PROJECT REPORT

ON

ETL IMPLEMENTATION FOR BIKE RIDE SHARE SYSTEM PROVIDER COMPANY

SUBMITTED TO

SAVITRIBAI PHULE PUNE UNIVERSITY

UNDER THE GUIDANCE OF

DR. SHWETA MESHRAM

SUBMITTED BY

ADITYA K. KULTHE (T.Y.MCA)

SEAT NO :- 15284

SAVITRIBAI PHULE PUNE UNIVERSITY

MASTER IN COMPUTER APPLICATION

MAHARASHTRA EDUCATION SOCIETY'S

INSTITUTE OF MANAGEMENT AND CAREER COURSES

(IMCC), PUNE-411038

2021-22

क्रियेवीण वाचाळता व्यर्थ आहे
महाराष्ट्र एज्युकेशन सोसायटी, पुणे
स्थापना १८६०

Maharashtra Education Society's

Institute of Management & Career Courses (IMCC)

Approved by AICTE and Recognized by Savitribai Phule Pune University, Pune

NBA Accredited MCA Program (Valid up to 30.06.2022)

IMCC Campus, 131, Mayur Colony, Kothrud, Pune 411038, Maharashtra, India | Ph.: 020-2546 3453 / 6271 / 73 | e-mail: info.imcc@mespune.in

Ref. No. MES IMCC / 006-A/ 2021 – 22/ **Date: / /**

CERTIFICATE

This is to certify that the Project Report entitled

"ETL Implementation for Bike Ride Share System Provider Company"

is prepared by

Aditya K Kulthe

M.C.A. Semester IV Course for the Academic Year 2021-22 at M.E. Society's Institute of Management & Career Courses (IMCC), Pune – 411038.
M.C.A Course is affiliated to Savitribai Phule Pune University.

To the best of our knowledge, this is original study done by the said student and important sources used by him/her have been duly acknowledged in this report.

The report is submitted in partial fulfillment of M.C.A Course for the Academic Year 2021-22 as per the rules and prescribed guidelines of Savitribai Phule Pune University.

Zirmite

Dr. Ravikant Zirmite
Head, Dept of MCA
MES IMCC

Dr. Santosh Deshpande
Director,
MES IMCC

CERTIFICATE

This is to certify that **Aditya K. Kulthe** has completed the project work entitled ***"ETL Implementation for Bike Ride Share System Provider Company"*** under my guidance. The report is submitted in partial fulfillment of M.C.A. Course for the Academic Year 2021-2022 as per the rules & prescribed guidelines of Savitribai Phule Pune University.

His/her work is found to be satisfactory and complete in all respects.

Dr. Shweta Meshram
(Internal Project Guide)

Acknowledgement

I present with pride and pleasure the project report on “ETL Implementation for Bike Ride Share System Provider Company” aimed to the supplement attachment as required under theregulation of the Savitribai Phule Pune University.

I am highly indebted to the project guide Dr.Shweta Meshram Professor IMCC for his guidance and constant inspection in each stage of project without which projectwork would not have taken this shape and form. I wish to offer most sincere thanks to Dr. Santosh Deshpande, Director IMCC for encouraging and providing with all the facilities. Thanks to Dr. Manasi Bhate, Deputy Director, MES' IMCC and Dr Ravikant Zirmite Sir, HOD IMCC who allotted us with enough time for the project & also thankyou. Finally, I am extremely thankful to teaching and non-teaching staff for their kind and whole co-operation.

Aditya K. Kulthe

Index

Chapter 1:- Introduction

1.1 Institute Profile : -

Institute of Management and Career Courses (IMCC) is a premier Management Institute, established in 1983 by Maharashtra Education Society (MES) for providing quality education and technical expertise at the Post Graduation Level in the Fields of Computers and Management. The Institute is recognized by SPPU under Section 46 of Pune University Act, 1974 and Section 85 of Maharashtra University Act, 1994 and Approved by AICTE New Delhi to conduct MCA and MBA programmes. The Institute is located at 131, Mayur Colony, Kothrud, Pune-411038 having 30,000 sq.ft. built area & totally independent campus.

IMCC is recognised as a Ph.D. Research Center under the Faculty of Management, SPPU. IMCC has 38 years standing & it is well-known for its conducive educational atmosphere. IMCC focuses on the all-round development of its students. Thus, apart from excellence in academics, students develop their inner potential by way of active participation in co- curricular & extra-curricular activities. IMCC has developed excellent rapport with Industry by way of Guest Lectures, Seminars, Workshops, Industrial Visits & Placements. The main motto of the Institute is to instill the concepts of total personality development in the students. The emphasis is laid on ‘Teacher Disciple Relationship’ in place of ‘Boss Subordinate’ relationship at their assignments.

The preamble of IMCC ``FACTA-NON-VERBA" lucidly means that the Institute produces the new breed of professionals, who’s deeds will speak and there could be no requirement of pomposity. The zooming enthusiastic, rational and excellent external endeavors are being imbibed in the students to prove their mettle. The conducive milieu of the Institute molds the budding managers to reveal in managing flexibility, integration, change and transformation. These ‘would be’ professionals are channelised in such a way to ‘orchestrate’ and deploy business and technological management skills in a synergistic manner to grab the tangible success. The faculty members put their relentless efforts in educating the students to synthesize business management acumen and technology insights in a creative manner.

1.2 Existing System Functionality: -

The current condition of the department of bike ride sharing system provider in analyzing and processing its data is described as follow:

Each information system has its own data. The data is in a flat file format (.csv), the data is scattered in each information system, redundancy data is still occurring, in the analysis and decision-making, the executives of the bike ride sharing systemprovider had to do a recap of data from each information system. For improving the performance of the company, bike ride share systems provider company executives must be provided with the accurate and up to date data to generate strategic decisions.

1.3 Business process understanding and specifications: -

1.3.1 Business Requirement Specifications:-

1.3.1.1

a) Required data to be filtered out from all data sources used in this project.

b) Dynamic Power BI dashboard to be developed for bike ride share system provider executives to make data driven business decisions.

1.3.1.2 Identify the dimensions, required attributes, measures, filter conditions, adjustments for KPIs going to be used in the Target system and its availability in the Source System. If any gaps suggest remediation of gaps

a) Dimensions :- start_station_name,end_station_name,usertype,gender,day_of_journey

b) Measures: - start_time,end_time

B) Filter Conditions: - Year wise (2017-2020) filter for cities like Boston, Chicago, New York, New Jersey, Washington DC in US City.

1.3.2 Business Rules Collections: -

Inspect each city file and apply preprocessing and transformation like handling duplicate values, missing values in data, converting columns in proper datatype format etc.

1.3.3 KPI’s: -

1.Total User Count
2.Popular Start Station
3.Popular Day For Journey
4.Most Frequent Tripduration(In Minute)

1.3.4 User Acceptance Criteria : -

- Ability to generate reports with little effort.
- Ability to get the aggregate report and drill down for further details.
- Ability to download data from a data warehouse and use it for further analysis.
- System reliability at all times.

.

1.4 Scope of the Project : -

The aim of bike ride share system provider is to be a significant player in its chosen markets. expertise, service and execution skills will differentiate the strategic business unit (SBU) from its peers. The project is required in order to help the decision makers of the company make an effective decision. Effective decisions are choices that move an organization closer to an agreed-on set of goals in a timely manner. Effective decision making is important at all organizational levels. Timely foundation and feedback information is needed as part of that effective decision-making. Therefore,we need to make business intelligence available throughout the organization.

1.5 Operating Environment :- Hardware & Software, Description of Tools / Technology to be used in the Target system: -

1.5.1.1 Operating systems used: -Windows 10.

1.5.1.2 RDBMS used to build databases: **- SQL** is a programming language that is used by most relational database management systems (RDBMS) to manage data stored in tabular form (i.e., tables). A relational database consists of multiple tables that relate to each other. The relation between tables is formed in the sense of shared columns.

