Dalhousie University

CSCI5408 Data Management and Warehousing Analytics

Assignment Report

Assignment 5

Submitted By :
Deeksha Behara :B00784704
Tushar Gupta :B00782699

Supervisor:
Suhaib Qaiser

July 17, 2018

Contents

1 Introduction

The motivation of the assignment is to learn data analytics using Microsoft Azure Data Lake. The tasks includes using a data analytics dashboard to learn real time analytics with pattern detection and explore big data analytics using the tools available on cloud.(https://github.com/bio33/Apache-Spark-Sentiment-Analysis/tree/master/Assignment

2 Experimental Setup - Task Description

The initial task implementation started off by creating an free trial account with Microsoft azure and downloading Tableau for desktop and using it to perform various visualization techniques to represent the data.

2.1 Tools, technologies and libraries used

List of tools,technologies and libraries used for the study are as follows:

1. Microsoft Azure Data Lake:

   This tool allows to explore the data and perform data analysis to perform batch streaming and interactive analysis.Microsoft azure assures scalable to perform massive parallel computing.

2. Tableau Desktop:

   Tableau[1] is a tool which assists in exploring and visualizing the data using different graphs provided by them.

   Capabilities of the tool:

   (a) Performing complex aggregations on the data.
   (b) Visualizing the data using appropriate graphs.
   (c) Allows the users to perform spatial joins on the data
   (d) Allows the users to construct geographical graphs
   (e) Can import data from different sources.

3 DataSet

The dataset[2] used for the assignment was extracted from traffic signal and open pedestrian catalog.The dataset contains the list of streets with associated volumes of the traffic and pedestrian. This data contains most recent 8 hours of peak vehicle and pedestrian volume counts at the traffic signals.

3.1 Dataset Description:

The dataset contains information like the street names, route information , geo-spatial information such as latitude and longitude and counts for the peak hour volumes for vehicle and pedestrians.

3.2 Preparing data for Analytics:

Before performing analysis on the data, the data was preprocessed. This included several steps like checking for nulls and punctuations and removing punctuation marks from the text. Additionally, the data checked for nulls and is replaced with suitable values.Certain useful information was also extracted from the data like the day and the year of the week for the given time stamp. Some of the other preprocessing steps included:

1. Trimmed Column Names
2. Removed digits from column Names
3. Removed Symbols from column Names
4. Converted XHL file to CSV

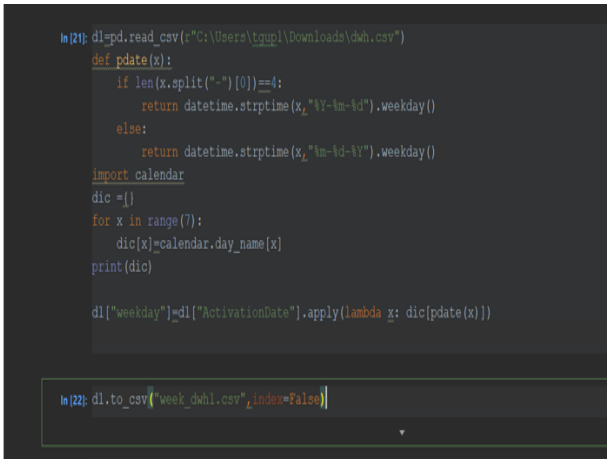

Figure 1: The query used to obtain the average volume of the vehicles for all the year provided in the dataset.

3.3 Data Loading:

Microsoft Azure provides an option to load the dataset from the local using the upload option (Data explorer) . This data can be accessed by creating a cursor in the U'SQL code.

4 Dashboard

The queries are built for each of the problems and the results are drawn using tableau.

