# Sri Lanka Institute of Information Technology (SLIIT)

A picture containing text, clipart

Description automatically generated

IT19957180

P.M.D.C.B Wijerathna

Y3.S2.04.WE

IT3021

DWBI Assignment1- Report

Contents

[Sri Lanka Institute of Information Technology (SLIIT) 1](#_Toc104537837)

[1. Data set selection 5](#_Toc104537838)

[2. Preparation of Data Source 7](#_Toc104537839)

[3. Solution Architecture 9](#_Toc104537840)

[4. Datawarehouse Design & Development 10](#_Toc104537841)

[5. ETL Development 13](#_Toc104537842)

[6. Accumulating fact tables 40](#_Toc104537843)

## Table of Figures

[Figure 1 : High Level ER Diagram of Dataset 6](#_Toc104537889)

[Figure 2 : High-Level Architecture 9](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537890)

[Figure 3 :Dimensional Model (Snowflake Schema) 12](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537891)

[Figure 4 :ETL Process High-Level Overview 13](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537892)

[Figure 5 :Staging Database Structure 14](#_Toc104537893)

[Figure 6 : DWH Database Structure 14](#_Toc104537894)

[Figure 7 :Overall Staging 17](#_Toc104537895)

[Figure 8 :Food Order database Structure 17](#_Toc104537896)

[Figure 9 :Data flow task StgOrders 18](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537897)

[Figure 10 : Event Handler 18](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537898)

[Figure 11 :Event Handler Properties 18](#_Toc104537899)

[Figure 12 :Staging Orders connection 19](#_Toc104537900)

[Figure 13 : Data flow Task for StgPayment 19](#_Toc104537901)

[Figure 14 :Staging Payment connection 20](#_Toc104537902)

[Figure 15 :Column Mapping StgPayment 21](#_Toc104537903)

[Figure 16 :Data flow task for Stg Payment 21](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537904)

[Figure 17 :Data flow task for Stg Store 22](#_Toc104537905)

[Figure 18 :Stg Store connection 22](#_Toc104537906)

[Figure 19 :Stg Store event handler 23](#_Toc104537907)

[Figure 20 :Stg Hubs Data flow task 23](#_Toc104537908)

[Figure 21 :Stg hub connection 24](#_Toc104537909)

[Figure 22 : Stg Channel data flow task 24](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537910)

[Figure 23 :Stg channel connection 25](#_Toc104537911)

[Figure 24 : Stg Channel event handler 26](#_Toc104537912)

[Figure 25 : Stg order dates data flow task 26](#_Toc104537913)

[Figure 26 : Stg Order dates connection 27](#_Toc104537914)

[Figure 27 : Stg order Dates event handler 28](#_Toc104537915)

[Figure 28 :DimStore Procedure 30](#_Toc104537916)

[Figure 29 :DimPayment Procedure 31](#_Toc104537917)

[Figure 30 :DimOrderDetails Procedure 31](#_Toc104537918)

[Figure 31 :DimHubs Procedure 32](#_Toc104537919)

[Figure 32 :DimDeliveries procedure 32](#_Toc104537920)

[Figure 33 :DimChannel Data Conversion 34](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537921)

[Figure 34 : DimChannel Data flow task 34](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537922)

[Figure 35 :DimDrivers SVD Data FLOW 35](#_Toc104537923)

[Figure 36:DimHub Data Flow Task 36](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537924)

[Figure 37 :Dim Stores Data flow task 36](#_Toc104537925)

[Figure 38 :Dim Stores Filling Nulls 37](#_Toc104537926)

[Figure 39 :Dim Order details data flow task 37](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537927)

[Figure 40 :Dim Deliveries Data flow task 38](#_Toc104537928)

[Figure 41 :Dim payments data flow task 38](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537929)

[Figure 42 :Fact Orders data flow task 39](https://d.docs.live.net/15fa1527d99d791c/Documents/IT19957180-Assignment_1.docx#_Toc104537930)

## Data set selection

Data Set: Delivery Center: Food & Goods orders in Brazil

Site: Kaggle

Source Link: <https://www.kaggle.com/datasets/nosbielcs/brazilian-delivery-center>

The dataset for assignment 01 was created using a real-world business entity called "Delivery Center." Because it is a real business, sensitive information such as the personal information of their drivers is not disclosed in detail, and just a few factors have been revealed to the public. This data set is about Delivery Center details. It consists of Food & Goods orders in Brazil. The data collection is made up of orders and deliveries processed by the Delivery Center between 2019 and 2022.