1.5.1.3 ETL tools used : - None

1.5.1.4 OLAP/ Data mining/ machine learning/ analytics tools used (Python/ Cognos, BO, etc.): -

Python (Pandas): - Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open-source data analysis/manipulation tool available in any language. It is already well on its way toward this goal.

pandas are well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data need not be labeled at all to be placed into a panda’s data structure

1.5.1.5 Data visualization tools (Power BI):-

Power BI: - Microsoft Power BI data visualization tool helps bike ride share system provider to provide data insights using data visualization to support decision making. However, analysts and IT developers shouldn't be the only ones with the ability and responsibility to slice data to uncover insights and trends. Company decision-makers need personalized and organized reports in their preferred formats. They should be able to look at the report, understand the data, draw a conclusion, and make informed decisions.

Chapter 2 : - Proposed System

The project is required in order to assist the management of the bike ride share system provider company in making better decisions using the historical data available within the organization. The business users (decision makers) lack the ability to access data easily when needed. In an attempt to address this shortcoming, several departments within the company find their own resources, use different data available and hire consultants to solve their individual short-term data needs. In many cases, the same data was extracted from the same source systems to be accessed by separate departments without any strategic overall information-delivery strategy. The management realized the negative effect the different sources of the data have on the reports presented by the managers as the lack of integration. Given the importance of the information for the company, the management was motivated to deal with the problem of data inconsistency by introducing a central data warehouse and to ensure that data is available to all users irrespective of their department. The data harmonization and the need for consistent and quality reports gave birth to the project of data warehouse in the company.

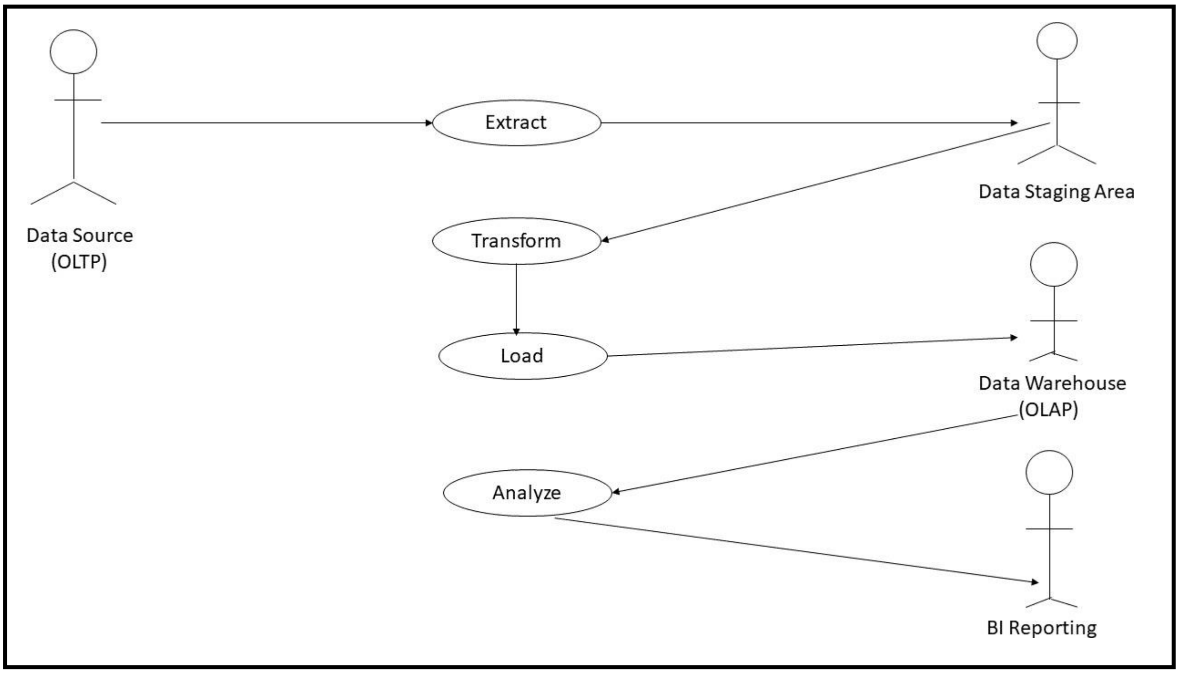
ETL can be accomplished in one of two ways. In some cases, businesses may task their developers with building their own ETL. However, this process can be time-intensive, prone to delays, and expensive. Most companies today rely on an ETL tool as part of their data integration process. ETL tools are known for their speed, reliability, and cost-effectiveness, as well as their compatibility with broader data management strategies. ETL tools also incorporate a broad range of data quality and data governance features. When evaluating an ETL tool, you'll want to consider the number and variety of connectors you'll need, as well as its portability and ease of use. You'll also need to determine if an open-source tool is right for your business, since these typically provide more flexibility and help users avoid vendor lock-in.

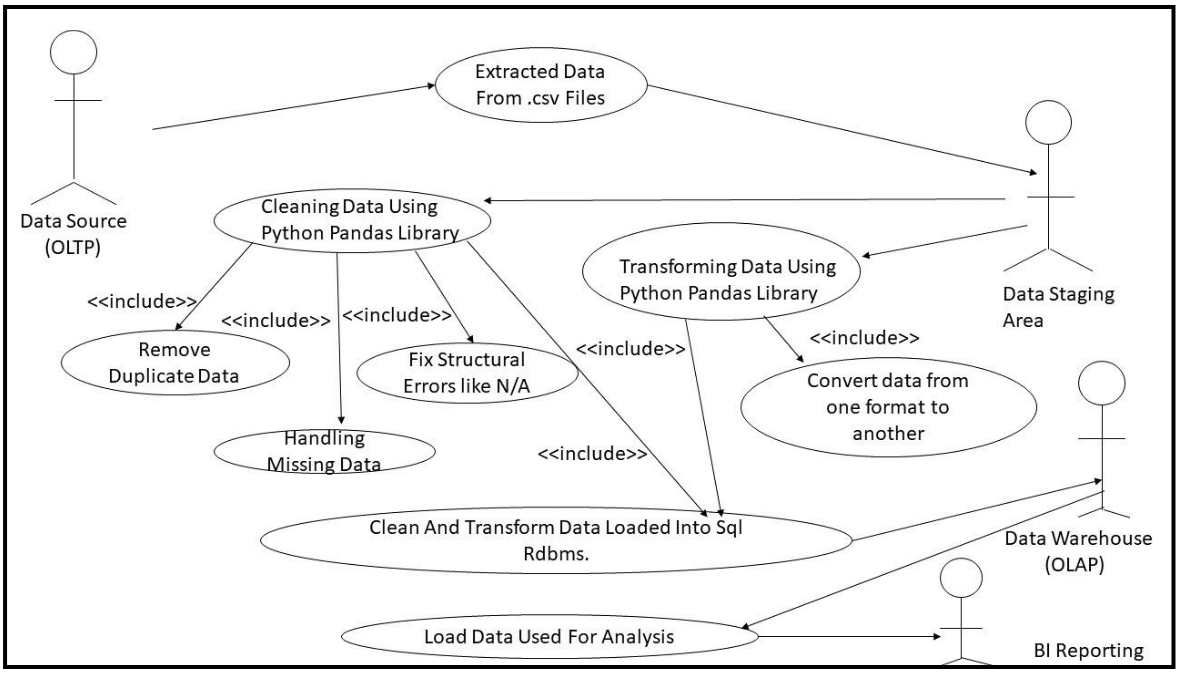
- Change in data formats over time.
- Increase in data velocity and volume.
- Rapid changes on data source credentials.
- Null issues.
- Change requests for new columns, dimensions, derivatives and features.
- Writing source specific code which tends to create overhead to future maintenance of ETL flows.

