IT3021 - DATA WAREHOUSING AND BUSINESS INTELLIGENCE

ASSIGNMENT 01

Submitted By:

Student Name: De Silva M.

Student Number: IT20207854

Batch: Y3.S1.WE.DS.04.02

DATE: 10/05/2022

Table of Contents

1. Data Selection and Introduction

The dataset used for this assignment is from a Bicycle Rental System in Los Angeles, California metropolitan area, USA. The original data has been edited and re-arranged to suit the requirements of the assignment. This dataset contains data from 1st January 2016 to 1st January 2017.

Scenario:

This system allows customers to rent a bike by paying a fee. Customers are given discounts based on their loyalty category(2 star, 3 star or 5 star). The system has multiple stations located within the state of California. Customers can rent and return the bike from the station closest to their start and destination. Details about the customers, the bicycles rented, stations used to borrow and return the bike and payments made are recorded by the rental company.

Hierachies

- Hierachies in the Customer dimension: State→ City→ Street

Link to the dataset:

https://www.kaggle.com/datasets/lukexun/lost-angeles-metro-bike-share

1.1. ER Diagram

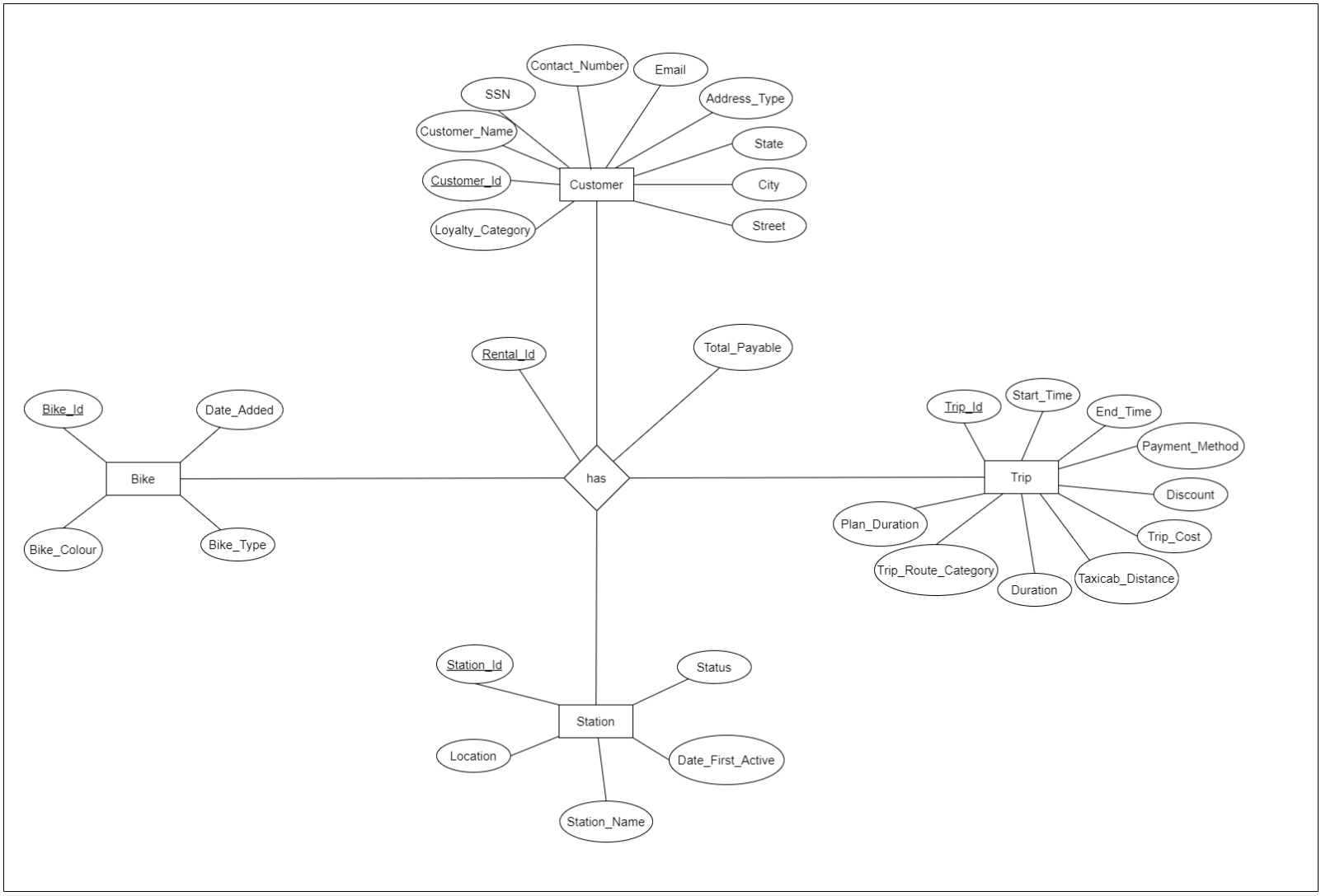

Figure 1: ER Diagram for Bicycle Rental System

2. Preparation of the Data Sources

Data from 3 data sources were used for this assignment.

