ETL Implementation For Bicycle Ride Share System Provider

By,

Aditya K Kulthe

MES Institute Of Management And Career Courses, Pune.

Mca :- III Div :- B

Sem :- VI

Roll No :- 1912024

Seat No :- 15284

Date :- 22/09/2022

Quote’

“If you torture the data long enough, it will confess to anything.”

- British Economist Ronald Coase

Project Scope

Due exponential growth in a data now a day it is fuel for company **to make data driven business decision**. In this project I would like **to implement ETL flow for bicycle ride share system provider company** to make data driven business decision.

The current condition of the department of bicycle ride sharing systems provider in analyzing and processing its data is 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 in data is still occurring in the analysis and decision-making.**
- The executives of the bicycle ride sharing system provider had to do a recap of data from each information system **for improving the performance** of the company by providing **the accurate and up to date data** to generate **strategic decisions**.

- The **aim** of bicycle ride share system provider is **to be a significant player in its chosen markets.**
- Their **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 to 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.

Project Workflow

1. **Extraction Phase :-** In this phase of project I extracted sensor generated open source bicycle ride share **(.csv files)** data for five different cities in US like Chicago, New York , New Jersey City, Washington DC, Boston. Data is from year 2017 to 2020 store on AWS S3 bucket to the system for ETL purpose.
2. **Transformation Phase :-** In this phase of project I perform all data preprocessing and cleaning operations **(Python panda library)** on all extracted bicycle ride share data. Data preprocessing include handling missing & duplicate data also transforming data in proper format so it is useful for query and analysis purpose.
3. **Load Phase :-** In this phase of project I load all transformed data in relational database system**(MySQL)** . The data which load into relational database system is useful for company to make data driven business decision using KPI's like

K P I – Key Performance Indicators

Popular times of travel :-

- Most common month

K P I – Key Performance Indicators

Popular station :-

- Most common start station

K P I – Key Performance Indicators

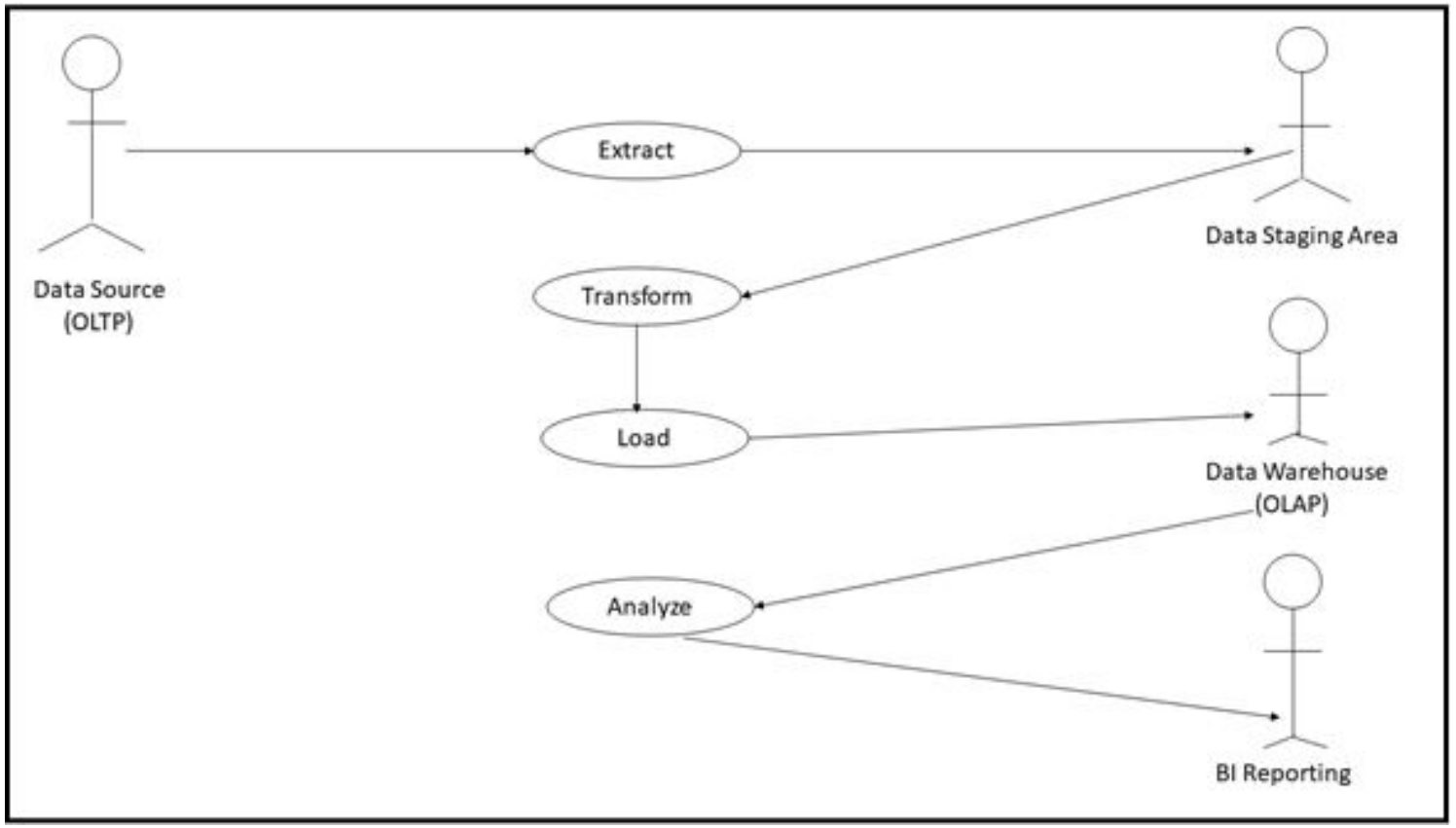
Trip duration :-

- Popular trip duration.