There are four csv, two excel, one txt and one data base backup for Order details files in the data set.

Graphical user interface, application

Description automatically generated

Channels: This dataset includes details on the sales channels (marketplaces) where our retailers' goods and food are sold.

Deliveries: This dataset contains data on deliveries performed by our delivery partners.

Drivers: This dataset has information about the partner deliverers. They stay in our hubs and every time an order is processed, they deliver it to consumers' homes.

Hubs: This dataset has information about Delivery Center hubs. Understand that Hubs are the distribution centers for orders and that's where deliveries come from.

Orders: This dataset has information about sales processed through the Delivery Center platform.

Payments: This dataset has information about payments made to the Delivery Center.

Orders: This dataset has information about the tenants. They use the Delivery Center Platform to sell their items (good and/or food) on marketplaces.

Orders Timeline: This dataset contains all the Order dates and delivery dates according to the order ID.

Diagram, schematic

Description automatically generated

Figure 1 : High Level ER Diagram of Dataset

## Preparation of Data Source

* 1. Data Quality and Preparation

**There are almost 400 000 thousand records in the primary dataset, all with 59 variables. However, when evaluating the dataset's overall data quality, most of the variables had null values. As a result, null values were eliminated from the original dataset and separated into several sources. In addition, further data was contributed to the dataset with the purpose of improving their overall quality.**

* 1. Data Source Details

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table Name | Column Name | Data Type | Link Table | Link Column |
| Channel | Channel\_id  Channel\_name  Channel\_type | Int  Varchar (50)  Varchar(50) |  |  |
| DeliveryDriver | Driver\_id  Driver\_order\_id  Delivery\_distance\_meters  delivery\_status  driver\_modal  driver\_type  Email Address  FirstName LastName  Phone Number  Phone Number | Int  Int  Varchar(50)  Varchar(50)  Varchar(50)  Varchar(50)  Varchar(50)  Varchar(50)  Varchar(50)  Varchar(50) | order | Delivery\_order\_id |
| Hubs | hub\_id  hub\_name  hub\_city  hub\_state | Int  Varchar(50)  Varchar(50)  Varchar(50) | store | Hub\_id |
| OrderDates | Order\_id  Order\_date  Deliver\_date  Order\_time\_hour  Order\_time\_minute  Deliver\_time\_hour  Deliver\_time\_minute | Int  Datetime  Datetime  Int  Int  Int  int | order | Order\_id |
| Orders | Order\_deliver\_id  Order\_id  Store\_id  Channel\_id  Payment\_id  Order\_status  Order\_amount  Order\_delivery\_fee  Order\_delivery\_cost | Int  Int  Int  Int  Int  Varchar(50)  Money  Money  Money | DeliveryDriver  Store  Channel  Payment | Delivery\_order\_id  Stoe\_id  Channel\_id  Payment\_order\_id |
| Payments | Payment\_id  Payment\_method  Payment\_order\_id  Payment\_fee  Payment\_amount  Payment\_status | Int  Varchar(50)  Int  Money  Money  Varchar(50) | order | Order\_id |
| Stores | Store\_id  Hub\_id  Store\_name  Store\_segment  Store\_plan\_price | Int  Int  Varchar(50)  Varchar(50)  money | hub | Hub\_id |

## Solution Architecture

**The diagram below shows the overall architecture of the Delivery Center: Food & Goods orders in Brazil Datawarehouse and Business Intelligence Solution that is being implemented. Data Sources, ETL (Extract, Transform, Load), Storage Layer Components, and Data Consumption are the four components that make the architecture.**

* Data Sources, it comprises of structured data in the format of text, SQL and excel. The database files are stored in SQL Server DB, and other formats are stored in the local folder.
* ETL, it is performed at two instances, first instance when extracting data from the sources and loading it to the Staging Layer and in second instance when performing extraction, and transformation on Staging Layer to load data into Datawarehouse Layer.
* Storage Layer, there are two main layers of storage, intermediate and core. Staging tables are implemented in intermediate layer, while the Datawarehouse and OLAP Server are implemented in core layer.
* Data Consumption, primary focus of this component is the analysis of the data that is being stored in Datawarehouse and OLAP Server. Analysis can be performed inform of reporting, dashboard, and self-service BI.

****Diagram

Description automatically generated****

Figure : High-Level Architecture

## Datawarehouse Design & Development

* 1. Dimensional Model

Snowflake schema was selected to design the Datawarehouse schema for our dataset, considering the number of dimensional tables, fact tables and the relationship among these table, in addition to the aspects such as performance, speed, and availability.

