Module Title: B9DA102 Data Storage Solutions for Data Analytics

Individual Report

for

Datawarehouse

**BY:**

**Akshay Kalmetar (10525530)**

**Introduction**

We were given the project to create a warehouse and provide insights using meaningful visualization. Unfortunately, this was not an easy task since Kaggle dataset was not supposed included as database source. Finally, we came across a dataset named “OurAirports” and we decided to go with this dataset.

“OurAirports” dataset represented various airports across different continent and its details. While designing the Star Schema for the warehouse I addressed an issue that some crucial details such as Date (Period), Flights flying to the airport and other require details such as flight arrival and departure time were missing. Thus, we decide to scrap this dataset and research further on more detailed dataset. Our vision of our study was clear, we wanted to provide our study on dataset related to aviation industry.

So finally, we came across a dataset **“Reporting Carrier On-Time Performance”** that was provided by **Bureau of Transportation Statistics (BTS**). We downloaded the dataset for following URL: <https://www.transtats.bts.gov/DL_SelectFields.asp>. This dataset contains detailed information about the Airline and its actual departure and arrival times.

This dataset consists of all the flights coming to and going from New York for the year 2018.This study provides analysis of the various factors responsible for delays in flight for different airlines. BTS provides this dataset on monthly basis. So, we created a database named “Airline” in SQL and merged all the csv file for different months into one single table. Further depending upon the available columns and details provided by BTS we identified our Dimension and Fact table and created “Airline\_DW” data warehouse and Star Schema.

**ETL Process**

After creating Datawarehouse, we extracted the details using ETL process. Microsoft .Net Framework SSIS package allows us to extract data from various sources such as SQL database, Excel and many more and allows to store data into required database destination. During ETL process I came across a situation where few important details such as Airline Name, Cancellation Cause, Carrier Name and Airport name was missing in our dataset. For this details BTS provide separate lookup dataset file. So, I created an ETL process which had multiple sources:

* Airline Dataset source in SQL
* Lookup EXCEL file provided by BTS (converted csv to excel format)

Also, the data was in Unicode 8 format, for this I added Data conversion control from SSIS toolbox to convert data into required format. I also used lookup control to add surrogate keys of Dimensions table into our FACT table as foreign Key. Thus, all the records were successfully inserted into our dimension and fact table.

**Visualisation using R**

For visualisation in R, I used various libraries such as RODBC, Dplyr, ggplot2, reshape2, plotly, viridis, hrbrthemes. In graphical visualisation I focused more on aesthetic of the graph so that plotting would be clear and concise, R library plotly, viridis, hrbrthemes helped me to achieve this.

In R visualisation, I created Bar Graph to plot number of flights cancelled against each month. After plotting the graph, I found that most number flights were cancelled on the month of March for year 2018. After researching online, for such huge number of cancelled flights, on CNBC news website I found out that due to Barrage of Snowstorm on the month of March flights were cancelled. It was a great experience to find out that our analysis and data on which I was providing insights were correct.

**Visualisation using SSRS**

Microsoft .Net Framework Reporting Services (SSRS) allows us to generate reports. Using this I created following report:

* Airline Cancellation Report for year 2018
* Airline Cancellation Monthly Report

While creating this report I learnt to write SQL Query using PIVOT Operator. SQL Server PIVOT operator rotates a table-valued expression. It turns the unique values in one column into multiple columns in the output and performs aggregations on any remaining column values.

**Neo4J Graph Database**

For any airline, efficiency is the key: delayed or cancelled flights often lead to unhappy customers. Planning flights is complex, and it requires a deep analysis of flight and airport data. In Neo4j I created graph to address delay problem and their cause. For creating graph database, I first visualised the require relationship and then created Cipher queries for the same.

**Conclusion**

It was a great learning experience working on flight dataset and provide valuable insight. I had prior knowledge in .NET and SQL, but I have never worked on graph database. This was a great opportunity to work on graph database and learning Cipher queries. I also learned graph plotting in R, SSIS and SSRS on which I have never worked earlier. In the end I would like to thank Shazia mam for guiding and teaching various aspects of data warehouse because of which we as a team were able to complete our assignment.