Combining all the above challenges compounds with the number of data sources, each with their own frequency of changes. All the above challenges are handled by ETL tool like python pandas module.

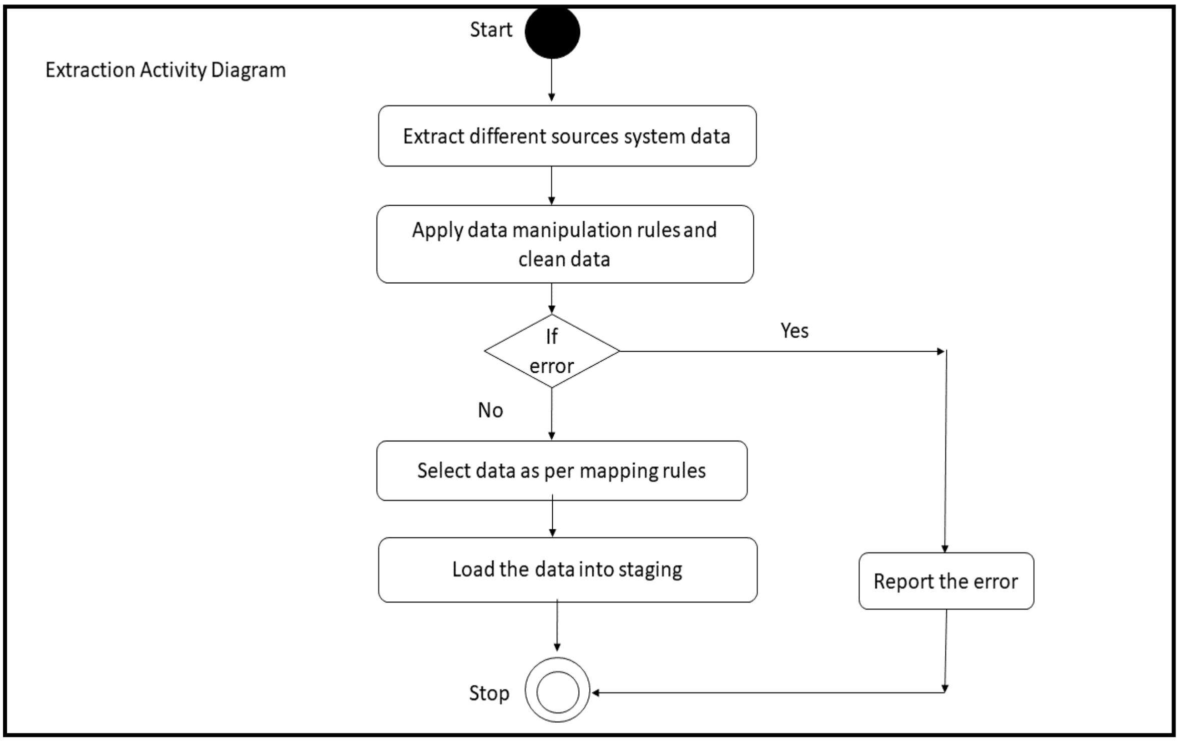
Chapter 3: -Analysis and Design

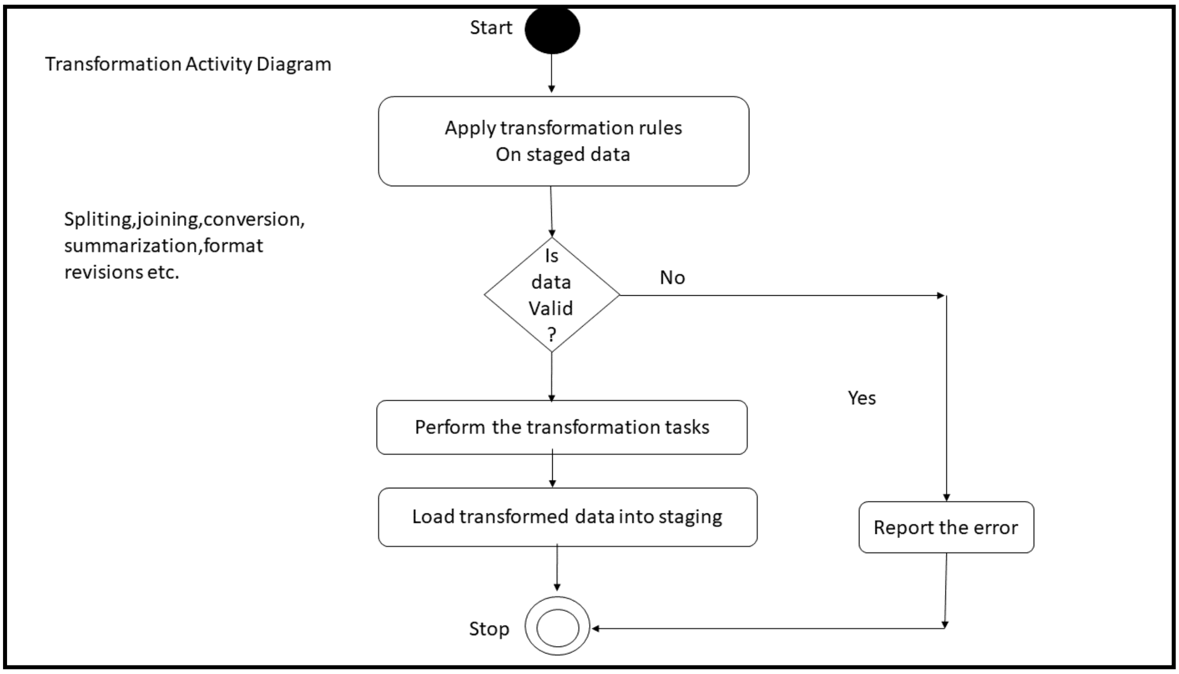
3.1 Use Case Diagram: -



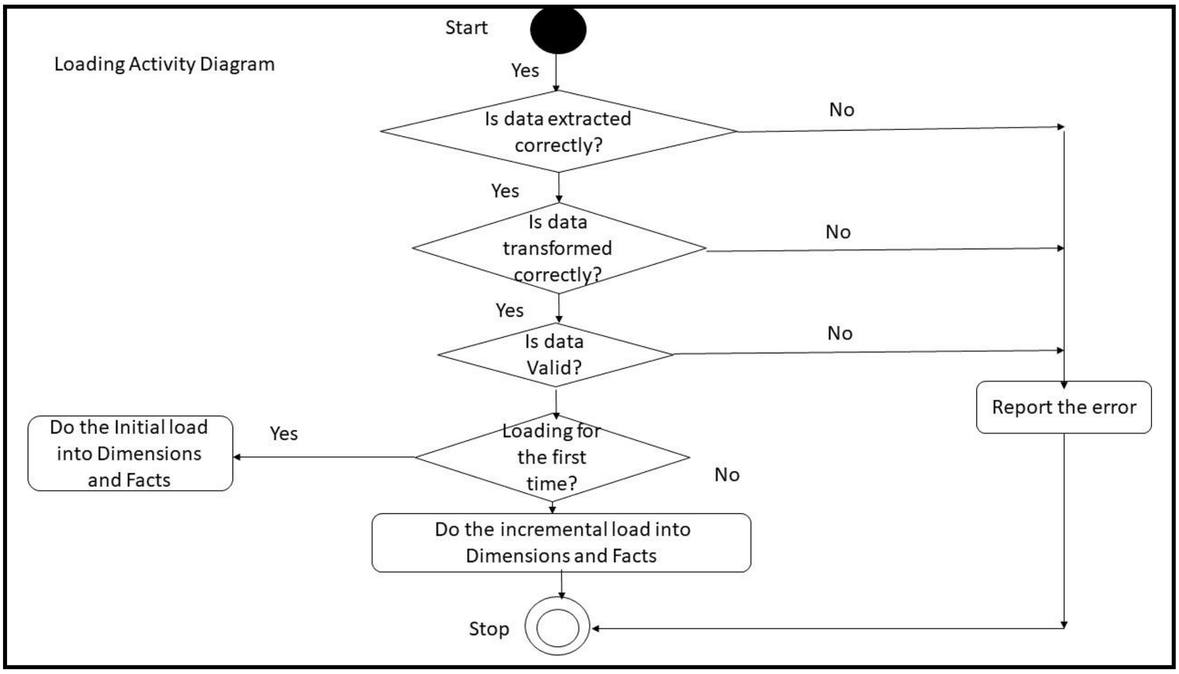


3.2 Activity diagram to demonstrate Process flow (execution of ETL process): -

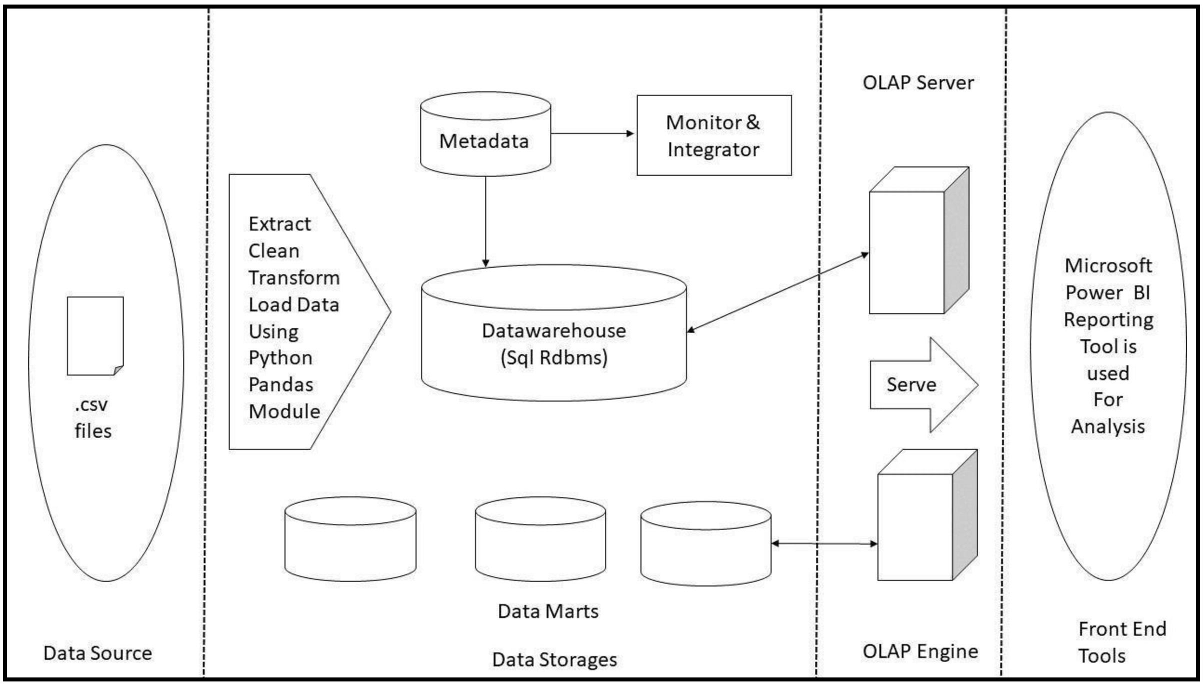

Start
Extraction Activity Diagram
Extract different sources system data
Apply data manipulation rules and clean data
If error
Yes
No
Select data as per mapping rules
Load the data into staging
Report the error
Stop


Start
Transformation Activity Diagram
Apply transformation rules On staged data
Spliting,joining,conversion, summarization,format revisions etc.
Is data Valid ?
No
Yes
Perform the transformation tasks
Load transformed data into staging
Report the error
Stop

`


Loading Activity Diagram
Start
Yes
Is data extracted correctly?
No
Yes
Is data transformed correctly?
No
Yes
Is data Valid?
No
Report the error
Do the Initial load into Dimensions and Facts
Yes
Loading for the first time?
No
Do the incremental load into Dimensions and Facts
Stop

3.3 Design of Target system (Elaborate the tiers of DW architecture in the Target System): -


OLAP Server
Metadata
Monitor & Integrator
Extract Clean Transform Load Data Using Python Pandas Module
Datawarehouse (Sql Rdbms)
.csv files
Serve
Microsoft Power BI Reporting Tool is used For Analysis
Data Marts
Data Source
Data Storages
OLAP Engine
Front End Tools

1. **Bottom Tier**
2. **Middle Tier**
3. **Top Tier**

- **Bottom Tier (Data sources and data storage): -**

  The bottom Tier usually consists of data sources. In project i used historical data **'.csv flat files '**from the public website for major cities in the U.S i.e Chicago, New York, Washington DC, New Jersey, Boston.

- **Middle Tier: -**

  The middle tier is an OLAP server that is typically implemented using either: A relational OLAP (ROLAP) model (i.e., an extended relational DBMS that maps operations from standard data to standard data); or A multidimensional OLAP (MOLAP) model (i.e., a special purpose server that directly implements multidimensional data and operations).In project SQL Rdbms is used as ROLAP server.

- **Top Tier: -**

  The top tier is a front-end client layer, which includes reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, etc.). In the project Microsoft Power BI is used as reporting tool.

Data Mart: -

