BIKE SHARE TREND ANALYSIS

Submitted by: Bhakti Raichura, Khushee Thakker, Shreya Srirama, Varsha Srinivasan

Abstract: It is estimated that there are more than 2 billion bikes used in the world and this number is constantly growing. It is also known that over 50 percent of the world's population knows how to use a bike. Owing to social and health benefits, people are now moving from car-sharing to the bike-sharing model and even further, e-bikes. The e-bike market is estimated to be valued at over $20B and continues to grow by the year 2030. In such a world where there are several bike and e-bike riders, our focus is to analyze the users' usage of bikes, their dependability on the weather conditions, and correlations with other nodes such as stations, trips, users' age groups. The data analysis is done in MySQL and Python, AWS for data warehousing and ETL and finally data visualization is done in Tableau. Through our trend analysis, we were able to gain insights into e-bike ride patterns and deduce suggestions to increase profitability for any e-bike-sharing company like Lime, Lyft etc.

Key Terms: Data warehouse, Data models, ETL process, Data Normalization, Visualization, Database, Datalakes

Motivation

In the world of increasing global temperature, bike riding not only provides health benefits but shifting trips from cars to bikes also helps reduce congestion, air pollution, and CO2 emissions. There are immense benefits of making this transition from driving cars to riding bikes. Through this project, we aim to highlight some of those benefits and highlight the patterns that can boost up the sales of such bikes. Our motive is to encourage not only the youth, but everyone to switch to riding bikes and e-bikes. Also, post 2020, the world saw an immense shortage of chips used in major vehicles creating a downfall for the automobile industry making the common man switch to riding bikes. We take this opportunity as helping the companies and manufacturers of such bike and e-bikes by identifying trends in the ride sharing of a typical user.

There are several studies done on bike sharing systems and predicting the usage of how external factors influence the patterns of e-bike ride sharing in people.

Literature Survey

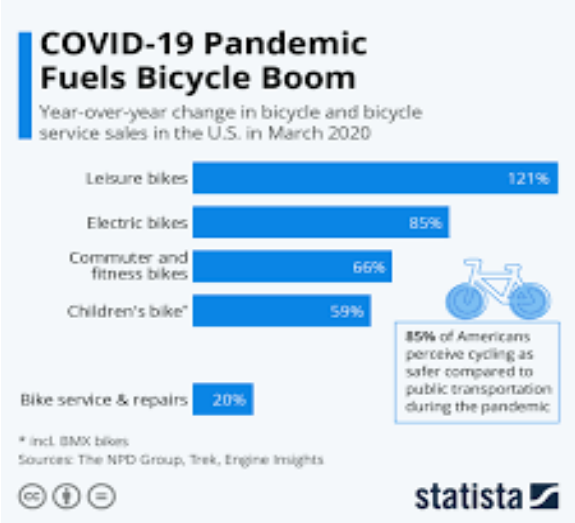
Nankervis's study shows the weather effects (long and short terms) on bicycle commute for tertiary students in Australia and Melbourne. The dataset focuses on students who are young and healthy, who commute using bicycles to their universities. The dataset includes various facets like places the students commuted, distance covered, riding time, % of riders based on months and semester weeks, etc. and helps identify the trends in seasonal behavior.

LarsBöcker, MartinDijst and JanFaber's research helps identify effect of weather conditions affecting choices of transport mode, travel experiences along outdoor thermal perceptions. Insights of users' emotional experiences influence travel behavior in terms of public and active mode transports are highlighted in this study

Elliot Fishmana, Simon Washington, Narelle Haworth, Angela Watsonc's study on the factors affecting membership in Australia's bike-sharing plan shows findings in the prospective increase of potential users in the future and the reliability of the transport system using a logistic regression model. It draws conclusion on several external and internal factors which decide whether or not a user will purchase membership to a bike share model.

Industry influence

The current automobile industry is at an all-time cost high. This is owing to the problems in the supply chain process post Covid-19. Due to this, people have switched gears to riding bikes and e-bikes and the below chart shows trends in purchase patterns of bikes across the United States. Studies have shown that people felt more comfortable riding bikes during the pandemic than using public transport



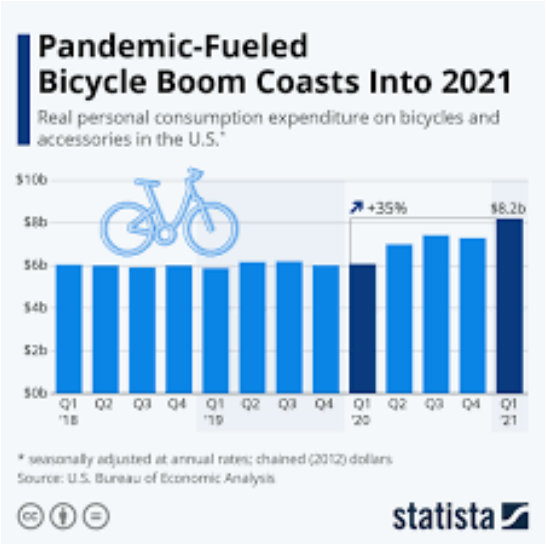


Fig 1

Fig 2

Fig 1. shows how the sales of bikes have boomed during Covid-19. It also depicts how electric bike sales have increased in the US starting March 2020. Another depiction in Fig 2. shows quarter over quarter growth in bike riding in 2021.

Several statistics and surveys show a booming demand for e-bikes in the coming year, estimatingET the industry to be over $60B by the year 2030.

Project Flow and design

We used multiple tools and technologies in the project.

Flow 1:

- The raw csv files were loaded to Talend (ETL Tool) where we performed normalization and cleaned the data.
- The schema was created in a MySQL database from the Mysql connector in python and then this transformed data was loaded into it.

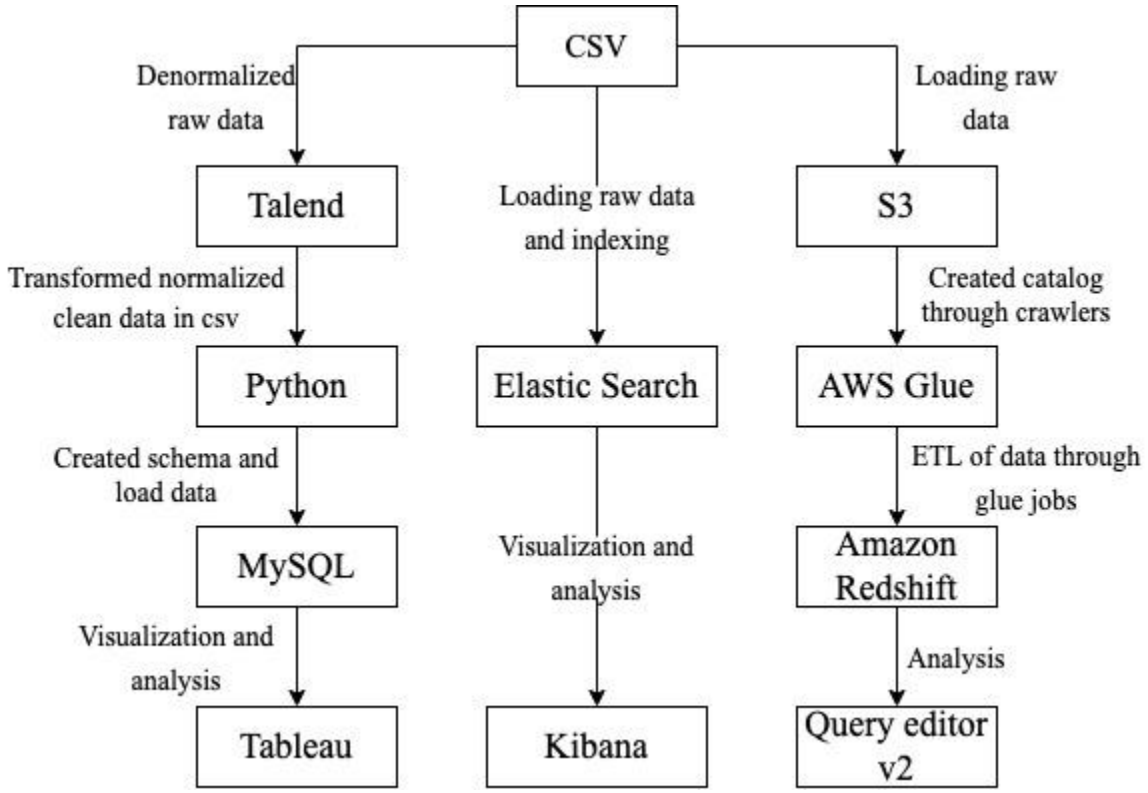
- And then we did analysis using various SQL queries in MySQL database
- Visualization of this analysis was done by connecting MySQL to Tableau