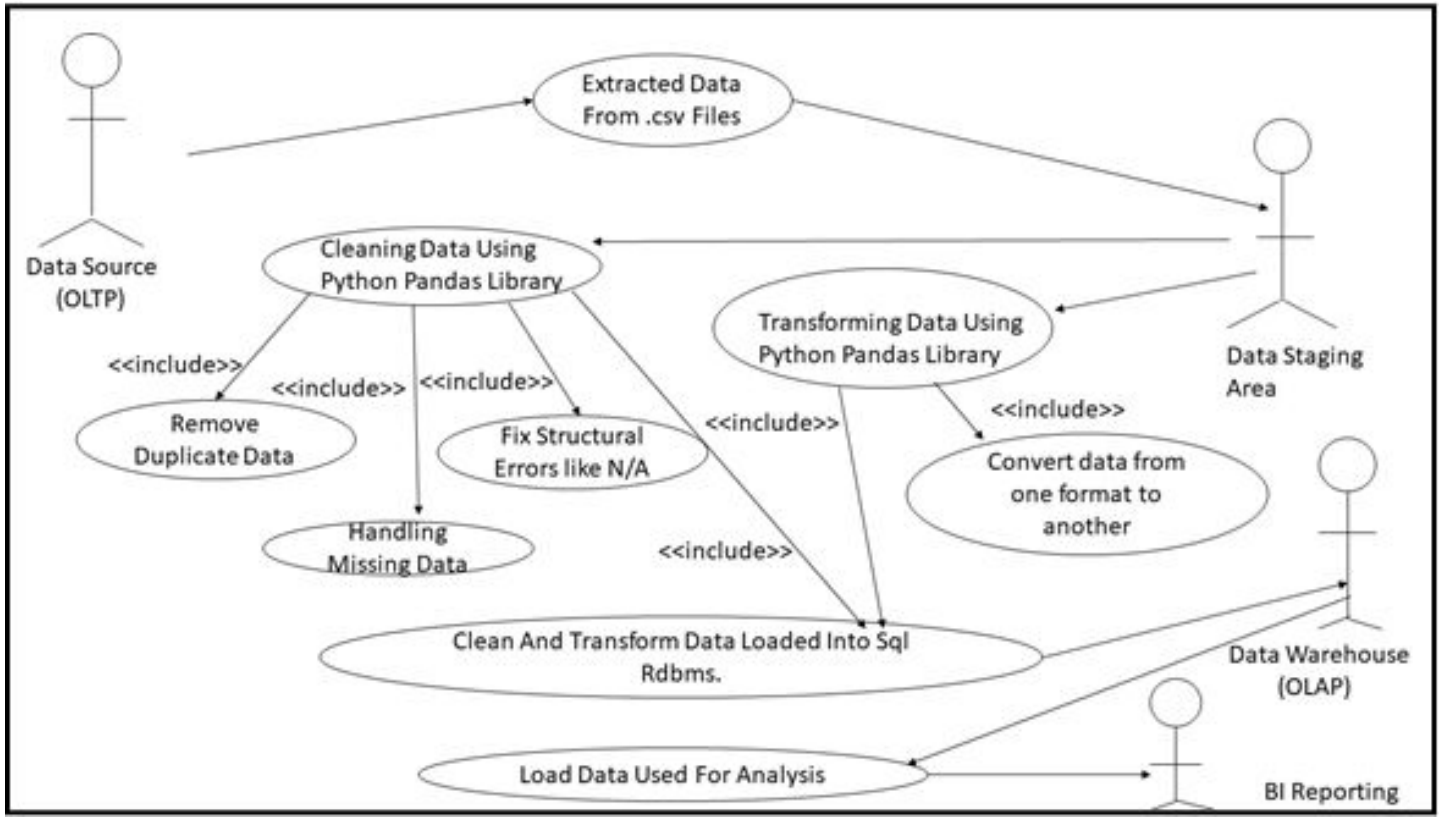
User info :-

- Counts of each user type
- Counts of each gender