1. **Database - SLIIT_Bike_SourceDB:**

- dbo.Bike – contains details about the bikes available for rental
- dbo.Station – contains details about stations that the customers can rent and return bikes at
- dbo.Trip – contains details about the trips the rented bikes were used for.

2. **Text Files:**

- Customer.txt – contains details about customers who have rented bikes
- Customer_Address.txt – contains the addresses of the customers

3. **Excel files:**

- Bike_Rentals.xlsx – contains details about each bike rental

3. Solution Architecture

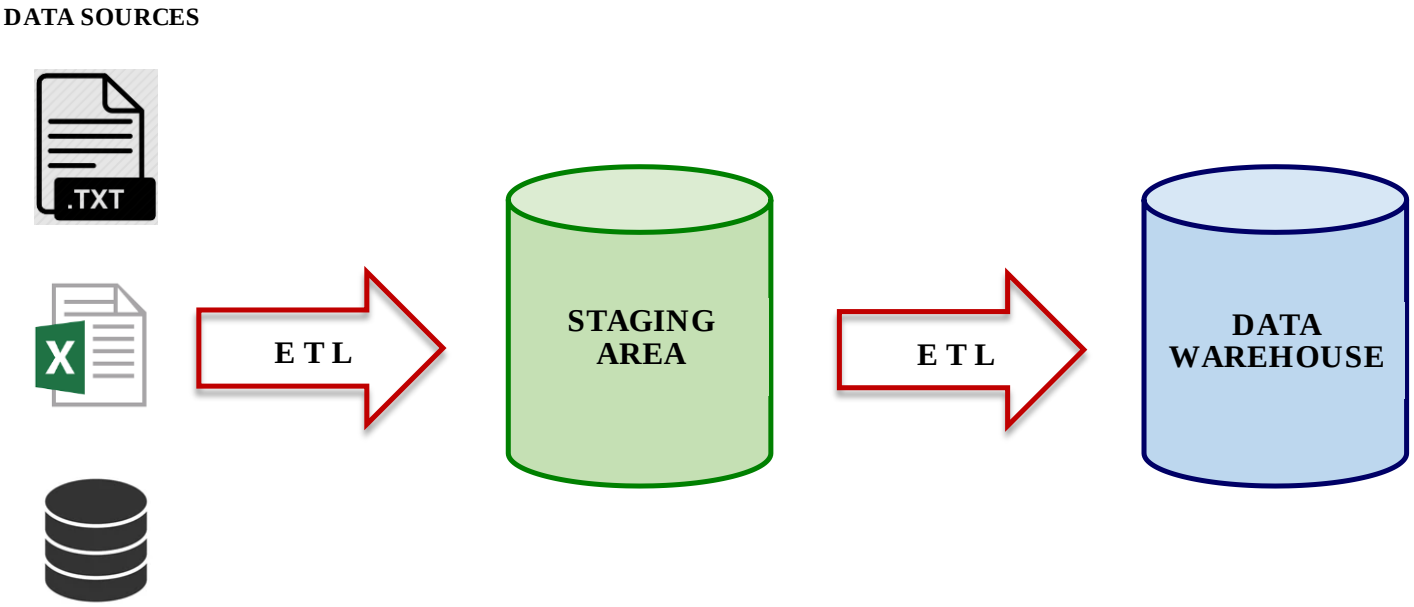

Figure 2: Architectural Diagram for Bike Rental Data Warehouse

- **Data Sources:**

This is where the data included in the data warehouse comes from. There can be various data sources such as databases, text files, excel files and csv files. Three data sources; a database, text files and an excel file are used for this solution

- **Staging Area:**

The staging area is an intermediate location between the data sources and the data warehouse. It is a temporary storage area where the data from data sources is gathered into prior to being loaded into the data warehouse. The data staging process imports data from the sources, changes it and produces integrated, cleaned data, and stages it for loading into the data warehouse.

In this solution, the staging area is implemented in the form of a database; 'SLIIT_Bike_Staging'.

- **ETL Process**:

ETL, which stands for Extract, Transform and Load, is a data integration process that combines data from multiple data sources into a single, consistent data store, transforms it in the staging area and finally loads it into a data warehouse.

The first step of the ETL process is Extraction. In this step, data from various source systems is extracted into the staging area.

The second step of the ETL process is Transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes such as Filtering, Cleaning, Sorting and Joining.

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse.

- **Data Warehouse:**

A data warehouse is a type of data management system that centralizes and consolidates large amounts of data from multiple sources to enable and support business intelligence (BI) activities such as analytics, reporting and data mining. Data warehouses often contain large amounts of historical data.