A data mart contains a subset of corporate-wide data that is important to a specific group of users. The scope is limited to specific selected subjects. For example, a marketing data mart may limit its topics to customers, goods, and sales. The data contained in the data warts are summarized. Data warts are typically applied to low-cost departmental servers that are Unix/Linux or Windows-based. The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years. However, it can be in the long run, complex integration is involved in its design and planning were not enterprise-wide.

3.4 Database schema / Table specifications of Target system:-

Database name: - us_cities

Table 1:- boston_2017

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birth_year	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 2:- boston_2018

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birth_year	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 3:- boston_2019

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birth_year	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 4:- boston_2020

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birth_year	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 5:- chicago_2017

Field Name	Data Type
start_station_name	text
to_station_name	text
usertype	text
gender	text
birthyear	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 6:- chicago_2018

Field Name	Data Type
start_station_name	text
to_station_name	text
usertype	text
gender	text
birthyear	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 7:- chicago_2019

Field Name	Data Type
start_station_name	text
to_station_name	text
usertype	text
gender	text
birthyear	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 8:- chicago_2020

Field Name	Data Type
start_station_name	text
to_station_name	text
usertype	text
gender	text
birthyear	int
journey_date	date
day_of_journey	text
start_time	time
stop_time	time
tripduration_in_min	int

Table 9:- new_jersey_2017

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 10:- new_jersey_2018

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 11:- new_jersey_2019

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 12:- new_jersey_2020

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 13:- new_york_2017

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 14:- new_york_2018

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 15:- new_york_2019

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 16:- new_york_2020

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
birthyear	int
gender	int
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 17:- washington_dc_2017

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 18:- washington_dc_2018

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 19:- washington_dc_2019

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

Table 20:- washington_dc_2020

Field Name	Data Type
start_station_name	text
end_station_name	text
usertype	text
journey_date	date
day_of_journey	text
start_time	time
end_time	time
tripduration_in_min	int