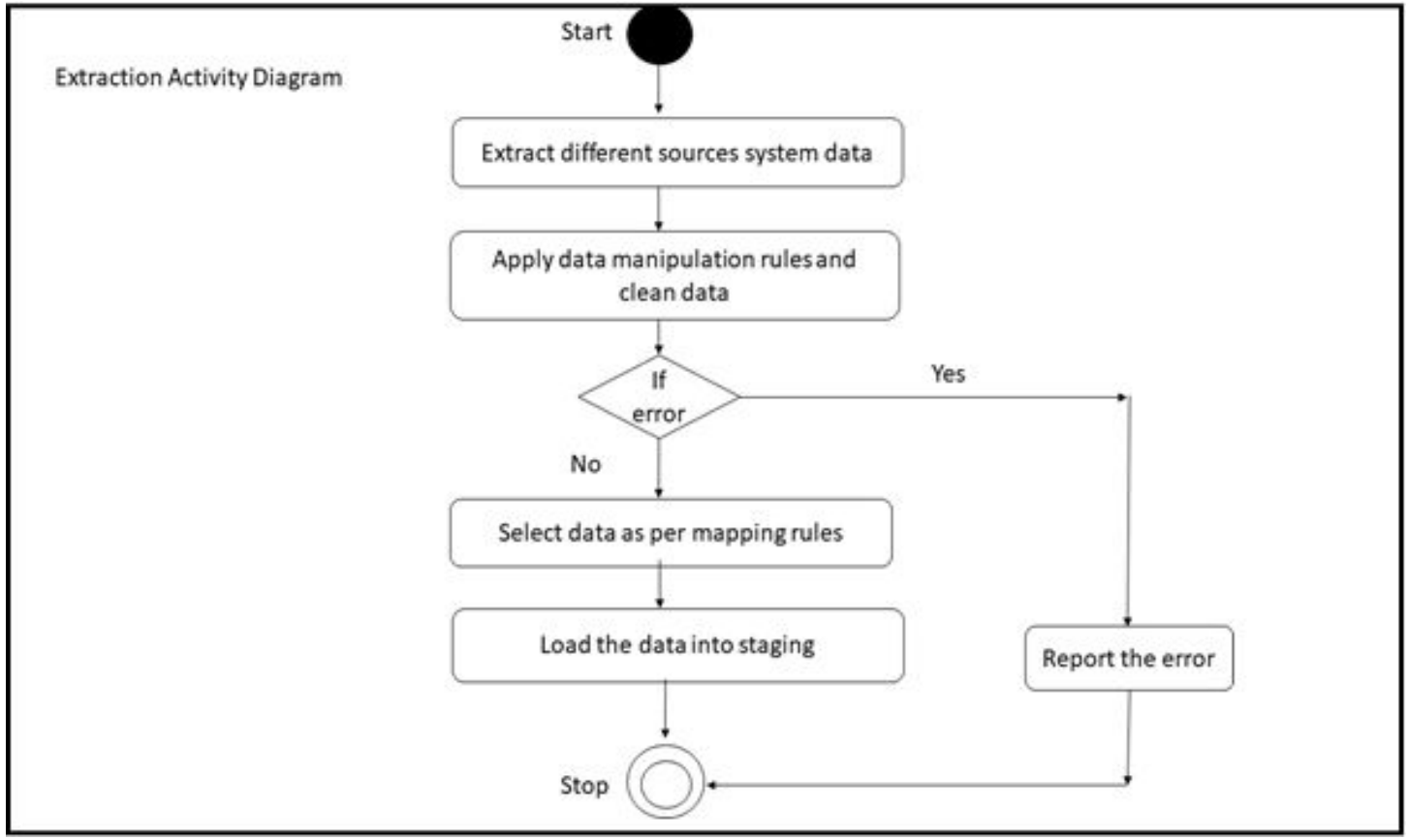
Use Case Diagram



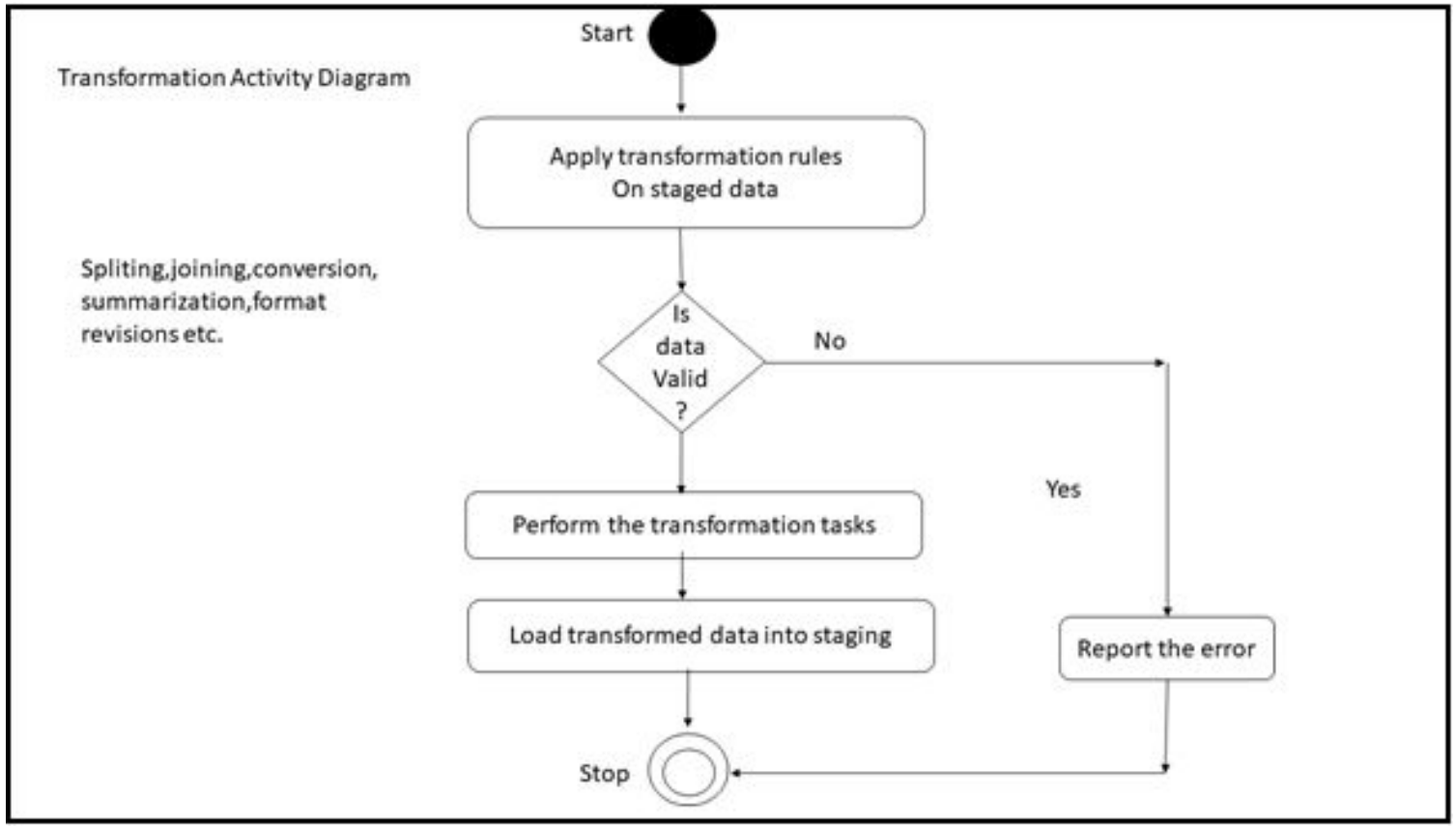
Business Use Case Diagram