Flow 2:

- The raw dataset was loaded to S3 bucket (Datalake)
- The data was crawled to AWS Glue (ETL Tool) where we performed transformation of this data to start schema.
- In Amazon Redshift (Analytical tool) where we created our schema and performed analysis of this data in query editor v2 using various SQL queries

Flow 3:

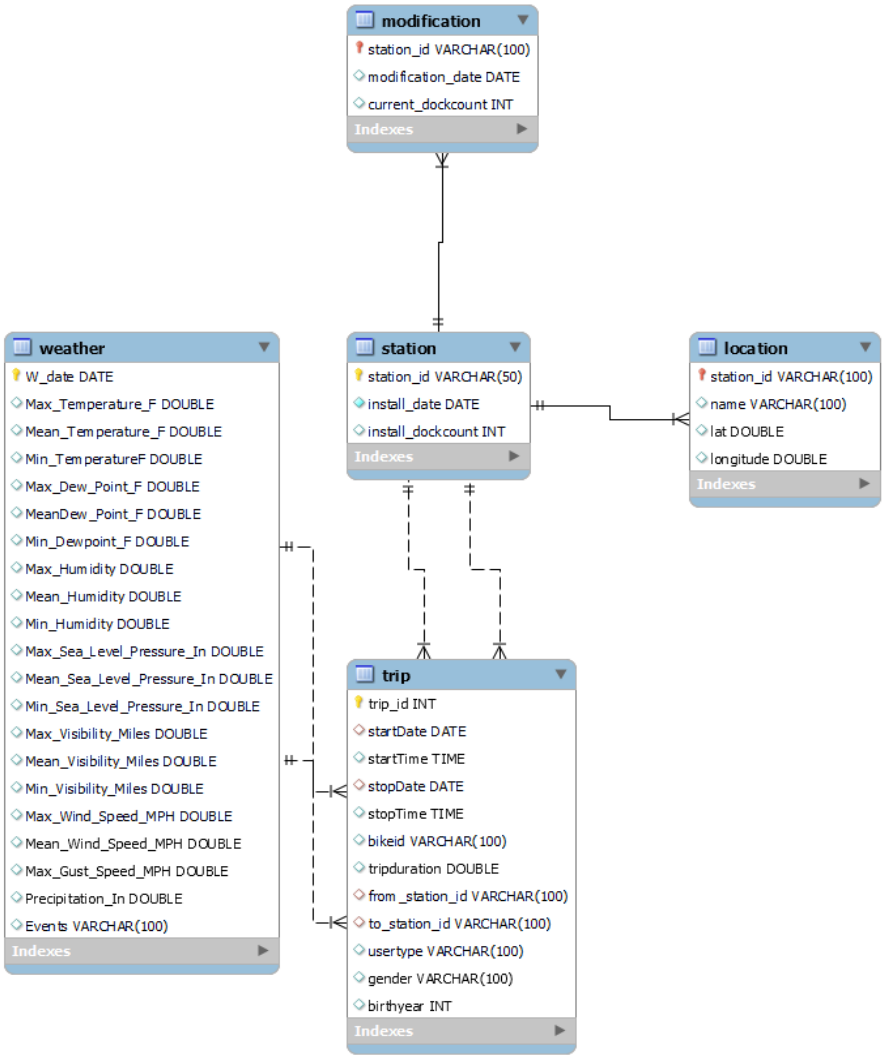
- Loaded the raw data into Elastic search (NoSQL database) and provided indexing
- The visualization and analysis of this data was performed using Kibana



Data Model

a) ER diagram

The below Entity- Relation diagram represents 5 tables used in the project and their attributes and relationships. The tables used are trip, station, weather, modification and location



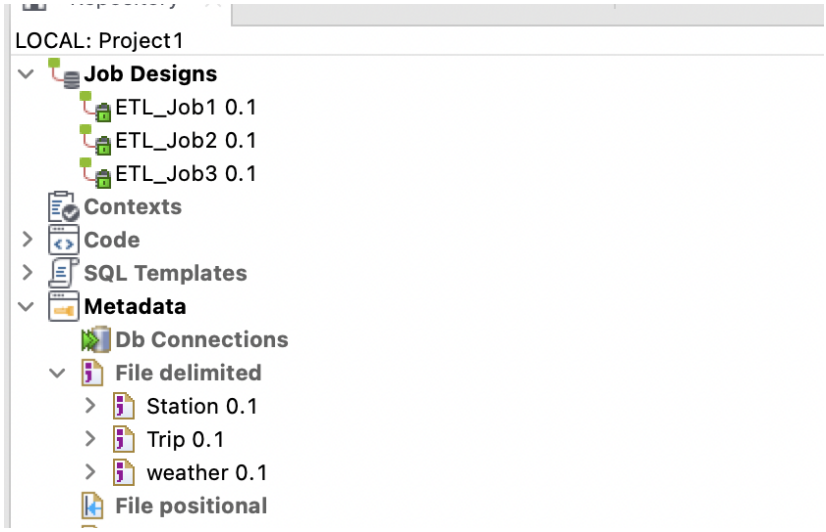
Datawarehouse:

ETL
Talend

Raw dataset contains 3 files

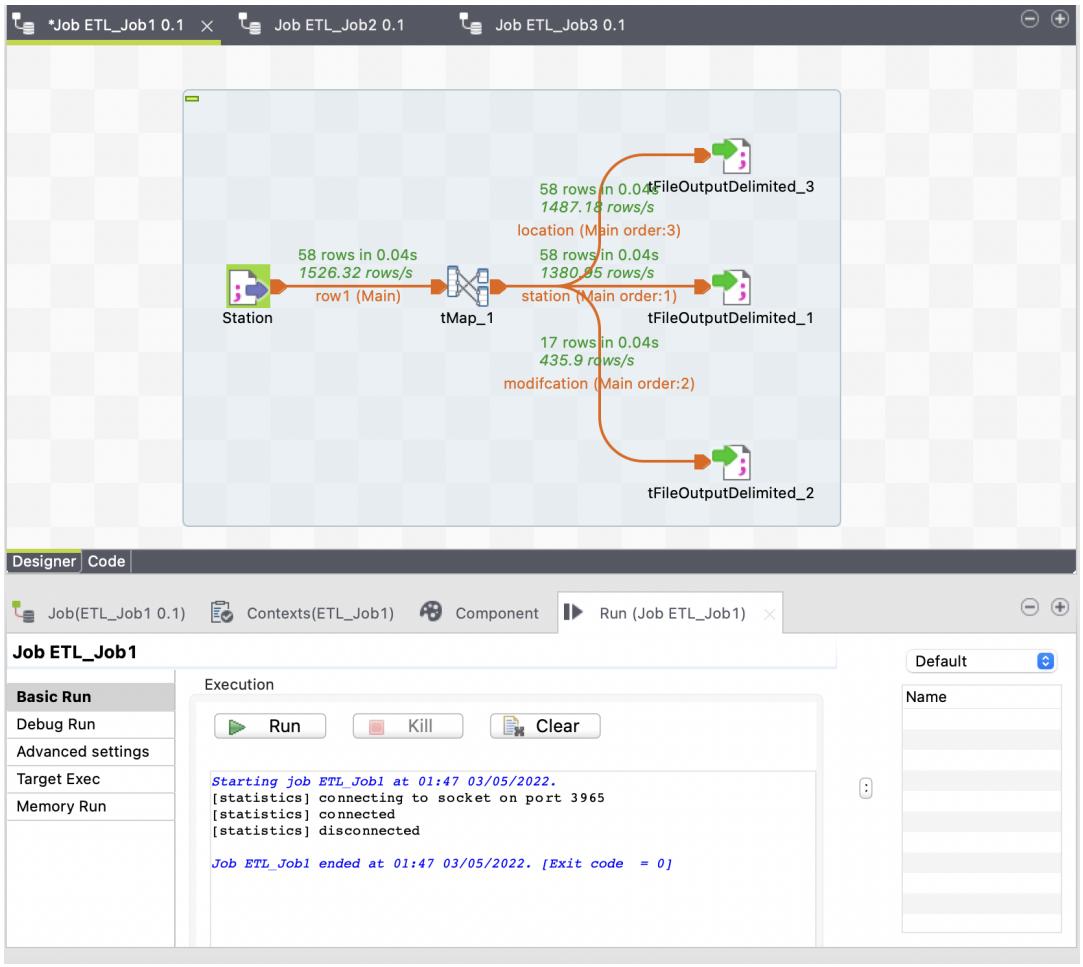
- ➔ Station.csv
- ➔ Trip.csv
- ➔ Weather.csv

