Techno India NJR Institute of Technology, Udaipur

FINAL PROJECT REPORT

on

"MOBILE SERVICE ANALYSIS"

Session 2017-2018

Submitted By:

- Erum Sanwari
- Falguni Gandhi
- Nitesh Vaishnav

IV Year,

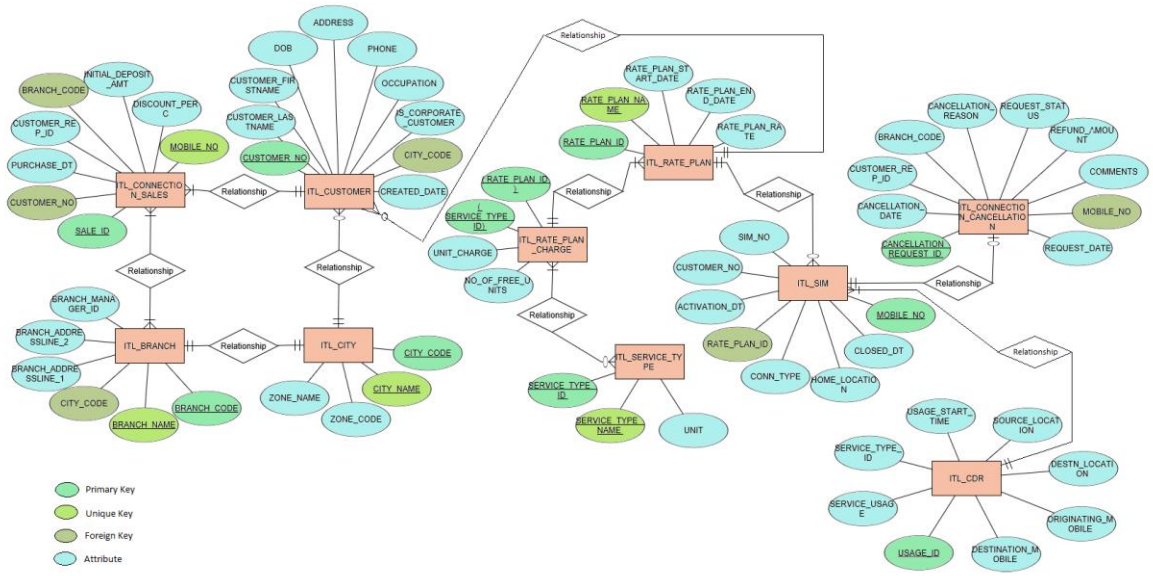
Computer Science & Engineering

TABLE OF CONTENTS:

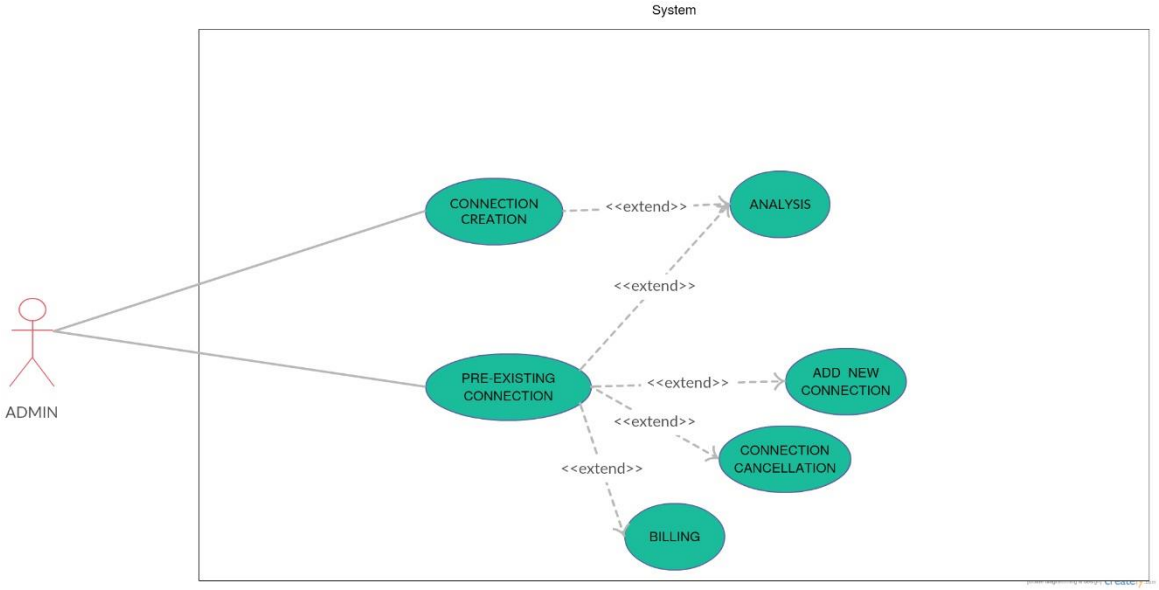
OBJECTIVE:

The objective of this assignment is to design and implement a fully operational data model for maintaining the OLAP DWH. Various Key Performance Indicators (KPIs) have been provided which are to be used for analysis and then reports be generated based on those analysis which give a proper visualization of the performance of the organization.