There are mainly eight dimensional tables, DimDate, DimChannel, DimDeliveries, DimDrivers, DimHubs, DimOrderDetails, DimPayments, DimStore and one fact table FactOrder. All the dimensional tables are linked with fact table, except DimDrivers and DimHubs, as it is linked with DimDeliveries with “driver\_id” and DimHubs with “hub\_id” foreign key constraint.

Dimension:

|  |  |
| --- | --- |
| Dimension Name | Type |
| DimDate | Static/Date/Role Playing |
| DimChannel | Role Playing |
| DimDeliveries | Parent-Child Dimension |
| DimDrivers | Parent-Child Dimension /Slowly Changing Dimension |
| DimHubs | Parent-Child Dimension |
| DimOrderDetails | Role Playing |
| DimPayments | Role Playing |
| DimStore | Parent-Child Dimension /Role Playing |

Fact Table:

|  |  |
| --- | --- |
| Fact Table Name | Type |
| FactOrders | Accumulating Fact Table |

Hierarchies:

* DimHub hierarchical breakdown from state > city
* DimDelivery is mapped as hierarchical dimension of DimDriver
* DimStore is mapped as hierarchical dimension of DimHub
* DimDate hierarchical breakdown from Year > Quarter > Month > Date

Additional Fields:

Additional fields were added to all tables in the Datawarehouse layer, as shown below. These fields are derived attributes, with the date field getting its values from the system date and time.

The DimDriver has two additional date field “start\_date and end\_date” as it is classified as slowly changing dimension. Further, surrogate keys have been added to all the dimensional tables, which is auto incrementing, and it’s used to map with tables.

Assumption(s)

* No assumption made during the Datawarehouse Design & Development Stage

Diagram, schematic

Description automatically generated

Figure 3 :Dimensional Model (Snowflake Schema)

## ETL Development

* 1. ETL Process Overview

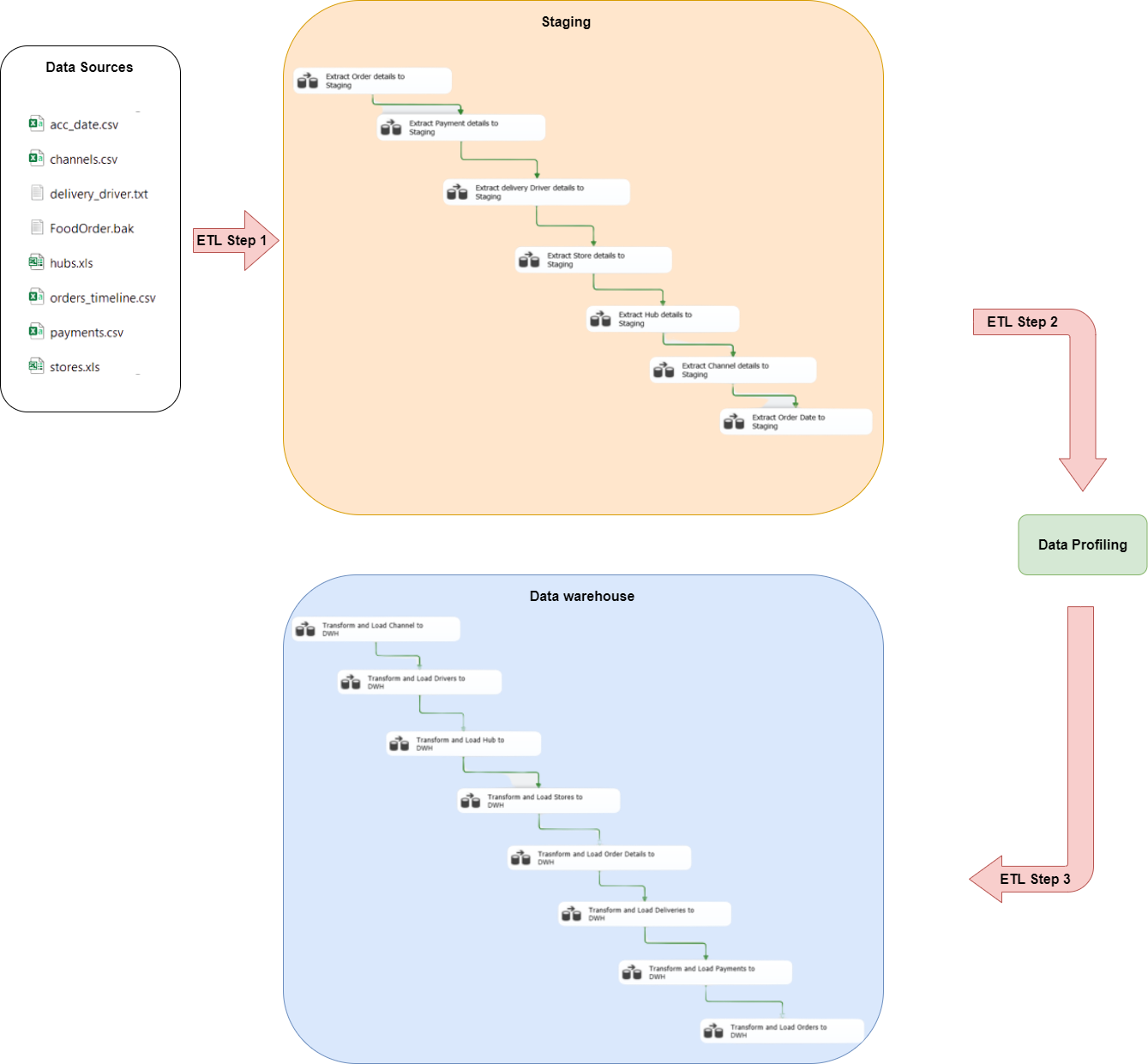
The diagram below illustrates the ETL development process in a high level along with the order of execution within each layer staring from Staging.

