Logo, company name

Description automatically generated

**Data Warehousing & Business Intelligence**

Assignment I

2022

Submitted By: Subasinghe S.S.

IT20273712

**Table of Contents**

[1. Data set selection and Introduction 3](#_Toc103432230)

[2. Preparation of Data Sources 6](#_Toc103432231)

[3. Solution Architecture 8](#_Toc103432232)

[4. Data warehouse design and development 9](#_Toc103432233)

[5. ETL Development 10](#_Toc103432234)

[7. ETL development – Accumulating fact tables 20](#_Toc103432235)

# Data set selection and Introduction

‘LA RESTAURANTS AND MARKET HEALTH DATA’ is a collection of transactional data which is used as the source data set here. The following is the link to the original data set:

<https://www.kaggle.com/datasets/cityofLA/la-restaurant-market-health-data?select=restaurant-and-market-health-violations.csv>

Modifications were made accordingly to the data set derived from the source. This data set reflects the restaurants' and market's health data in Los Angeles, California. The derived data set ‘HV’ (Health Violations) mainly focuses on the violations that happened in restaurants within the range of 2018 to 2020.

In this Data set Owner has many facilities and owner is considered as a place where lot of facilities (Target, Subway etc.) are located. These facilities are rated by taking records on Violations and given a score and fine for the relevant facility.

## Description of the Data Set

|  |  |  |  |
| --- | --- | --- | --- |
| **Table name** | **Column name** | **Data type** | **Description** |
| Violations | ViolationID | nvarchar(50) | Contains the Details of Violations associated with the rated facilities. |
| ViolationDescription | nvarchar(100) |
| ViolationStatus | nvarchar(50) |
| Facility | FacilityID | nvarchar(50) | Contains the Details of the facilities Owned by the Owners  This contains the hierarchy where one City (FacilityCity) has many facility addresses (FacilityAddress). |
| FacilityName | |  |  | | --- | --- | | nvarchar(200) | FacilityName | |
| FacilityCity | |  | | --- | | nvarchar(50) | |
| FacilityAddress | nvarchar(200) |
| OwnerID | nvarchar(50) |
| Owner | OwnerID | nvarchar(50) | Contains the details of the Owners |
| OwnerName | nvarchar(50) |
| OwnerPhoneNumber | numeric(18,0) |
| OwnerEmail | nvarchar(50) |
| Rating | RatingID | nvarchar(50) | Details of the ratings associated with the facilities and violations.  This contains the score for a given rating and a fine for that rating as measurable values |
| FacilityID | nvarchar(50) |
| ViolationID | nvarchar(50) |
| ActivityDate | datetime |
| Score | int |
| Fine | int |

## ER Diagram

Diagram

Description automatically generated

This diagram shows the connection between the entities in the data set.

# Preparation of Data Sources

A database named Health Violations (HV) was created including the database, csv and txt source files.

* Dbo.Violations
* Owner.txt
* Facility.csv
* Rating.txt

A new database named HV\_DW was built for the data warehouse which contains the dimensions and the fact table.

* DimViolation
* DimDate
* DimOwner
* DimFacility
* FactRating

The HV\_Staging database was created to extract and load the data to the database.

A script file was used to create the DimDate relation in the Data Warehouse.

## Class Diagram using Data sources

# Solution Architecture

Diagram

Description automatically generated

**Data Sources**: locations where the Data is needed for DB coming from, in this scenario primary data source is database and others are csv file, and txt files

**ETL**: Extract-Transform-Load is an ETL standard. It is the process of transferring data from one or more sources into a destination system that has a different representation of the data than the source (s).

**Staging**: As explained next step is staging the source data set. After the staging layer the below mentioned staging tables are created:

1. stgViolations
2. stgOwner
3. stgFacility
4. stgRating

**Data warehouse**: Following staging, the staging database's contents will be used as sources for the transformation process. Data is transformed and loaded into tables in the Datawarehouse database.

# Data warehouse design and development

## A picture containing text, indoor, screenshot Description automatically generatedData Warehouse Schema

Snowflake schema is used to design the Datawarehouse design. There is one fact table as transactions and four dimensions including the Date dimension.

Assumptions:

DimOwner is considered as a slowly changing dimension. OwnerName (Here the Owner is a restaurant or a marketplace which contains lot of facilities like Target, Subway etc.) as a historical attribute and OwnerPhoneNumber and the OwnerEmail are taken as changing attributes.

# ETL Development

## **Extraction from Source database to staging area**

Data was extracted from the sources in the first step (DB source, CSV file & text files). Data was extracted from the source to the staging table using a data flow job for each extraction. A truncate table was then generated for every staging table. At the conclusion, all the data flow jobs were merged as follows:

Graphical user interface, text, application, chat or text message

Description automatically generated

Below are screenshots of all the data sources that were staged and the truncate tables that were created:

Graphical user interface, website

Description automatically generatedStaging Violations table

Staging Owner table

Graphical user interface

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceStaging Facility table

Staging Rating Table

Graphical user interface, application

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceAfter following the above steps and executing

Next step is data profiling, and it is done as shown below:

## **Graphical user interface Description automatically generatedData Profiling**

Every staging table is profiled and saved in a specific location.

## **Data warehouse Design and Development**

Load Data to the Violation Dimension

A screenshot of a computer

Description automatically generated with medium confidenceFirst Data was loaded from the Violation staging table (stgViolations)to the Violation Dimension (DimViolation).

Stored procedure used for the DimViolation can be found below:

Text

Description automatically generated

Load Data to the Owner Dimension

As mentioned earlier under assumptions, Owner was considered as a slowly changing dimension.

The below mentioned columns were set as changing attributes:

1. OwnerPhoneNumber (Phone number of the Owner)
2. Owner Email (Email of the Owner)

Owner Name was taken as a historical attribute. (Ex: Target, Subway)

After extracting data from the Owner staging table, then after replacing the null ownerEmail values with a ‘N’ and as it was identified as a slowly changing dimension, it was connected as shown below and loaded data to the Owner dimension table.

Graphical user interface, application

Description automatically generated

1. Extracted from Customer Staging table
2. Sorted by Customer ID
3. Making the dimension a Slowly Changing Dimension
4. Loaded to Customer Dimension

Load Data to the Facility Dimension

Graphical user interface, application

Description automatically generatedAfter merging the Facility staging table with the Owner dimension table, data was loaded from the Facility staging table to the Facility dimension. Before loading, both the Facility staging table and the Owner dimension were sorted by OwnerID and then merge joined to extract Facility details from the Facility staging table and Owner surrogate key (OwnerSK) from the Owner dimension.

Graphical user interface, text, application, email

Description automatically generatedStored procedure used for the DimFacility can be found below:

After loading all the dimensions, lastly data is loaded to the fact table. The below steps were followed:

After extracting the Data from the Rating staging, Next to join relevant dimension tables with the FactRating table, Surrogate keys which are required for ‘ViolationID’, ‘FacilityID’, ‘ActivityDate’ are taken using Lookup component and then insert Data to the FactRating table

Graphical user interface, application

Description automatically generated

After loading data to all the dimensions and the fact table:

Graphical user interface, text, application, chat or text message

Description automatically generated

# ETL development – Accumulating fact tables

Graphical user interface, application

Description automatically generated

A print screen of the fact table can be found below:

Graphical user interface, text, application

Description automatically generated with medium confidence