So we created 3 metadata schema in Talend and respective ETL jobs to process each of these datasets

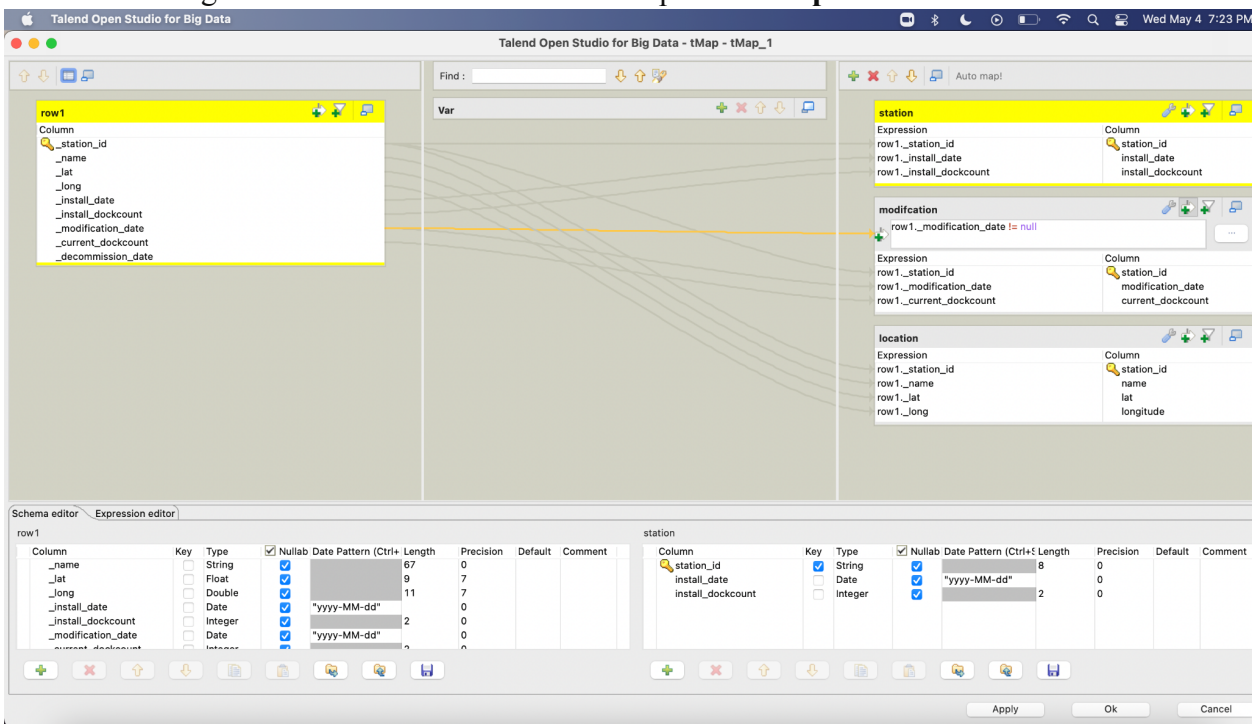


ETL_Job1:

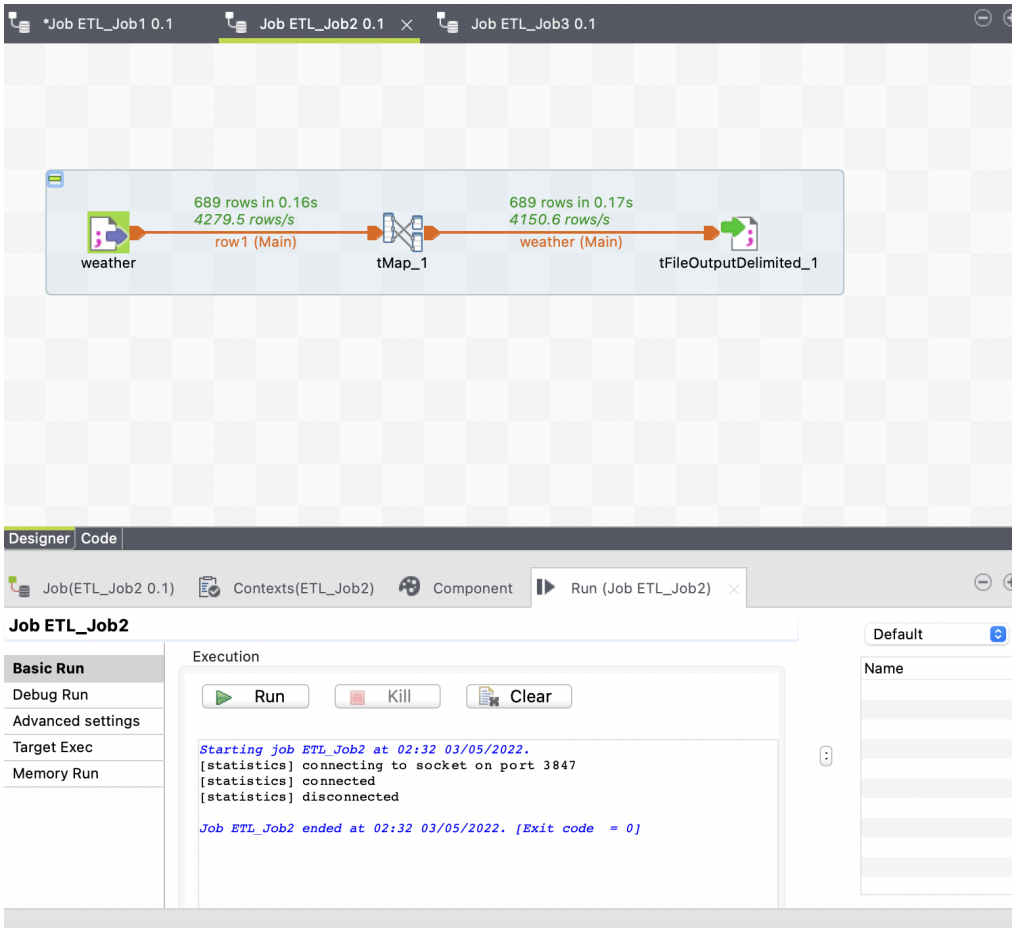
In this job our goal was to normalize the station data as there were insert and delete anomalies. So we split our dataset into 3 new tables as shown in the below snippet.