3.5 Details of Source & Targets of mapping in the database: -

Chicago:-

Source System	Target System
start_station_name	start_station_name
to_station_name	to_station_name
usertype	usertype
gender	gender
birthyear	birthyear
journey_date	journey_date
day_of_journey	day_of_journey
start_time	start_time
stop_time	stop_time
tripduration_in_min	tripduation_in_min

Boston :-

Source System	Target System
start_station_name	start_station_name
to_station_name	to_station_name
usertype	usertype
gender	gender
birthyear	birthyear
journey_date	journey_date
day_of_journey	day_of_journey
start_time	start_time
stop_time	stop_time
tripduration_in_min	tripduration_in_min

New Jersey :-

Source System	**Target System**
start_station_name	start_station_name
end_station_name	end_station_name
usertype	usertype
birthyear	birthyear
gender	gender
journey_date	journey_date
day_of_journey	day_of_journey
start_time	start_time
end_time	end_time
tripduration_in_min	tripduration_in_min

New York :-

Source System	**Target System**
start_station_name	start_station_name
end_station_name	to_station_name
usertype	usertype
birthyear	birthyear
gender	gender
journey_date	journey_date
day_of_journey	day_of_journey
start_time	start_time
end_time	end_time
tripduration_in_min	tripduration_in_min

`

Washington DC :-

Source System	**Target System**
start_station_name	start_station_name
end_station_name	to_station_name
usertype	usertype
journey_date	journey_date
day_of_journey	day_of_journey
start_time	start_time
end_time	end_time
tripduration_in_min	tripduration_in_min

3.6 Details of Load (Full/Incremental etc.):-

There are two primary methods to load data into a warehouse: -

- **Full load: -** with a full load, the entire dataset is dumped, or loaded, and is then completely replaced (i.e., deleted and replaced) with the new, updated dataset. No additional information, such as timestamps, is required.

  For example, take a store that uploads all of its sales through the ETL process in a data warehouse at the end of each day. Let's say 5 bike rides were made on a Monday, so that on Monday night a table of 5 records would be uploaded. Then, on Tuesday, another 3 bike rides were made which need to be added. So on Tuesday night, assuming a full load, Monday's 5 records as well as Tuesday's 3 records are uploaded – an inefficient system, although relatively easy to set up and maintain. While this example is overly simplified, the principle is the same.

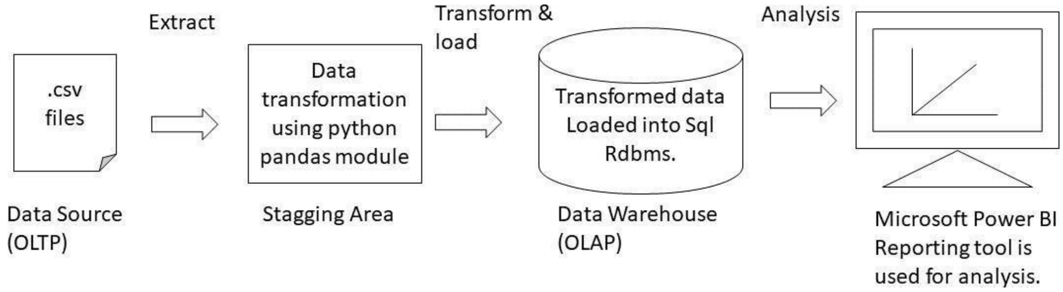
- **Incremental load:** -with an incremental load applying ongoing changes as when needed periodically. Only the difference between the target and source data is loaded through the ETL process in the data warehouse. There are 2 types of incremental loads, depending on the volume of data you're loading; streaming incremental load and batch incremental load.

  Following the previous example, the station that made 3 bikes rides on Tuesday will load only the additional 3 records to the table, instead of reloading all records. This has the advantage of saving time and resources, but increases complexity.

  Incremental loading is of course much faster than a full load. The main drawback to this type of loading is maintainability. Unlike a full load, with an incremental load you can't re-run the entire load if there's an error. In addition to this, files need to be loaded in order, so errors will compound the issue as other data queues up.

Today, organizations are moving away from processing and loading data in large batches, preferring – or driven by business needs – to process in real-time using stream processing, meaning that as client applications write data to the data source, data is treated, transformed and saved to the target data store. Today, tools exist to enable this process. This eliminates many of the drawbacks of traditional processing and loading, increases speed and decreases complexity.

3.7 Design of ETL schema/strategy :-



Extract: -In this phase I used historical data **'.csv flat files '**from the public website for major cities in the U.S i.e Chicago, New York, Washington DC, New Jersey City, Boston to design Datawarehouse and bi system.

Cleansing: -The cleansing stage is crucial in a data warehouse technique because it is supposed to improve data quality. **Python(Pandas Library)** script has been

used for cleaning activity on datasets like removing duplicate data, handling missing data, fix structural errors like N/A ,not applicable .

Transform: - It converts records from its operational source format into a particular data warehouse format. **Python (Pandas Library)** script has been used for transformation of datasets.

Load: - The Load is the process of writing the data into the target database. During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. **Relational database SQL** has been used to load the data.

Chapter 4 :- Design of strategy for Visualization

It’s no secret that data visualization is very important right now.And it should be from the public sector to nonprofits and private enterprise, data and how it is used is remaking our world. Perhaps nowhere is data more popular than data visualization, where the design-minded dedicate themselves to charting just about anything. All of which begs the inevitable question, is data visualization a passing fad, geared to dress up otherwise unappealing information? Or does it underscore our deeper need for people to simplify information, organize it and make it more accessible so that we can increase our understanding of the world around us?

Start by Asking One Fundamental Question -

Every data visualization challenge is different. Some are simple, geometric abstractions like pie charts, scatter diagrams or area charts that help clarify phenomena from a larger data set. Other visualizations such as narrative infographics are less about a purely objective look at the data, taking an illustrative approach to promote greater understanding of a complex process or system. Whatever the case may be, it’s helpful to start starting with one simple question: “Why is this information meaningful to the audience?” Because ultimately, it’s about meeting the audience where they are. As communicators and designers, we need to take the complexity of the numbers, process or ecosystem and distill it into something engaging and meaningful for the audience.

Constructive’s process for designing with data is rooted in the strategic approach we take to designing for any medium or content, which like many design processes, consists of four phases: Research, Strategy, Exploration, and Execution. For data design, the key is to be not just willing, but eager to dig into the numbers and make sense of them. To embrace their complexity so that we can work with our clients to find effective ways to present them so that they can use their data to increase understanding, strengthen their brand narrative, and ultimately motivate audiences to action.

Once we've started with this simple question and used it to contextualize our thinking and inform our decisions, here's how we tailor our design process to create effective communications and experiences with data.

Step 1: Design Research

As with all research, effective design research requires that we ask the right questions. So, start by taking a step back from the data itself and discuss the project with stakeholders to identify broader goals for the work. What information are we trying to communicate? How does it fit into achieving broader organizational goals? Who are our audiences and what are their levels of expertise with our subject matter? What's their level of data literacy and visual literacy?

Once you have answers to top-level strategic questions like these you can dig deeper into researching to find answers to data-specific questions: Is our data currently organized in a way that's easy to understand? What are the top-level takeaways we need to extract and elevate? What are the different types of visualizations we can use to to communicate our ideas? How is the data structured, and how flexibly can we apply it? The goal of this research is to immerse ourselves—to understand our organizational goals and our audience, absorb everything we can, and then approach designing solutions in ways that align the interests of the brand and the audience.