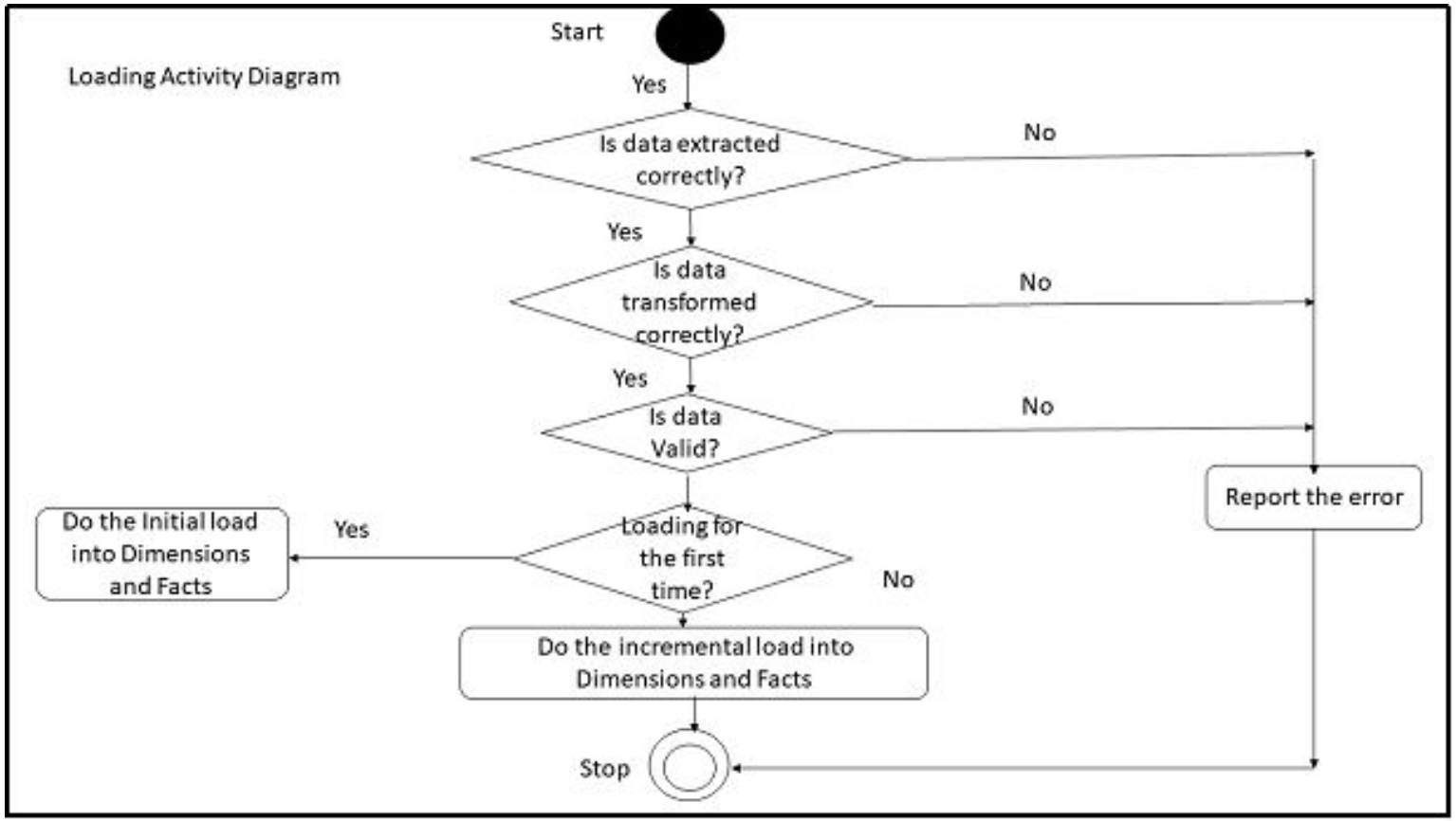
Activity Diagram (Extract)



Activity Diagram (Transform)



Activity Diagram (Load)