HIGH LEVEL DESIGN:



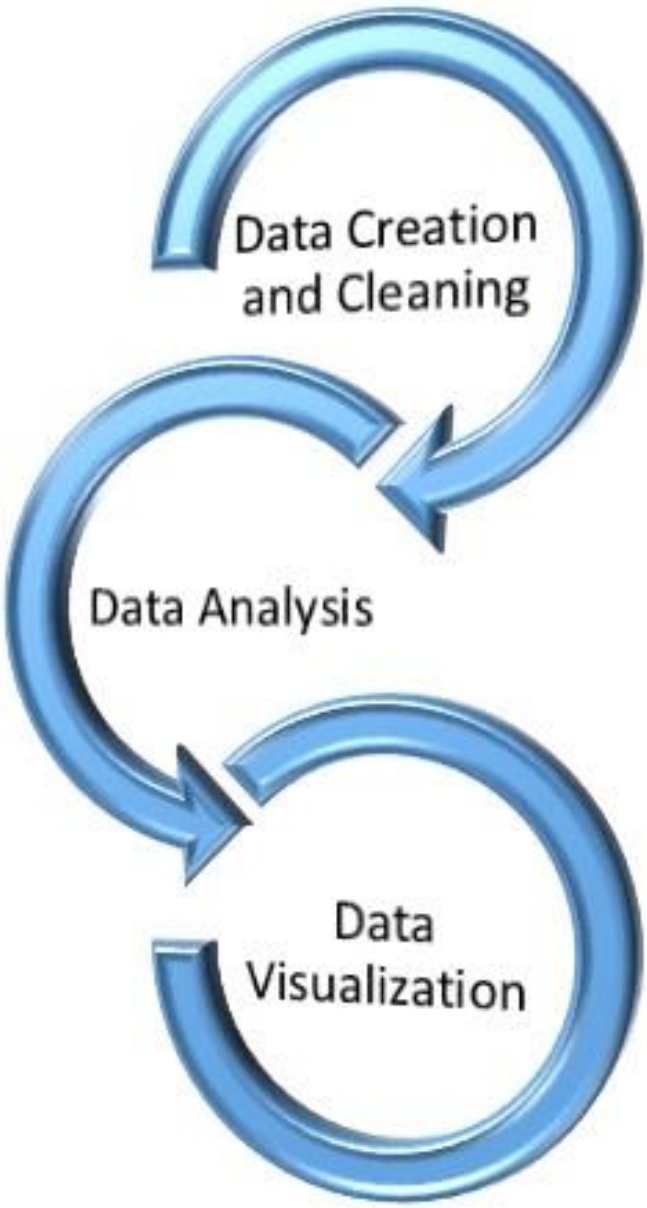
Entity Relationship Diagram



Use Case Diagram

DEVELOPMENT CYCLE:

The project development cycle typically consists of 3 stages:



DETAILED ANALYSIS:

1. Data Creation and Cleaning:

- By using the given schema, a total of 10 tables, each with around 7-8 cardinality were created.
- The data in the tables was initially maintained in excel sheets.
- Random data was generated using python scripts.
- A data set of thousand records was used to implement and test the analysis queries and obtain appropriate outputs.
- It was then cleaned and filtered for appropriate analysis fields.
- The data was aggregated and a final database containing the required data for analysis was prepared.

2. Data Analysis:

- The data was uploaded to HDFS in Hadoop.
- For query processing, Hive is used as data access tool.
- The hive script submitted as analysis document, consists of queries that are performed in order to fulfill KPI requirements.
- Hadoop used for this project runs on Ubuntu 14.04.
- The commands “start-dfs.sh” and “start-yarn.sh” are used to run datanode, namenode, secondary namenode, resource manager and node manager.
- Hive queries retrieve data from HDFS, produce results that either get stored on HDFS or local file system.