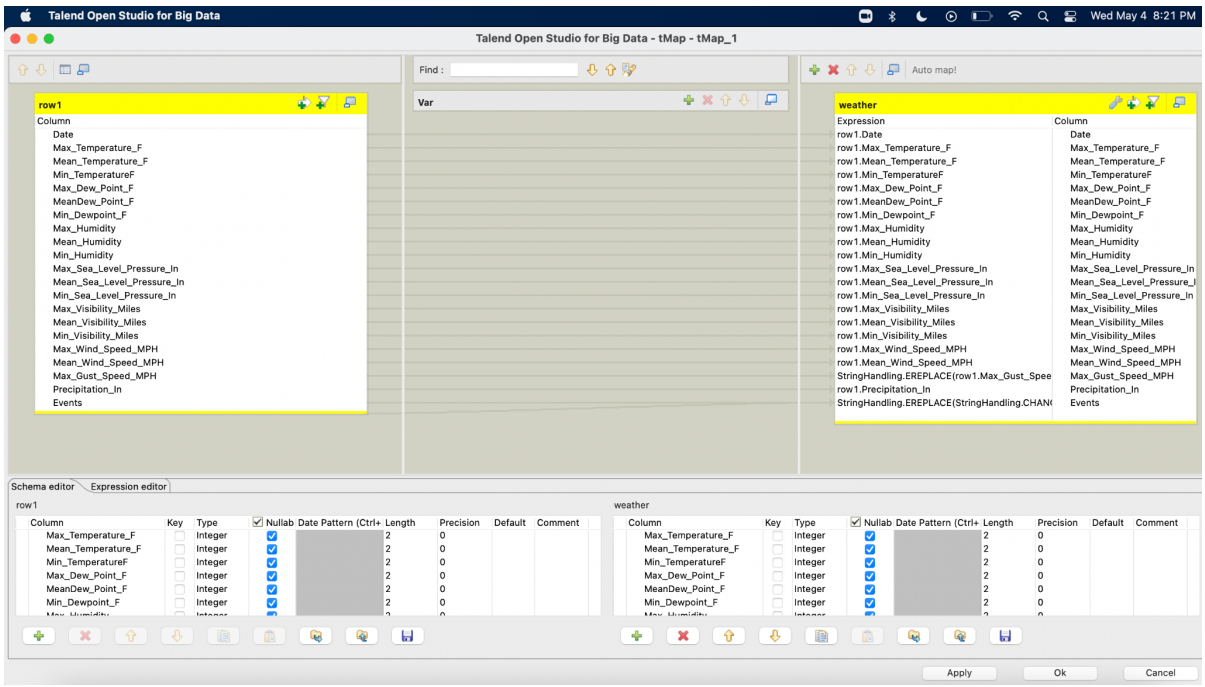


To achieve this goal we used Talend's advance component **tMap.**

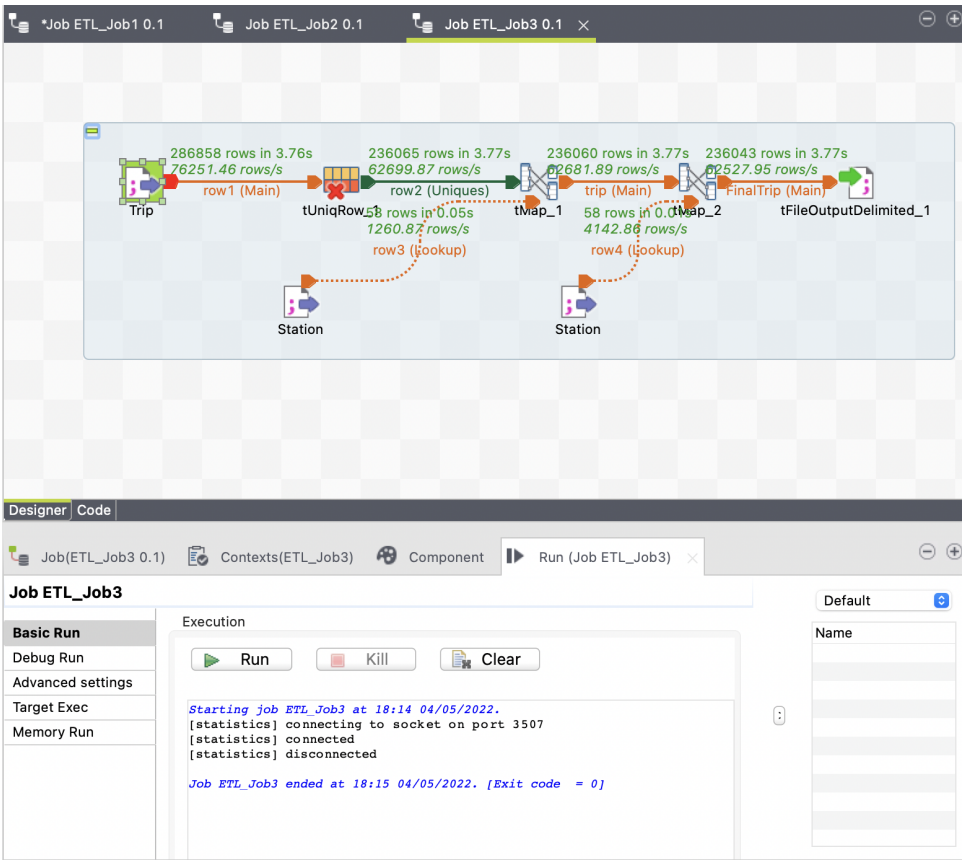


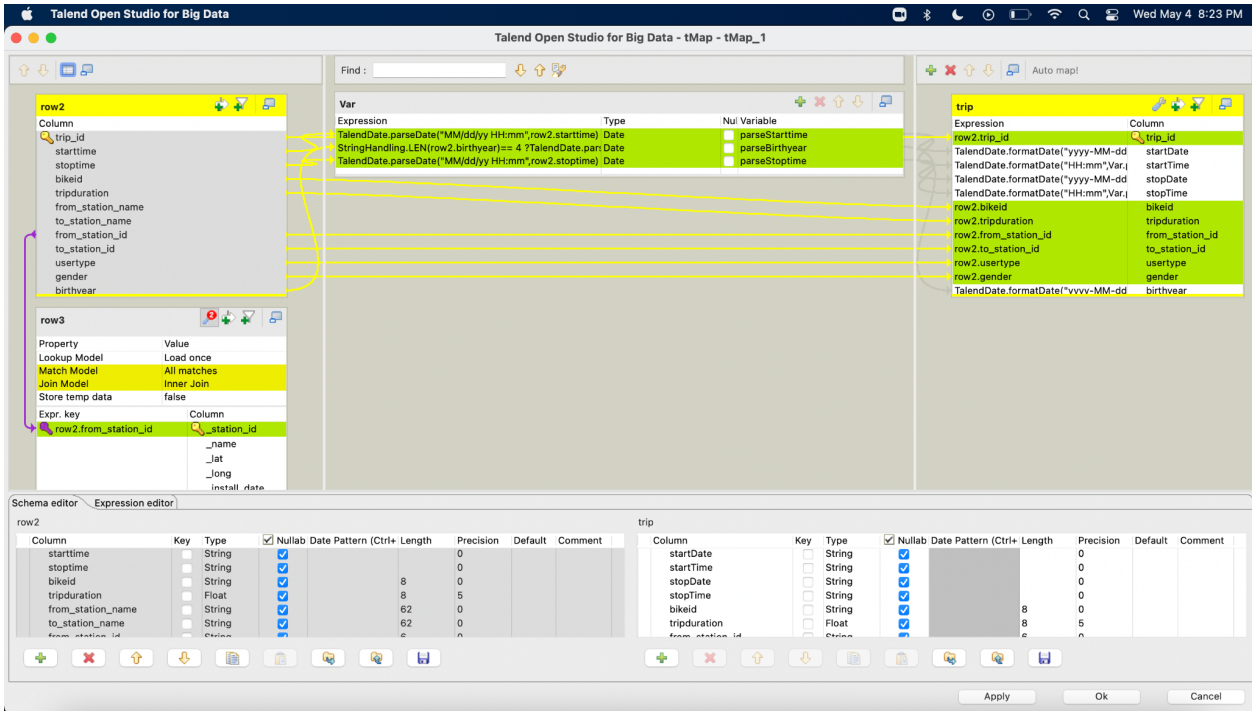
ETL_Job2
In this job our goal was to clean the weather dataset