The data warehouse in this solution is the database ‘SLIIT_Bike_DW’.

4. Data Warehouse Design & Development

Diagram depicting the relationships between the dimensions

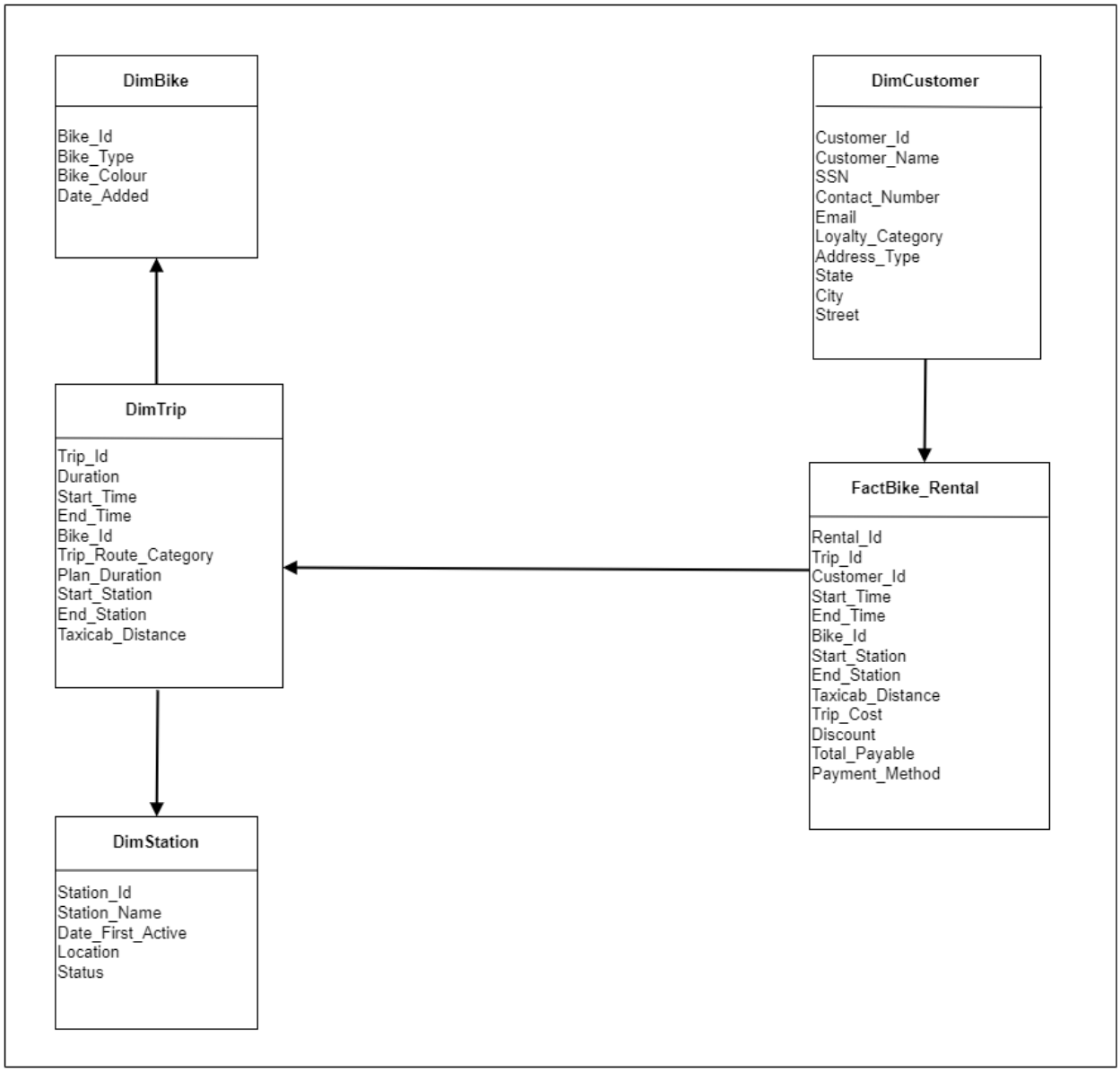

Figure 3: Diagram showing the relationships between the dimensions in the data warehouse

By studying the relationships between the dimensions in the data warehouse shown in figure 3, it is clear that the most suitable schema for the data warehouse is the Snowflake Schema.