4.1 Query1

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".

```
@searchlog=
    EXTRACT
        Tcs int,
        Main string,
        MidblockRoute string,
        Side_1_Route string,
        Side_2_Route string,
        Activation_Date DateTime,
        Latitude float,
        Longitude float,
        Count_Date DateTime,
        Peak_Hr_Vehicle_Volume long,
        Peak_Hr_Pedestrian_Volume long
FROM "/files/book1.csv"
USING Extractors.Csv(encoding: Encoding.UTF8,skipFirstNRows:1);

@rs1=
    SELECT Main, AVG(Peak_Hr_Vehicle_Volume) AS Peak_Hr_Vehicle_Volume_avg
    FROM @searchlog
    GROUP BY Main ;


OUTPUT @rs1
TO "output/temp.csv"
 USING Outputters.Csv(outputHeader: true);
```

Figure 2: The query used to obtain the average volume of the vehicles for all the year provided in the dataset.

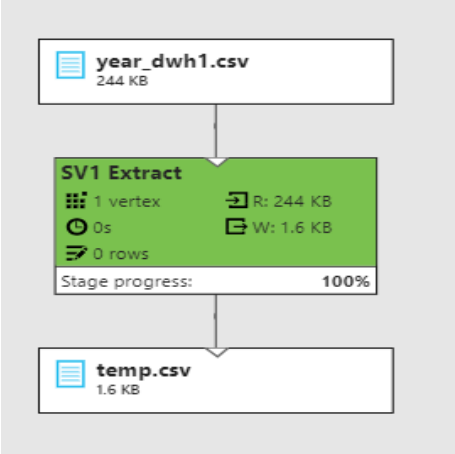

Figure 3: The picture depicting the flow of query execution for query 1

0	1
Main	Peak_Hr_Vehicle_Volume_avg
ADELAIDE ST E	9163
ADELAIDE ST W	8318
ALBION RD	16264
ALLIANCE AVE	8700
ALNESS ST	7076
ANNETTE ST	9363
ASHTONBEE RD	7898
ATTWELL DR	6957
AVENUE RD	20692

Figure 4: The results obtained for query1

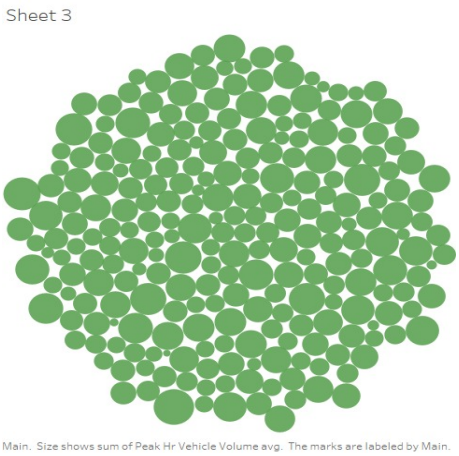

Figure 5: Visualisation done for Query 1 depicting the relationship between the streets and the calculated values of the average volumes.

4.2 Query 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).

```
@searchlog=
    EXTRACT
        Tcs int,
        Main string,
        MidblockRoute string,
        Side1Route string,
        Side2Route string,
        ActivationDate DateTime,
        Latitude float,
        Longitude float,
        CountDate DateTime,
        PeakHrVehicleVolume long,
        PeakHrPedestrianVolume long
FROM "dwh.csv"
USING Extractors.Csv(encoding: Encoding.UTF8,skipFirstNRows:1);

@rs1=
    SELECT Main,ActivationDate,(PeakHrVehicleVolume+PeakHrPedestrianVolume) AS total_traffic
    FROM @searchlog
    WHERE ActivationDate > DateTime.Parse("2010/01/01")
    ORDER BY total_traffic DESC
    FETCH 10 ROWS;


OUTPUT @rs1
TO "output/temp.csv"
 USING Outputters.Csv(outputHeader: true);

```

Figure 6: The query that returns the results of the 10 traffic locations that were the busiest during the peak hours.