Figure 4 :ETL Process High-Level Overview

* 1. ETL Development Process
     1. Environment and Data Sources
* The Data warehouse Solution is design and implement using MSSQL Server. In here two additional databases ware developed for staging as “Assignment1\_Staging” and data warehouse process as “Assignment1\_DWH”. The Datawarehouse layer will extract, process, and load data from the staging layer, whereas the Staging layer will extract, transform, and load data from the source files.
* The Staging database tables are not pre-defined, because in staging layer it will be auto generated through the ETL process executed by SSIS.
* The Data warehouse tables are predesigned, total nine tables ware created .one for Fact table and rest of the tables are dimensions.

|  |  |
| --- | --- |
| Table  Description automatically generated with low confidence  Figure 5 :Staging Database Structure | Table  Description automatically generated with medium confidence  Figure 6 : DWH Database Structure |

* “DimDate” is static and loaded manually, the values for the table were generated by using “Date Master” query in SQL.

Graphical user interface, application, table

Description automatically generatedText

Description automatically generated

* + 1. Staging

A new SSIS package was created for this process called “StgAssignment” in a Data Tool application.

Timeline

Description automatically generated

Figure 7 :Overall Staging

* + - * 1. Extract & Load Order Details to the Staging

In the file sources Order details are stored in a FoodOrder.bak. after restoring this database backup file, its structure is showed in below figure.

Graphical user interface, application

Description automatically generated with medium confidence

Figure 8 :Food Order database Structure

Diagram

Description automatically generatedIn the same package, a control task was created. withing the control task the data flow was designed along with an event handler as onPreExecute.

Figure 9 :Data flow task StgOrders

Graphical user interface, text, application

Description automatically generated

Figure 10 : Event Handler

* Same as other Data flow task also have Event Handlers

Graphical user interface, application

Description automatically generated

Figure 11 :Event Handler Properties

Graphical user interface, application

Description automatically generated

Figure 12 :Staging Orders connection

* + - * 1. Extract & Load Payment Details to the Staging

The source file type for Payment details is CSV. According to the data type, to Extract Data from CSV, the best option is Flat file source.

A picture containing diagram

Description automatically generated

Figure 13 : Data flow Task for StgPayment

Graphical user interface, application

Description automatically generated

Figure 14 :Staging Payment connection

* + - * 1. Extract & Load Delivery Driver Details to the Staging

The File type of the Source is txt. Therefore, most suitable data flow task is flat file source. Because of the file format, need to clarify delimiter when creating the connection manager.

Table

Description automatically generated

Figure 15 :Column Mapping StgPayment

Diagram

Description automatically generated

Figure 16 :Data flow task for Stg Payment

* + - * 1. Extract & Load Store Details to the Staging

Store Details are contained in Excel file. Extract all the data using Excel Source in Data flow to OLED DB destination source.

Diagram

Description automatically generated

Figure 17 :Data flow task for Stg Store

Graphical user interface, application

Description automatically generated

Figure 18 :Stg Store connection

Graphical user interface, text, application

Description automatically generated

Figure 19 :Stg Store event handler

* + - * 1. Extract & Load Hub Details to the Staging

Hub Details are contained in Excel file. Extract all the data using Excel Source in Data flow to OLED DB destination source.