3. Data Visualization:

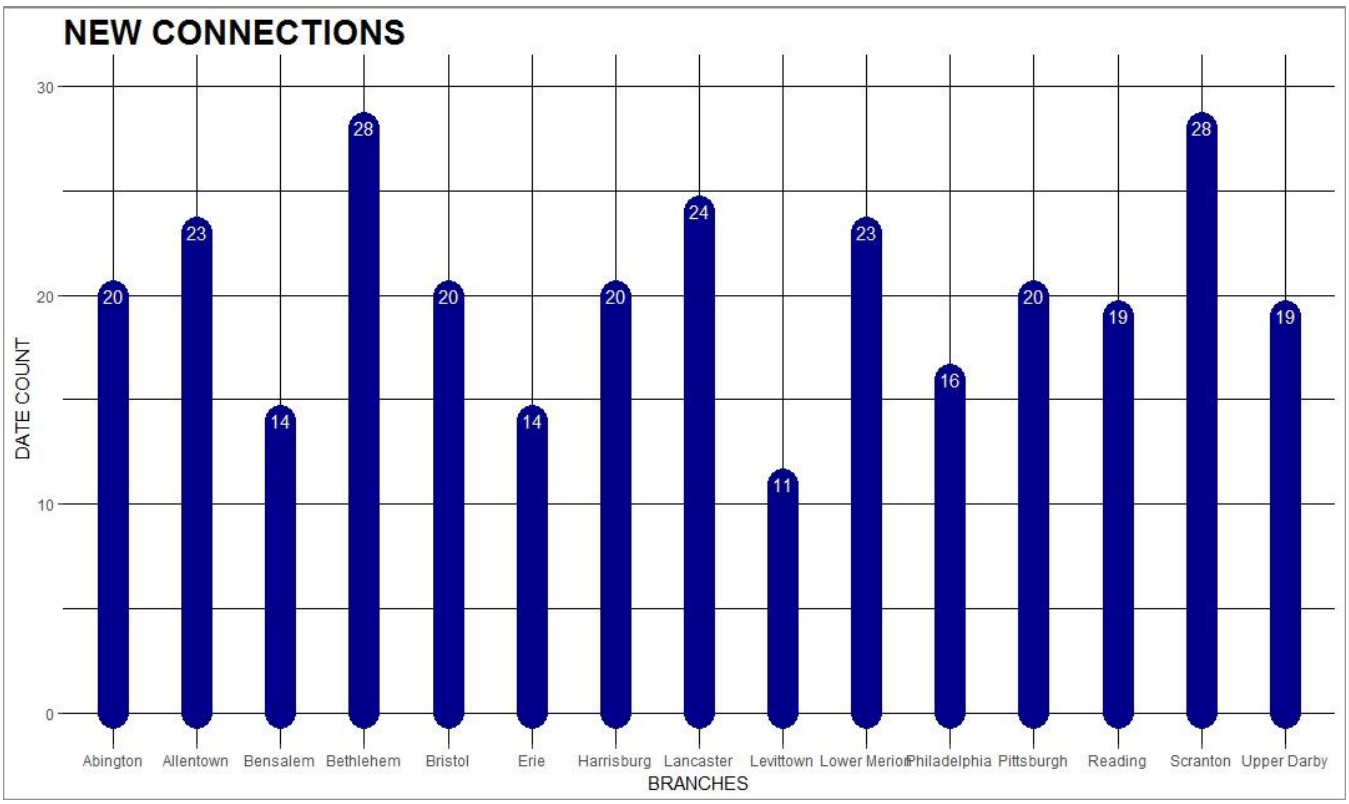
- Hive tables so obtained after successful query execution are exported from hive as CSV files.
- These are then imported into RStudio for the purpose of visualization.
- Each CSV files are imported into RStudio and visualization is performed using ggplot2 package.

VISUALIZATIONS:

Visualization of analyzed data is done in R Studio. The data obtained after query processing was imported in the R Studio, and plotting was done through ggplot() package.

