**SRI LANKA INSTITUE OF**

**INFORMATION TECHNOLOGY**



**ID No: IT20237486**

**Name: Gunarathne H.M.Y.B**

**Batch: DS weekday**

**Assignment: 01**

Table of Contents

[01).Data set selection 3](#_Toc26268)

[ER Diagram 3](#_Toc26269)

[02).Preparation of Data Sources 4](#_Toc26270)

[03).Solution architecture 5](#_Toc26271)

[04). Data warehouse design & development 6](#_Toc26272)

[05).ETL Development 7](#_Toc26274)

[Data Extraction 7](#_Toc26275)

[5.1 Product Subcategory Data from Source to Staging 7](#_Toc26276)

Contents

[01). Data set selection 3](#_Toc103708636)

[ER Diagram 3](#_Toc103708637)

[02).Preparation of Data Sources 4](#_Toc103708638)

[03). Solution architecture 5](#_Toc103708639)

[04). Data warehouse design & development 6](#_Toc103708640)

[05). ETL Development 7](#_Toc103708641)

[Data Extraction 7](#_Toc103708642)

[5.1 Product Subcategory Data from Source to Staging 7](#_Toc103708643)

[5.2 Invoice Data from Source to Staging 8](#_Toc103708644)

[5.3 Customer Data from Source to Staging 9](#_Toc103708645)

[Overall control flow 10](#_Toc103708646)

[Data Profiling 11](#_Toc103708647)

[Data Transforming and loading 12](#_Toc103708648)

[b) Load Slowly changing Dimensions 12](#_Toc103708649)

[5.11 Customer Data from Staging to Data Warehouse 12](#_Toc103708650)

[5.12 Product Subcategory Data from Staging to Datawarehouse 13](#_Toc103708651)

[5.15 Creation of Date Dimension 14](#_Toc103708652)

[c)Load Fact Table 16](#_Toc103708653)

[Overall ETL Transformation 17](#_Toc103708654)