A picture containing line chart

Description automatically generated

Figure 20 :Stg Hubs Data flow task

Graphical user interface, application

Description automatically generated

Figure 21 :Stg hub connection

* + - * 1. Extract & Load Channel Details to the Staging

The source file type for Channel details is CSV. According to the data type, to Extract Data from CSV, the best option is Flat file source.

A picture containing chart

Description automatically generated

Figure 22 : Stg Channel data flow task

Graphical user interface, application

Description automatically generated

Figure 23 :Stg channel connection

Graphical user interface, text, application

Description automatically generated

Figure 24 : Stg Channel event handler

* + - * 1. Extract & Load Order Date Details to the Staging

The source file type for Order date details is CSV. According to the data type, to Extract Data from CSV, the best option is Flat file source.

A picture containing text

Description automatically generated

Figure 25 : Stg order dates data flow task

Graphical user interface, application

Description automatically generated

Figure 26 : Stg Order dates connection

Graphical user interface, text, application

Description automatically generated

Figure 27 : Stg order Dates event handler

* + 1. Data Profiling

Once the Step 01 was completed successfully, with an objective of assessing the quality, nature and null values of the data loaded to the staging layer Data Profiling was performed on all the tables.

Chart

Description automatically generated with low confidence

Graphical user interface, application, table

Description automatically generated

Graphical user interface, table

Description automatically generated

* Only contain null values StgOrders and StgStores tables.
  + 1. Data Extraction, Transformation, and Ingestion from Staging to Data warehouse
* In contrast to Step 01, this is the final step of the Datawarehouse solution; Step 03 will implement data transformation when extracting from the staging tables. In Step 0, the tables in the Datawarehouse layer were already designed, also the values for the “DimDate” table have been loaded.
* Before the design of control flow in SSIS, first a stored procedure was generated in the SQL server. The purpose of this, is to update the record that is already existing when data are extracted and add if it’s completely a new record. The record will be identified via the surrogate key.
  + - 1. Stored Procedures

Graphical user interface, text, application, email

Description automatically generated

Figure 28 :DimStore Procedure

Graphical user interface, text, application, email

Description automatically generated

Figure 29 :DimPayment Procedure

Graphical user interface, text, application, email

Description automatically generated

Figure 30 :DimOrderDetails Procedure

Graphical user interface, text, application, email

Description automatically generated

Figure 31 :DimHubs Procedure

Graphical user interface, text, application, email

Description automatically generated

Figure 32 :DimDeliveries procedure

* + - 1. Data Flow

After the execution of the stored procedure, a new SSIS package “Assignment\_Load\_DWH” was created for this process in Data tools. Comparable to Step 01, control flows were created for each table extraction from staging layer, and within the control task the data flow was designed.

Timeline

Description automatically generated

Transform and Load Channel to DWH

Graphical user interface, application, Word

Description automatically generatedDiagram

Description automatically generatedData Conversion flow convert varchar to nvarchar and integer values.

Figure 33 :DimChannel Data Conversion

Figure 34 : DimChannel Data flow task

Transform and Load Drivers to DWH

Diagram

Description automatically generated

Figure 35 :DimDrivers SVD Data FLOW

Diagram

Description automatically generatedTransform and Load Hub to DWH

Figure 36:DimHub Data Flow Task

Transform and Load Stores to DWH

According to the data profile, additionally cleaned and replaced the null values to the “0” value in Store\_plan\_price

Diagram

Description automatically generated

Figure 37 :Dim Stores Data flow task

Graphical user interface, text, application

Description automatically generated

Figure 38 :Dim Stores Filling Nulls

Text

Description automatically generatedTransform and Load Order Details to DWH

Figure 39 :Dim Order details data flow task

Transform and Load Deliveries to DWH

Diagram

Description automatically generated

Figure 40 :Dim Deliveries Data flow task

Transform and Load Payments to DWH

Diagram

Description automatically generated

Figure 41 :Dim payments data flow task

Transform and Load Orders to DWH

Diagram

Description automatically generated

Figure 42 :Fact Orders data flow task

## Accumulating fact tables

* 1. Extended Fact Table

Graphical user interface, text, application

Description automatically generated

* 1. Prepared Dataset

Graphical user interface, application, table, Excel

Description automatically generated

* 1. ETL Pipeline

Timeline

Description automatically generated

* 1. Result

Graphical user interface, application, table

Description automatically generated