DALHOUSIE UNIVERSITY

Faculty of Computer Science

CSCI 5408 Data Management, Warehousing Analytics

Assignment - 5

Using Data Lake Analytics to capture and analyze real time traffic data set and summarize findings

Submitted To

Suhaib Qaiser

Submitted By

Aman Tewary (B00782684)

Nikhil Kamath (B00779777)

Submitted On
July 17, 2018

Table Of Content

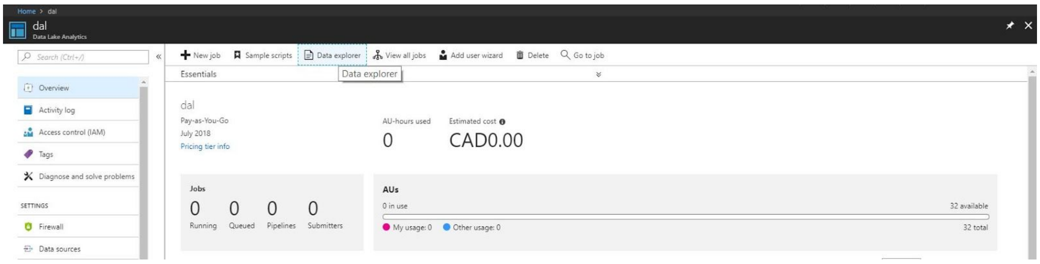
Tool Selection

In this assignment we were required to analyze and visualize data provided by Toronto Transportation Services [1]. To complete the required task effectively, we used Microsoft Azure Data Lake Services & Microsoft Power BI which is a business analytics service. We choose Azure Data Lake for data analysis and transformation because it provides storage repository that can hold significantly large amount of raw data in its original format. Furthermore, it provides analytical services which can run data transformation and processing programs in U-SQL [2] (and other analytical programming languages) over petabytes of data. It also allows to run parallel jobs with pay per job option. We chose Power BI for visualizing data because it provides interactive visualizations with self-service business intelligence capabilities. Power Bi supports Data Analysis Expression (DAX) which is a library of functions and operators that can be combined to build formulas [1]. We utilized DAX for data cleaning and data manipulation. Power BI also supports real-time filtering of data which we utilized to provide better insight on the given dataset.

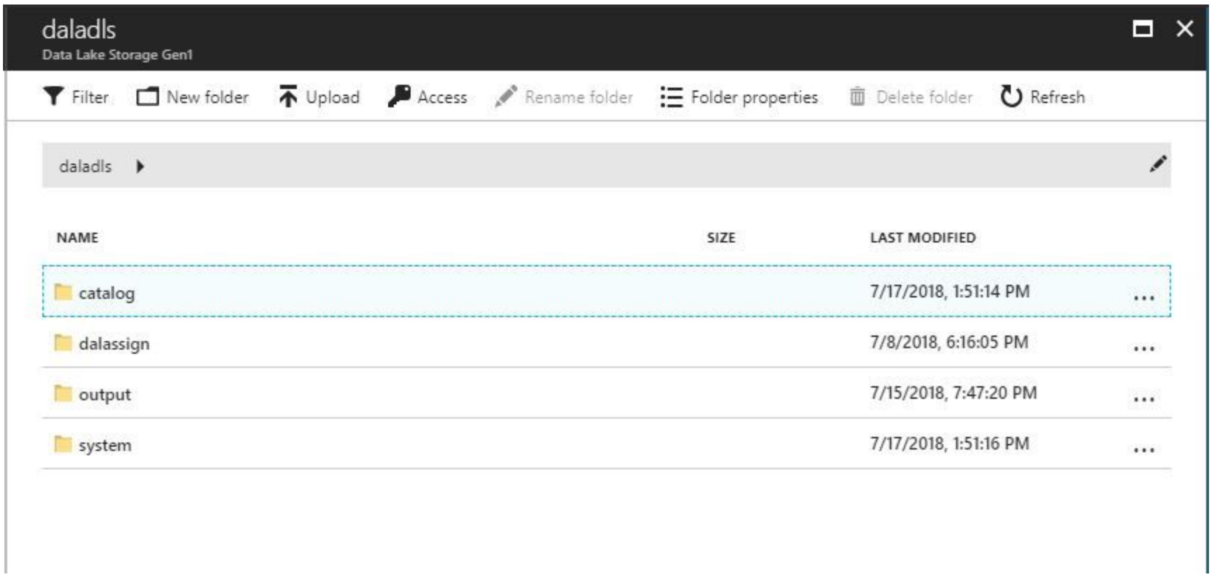
Data Loading

MICROSOFT AZURE DATA LAKE

Step 1: In your data lake dashboard, Select Data Explorer.