Data Warehouse Schema – Snowflake Schema

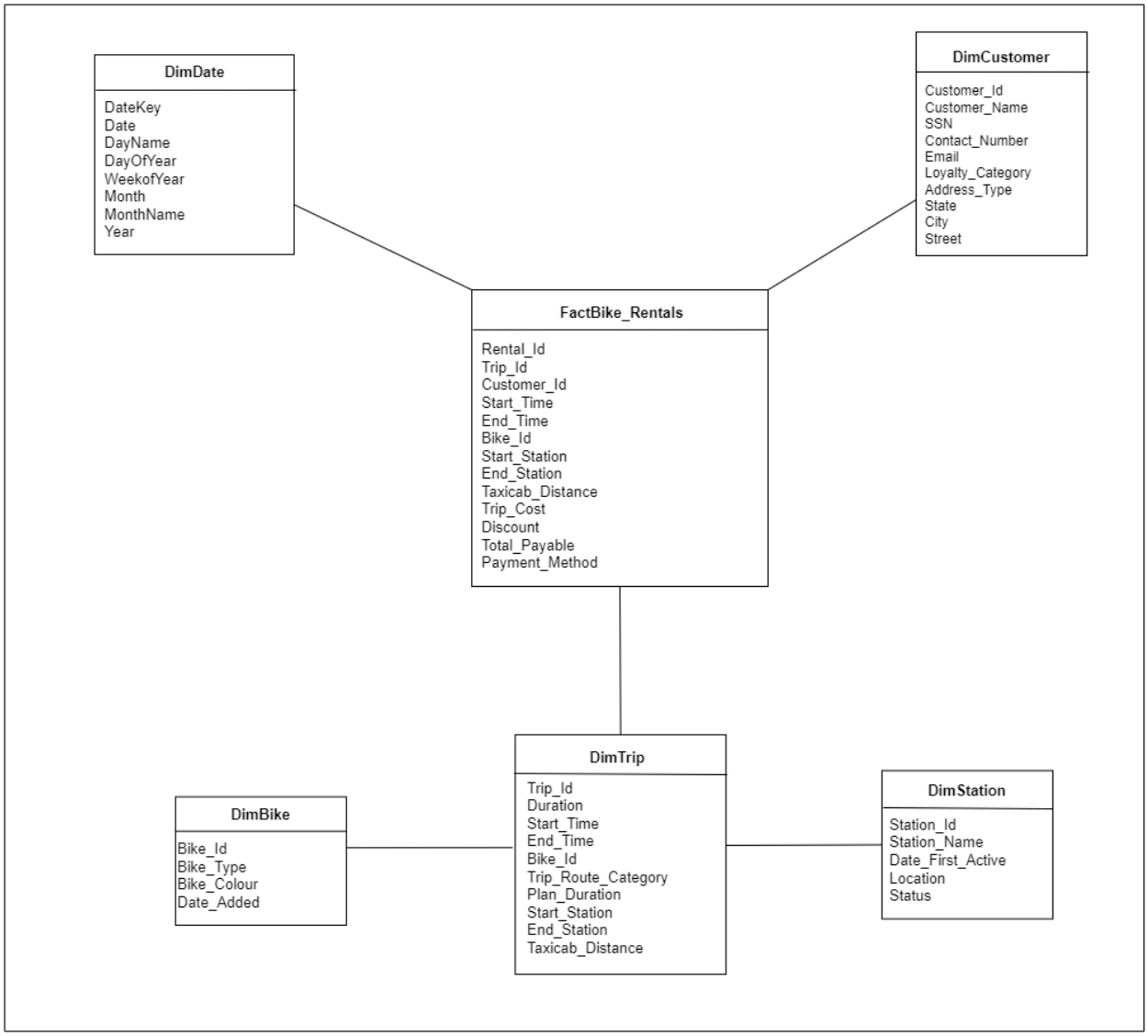

Figure 4: Data Warehouse Schema

Figure 4 above, shows the schema for the Bicycle Rental Data Warehouse.

5. ETL Development

Step 1: Extracting the data from multiple sources into the staging area

The data from the database 'SLIIT_Bike_Source_DB', text files 'Customer.txt' and 'Customer_Address.txt' and the excel file 'Bike_Rentals.xlsx' is extracted into the staging database 'SLIIT_Bike_Staging'.

An SSIS package(SLIIT_Load_Staging.dtsx) is created and data is loaded into the staging database by establishing connections between the data sources and the staging database.

Event handlers are used to ensure that the data in each table is truncated before loading the data. This is done to ensure that duplicates of the same records are not loaded repeatedly into the staging database each time the package is executed.