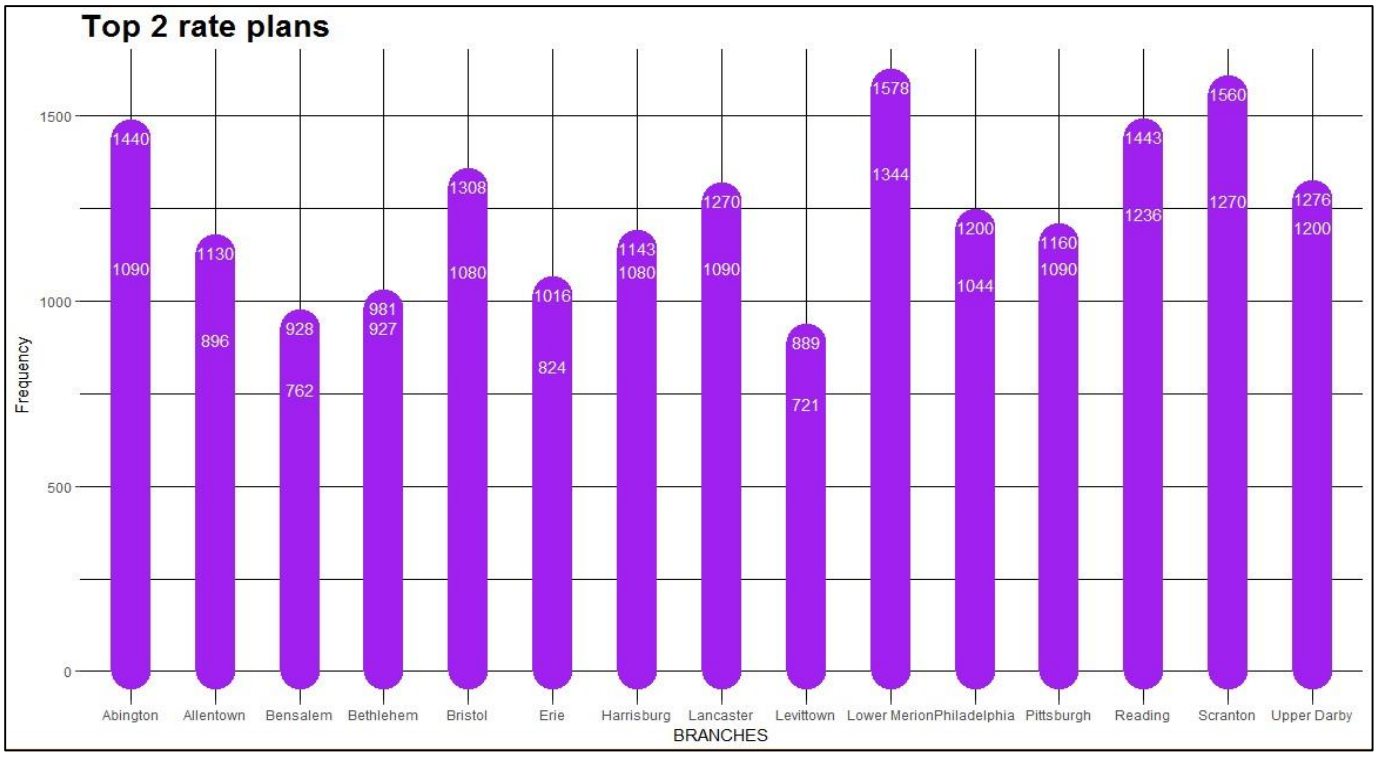
Following visualizations are in order of the various queries processed earlier:

1. ALL NEW CONNECTIONS ADDED:

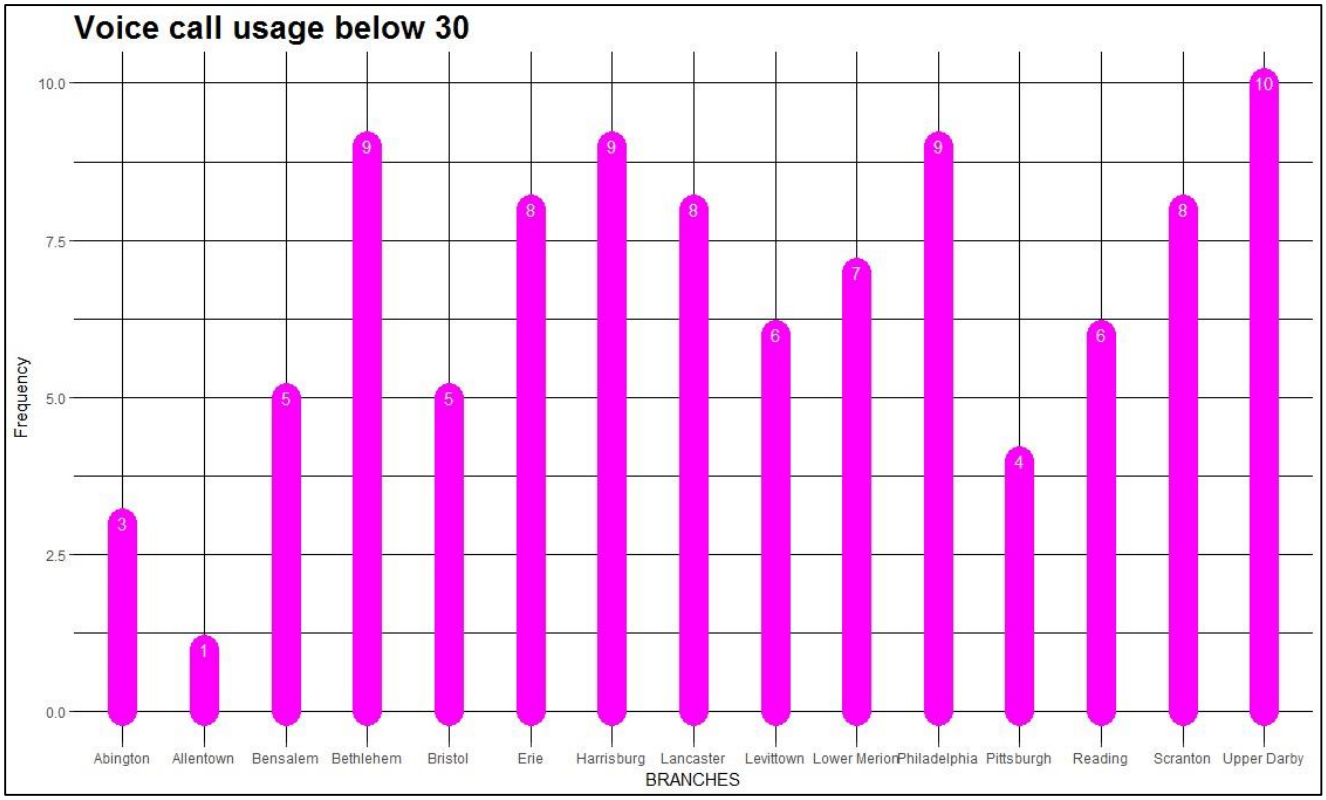


The plot above shows all the new connections added after 21st October 2016. The values in the bars show the number of new connections added in each branch.

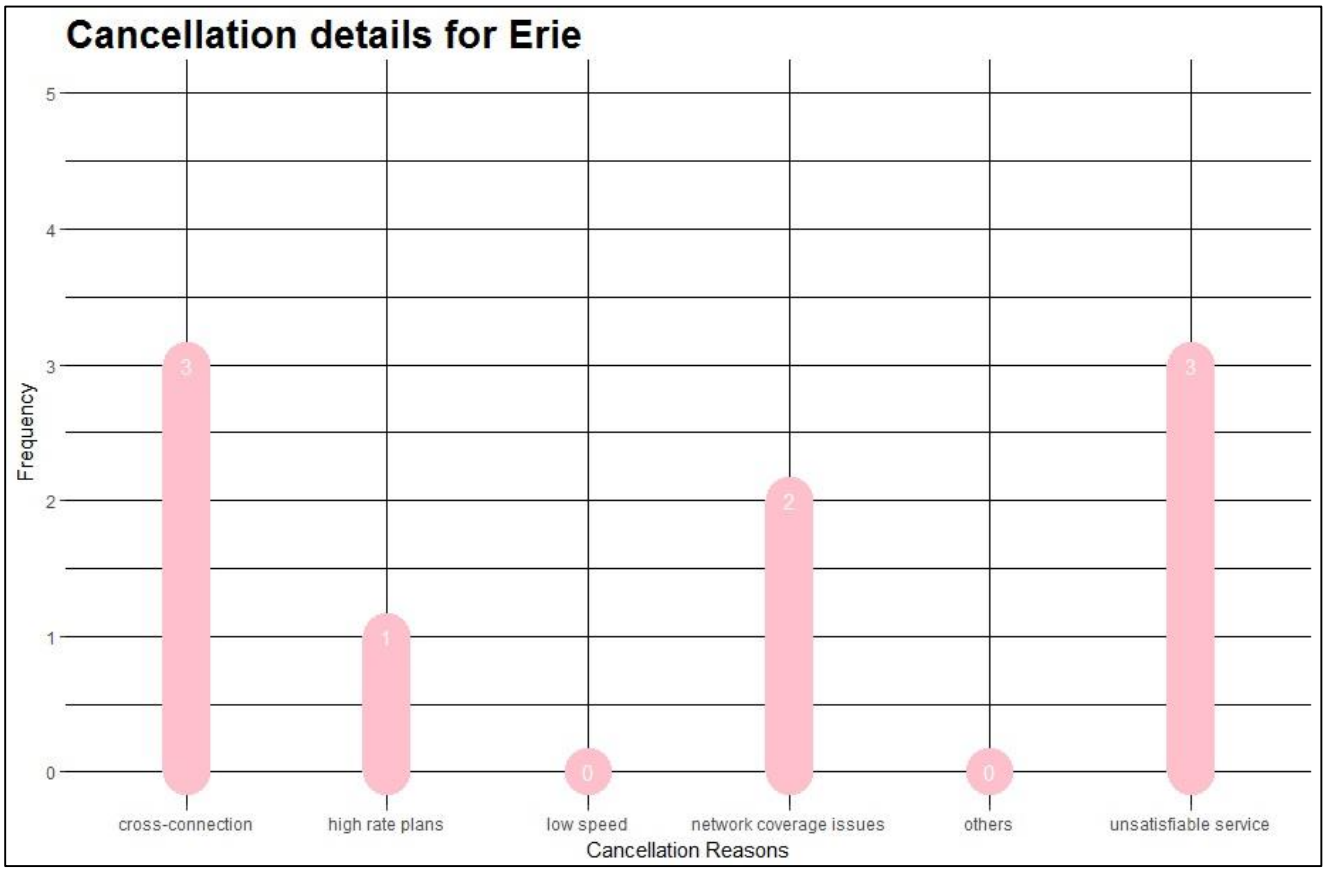
2. TOP 2 RATE PLANS:

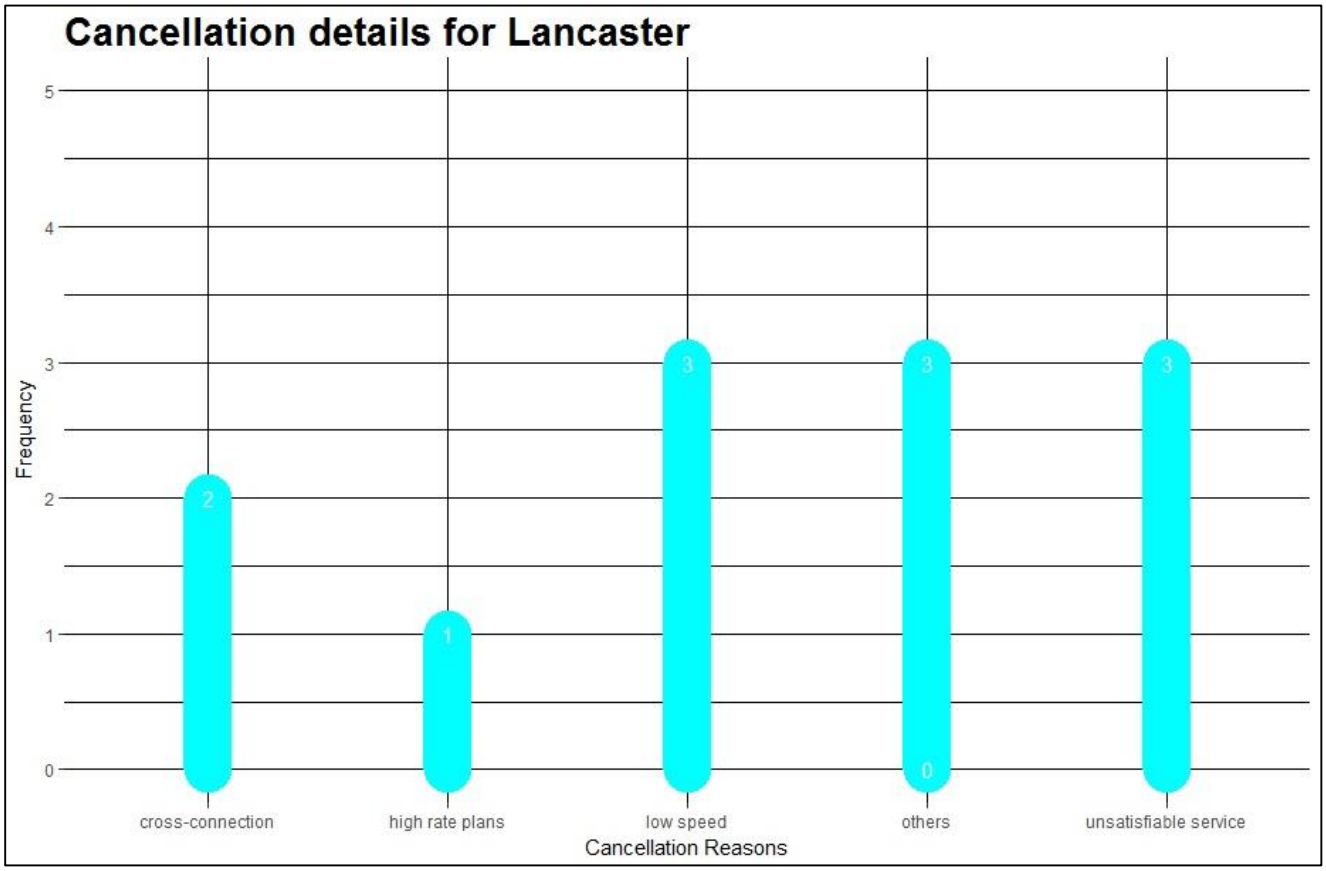


3. VOICE CALL USAGE BELOW 30:

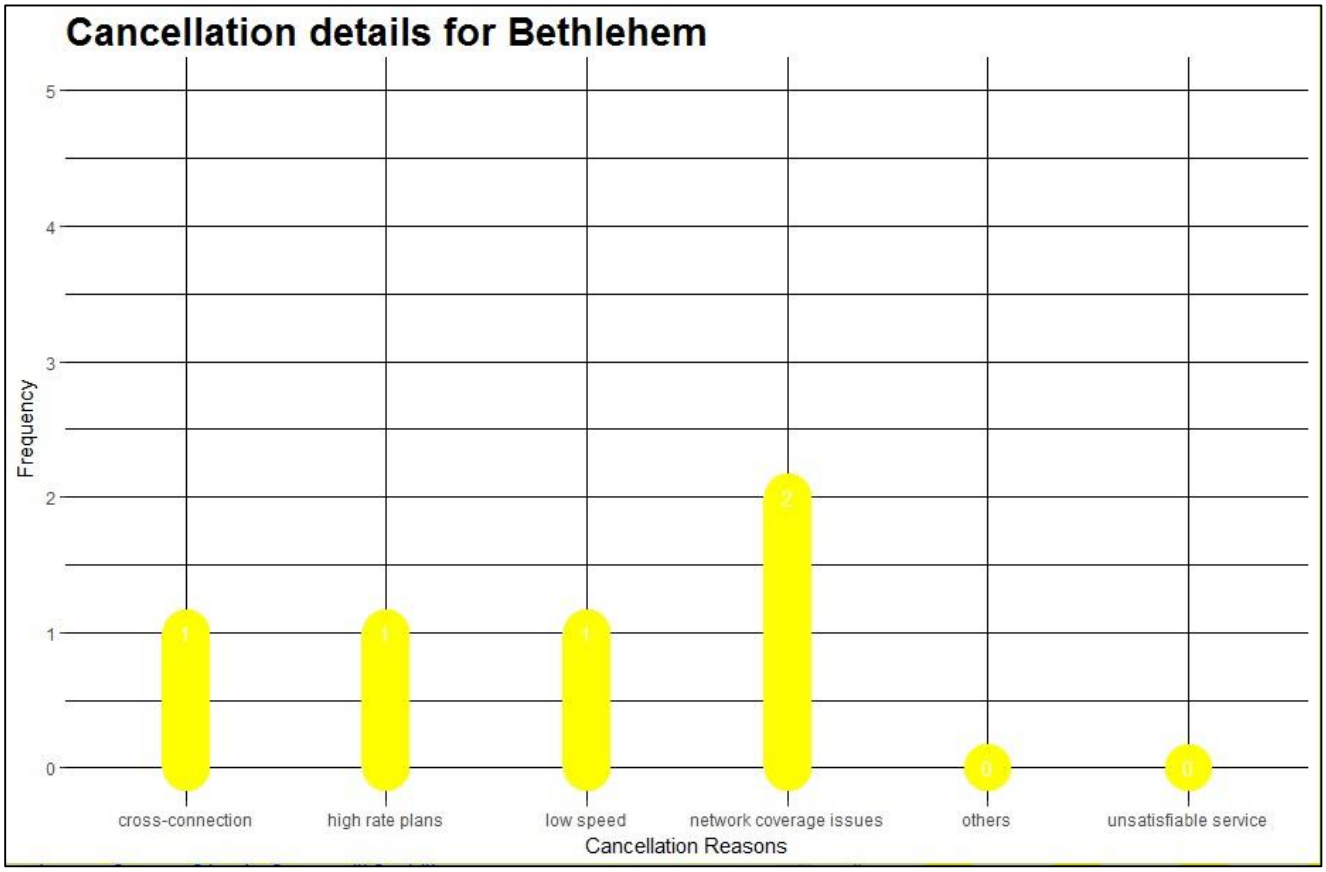


4. CANCELLATION DETAILS:

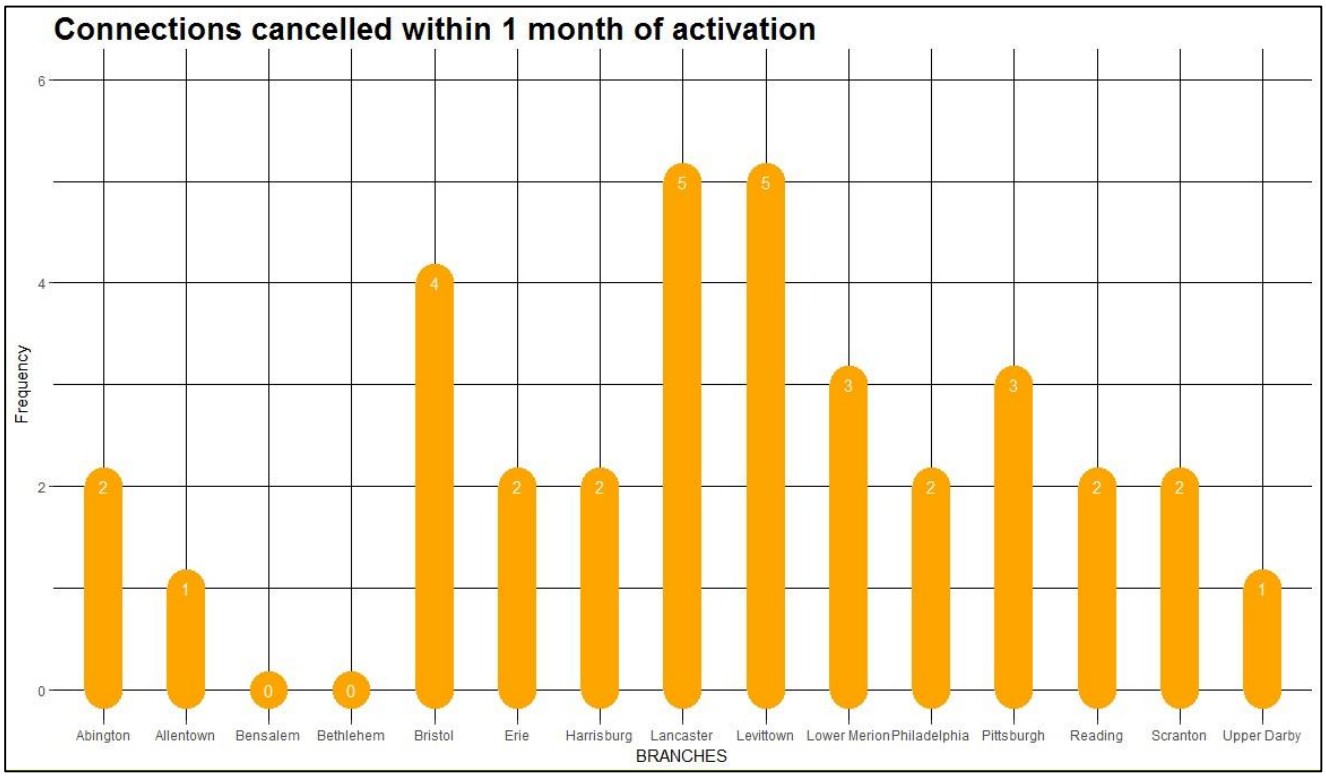





Cancellation details for Pittsburgh
Frequency
0
1
2
3
4
5
1
0
1
1
1
1
cross-connection
high rate plans
low speed
network coverage issues
others
unsatisfiable service
Cancellation Reasons


Cancellation details for Bethlehem
Frequency
0
1
2
3
4
5
1
1
1
2
0
0
cross-connection
high rate plans
low speed
network coverage issues
others
unsatisfiable service
Cancellation Reasons

5. CONNECTIONS CANCELLED WITHIN 1 MONTH OF ACTIVATION:



The above visualizations are performed on reports generated by KPIs in Hive. Frequency of occurrence is calculated and plotted against Y-axis and the other parameters are plotted against X-axis.

CONCLUSION:

We have successfully designed the data model which takes in structured data, performs query processing on that data, and generates resultant data in structured format which is then visualized and reports are generated.

FUTURE SCOPE:

- More KPIs can be generated based on actual data and requirements of analysis.
- The data set used could be real time data, so that we could get more variance.
- Visualizations could be more detailed and versatile.