Design Of Target System

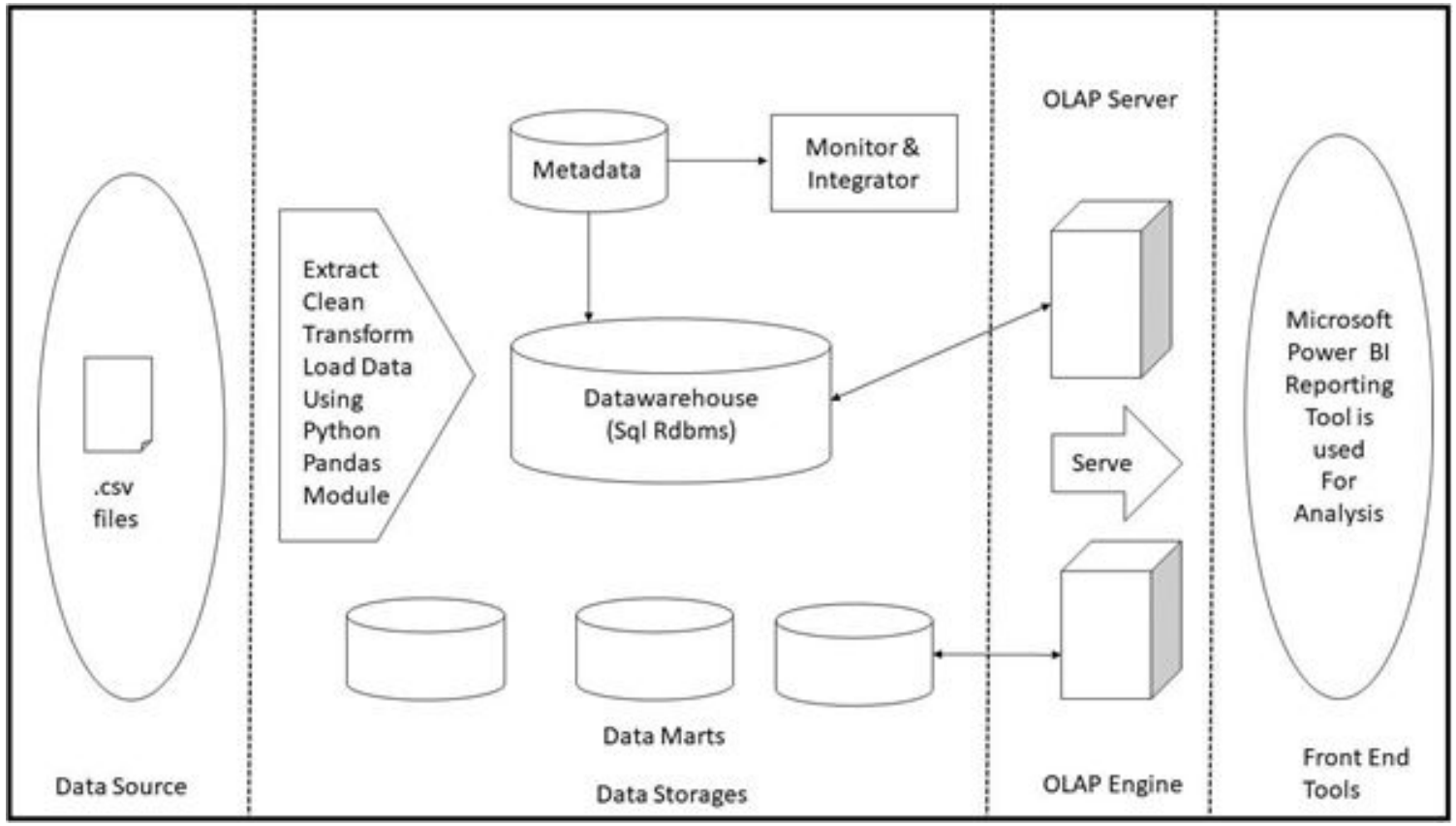


Table Specifications

Table 1

Chicago

Field Name	Data Type
trip_id	int
start_time	datetime
end_time	datetime
bikeid	int
tripduration	time
from_station_id	int
from_staion_name	varchar(20)
to_station_id	int
to_station_name	varchar(20)
usertype	char(20)
gender	char(20)
birthyear	year

Table 2

New York

Field Name	Data Type
tripduration	time
starttime	datetime
stoptime	datetime
start station id	int
start station name	varchar(20)
start station latitude	float
start station longitude	float
end station id	int
end station name	varchar(20)
end station latitude	float
end station longitude	float
bikeid	int
usertype	char(20)
Birth year	year
gender	int

Table 3

Washington DC

Field Name	Data Type
Duration	time
Start date	datetime
End date	datetime
Start station number	int
Start station	char(20)
End station number	int
End station	char(20)
Bike number	varchar(20)
Member Type	char(20)

Table 4

New Jersey City

Field Name	Data Type
tripduration	time
starttime	datetime
stoptime	datetime
start station id	int
start station name	varchar(20)
start station latitude	float
start station longitude	float
end station id	int
end station name	varchar(20)
end station latitude	float
end station longitude	float
bikeid	int
usertype	char(20)
birth year	year
gender	int

Table 5

Boston

Field Name	Data Type
tripduration	time
starttime	datetime
stoptime	datetime
start station id	int
start station name	varchar(20)
start station latitude	float
start station longitude	float
end station id	int
end station name	varchar(20)
end station latitude	float
end station longitude	float
bikeid	int
usertype	char(20)
birth year	year
gender	int