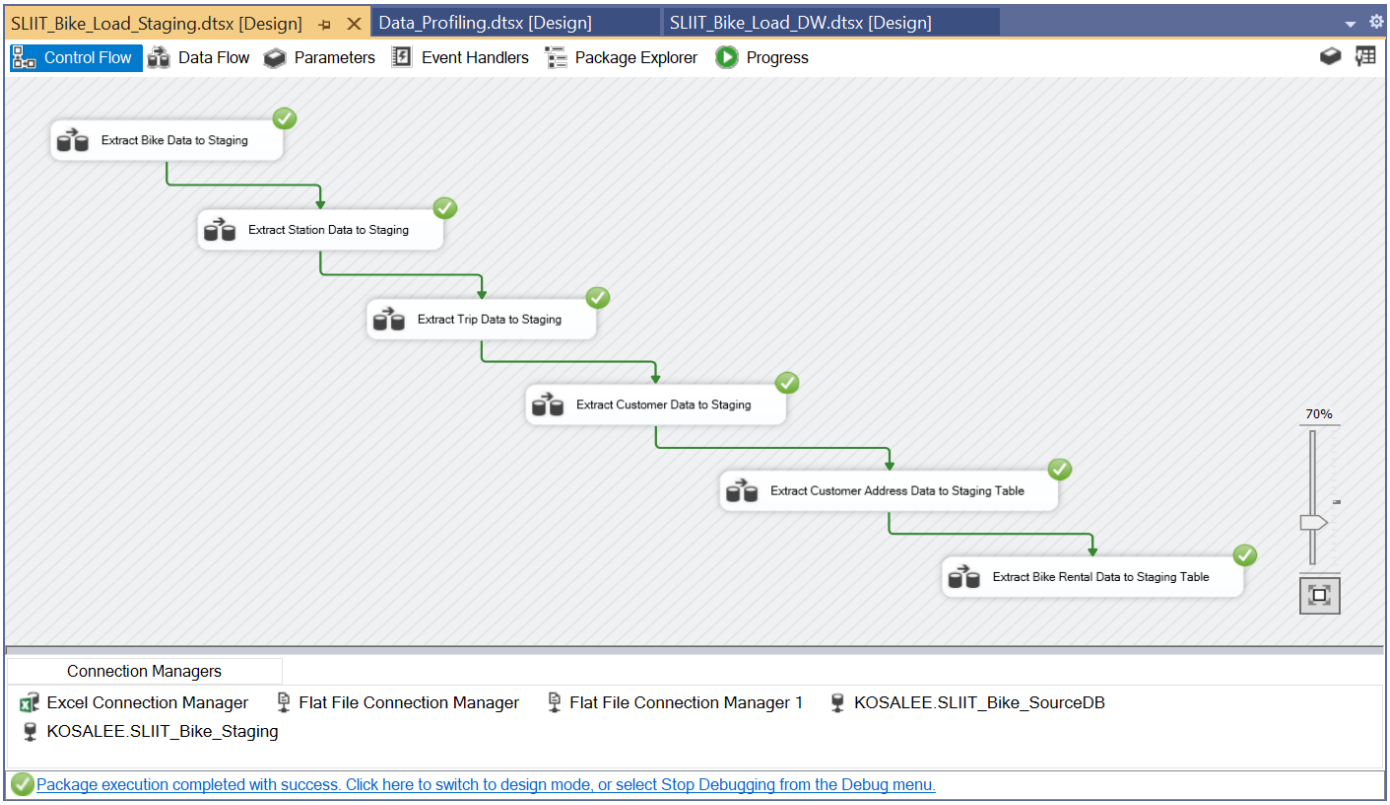

Figure 5: Extracting data into the staging database

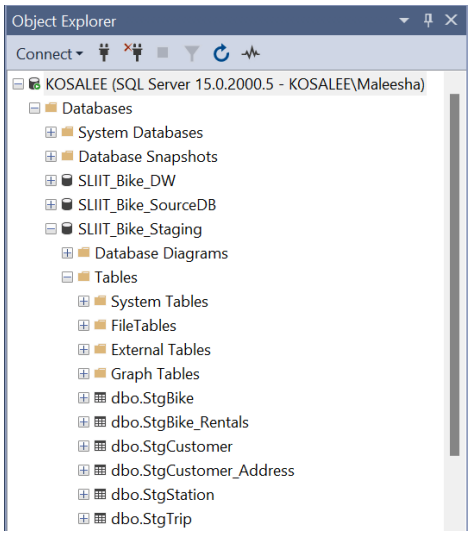

Figure 6: Staging database

Step 2: Data Profiling

A separate SSIS package(Data_Profiling.dtsx) is used for data profiling. Data profiling is used to analyze the data in the staging area and determine what type of transformations needed to be performed on the before loading into the data warehouse.