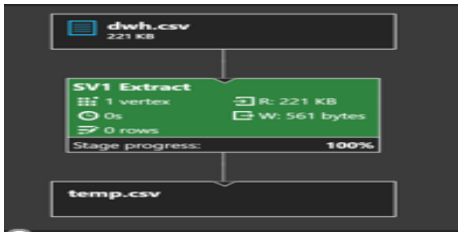

Figure 7: The Flow designed for the Query 2

Main	ActivationDate	total_traffic
VICTORIA PARK AVE	2015-06-16T00:00:00.0000000	31452
STEELES AVE W	2012-08-01T00:00:00.0000000	29596
WARDEN AVE	2010-11-20T00:00:00.0000000	26794
BAY ST	2014-03-20T00:00:00.0000000	24423
STEELES AVE W	2014-03-12T00:00:00.0000000	23875
SHEPPARD AVE E	2011-02-16T00:00:00.0000000	22753
KEELE ST	2011-05-25T00:00:00.0000000	21161
BATHURST ST	2011-12-16T00:00:00.0000000	21291
VICTORIA PARK AVE	2013-07-11T00:00:00.0000000	20899
RENFORTH DR	2015-03-31T00:00:00.0000000	20165

Figure 8: Histograms created for the permits for the given permit_type.

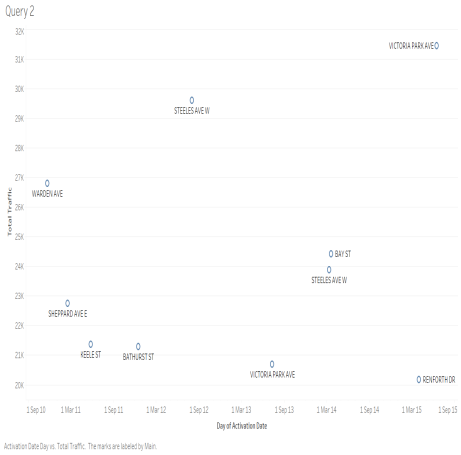

Figure 9: Visualisation representing the query results retrieved above (Query 2).

4.3 Query 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.

```
DECLARE @in  string = "/files/year_dwh1.csv";
DECLARE @out string = "/output/out3.csv";

@searchlog=
    EXTRACT
        Tcs int,
        Main string,
        MidblockRoute string,
        Side1Route string,
        Side2Route string,
        ActivationDate DateTime,
        Latitude float,
        Longitude float,
        CountDate DateTime,
        PeakHrVehicleVolume long,
        PeakHrPedestrianVolume long,
        weekday int,
        year int,
        VehiclePedestrianSum int
FROM "/files/year_dwh1.csv"
USING Extractors.Csv(encoding: Encoding.UTF8,skipFirstNRows:1);

@rs1=
    SELECT year, SUM(PeakHrVehicleVolume+PeakHrPedestrianVolume) AS traffic
    FROM @searchlog
    GROUP BY year ;

OUTPUT @rs1
TO "output/result3.csv"
 USING Outputters.Csv(outputHeader: true);
```

Figure 10: The query that returns the results of the traffic in the last few years.

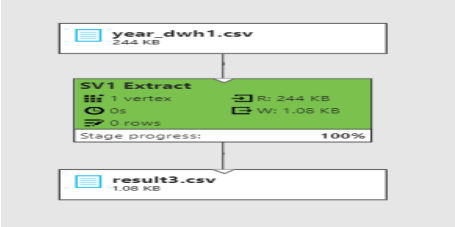

Figure 11: The Flow designed for the Query 3

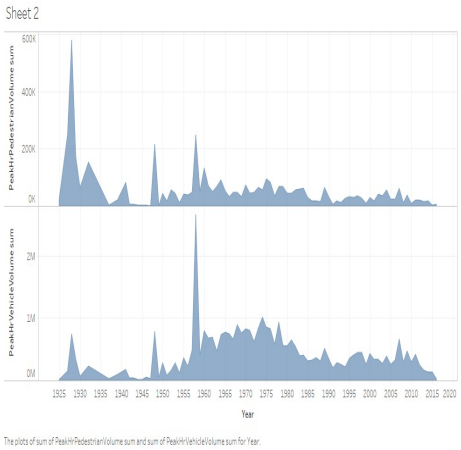

Figure 12: Individual traffics for pedestrian and vehicle.