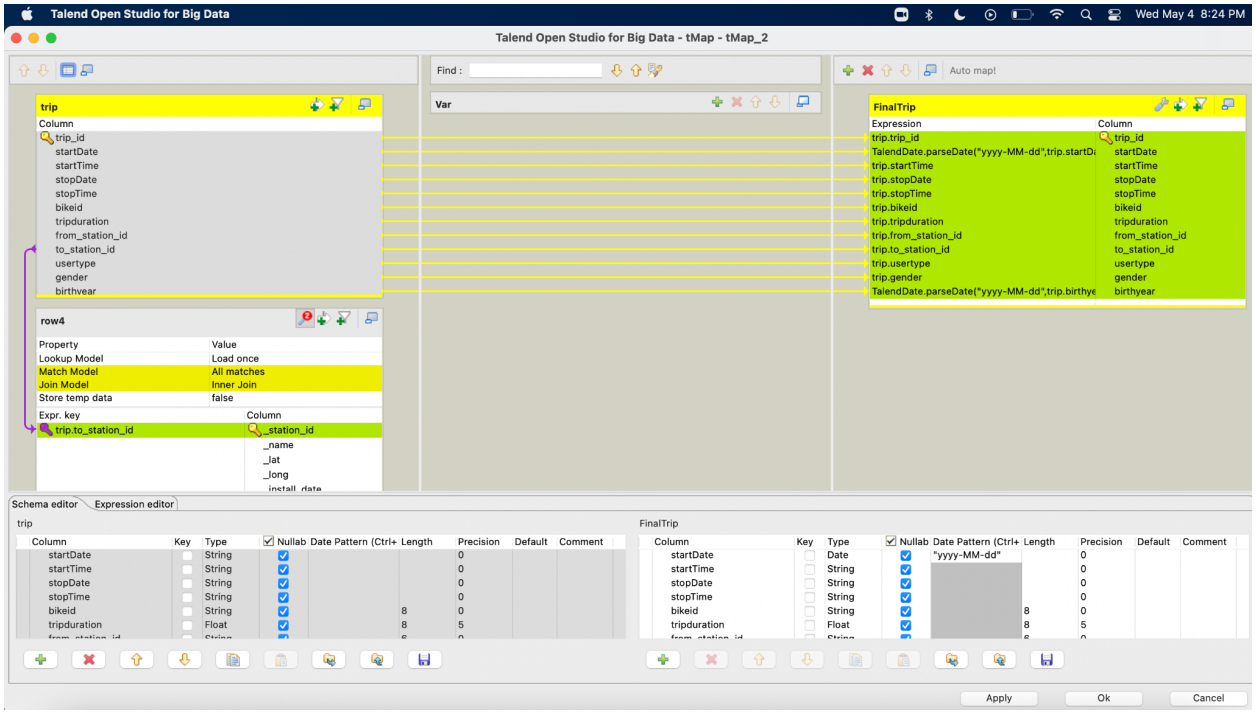

*Job ETL_Job1 0.1
Job ETL_Job2 0.1
Job ETL_Job3 0.1
689 rows in 0.16s
4279.5 rows/s
row1 (Main)
689 rows in 0.17s
4150.6 rows/s
weather (Main)
weather
tMap_1
tFileOutputDelimited_1
Designer
Code
Job(ETL_Job2 0.1)
Contexts(ETL_Job2)
Component
Run (Job ETL_Job2)
Job ETL_Job2
Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run
Execution
Run
Kill
Clear
Starting job ETL_Job2 at 02:32 03/05/2022.
[statistics] connecting to socket on port 3847
[statistics] connected
[statistics] disconnected
Job ETL_Job2 ended at 02:32 03/05/2022. [Exit code = 0]
Default
Name


Talend Open Studio for Big Data
Wed May 4 8:21 PM
Talend Open Studio for Big Data - tMap - tMap_1
row1
Column
Date
Max_Temperature_F
Mean_Temperature_F
Min_TemperatureF
Max_Dew_Point_F
MeanDew_Point_F
Min_Dewpoint_F
Max_Humidity
Mean_Humidity
Min_Humidity
Max_Sea_Level_Pressure_In
Mean_Sea_Level_Pressure_In
Min_Sea_Level_Pressure_In
Max_Visibility_Miles
Mean_Visibility_Miles
Min_Visibility_Miles
Max_Wind_Speed_MPH
Mean_Wind_Speed_MPH
Max_Gust_Speed_MPH
Precipitation_In
Events
Find :
Var
Auto map!
weather
Expression
row1.Date
row1.Max_Temperature_F
row1.Mean_Temperature_F
row1.Min_TemperatureF
row1.Max_Dew_Point_F
row1.MeanDew_Point_F
row1.Min_Dewpoint_F
row1.Max_Humidity
row1.Mean_Humidity
row1.Min_Humidity
row1.Max_Sea_Level_Pressure_In
row1.Mean_Sea_Level_Pressure_In
row1.Min_Sea_Level_Pressure_In
row1.Max_Visibility_Miles
row1.Mean_Visibility_Miles
row1.Min_Visibility_Miles
row1.Max_Wind_Speed_MPH
row1.Mean_Wind_Speed_MPH
StringHandling.EREPLACE(row1.Max_Gust_Spee
row1.Precipitation_In
StringHandling.EREPLACE(StringHandling.CHAN(
Schema editor
Expression editor
Key
Type
Nullab
Date Pattern (Ctrl+
Length
Precision
Default
Comment
Integer
Apply
Ok
Cancel

ETL_Job3


*Job ETL_Job1 0.1
Job ETL_Job2 0.1
Job ETL_Job3 0.1
286858 rows in 3.76s
76251.46 rows/s
row1 (Main)
Trip
236065 rows in 3.77s
62699.87 rows/s
row2 (Uniques)
tUniqRow_1
236060 rows in 3.77s
62681.89 rows/s
trip (Main)
tMap_1
236043 rows in 3.77s
62527.95 rows/s
FinalTrip (Main)
tMap_2
tFileOutputDelimited_1
58 rows in 0.05s
1260.87 rows/s
row3 (Lookup)
58 rows in 0.01s
4142.86 rows/s
row4 (Lookup)
Station
Station
Designer
Code
Job(ETL_Job3 0.1)
Contexts(ETL_Job3)
Component
Run (Job ETL_Job3)
Job ETL_Job3
Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run
Execution
Run
Kill
Clear
Starting job ETL_Job3 at 18:14 04/05/2022.
[statistics] connecting to socket on port 3507
[statistics] connected
[statistics] disconnected
Job ETL_Job3 ended at 18:15 04/05/2022. [Exit code = 0]
Default
Name


Talend Open Studio for Big Data
Wed May 4 8:23 PM
Talend Open Studio for Big Data - tMap_1
row2
Column
trip_id
starttime
stoptime
bikeid
tripduration
from_station_name
to_station_name
from_station_id
to_station_id
usertype
gender
birthyear
row3
Property
Value
Lookup Model
Load once
Match Model
All matches
Join Model
Inner Join
Store temp data
false
Expr. key
Column
row2.from_station_id
_station_id
_name
_lat
_long
Find :
Var
Expression
Type
Nul
Variable
TalendDate.parseDate("MM/dd/yy HH:mm",row2.starttime)
Date
parseStarttime
StringHandling.LEN(row2.birthyear)== 4 ?TalendDate.par
Date
parseBirthyear
TalendDate.parseDate("MM/dd/yy HH:mm",row2.stoptime)
Date
parseStoptime
Auto map!
trip
Expression
Column
row2.trip_id
trip_id
TalendDate.formatDate("yyyy-MM-dd
startDate
TalendDate.formatDate("HH:mm",Var.
startTime
TalendDate.formatDate("yyyy-MM-dd
stopDate
TalendDate.formatDate("HH:mm",Var.
stopTime
row2.bikeid
bikeid
row2.tripduration
tripduration
row2.from_station_id
from_station_id
row2.to_station_id
to_station_id
row2.usertype
usertype
row2.gender
gender
TalendDate.formatDate("yyyy-MM-dd
birthyear
Schema editor
Expression editor
row2
trip
Column
Key
Type
Nullab
Date Pattern (Ctrl+
Length
Precision
Default
Comment
starttime
stoptime
bikeid
tripduration
from_station_name
to_station_name
startDate
startTime
stopDate
stopTime
String
Float
Apply
Ok
Cancel



AWS Glue

Analysis Using Database Techniques

We connected MySQL with Python to import cleaned data into the MySQL database.

```
In [53]: import mysql.connector

In [54]: import pandas as pd

In [55]: db=mysql.connector.connect(
             host='localhost',
             user='root',
             passwd='[redacted]')
```

Created a ‘Cycle’ database for bike sharing analysis.

```
In [44]: ##Creating a new database

mycursor.execute("CREATE DATABASE cycle")
db=mysql.connector.connect(
    host='localhost',
    user='root',
    passwd='root',
    database='cycle')

mycursor=db.cursor()
```

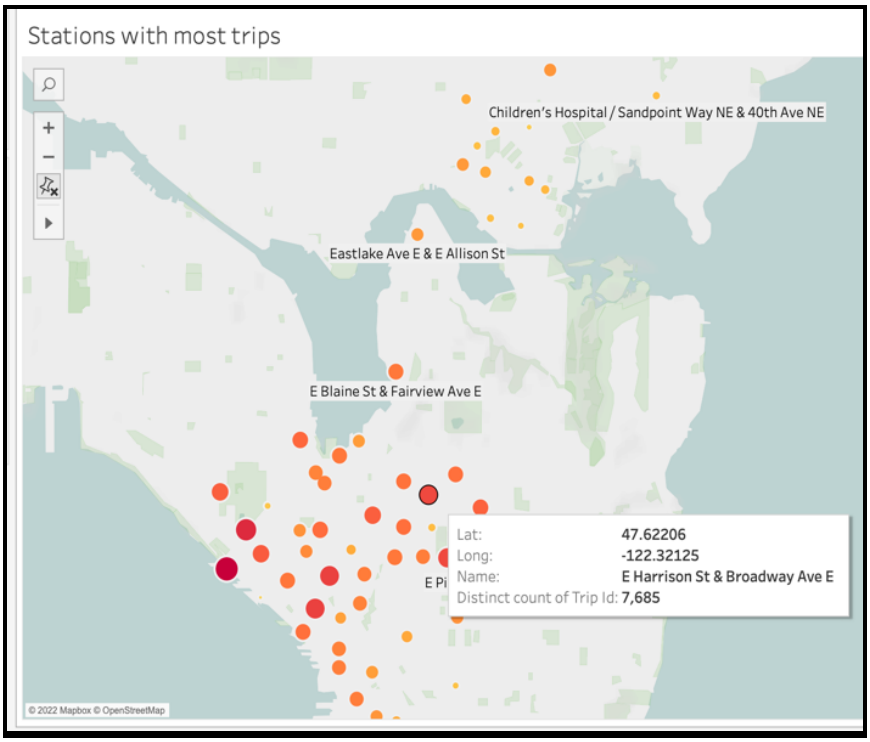
RDBMS for Analysis:

MySQL is a relational database and uses structured query language.
We used MYSQL for storing and modeling our data. We can fetch data using simple queries and perform useful analysis.
Some interesting analysis we performed using MYSQL:

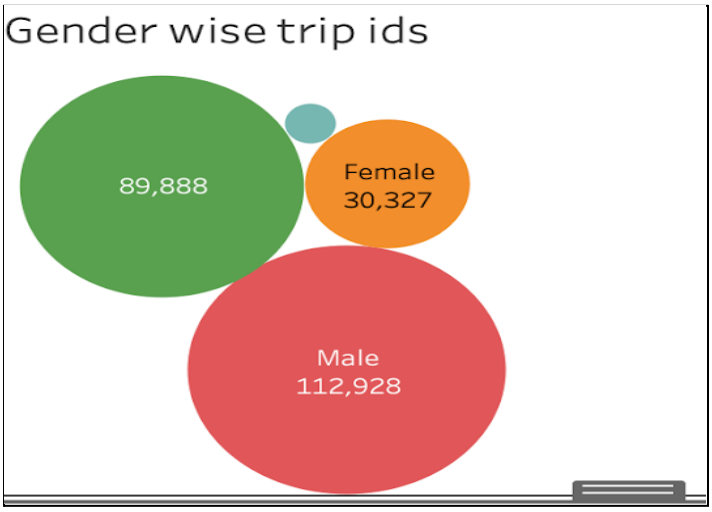
Data Visualization

To represent our analysis in a diagrammatic format, we used data visualization using Tableau interactive dashboard. Some of the visuals we derived from our analysis are shown below

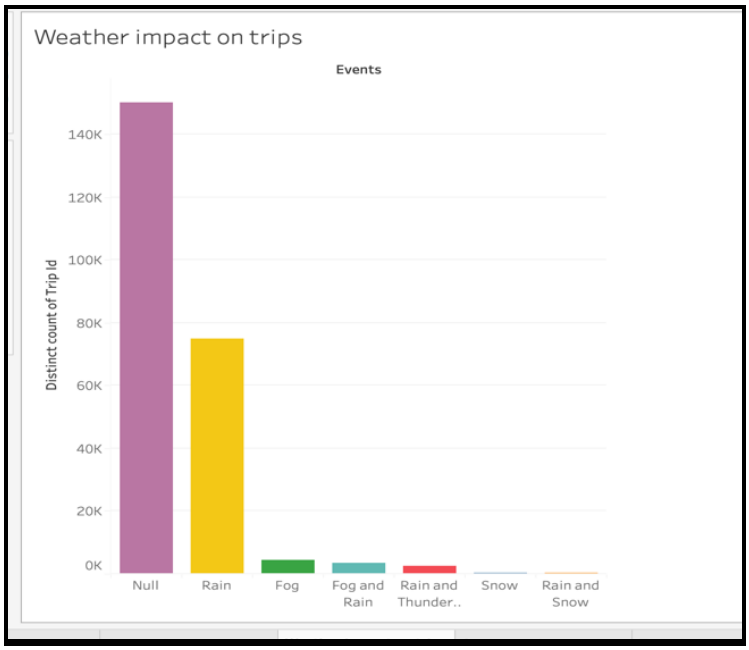
- Busiest stations- stations having highest demands for bikes- The bigger circles with darker red indicate business/high demand



- Gender wise rides taken- The below bubble chart indicates that there are more males than females who have taken rides. The green bubble indicates the user who have not defined their gender or are not members on the ride sharing platform

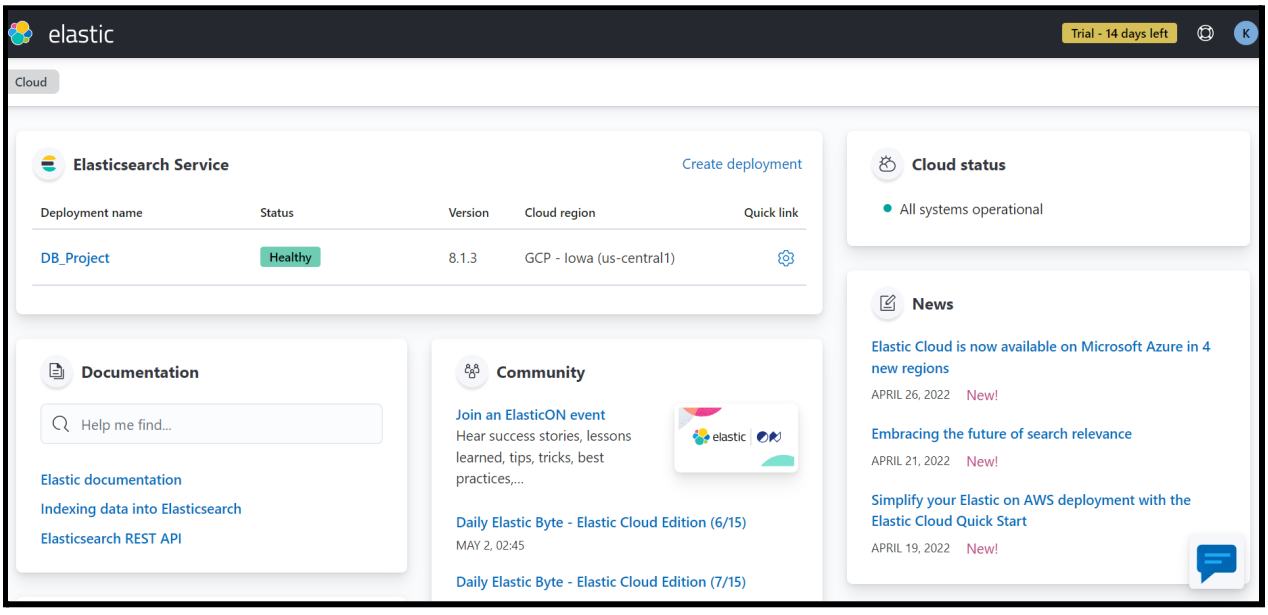


- Impact of weather on trips- We can see

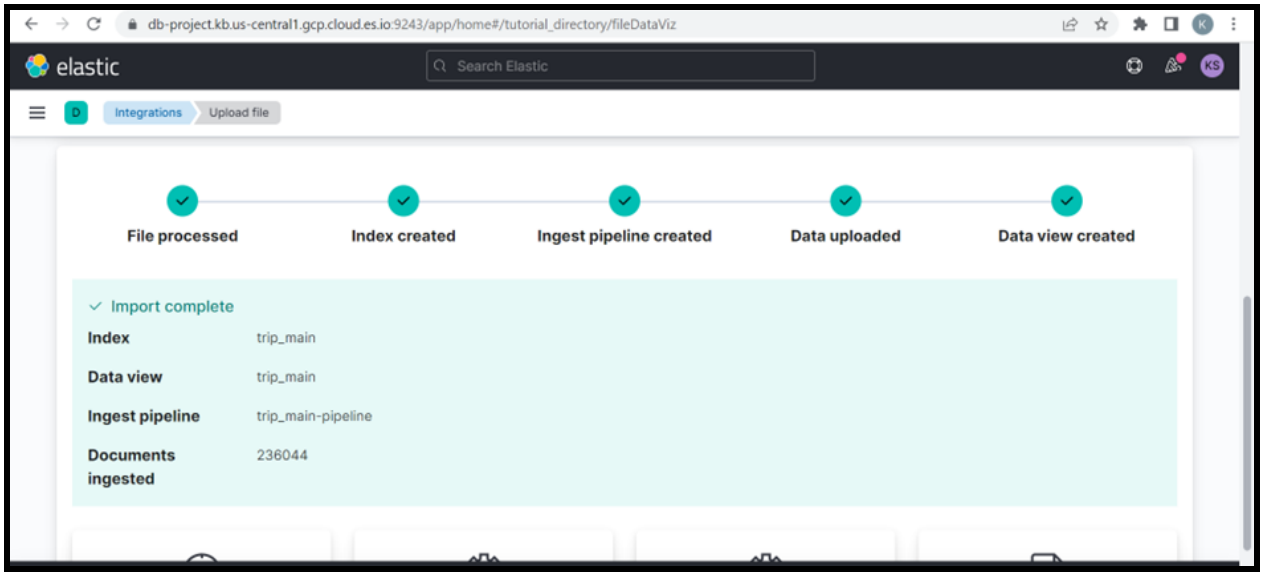


New Database:
ElasticSearch: Elastic search is a document oriented NoSQL database. It allows us to store a huge volume of data and search and analyze it effortlessly.