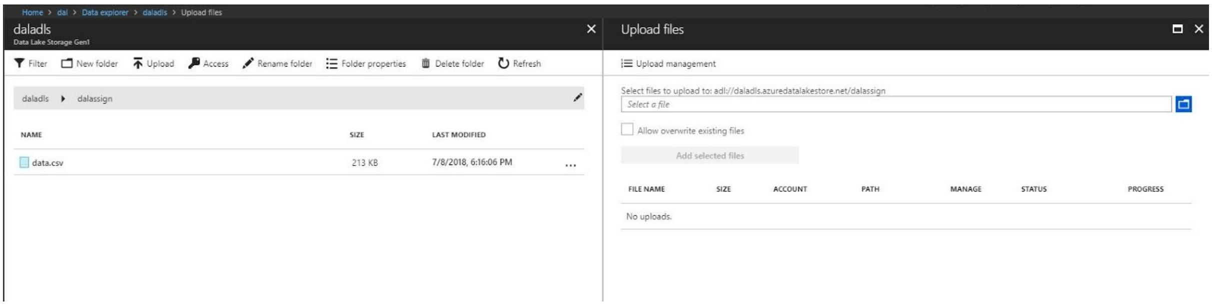


Step 2: In "Data Explorer", create two folders, "dalassign" and "outputs". The "dalassign" folder will contain raw data and the "output" folder will contain the outputs of the analytical jobs.



NAME	SIZE	LAST MODIFIED
catalog		7/17/2018, 1:51:14 PM
dalassign		7/8/2018, 6:16:05 PM
output		7/15/2018, 7:47:20 PM
system		7/17/2018, 1:51:16 PM

Step 3: Inside "dalassign" folder, select "Upload" from toolbar and choose the file you want to upload from your system. We will use this data for our analysis.



NAME	SIZE	LAST MODIFIED
data.csv	213 KB	7/8/2018, 6:16:06 PM

Select files to upload to: adl://daladls.azuredatalakestore.net/dalassign

Select a file

□ Allow overwrite existing files

Add selected files

FILE NAME	SIZE	ACCOUNT	PATH	MANAGE	STATUS	PROGRESS
No uploads.						

Dataset Manipulation

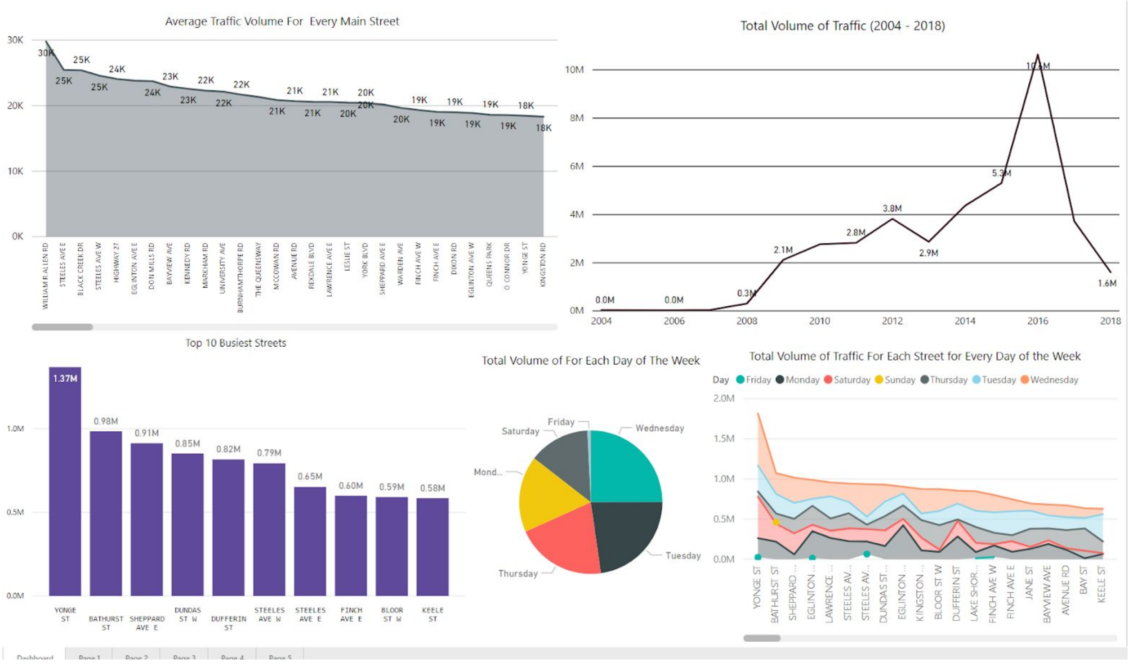
We didn't carried out any data cleaning steps. However, we had to carry out following steps to avoid syntax errors in U-SQL.