# 01). Data set selection

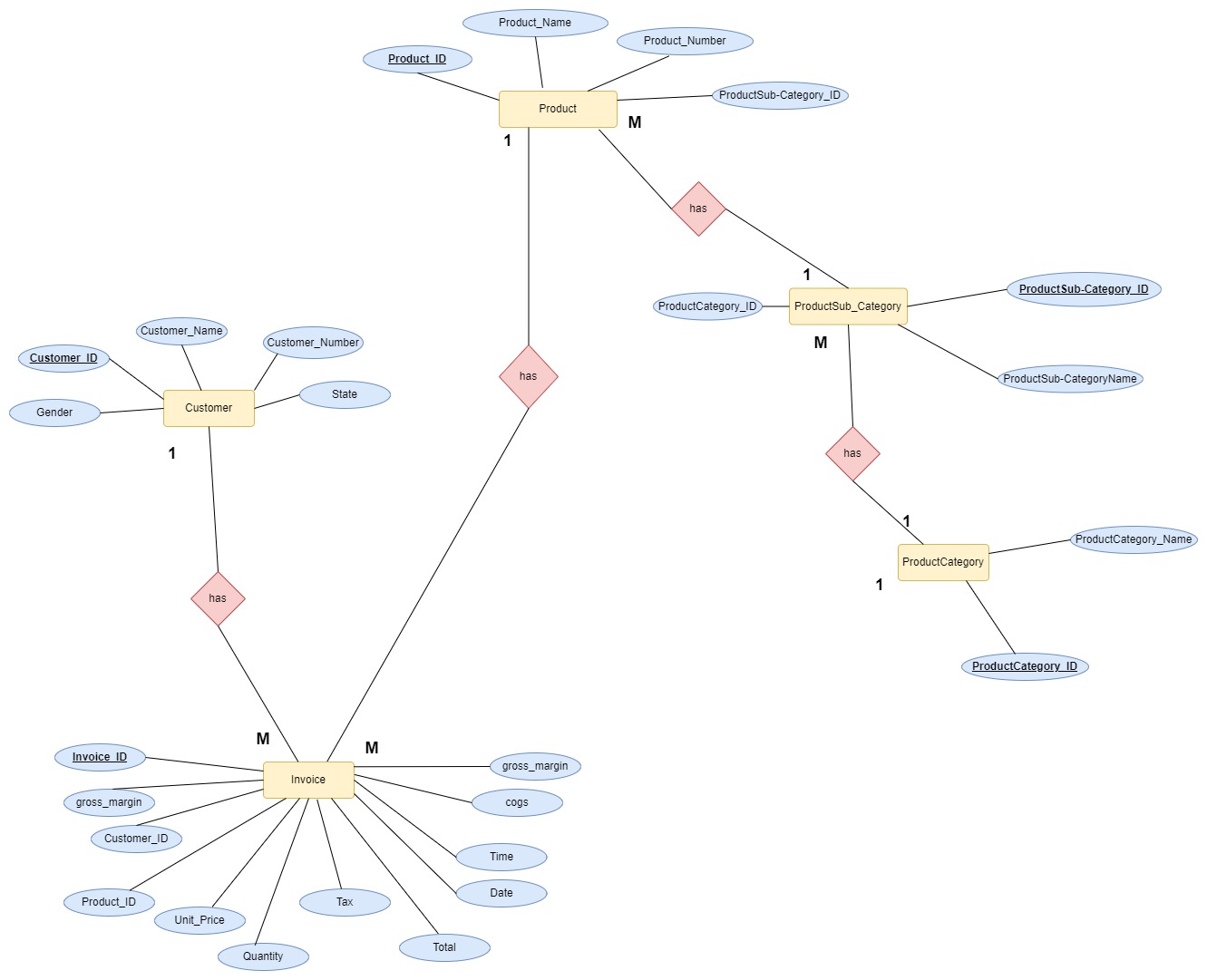
Data set : Supermarket sales

Source : Kaggle

Link to the source: <https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales>

The dataset contains the supermarket sales information. The data set consists of five files: three csv file, one excel file and one text files (Necessary modifications has been done in order to meet the requirements).

### ER Diagram



*Figure 1.0-ER diagram*

# 02).Preparation of Data Sources

The dataset was originally in the form of one excel file. Data in the file has been separated into 5 different files of type Excel,CSV and text.

|  |  |
| --- | --- |
| **Table** | **File type** |
| Product | Csv file(.csv) |
| Product Category | Csv file(.csv) |
| Product Sub Category | Csv file(.csv) |
| Invoice | Excel file(.xlsx) |
| Customer | Text file(.txt) |

*Figure 1.1-sample source table creation*

Similarly other tables has also been created and then the tables has been exported in relevant file types.

Final set of Sources:

Graphical user interface, application

Description automatically generated with medium confidence

# 03). Solution architecture

Diagram

Description automatically generated

*Figure 2.0-Solution architecture*

As can be seen in the figure 4 different resource types has been used to extract data to staging. Staging layer has been used to have all the tables in a single location as in the below figure.

A picture containing graphical user interface

Description automatically generated

*Figure 3.0-Staging*

The tables at the staging are then profiled and after performing a rich set of ETL tasks, data is loaded to the data warehouse where from that several reporting tools and analyzing tools can use data for reporting mining and analyzing.

# 

# 04). Data warehouse design & development

The data warehouse is designed as a snowflake schema with one fact table and five-dimension table.

Diagram, schematic

Description automatically generated

*Figure 4.0-Datawarehouse design*

# 05). ETL Development

As the first step data has been extracted from sources to staging area. Data flow task has been used for every extraction.

## Data Extraction

### 5.1 Product Subcategory Data from Source to Staging

5.1.1 Data Flow Diagram

Description automatically generated

*Figure 5.1.1-Product Subcategory data flow*

5.1.2 Event handler

Before executing ‘Extract Product Subcategory to Staging’ existing data in the staging layer has been truncated. A screenshot of a computer

Description automatically generated with medium confidence

*Figure 5.1.2.-Product Subcategory Event handler*

### 5.2 Invoice Data from Source to Staging

5.2.1 Data Flow

|  |  |
| --- | --- |
|  |  |

*Figure 5.2.1-Invoice data flow*

5.2.2 Event handler

Before executing ‘Extract Invoice to Staging’ existing data in the staging layer has been truncated.

A picture containing graphical user interface

Description automatically generated

*Figure 5.2.2.-Invoice Event handler*

### 5.3 Customer Data from Source to Staging

5.3.1 Data Flow

|  |  |
| --- | --- |
|  |  |

*Figure 5.3.-Customer data flow*

5.3.2 Event handler

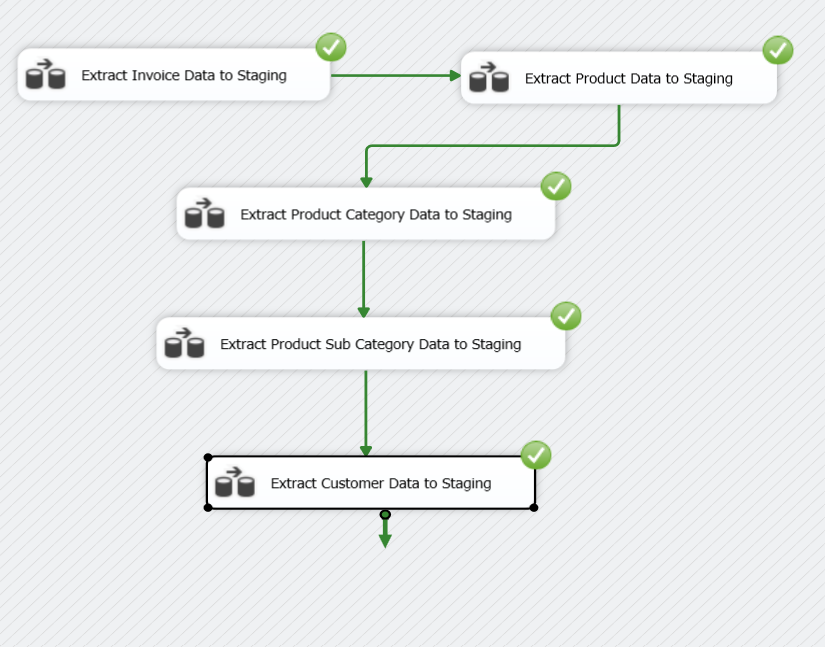
Before executing ‘Extract Customer to Staging’ existing data in the staging layer has been truncated.