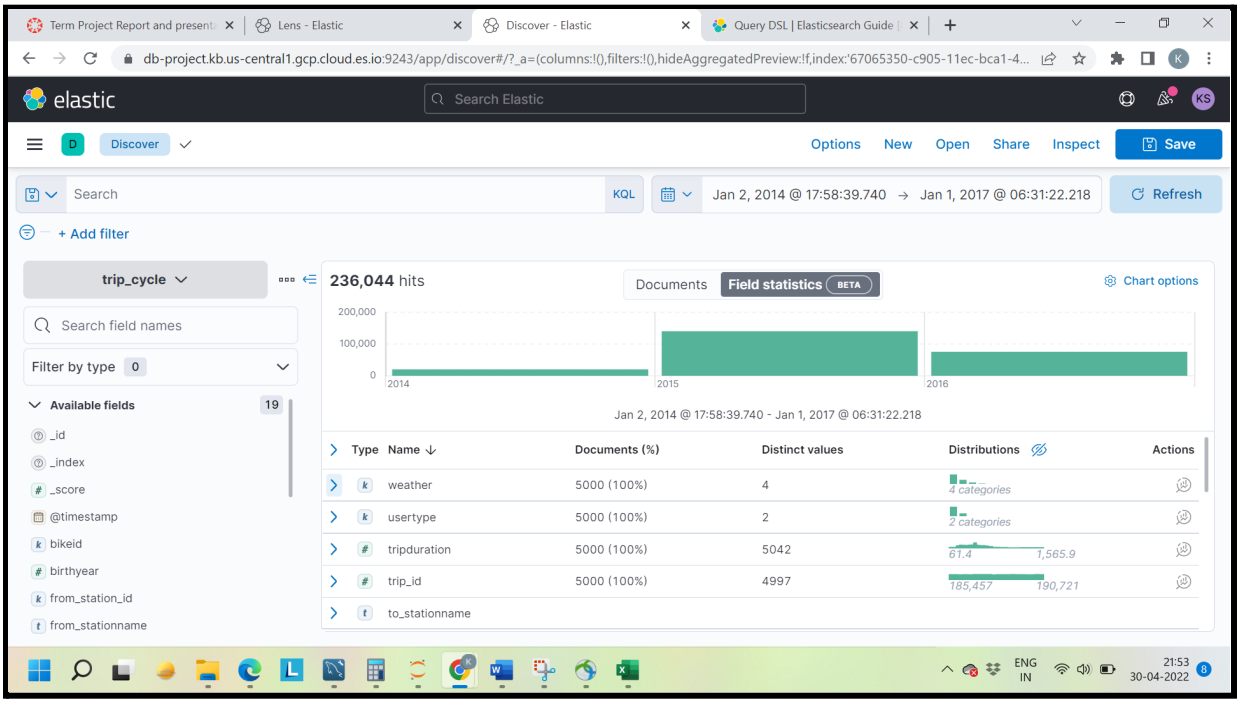
Create a deployment: Created DB_Project deployment in elastic search web browser.

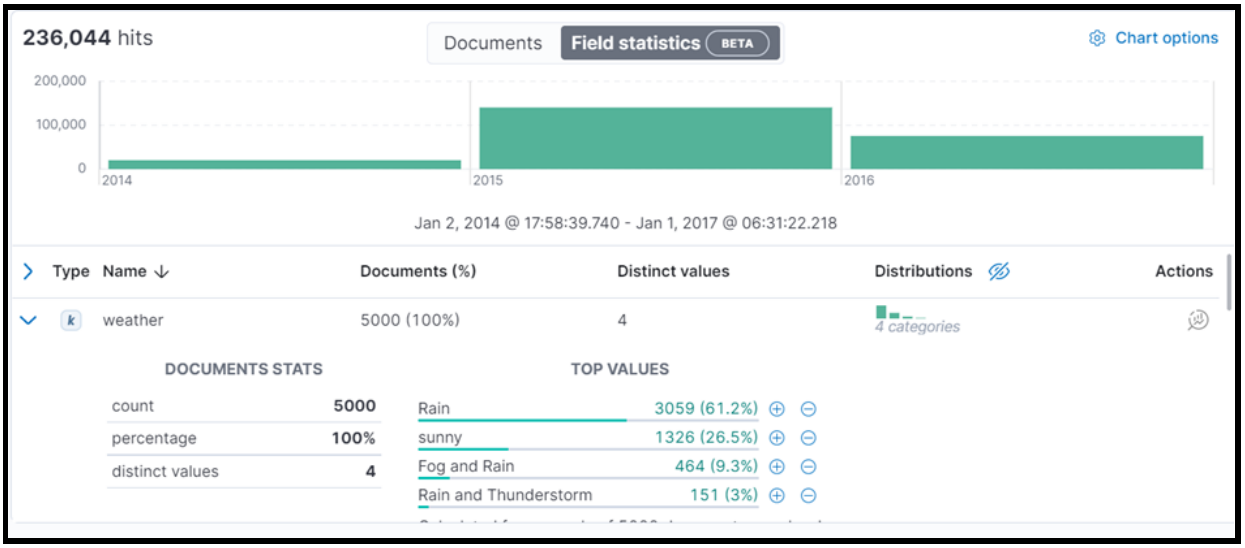


We uploaded denormalized data ‘tripmain’ which was in csv format.



It is very easy and quick to view and search the data using Discover in elastic . We can analyze details of our data. We can check distinct values in a column, check top values and view different analyses.



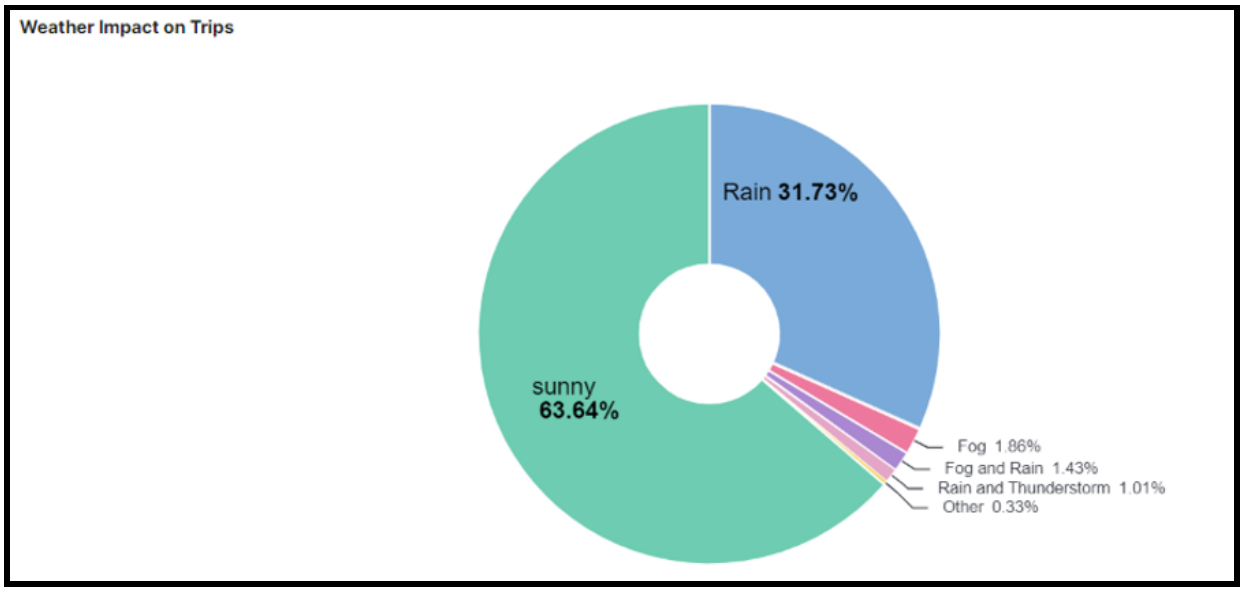


We represented our data as a graph using Graph in elastic. Elasticsearch uses the concept of indexing. It will collect data with similar characteristics. We can query against these indexes. In backend elastic search will have clusters and nodes. Cluster is a collection of similar nodes. Elasticsearch groups all nodes with similar characteristics. We can add clusters and relationships between them. We added our trip dataset, made clusters such as bikeid ,trip_id, startDate, stopDate,from_station_id, from_stationname, to_station_id, to_stationname, usertype, gender, birth year, weather and mapped them.