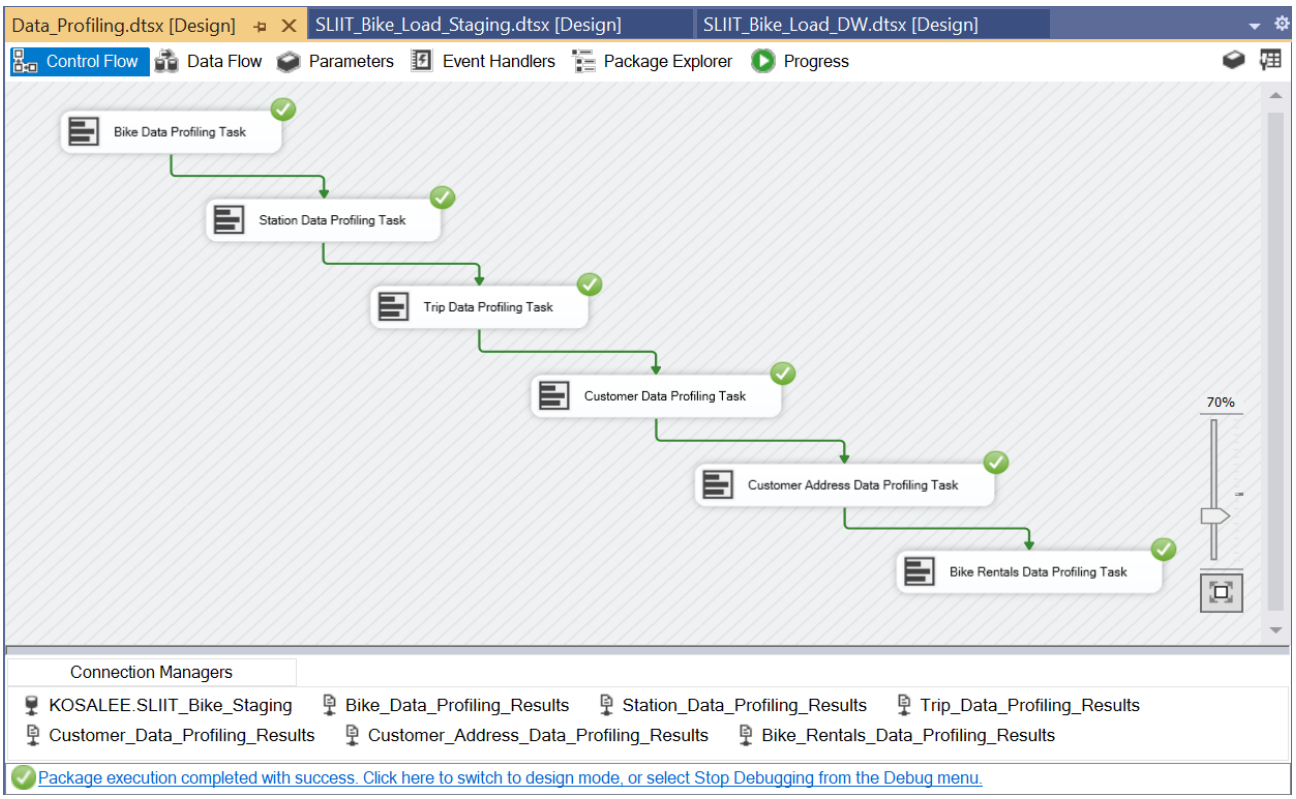

Figure 7: Execution of data profiling tasks

Step 3: Loading data into the data warehouse

3.1 Loading data into the dimension tables:

- A dimension table stores attributes that describe the objects in a fact table.

- If multiple staging tables are combined to form one dimension the data must be sorted and combined using a sorting column.

  Eg: Customer staging table and Customer_Address staging table combine to form the Customer dimension. Therefore, data in the two staging tables are sorted and combined using Sort and Merge Join components. The sorting column used to combine the data is Customer_Id field.

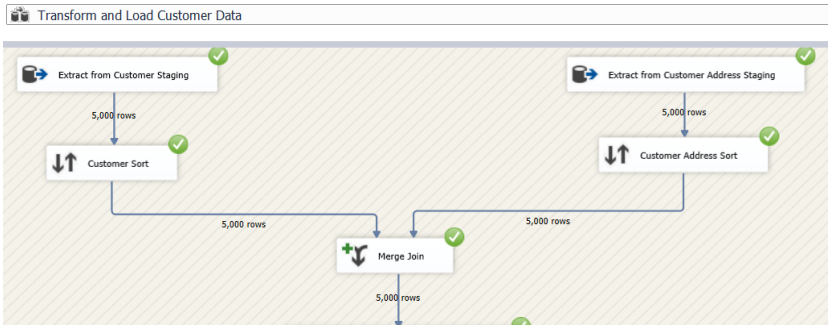

Figure 8: Combining two staging tables to form the Customer Dimension

- Null values in the staging tables can be replaced by a preferred value when loading into the data warehouse.

- Steps can be taken to handle slowly changing dimensions(SCD).

3.2 .Loading data into the fact table

- Fact tables store the measures of interest (facts) and foreign keys to dimension tables.
- When loading data into fact tables, Lookup components are used to add the Surrogate keys from dimension tables into the fact table.