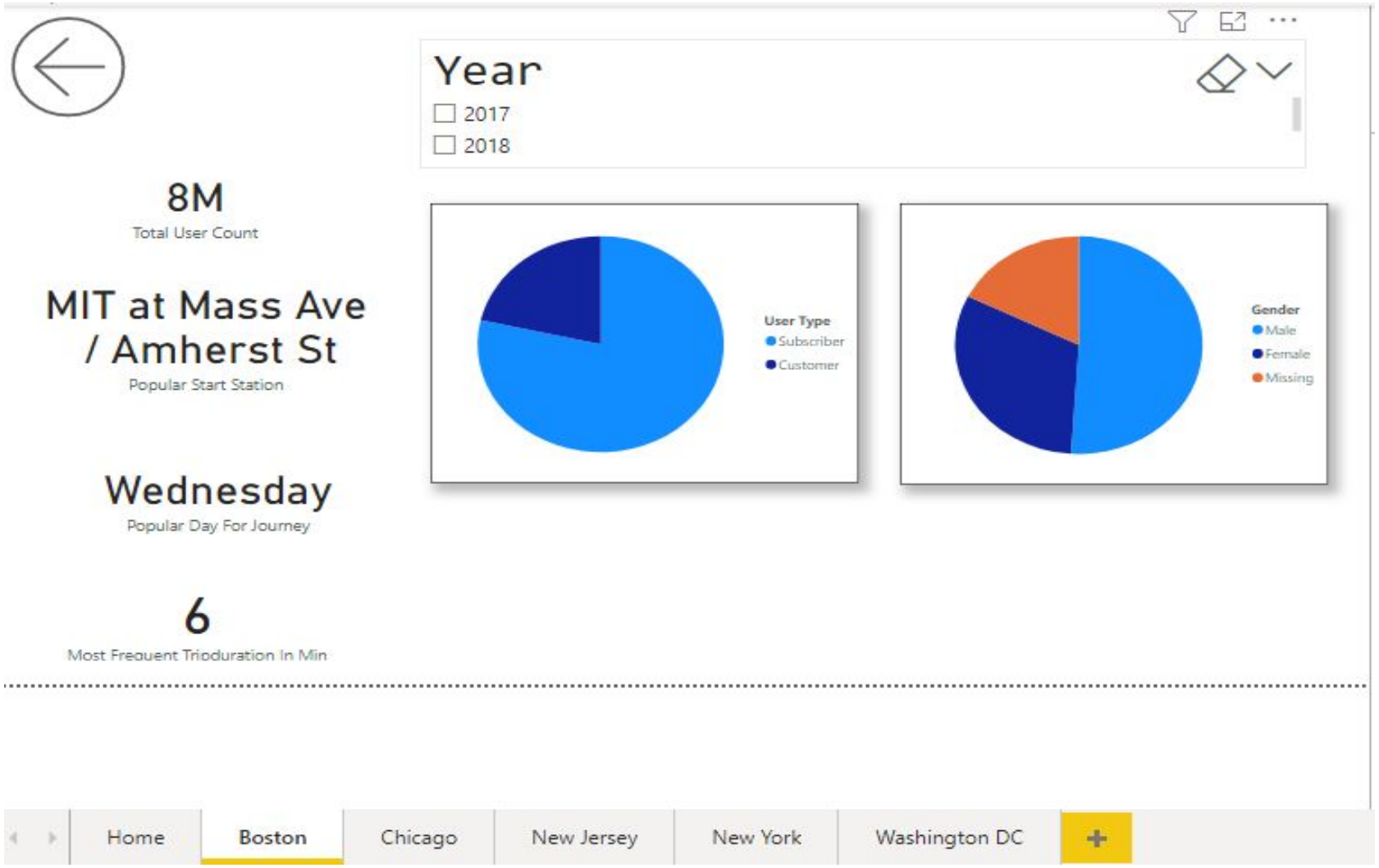
Technology

- **Operating System :-** Windows 10 Professional
- **RDBMS :-** MySQL
- **Language :-** Python *(Panda library for data ETL purpose)*
- **BI Tool :-** Microsoft Power BI *for reporting purpose*

Power BI Dashboard

Bike Ride Analysis

- Boston
- Chicago
- New Jersey
- New York
- Washington DC


Year
2017
2018
8M
Total User Count
MIT at Mass Ave / Amherst St
Popular Start Station
Wednesday
Popular Day For Journey
6
Most Frequent Triduration In Min
User Type
Subscriber
Customer
Gender
Male
Female
Missing
Home
Boston
Chicago
New Jersey
New York
Washington DC