Graphical user interface, text

Description automatically generated

*Figure 5.3.1.-Customer Event handler*

### Overall control flow

**

*Figure 5.8.-source to staging control flow*

### Data Profiling

Before Loading staging tables to the data warehouse data must be enriched to obtain the most suitable data for analyzing. Data profiling has been done in order to identify what need to be corrected in ETL process in order to meet this requirement.

Each and every table at staging is profiled and stored in a specific file location.

Graphical user interface, text, chat or text message

Description automatically generated

*Figure 5.9.1-profiling diagram*

## Data Transforming and loading

### b) Load Slowly changing Dimensions

### 5.11 Customer Data from Staging to Data Warehouse

Diagram

Description automatically generated

### 5.12 Product Subcategory Data from Staging to Datawarehouse

Product Subcategory data has been loaded to DimProductSubcategory

Timeline

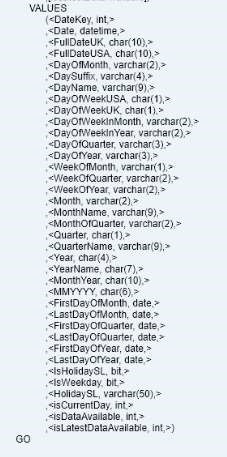
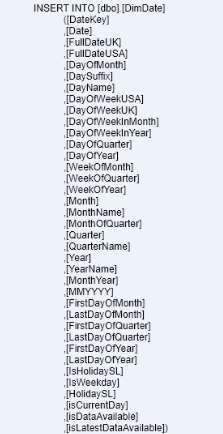
Description automatically generated

*Figure 5.12.1-load to DimProductSubcategory*

### 5.15 Creation of Date Dimension



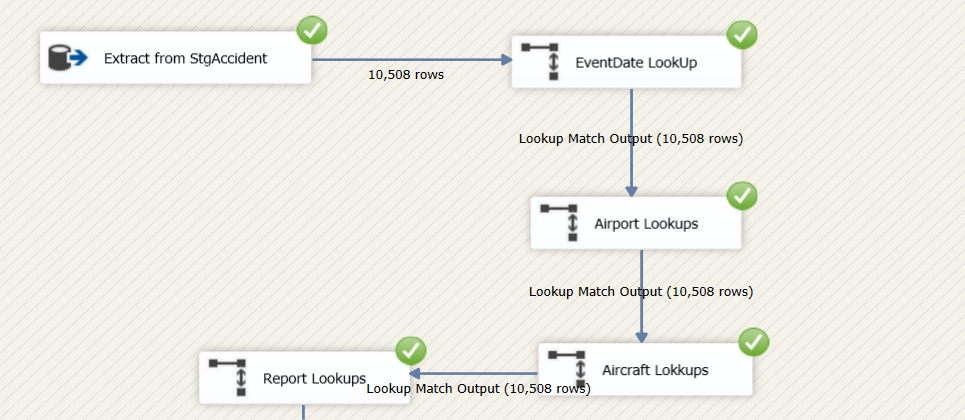
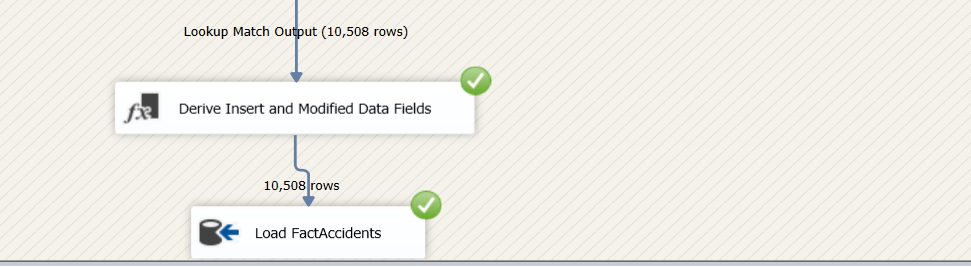
*Figure 5.14.1-Load to DimDate*



Query used to create and load data to date dimension is listed above. Date dimension is assumed to tally with reservation status date in booking table

### c)Load Fact Table

TransactionStaging table and some of the coulmns in BookingStaging table is merged in order to make the fact table. RoomStaging is loaded and merged to obtain RoomId. All required surrogate keys has been loaded to data warehouse after a lookup through alternate keys in dimension tables.

*Figure 5.16.1-FactAccident ETL*

### Overall ETL Transformation

Diagram

Description automatically generated

*Figure 5.17.1-overall ETL to data warehouse*

*Thank you!*