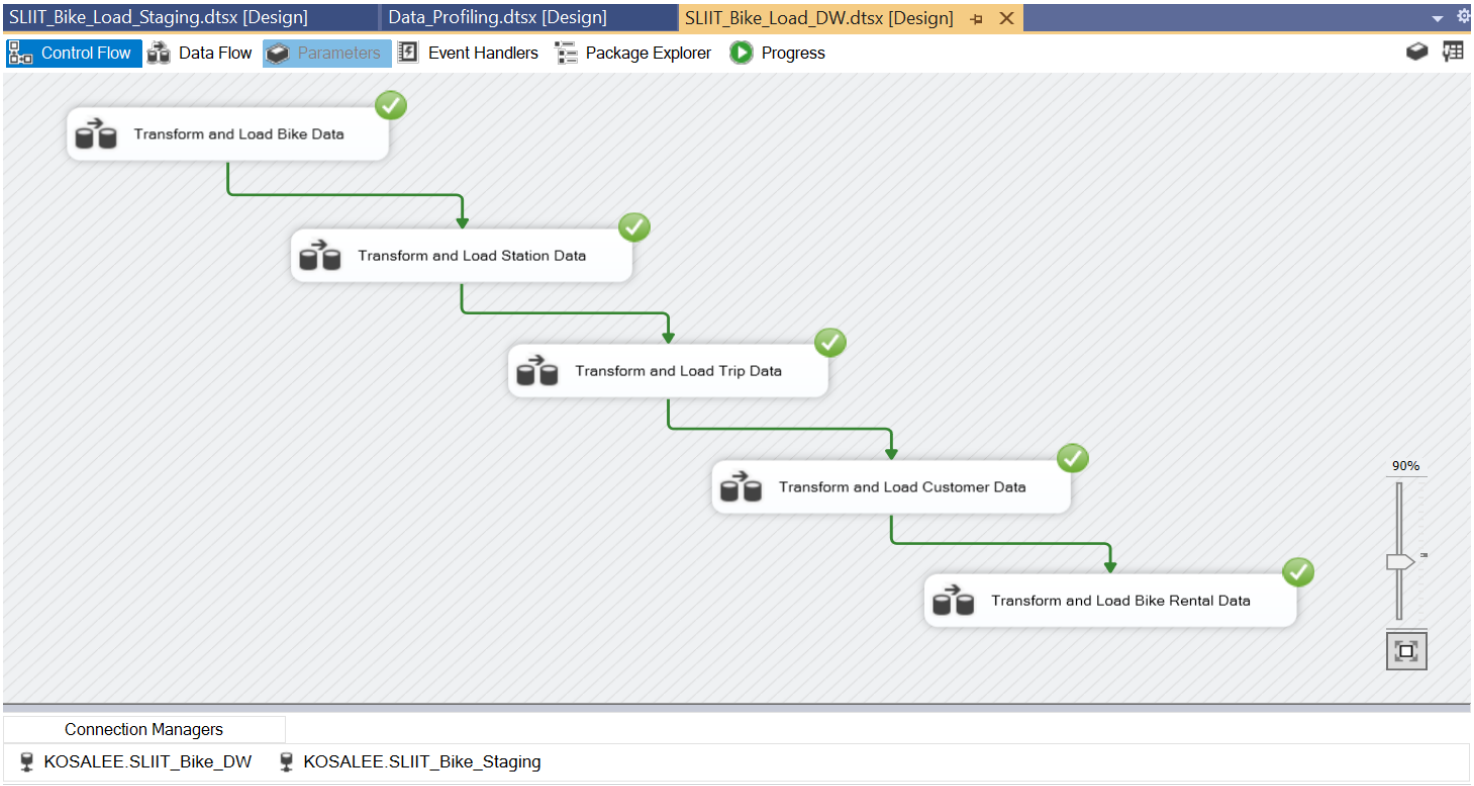

Figure 9: Loading data into the data warehouse

Handling Slowly Changing Dimensions

Customer dimension is a SCD with slowly changing attributes.

- Contact_Number, Email →Type 1 approach is used to handle. The old attribute value in the dimension row is overwritten with the new value.

- Address Type, State, City, Street → Type 2 approach is used to handle. A new record is added to the data warehouse with the updated values.