Created interactive visualization in Kibana, a feature of elastic search. Kibana Lens allows you to create a visualization by simply dragging and dropping elements.

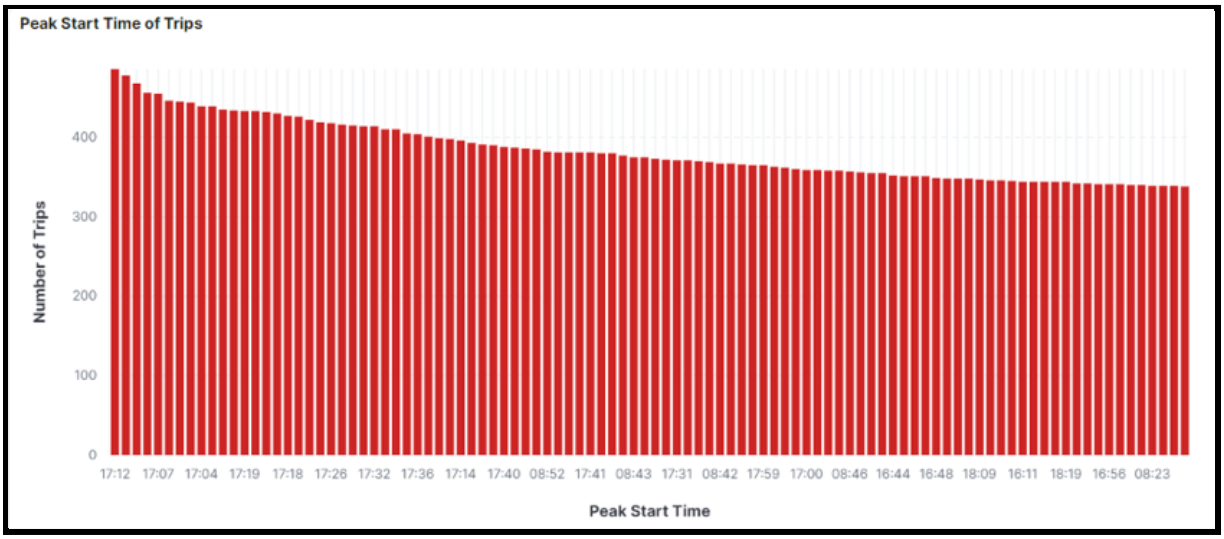
Does weather impact the number of trips taken?

We can evidently observe that when the weather is sunny there are more people who prefer to ride a bicycle.



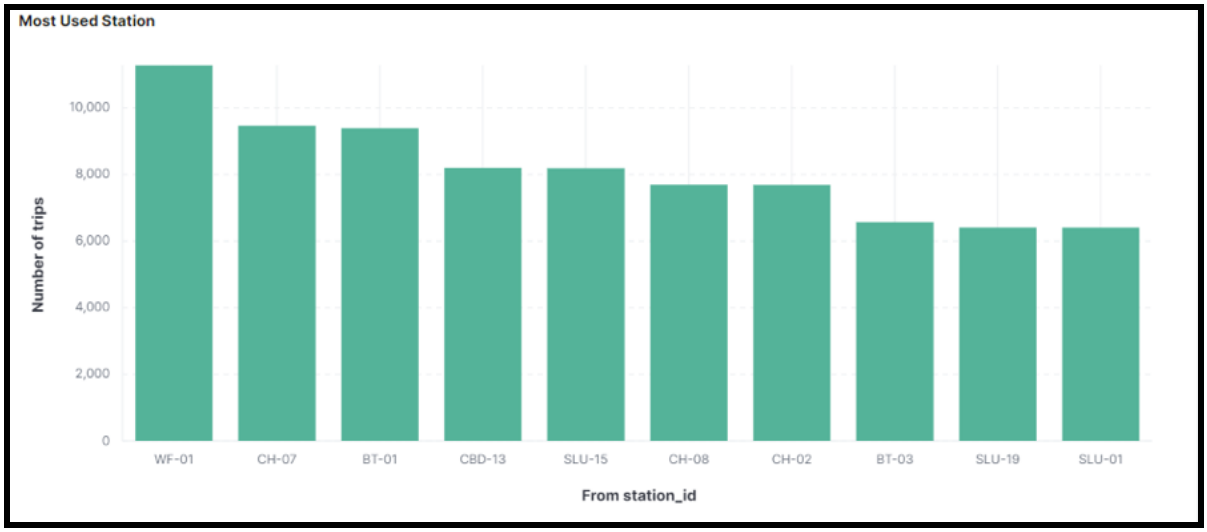
What is the most peak time for trips?

According to the graph below most trips were between 4.30pm to 5 pm. The company can increase the number of bicycles for rent during that time.



Which are the most popular stations to start the trip?

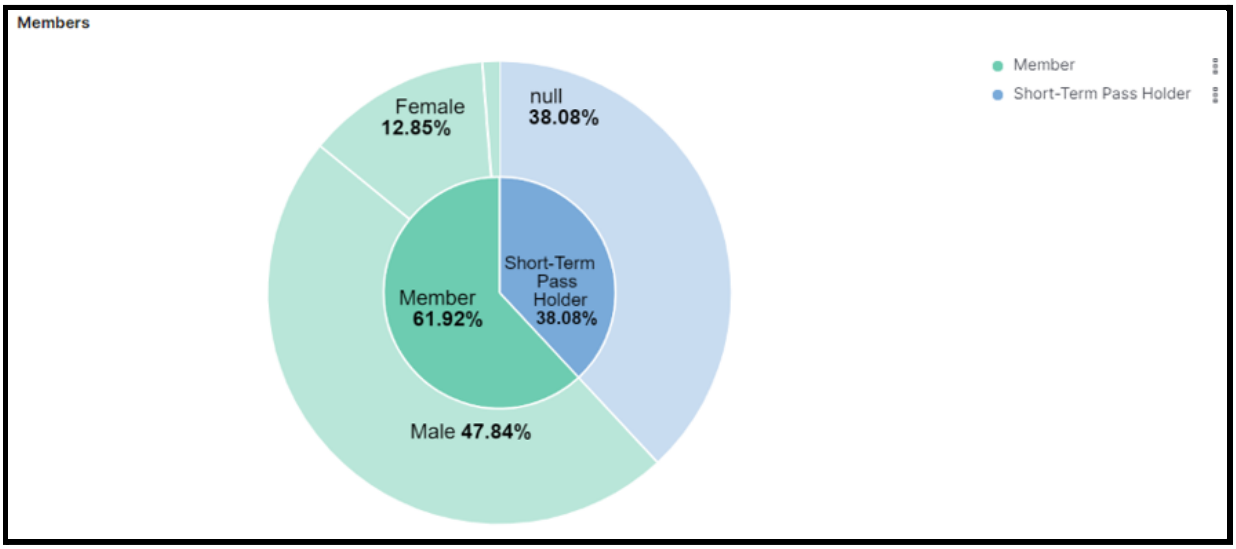
WF-01 is the most popular station from where the rides started.



Also the same can be understood by the graph. Using KQL(Kibana Query Language) we can filter our data. We queried ‘from_station_id : WF-01 and bikeid : * ‘. This displayed all the bike id's and the bike id’s which started their trip from the station ‘WF-01’ were connected to them. We can also see in the summary that 11274 bike ids were started from the station ‘WF-01’.

What are the different memberships and which gender rents more bikes?

There are two types of membership plans, a person can become a member or can be a short term pass holder. There are 61.92% of people who prefer being a member and the percentage of males enrolling for membership is more.



Key Findings