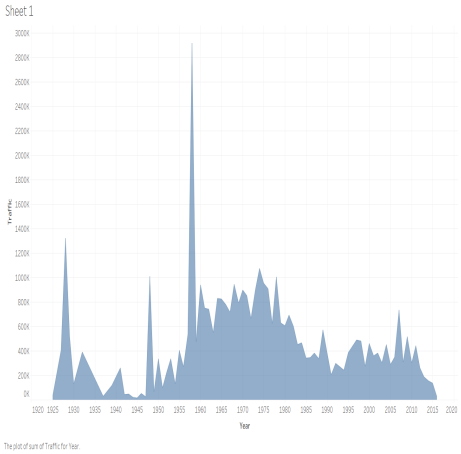

Figure 13: The cumulated sums of the individual traffics over the years.

0	1
year	traffic
1925	37733
1927	404210
1928	1321748
1929	516051
1930	139347
1932	380198
1937	20046
1939	114541
1941	230184

Figure 14: Total traffics for pedestrian and vehicle.

4.4 Query 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.

```
@searchlog=
    EXTRACT
        Tcs int,
        Main string,
        MidblockRoute string,
        Side1Route string,
        Side2Route string,
        ActivationDate string,
        Latitude float,
        Longitude float,
        CountDate string,
        PeakHrVehicleVolume long,
        PeakHrPedestrianVolume long,
        weekday string
FROM "week_dwh1.csv"
USING Extractors.Csv(encoding: Encoding.UTF8,skipFirstNRows:1);


@rs5 =
SELECT weekday,SUM(PeakHrVehicleVolume+PeakHrPedestrianVolume) AS traffic
FROM @searchlog
    GROUP BY weekday;
//    ORDER BY traffic DESC;

OUTPUT @rs5
TO "output/temp3.csv"
USING Outputters.Csv(outputHeader: true);

```

Figure 15: The query that returns the results of busiest day of the week with the highest volume of vehicle and pedestrian traffic.

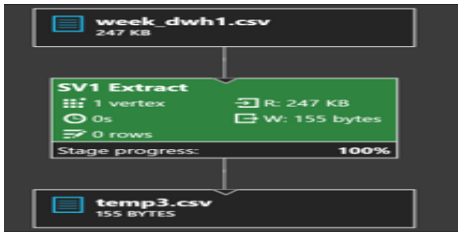

Figure 16: The Flow designed for the Query 4

	A	B
1	weekday	traffic
2	Friday	8234526
3	Monday	4636425
4	Saturday	2169649
5	Sunday	2043595
6	Thursday	8448622
7	Tuesday	6805217
8	Wednesd	8107550

Figure 17: The results for q4 are given in the diagram above.

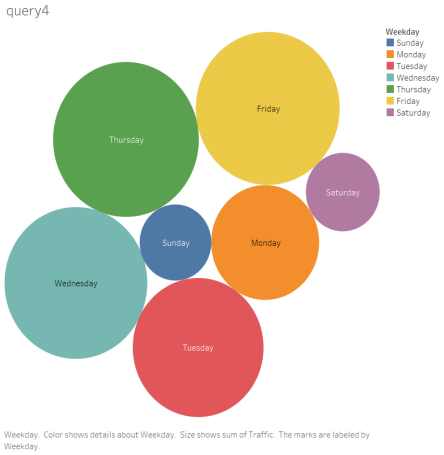

Figure 18: Visualisation representing the query results retrieved above (Query 4).