1. In our raw data, one of the field was TCS #. We converted it from all uppercase to title-case as in U-SQL column names cannot be all uppercase.
2. Most of the column names in our raw dataset had spaces between them. So we replaced spaces with underscore.
3. In our raw dataset, two column names were starting with a number - "8 Peak Hr Vehicle Volume" & "8 Peak Hr Pedestrian Volume". This is illegal syntax in C#. So we replaced numeric eight with word eight.

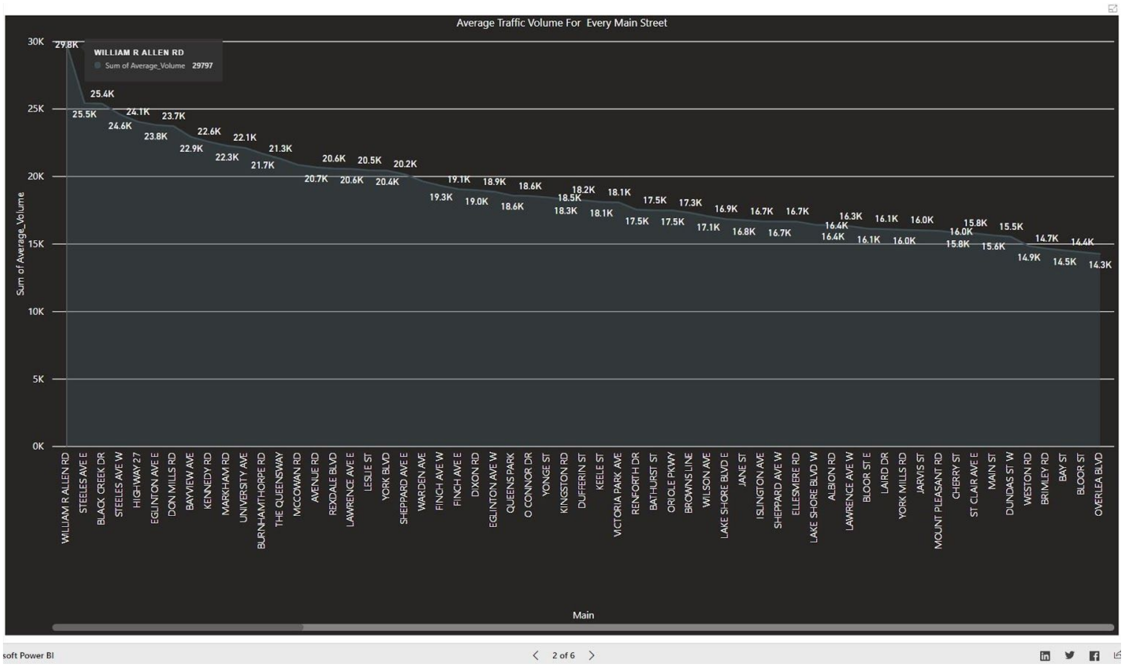
Dashboard

URL: [Link to Dashboard]()

