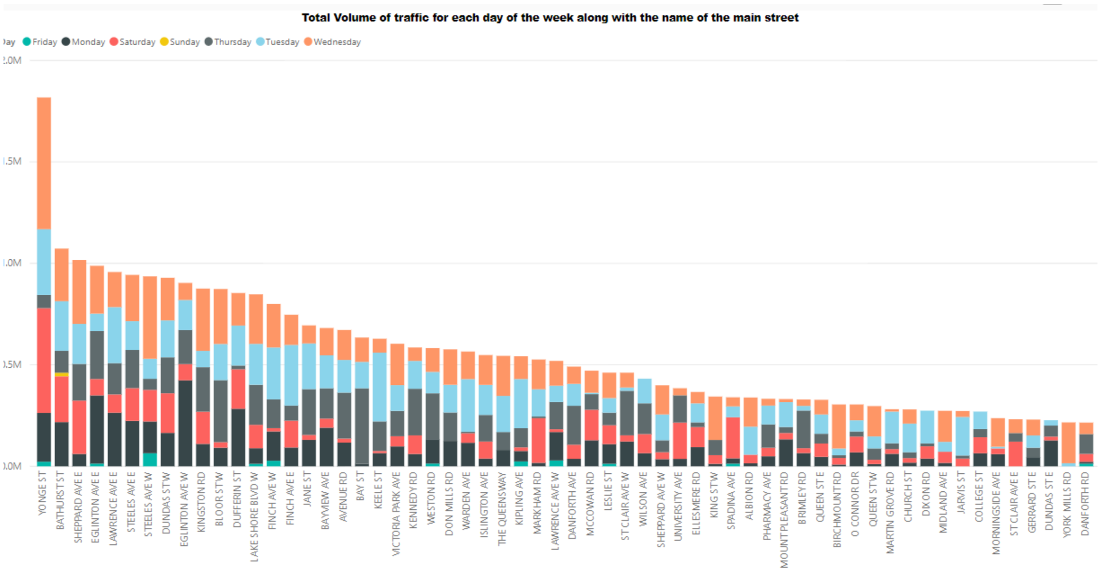


Fig 4.6 – Total Volume of traffic for each day of the week and the main street

6. Output:

The different patterns observed from above visualizations are:

The busiest street considering the vehicle as well as pedestrian traffic over all the years is the Steels Ave E. The sum of the traffic volume at Steels Ave E was more than 300K. While the second busiest street was Eglinton Ave E. with the sum of the traffic volume of almost 250K. This analysis was done for the year 2004 to 2018. But when we process the data for the last 5 years, it was surprising to see that Steels Ave E was 6th busiest street. It was Yonge Street which tops this list. Thus, it seems that Yonge street has been getting more volume of traffic over the past 5 years as compared to the Steels Ave E.

To dig it deeper, the sum of the traffic volume of the vehicle as well as the pedestrian was plotted against each year. It is evident from the figure that for years 2004-2008, the overall volume of traffic was less and steady. But after 2008, a steady increased in the traffic volume was observed until 2015. In the year 2016, there was a huge increase in the volume of traffic but after that, there is a sharp drop in traffic volume in the year 2017 and 2018.

Furthermore, each day of the was plotted (Figure 4.5) with the overall volume of the traffic on that week of the day for all the locations and for every year. It was observed that Wednesday seems to be the busiest day amongst all. Finally, to get deep insights similar plot (Figure 4.6) was generated but by grouping each day of the week and main streets. Again, the clear winner was Wednesday and busiest street on Wednesday as it was identified from figure 4.3 Yonge Street.

Thus, it seems Wednesday to be the busiest day and Yonge street seems to be the busiest street among all in recent years.

7. Code Submission:

The data was loaded in Data Lake while each output from the query is stored in CSV format was loaded in PowerBI.

GitHub Repo script for cleaning the data: Click here

In case, if the above provided link doesn't work then please copy paste the below URL provided to access GitHub repository.

https://github.com/dalalbhargav07/Data-WareHousing-Analytics/tree/master/Assignment%205

Dashboard URL: Click here

As a part of submission along with a dashboard URL, a power point presentation has been uploaded in the GitHub Repo which includes all the six plots.

Bibliography

[1] "data.csv," [Online]. Available: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#7c8e7c62-7630-8b0f-43ed-a2dfe24aadc9.