4.5 Query 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: Group By Main Street Name, Day of Week and calculate SUM(Vehicle Traffic + Pedestrian Traffic)]

```
DECLARE @in  string = "/files/year_dwh1.csv";
DECLARE @out string = "/output/out5.csv";
@searchlog=
    EXTRACT
        Tcs int,
        Main string,
        MidblockRoute string,
        Side1Route string,
        Side2Route string,
        ActivationDate DateTime,
        Latitude float,
        Longitude float,
        CountDate DateTime,
        PeakHrVehicleVolume long,
        PeakHrPedestrianVolume long,
        weekday int,
        year int,
        VehiclePedestrianSum long

FROM "/files/year_dwh1.csv"
USING Extractors.Csv(encoding: Encoding.UTF8,skipFirstNRows:1);

@rs1=
    SELECT Main, weekday,SUM(VehiclePedestrianSum) AS total_traffic
    FROM @searchlog
    GROUP BY Main,weekday
    ORDER BY total_traffic DESC
    FETCH 20 ROWS;
OUTPUT @rs1
TO "output/temp5.csv"
USING Outputters.Csv(outputHeader: true);
```

Figure 19: The query that returns the results of the day of the week that has been busiest with traffic.

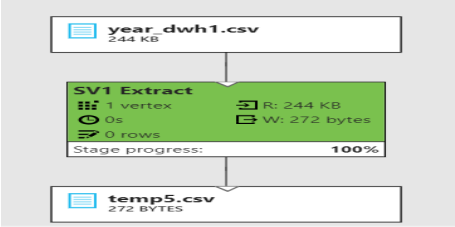

Figure 20: The Flow designed for the Query 5

0	1	2
Main	weekday	total_traffic
DUNDAS ST W	3	425123
STEELES AVE W	2	425075
EGLINTON AVE E	4	398125
YONGE ST	3	389999
YONGE ST	4	361785
EGLINTON AVE W	4	346176
YONGE ST	2	337492
YONGE ST	1	315001
STEELES AVE E	3	291525
BATHURST ST	4	284391
SHEPPARD AVE E	2	260534

Figure 21: The results for Query5 are given in the diagram above.

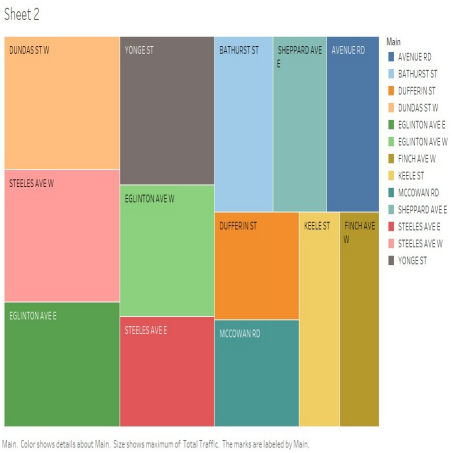

Figure 22: Visualisation representing the query results retrieved above (Query 5).

4.6 Output

1. For Query 1,based on the aggregated results, it was observed that the William street had the highest vehicle volume average traffic. The least traffic was observed at the Old finch avenue.

2. The busiest route in the past 10 years is Victoria Park Avenue with total traffic of 31500. Most of the routes have a similar amount of traffic within the range of 20,000-25000.

3. After aggregating the results by year, it was observed that the year 1958 had the highest traffic (both pedestrian and the vehicular traffic included). On the other hand, It was observed that the year 1945 had the least traffic.

4. Since the traffic is recorded during evening time , Wednesday-Friday have been observed with similar amount of traffic while weekends and early days of week have less traffic since usually people travel together in weekends while during weekdays they travel alone for work.

5. Based on the results obtained , It was observed that "DUNDAS st W" had the highest traffic for the third weekday which (Tuesday).

5 Conclusion

The data was explored and analytics was performed on the data using the data lake . Additionally visual analytics was performed on the data to represent the results obtained in the form of meaningful graphs.

References

[1] Tableau. (n.d.).TableauSoftware.(n.d.).Retrievedfromhttps://www.tableau.com/.

[2] Dataset. (n.d.).CityofToronto.(2018,July13).OpenDataCatalogue. Retrievedfromhttps://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#7c8e7c62-7630-8b0f-43ed-a2dfe24aadc9,.