Step 2: Design Strategy

Now we're ready to focus on developing a design strategy that follows through on our research. First, start by creating a communication strategy: define the purpose of the work based on our research and what factors will go into designing and delivering a meaningful experience? Where will our audience be? Will they have time to do a deep dive into our data or is it more of a high-level quick take? Develop a content strategy: how should our information be organized or edited to make it more accessible and understandable for our audience? Can we simplify ideas and concepts without compromising the integrity of our data? And finally, develop a design strategy: techniques can we use to create a more engaging presentation? How does the design of our brand impact design execution? And if it's an interactive visualization, how do we want audiences to interact with out data? And don't forget technology strategy: What systems are used to produce or store the data? What are the right digital tools to deliver them? The goal is to think holistically across multiple strategies that all influence how our data will be experienced and what that means for both our audience and our brand—taking each into account and finding the right mix to effectively balance them.

Step 3: Design Exploration

So we've got a good idea of what the issues and ideas are, why they're important, and how they're going to be used. We have a sense of what design approaches are likely to work best. It's finally time to start iterating on design concepts that make our ideas tangible! At Constructive, we prefer rapid prototyping to get as many ideas out there quickly—because especially when designing and building with data, the costs of going down the wrong path can be significant. For us, this means starting with pencil sketches to quickly get our ideas out so they can be discussed. For projects with more complex data visualizations or data tools, we usually create information architecture and wireframes to that provide a greater level of detail and interaction. Once we've got our structure, hierarchy, and functionality in place, we're ready to jump into look and feel, adding visual style by choosing the appropriate color, fonts, iconography, etc. The key here is to work from the outside in; start with basic, high-level concepting, collaborate to develop the structural elements; then gradually hone in on the details.

Step 4: Design Execution

As they say, without proper execution, even the best strategy is useless. So, making sure you have the right team in place to execute the project is critical—particularly for interactive visualizations that require integration with outside data systems and complex development. Execution for data design projects follows the same approach that all design work takes: design, refine, produce. In the case of designing data visualizations, it's incredibly important to emphasize content accuracy and legibility—especially because data design can require communicating complex information in tight spaces or with a lot of competing elements. We should also design for flexibility to extend our work across other parts of our communications and brand, so plan ahead and consider how visualizations might be exported or adjusted to be used in other mediums. And be sure you test your work! If we're working in print, reproduce it in as close to the final format as possible. And if it's interactive, make sure you allocate extra time to the QA process if there's much interactivity with your data.

Chapter 5 :- Drawbacks and Limitation Proposed Enhancement

Company require experienced data architect/engineer to build optimal data pipeline for real time live streaming data to do this scalable system architecture and distributed data processing platform is required. Require continuous maintenance and monitoring. Require compatible, powerful hardware/software to process huge amount of data.

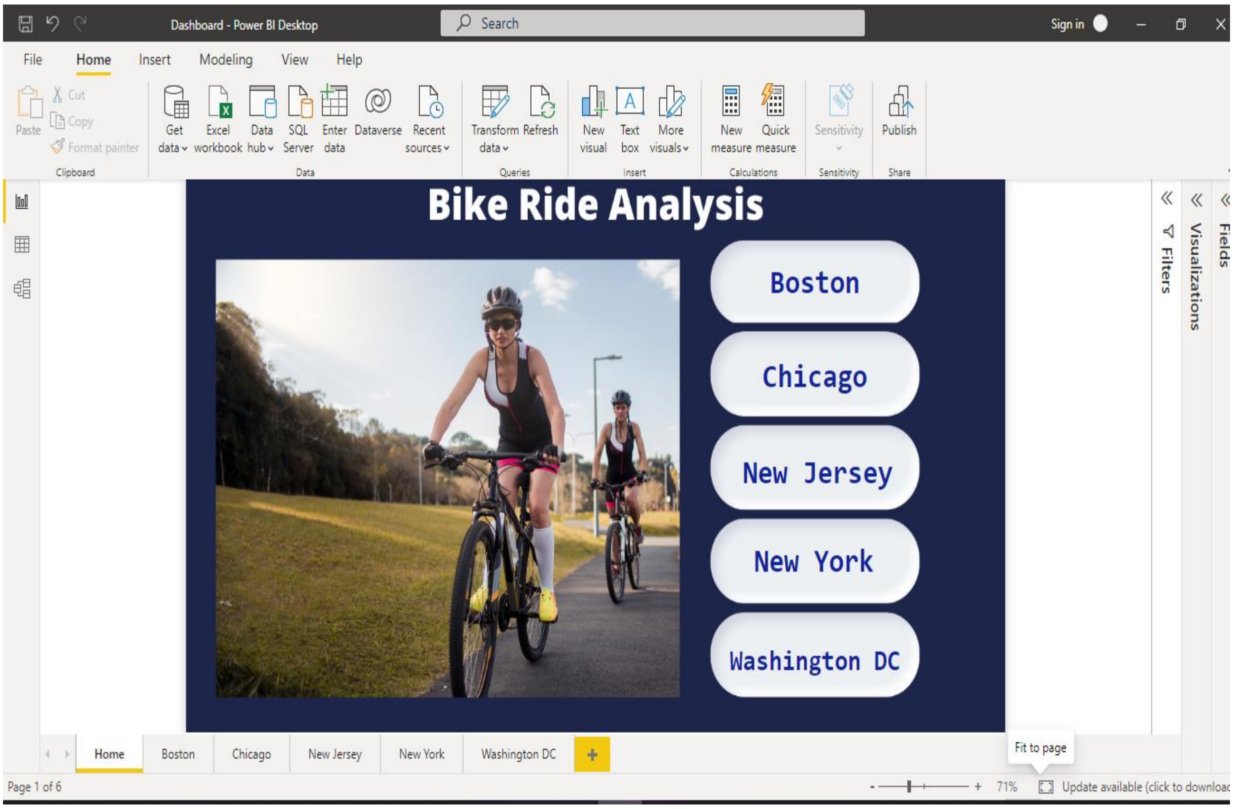
In future if company want to integrate AI features such as recommending place based on user history, chat bot to solve customer query, demand forecasting for specific location based on traffic in a day, path optimization to avoid traffic, increase customer retention rate by providing them attractive offers in their existing application which impact business revenue and profit we can use historical data which is stored in data warehouse.