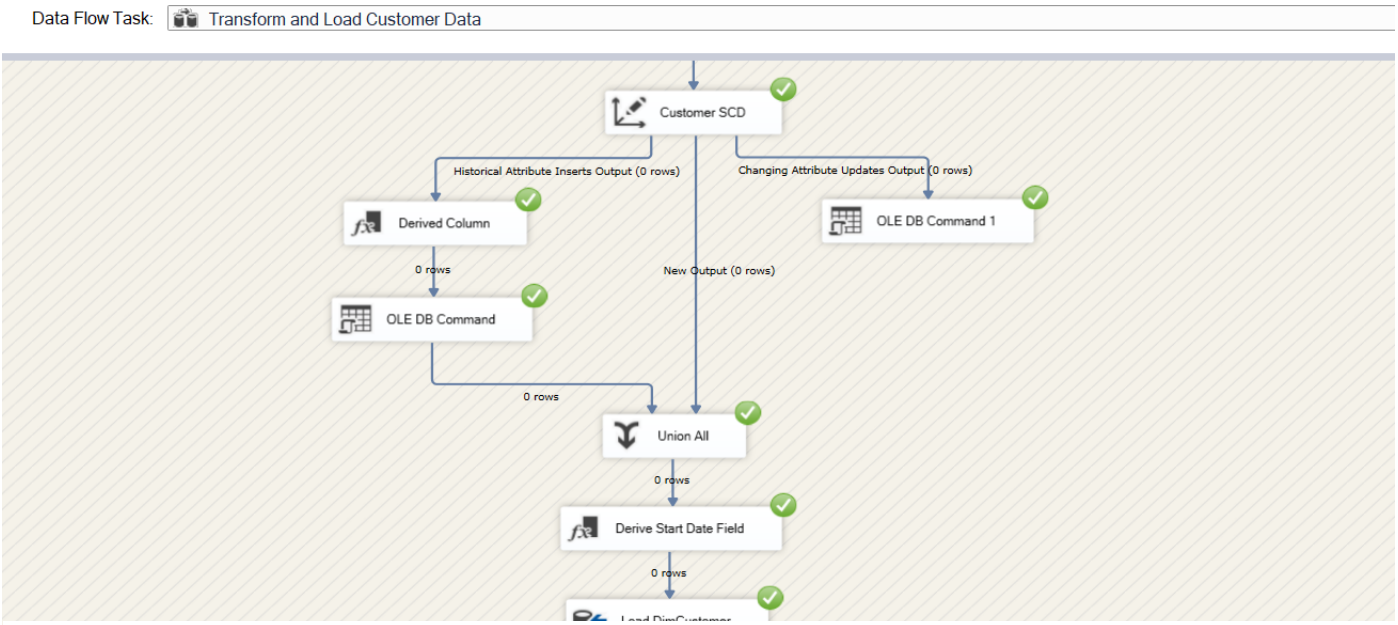

Figure 10: Handling slowly changing dimensions

7. ETL development – Accumulating Fact Tables

The fact table 'FactBike_Rentals' is extended with 3 columns;

1. accm_txn_create_time – this column is set to be equal to the current system date using a derived column and the GETDATE() function.

A separate sql table with the following structure is prepared and populated with data.

fact_table_natural_key (txn_id)	accm_txn_complete_time

2. accm_txn_complete_time – A new SSIS package 'Load_accm_txn_complete_time' is created. Data from the above table is imported into this column.

3. txn_process_time_hours – The time difference(in hours) between the above mentioned create_time and complete_time columns will be calculated using a Derived Column component and stored in this column.

Formula used for the calculation:
DATEDIFF("Hh",accm_txn_create_time,accm_txn_complete_time)

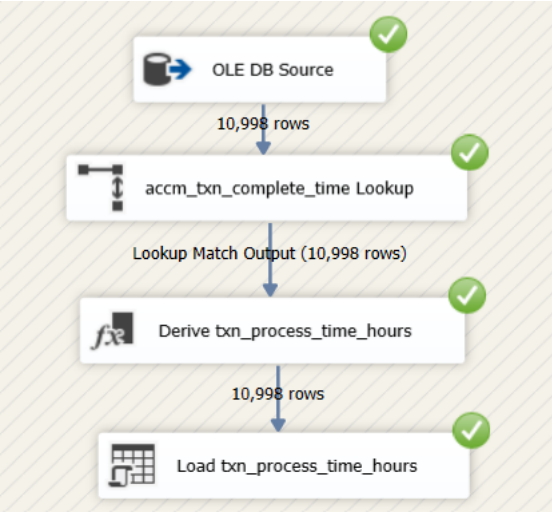

Figure 11: Implementation of 'Load_accm_txn_complete_time' SSIS package

Trip_Cost	Discount	Total_Payable	Payment_Method	Insert_Date	Modified_Date	accm_txn_create_time	accm_txn_complete_time	txn_process_time_hours
26.00	0.00	26.00	credit card	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-09 08:59:00	34
24.00	0.00	24.00	cash	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-11 12:20:00	86
18.00	0.00	18.00	cash	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-10 13:55:00	63
21.00	0.00	21.00	debit card	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-09 23:44:00	49
23.00	0.00	23.00	debit card	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-08 21:59:00	23
9.00	0.00	9.00	debit card	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-12 10:34:00	108
27.00	0.00	27.00	cash	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-11 18:22:00	92
15.00	0.00	15.00	cash	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-09 00:58:00	26
14.00	0.00	14.00	cash	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-11 11:51:00	85
13.00	0.00	13.00	cash	2022-05-07 22:47:50.173	2022-05-07 22:47:50.173	2022-05-07 22:47:00	2022-05-08 05:19:00	7

Figure 12: Fact table after completing Step 6