Year

- 2017
- 2018

15M

User_Count

Streeter Dr & Grand Ave

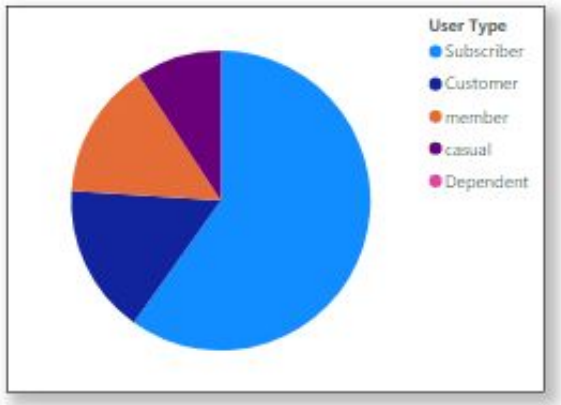
Popular_Start_Station

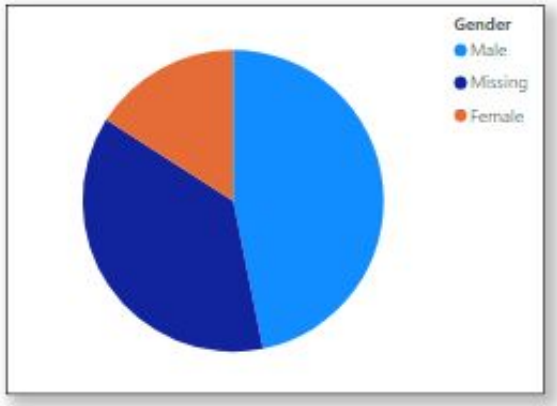
Thursday

Popular_Day_For_Journey

6

Most_Frequent_Tripduration_In_Min





Home | Boston | Chicago | New Jersey | New York | Washington DC

Year

- [ ] 2018
- [ ] 2019

58M

User-Count

Pershing Square North

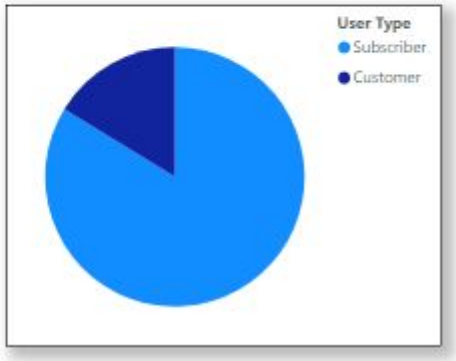
Popular-Start-Station

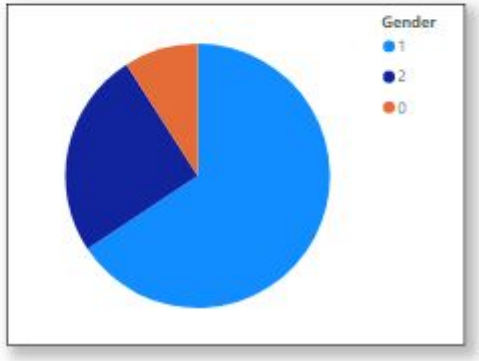
Wednesday

Popular-Day-For-Journey

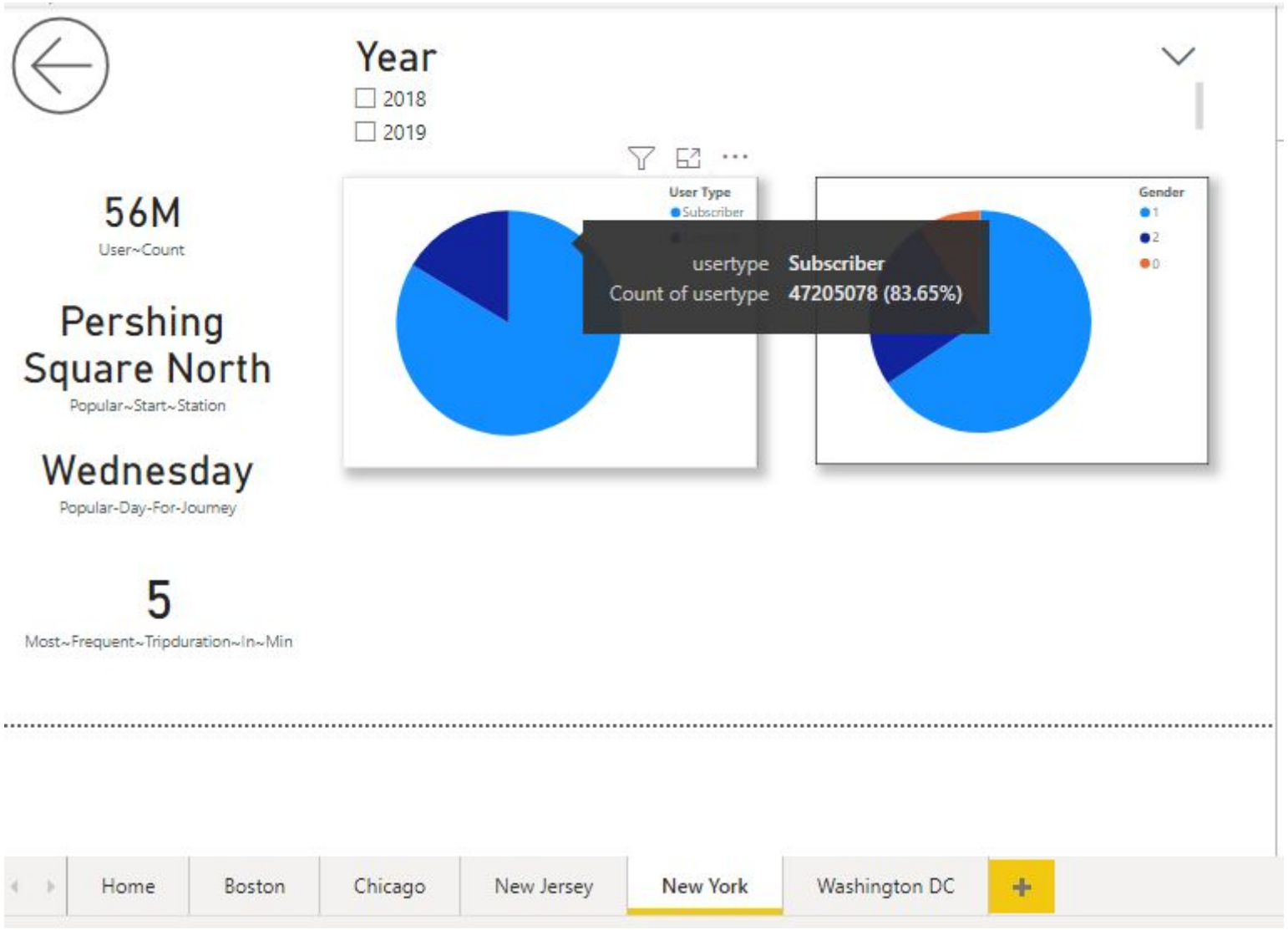
5

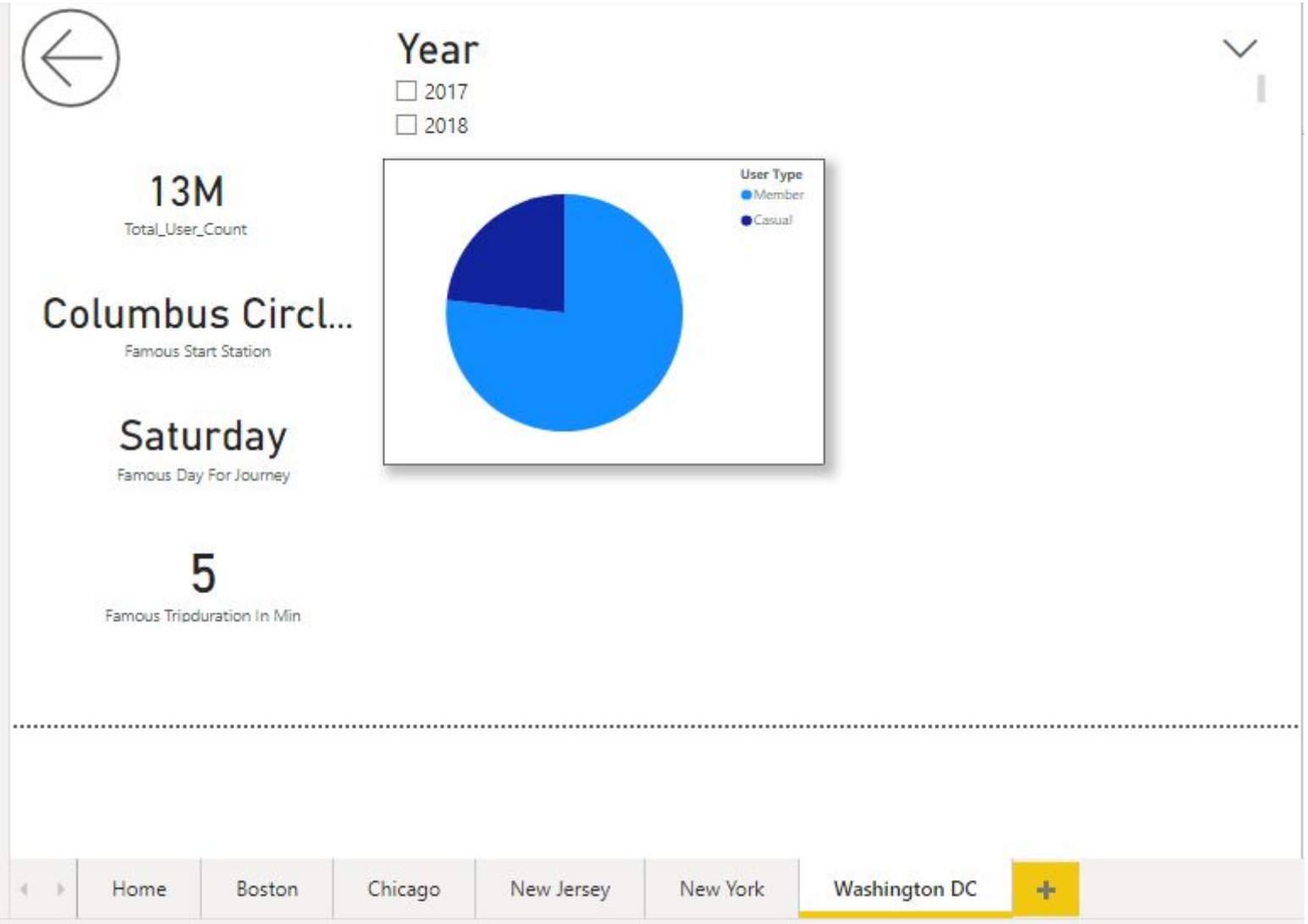
Most-Frequent-Tripduration-In-Min





Home | Boston | Chicago | New Jersey | New York | Washington DC | +


Year
2018
2019
56M
User~Count
Pershing Square North
Popular~Start~Station
Wednesday
Popular-Day-For-Journey
5
Most~Frequent~Tripduration~In~Min
User Type
Subscriber
usertype Subscriber
Count of usertype 47205078 (83.65%)
Gender
1
2
0
Home
Boston
Chicago
New Jersey
New York
Washington DC


Year
2017
2018
13M
Total_User_Count
Columbus Circl...
Famous Start Station
Saturday
Famous Day For Journey
5
Famous Tripduration In Min
User Type
Member
Casual
Home
Boston
Chicago
New Jersey
New York
Washington DC

Thank You!

Q & A