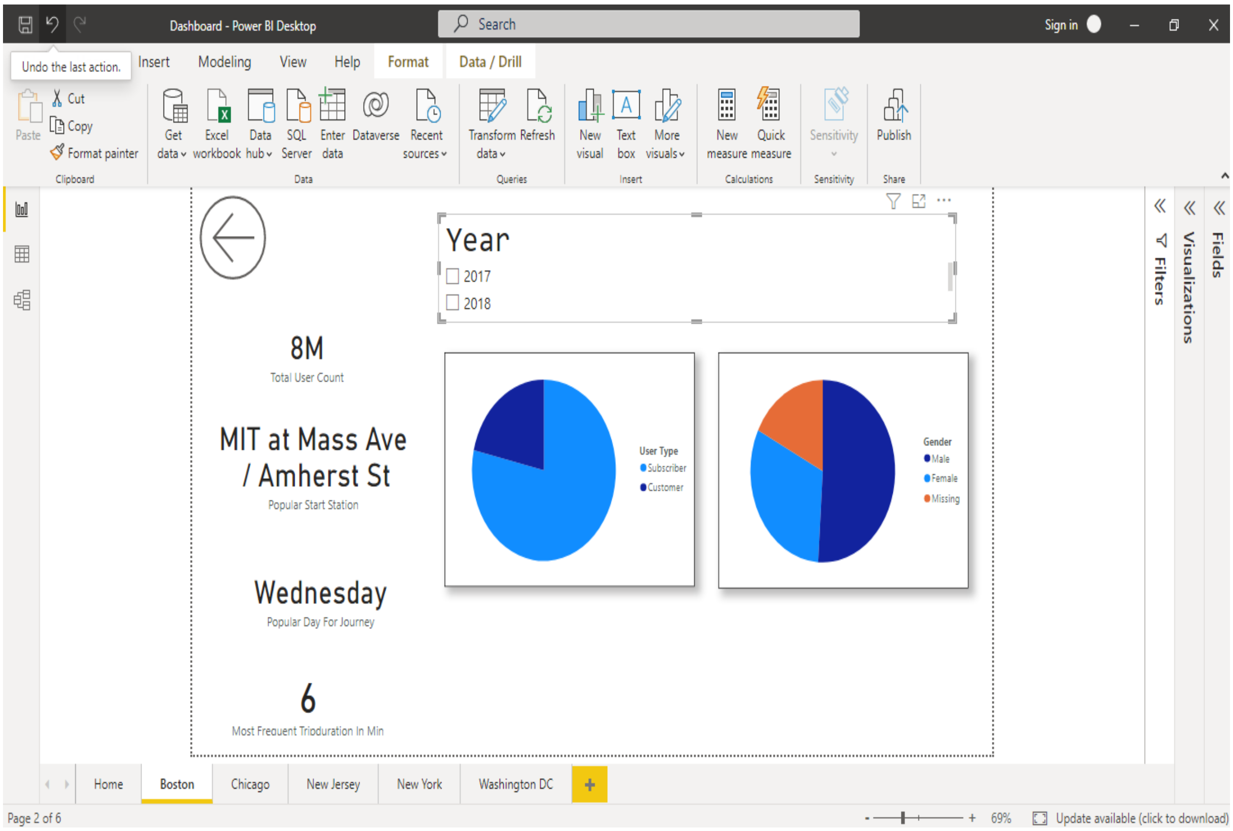
Chapter 6 :- Conclusion

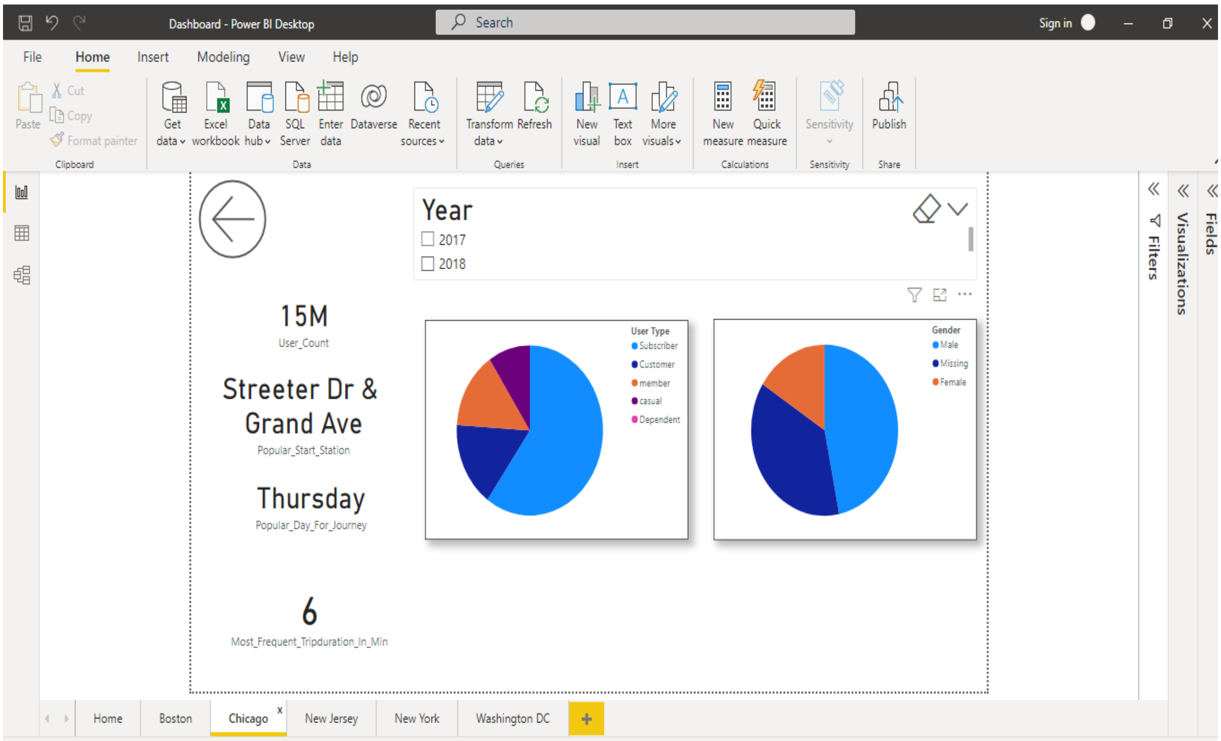
In order to be prepared for further development as integrated data warehouse at bike ride share company, this initial system is developed by considering dynamic growth of requirement. Requirement changing is accommodated without altering the structure of system as well as the programs that system used. It will only change the system configuration. ETL process for data warehouse at bike ride share company is arranged to be accomplished on the fly, it will not involve temporary table as an intermediary. Data is extracted, transformed, and loaded directly into database. This result in faster ETL process, but on the other hand, it will consume resource in operational system as well as data warehouse system.

Therefore, ETL process is scheduled when the operational system is not busy with its transaction. Transformation process does simple process, that is correcting data value. Having been transformed, data is loaded directly to database in data warehouse.