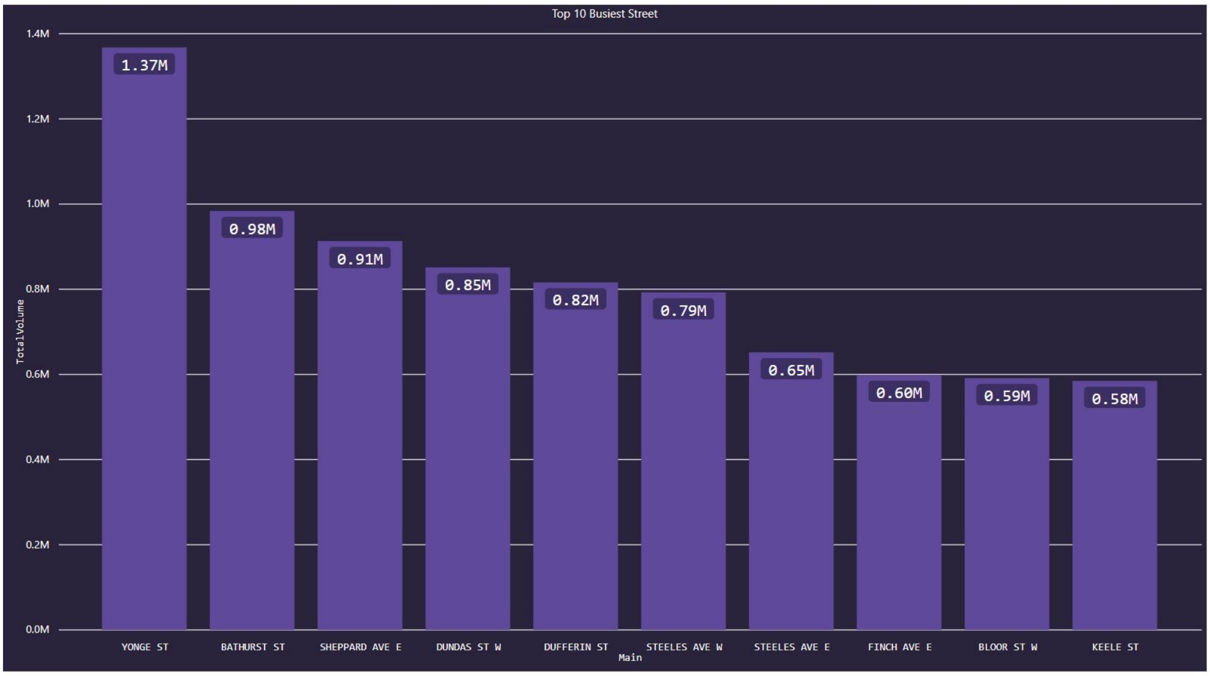
Dashboard:



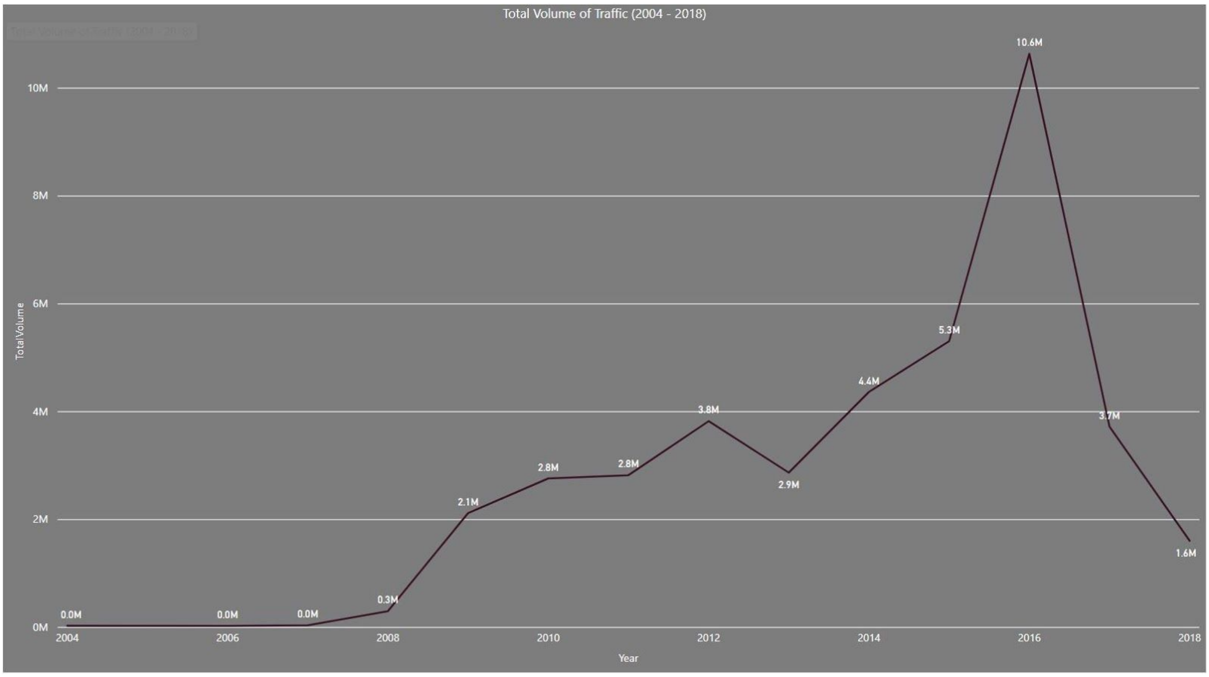
Solution 1:



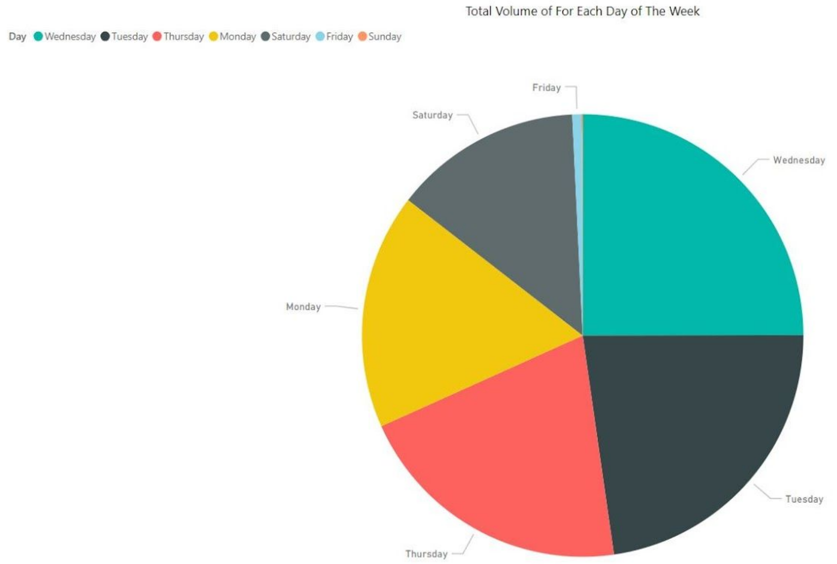
Solution 2:



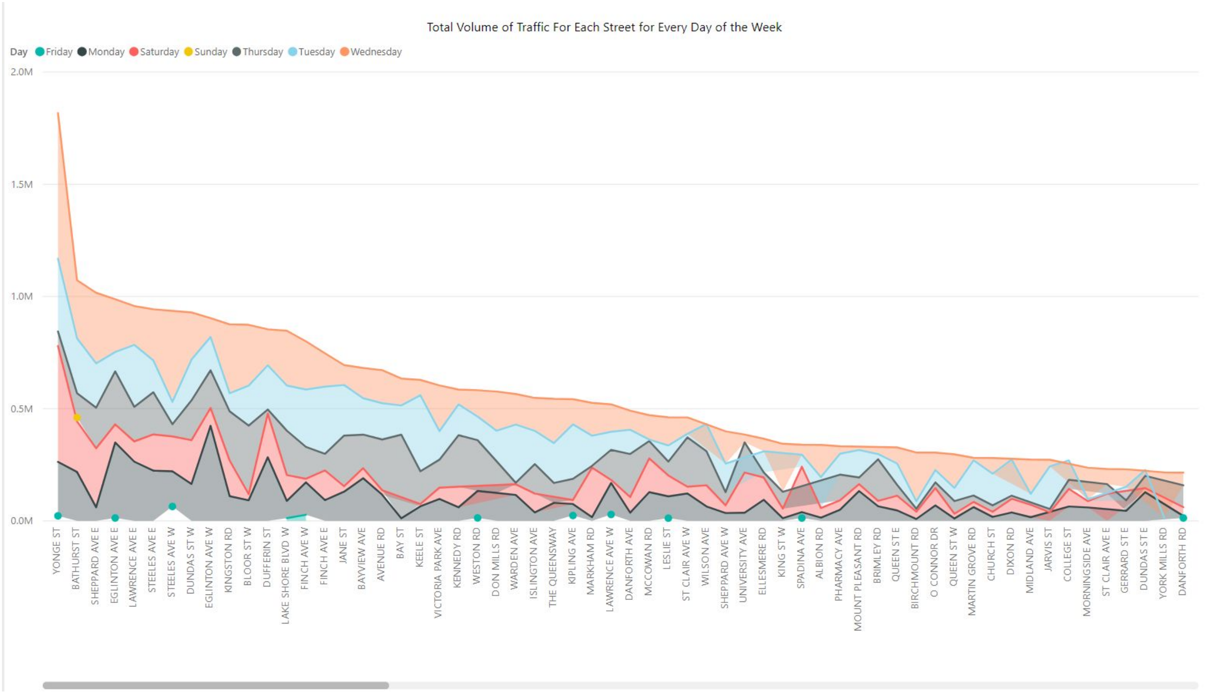
Solution 3:



Solution 4:



Solution 5:



Output

Query 1:

Values ranged from 1,081 (OLD FINCH AVE) to 29,797 (WILLIAM R ALLEN RD). The average volume of traffic decreased by 96% over the course of series, decreasing mainly in the final 133 streets. The largest single decline on a percentage basis occured in OLD FINCH AVE (-27%). However, the largest single decline on an absolute basis occured in STEELES AVE E (-4,343).

Query 2:

The minimum value is 584,742 (KEELE ST) and the maximum is 1.4 Million (rounded), a difference of 783,067. The distribution is positively skewed as the average of 815,138 is much greater than the median of 804,398. The total volume of traffic is fairly evenly distributed across all the streets.

Query 3:

Average of total volume of traffic is 2.9 million in 14 years. The minimum value is 28,616 (2006) and the maximum is 10.6 million (2016). Total Volume of traffic fell 98% over the course of all the series but increased in the final year. The largest single decline on a percentage basis occured in 2007 (-87%). However, the largest single decline on an absolute basis occured in 2015 (-5.3 million).

Query 4:

The average of total volume of traffic is 5.8 million across all seven days. The distribution ranges from 27,268 (Sunday) to 10.1 million (Wednesday), a difference of almost 10.1 million. Total volume of traffic is somewhat concentrated with four of the seven days (57%) representing 85% of the total.

Query 5:

The busiest day of the week is Wednesday. For Wednesday, the distribution ranges from 4,301 (DWIGHT AVE) to 648746 (YONGE ST), a difference of 644,445. The distribution is positively skewed as the average of 70,139 is much greater than the median 0f 27,504. Wednesday is somewhat concentrated with 46 of 144 streets (32%) representing 80% of the total. YONGE ST (648,756) is more than nine times bigger than the average across the 144 streets.

Code Submission

GITHUB URL: Assignment-5 Repository

DASHBOARD URL: Assignment-5 Dashboard

REFERENCE

[1] City of Toronto, “Open Data Catalogue,” City of Toronto, 13-Jul-2018. [Online]. Available: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#7c8e7c62-7630-8b0f-43ed-a2dfe24aadc9. [Accessed: 17-Jul-2018]

[2] MikeRys, “U-SQL Language Reference,” About Processes and Threads (Windows). [Online]. Available: https://msdn.microsoft.com/en-us/azure/data-lake-analytics/u-sql/u-sql-language-reference. [Accessed: 17-Jul-2018]