Output Screen (Dashboard) :-







Year

□ 2017

□ 2018

15M

User_Count

Streeter Dr & Grand Ave

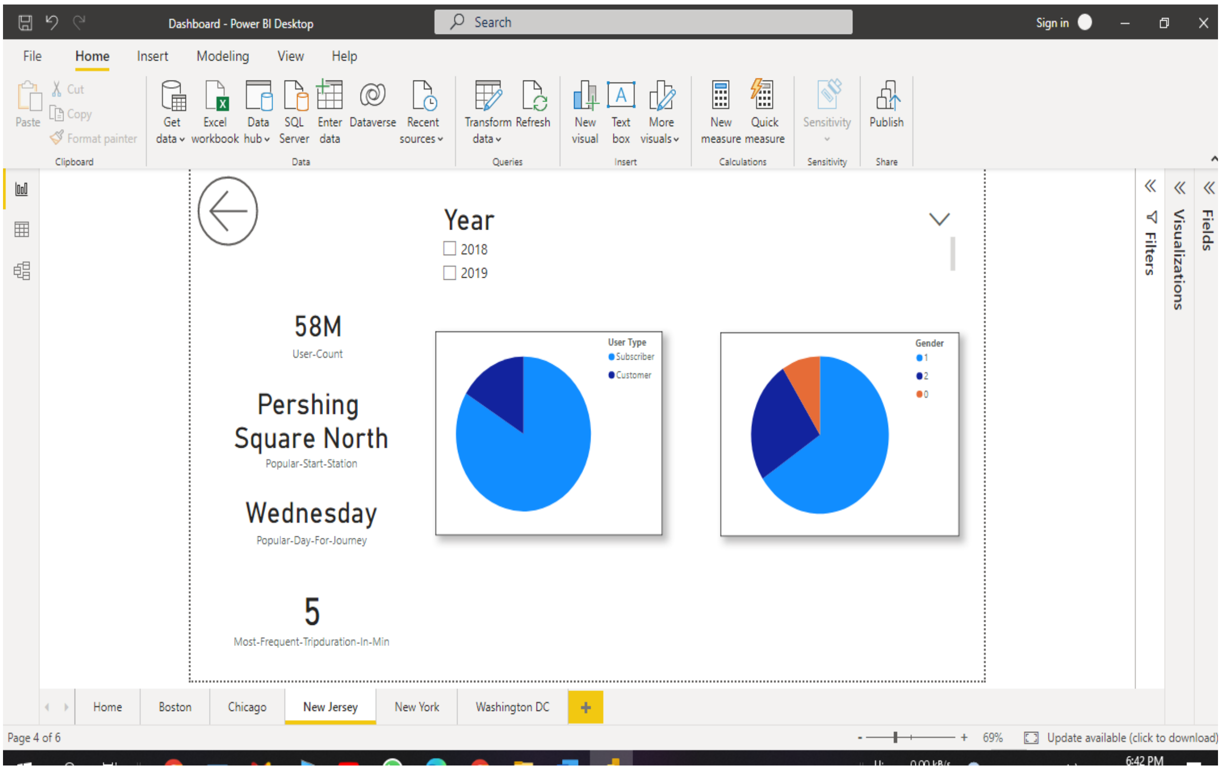
Popular_Start_Station

Thursday

Popular_Day_For_Journey

6

Most_Frequent_Tripduration_In_Min



Year

□ 2018

□ 2019

58M

User-Count

Pershing Square North

Popular-Start-Station

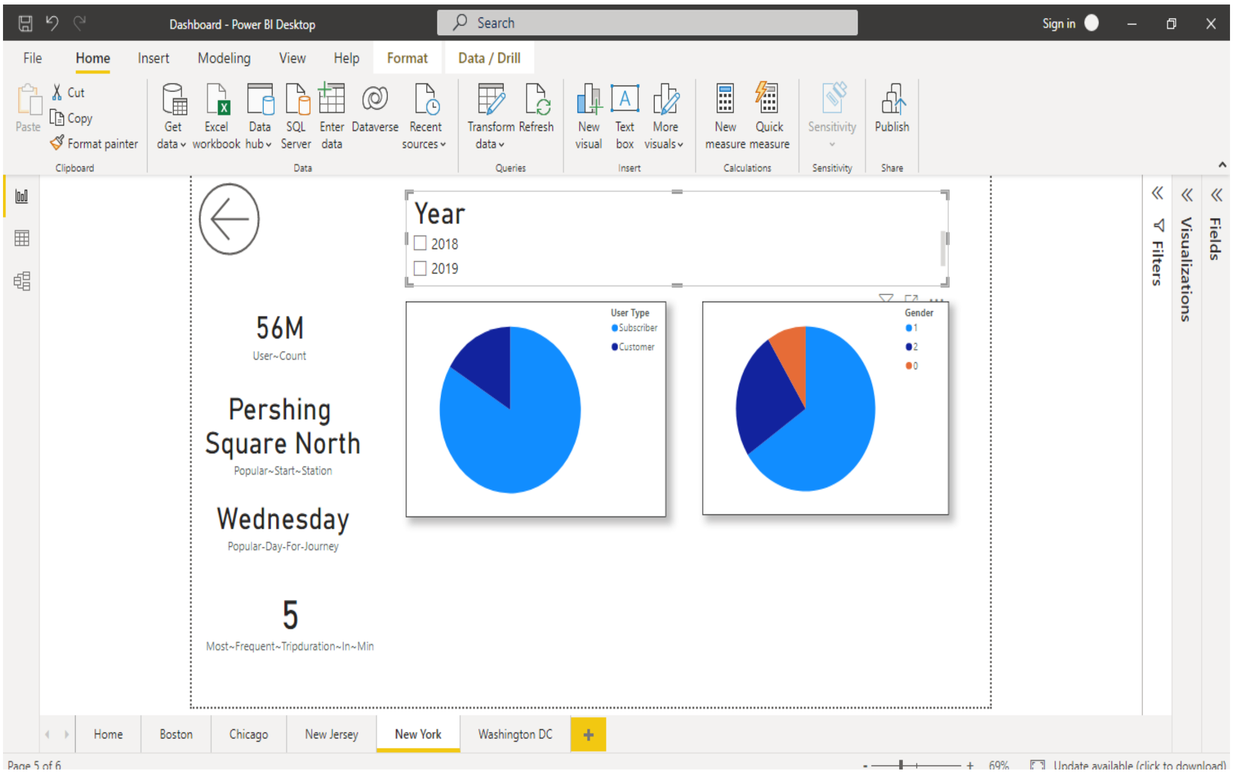
Wednesday

Popular-Day-For-Journey

5

Most-Frequent-Tripduration-In-Min

`

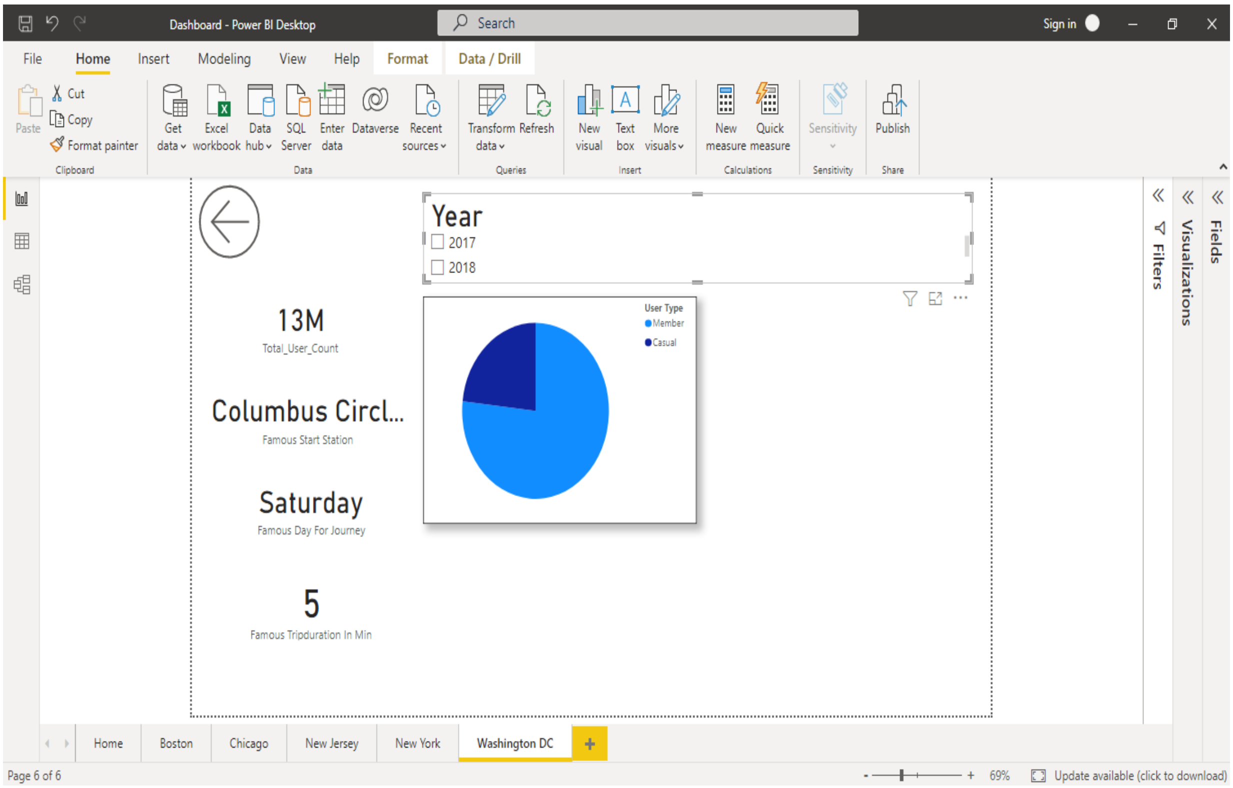


56M
User~Count

Pershing Square North
Popular~Start~Station

Wednesday
Popular-Day-For-Journey

5
Most~Frequent~Tripduration~In~Min



13M
Total_User_Count

Columbus Circl...
Famous Start Station

Saturday
Famous Day For Journey

5
Famous Tripduration In Min

`