Sri Lanka Institute of Information Technology

Data Warehousing and Business Intelligence (IT3021)

Assignment 1 – Year 3 Semester 2

ID: IT22358202

Name: B. N. P. Galwatta

Data set selection - Brazilian E-Commerce Public Dataset by Olist

This is a Brazilian ecommerce public dataset of orders made at Olist Store. Olist is the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels with a single contract. Those merchants can sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners.

The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. It contains 9 datasets: customers dataset, geolocation dataset, order items dataset, order payments dataset, order reviews dataset, orders dataset, products dataset, seller dataset and product name translations dataset.

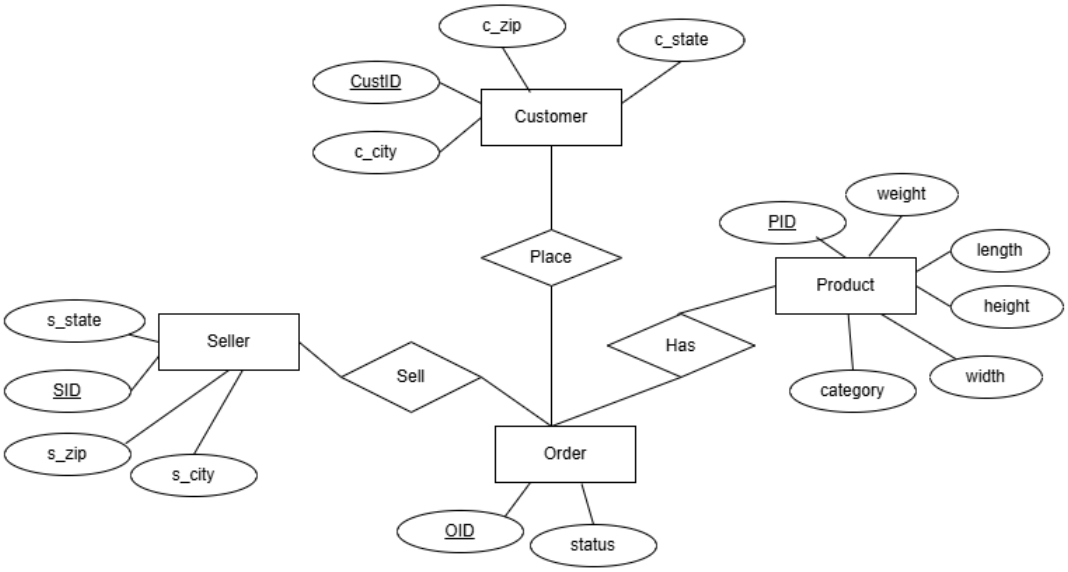
I will use the datasets below for the assignment:

Customers dataset - Information about the customers and their location.

Order dataset, order items dataset, order payments dataset – Order details.

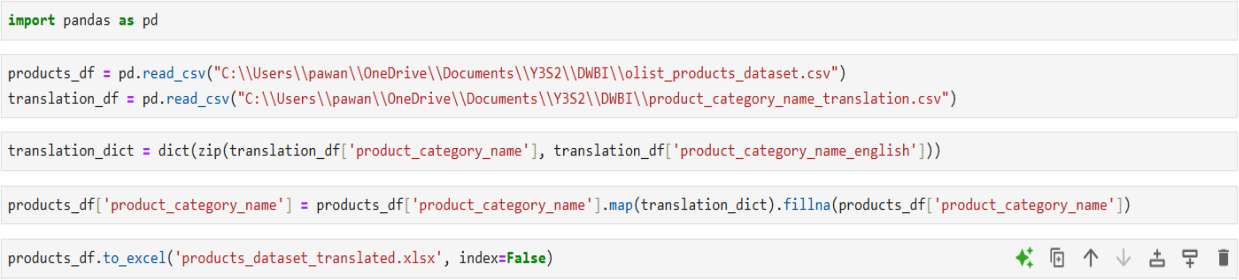
Products Dataset – Data about the products sold by Olist.

Sellers Dataset – Data about the sellers that fulfilled orders made at Olist.



Preparation of data sources

The initial step involved translating the product category names in the olist_products_dataset.csv file into English using the product_category_name_translation.csv file. This was accomplished using Python.



```python
import pandas as pd
```

```python
products_df = pd.read_csv("C:\\Users\\pawan\\OneDrive\\Documents\\Y3S2\\DWBI\\olist_products_dataset.csv")
translation_df = pd.read_csv("C:\\Users\\pawan\\OneDrive\\Documents\\Y3S2\\DWBI\\product_category_name_translation.csv")
```

```python
translation_dict = dict(zip(translation_df['product_category_name'], translation_df['product_category_name_english']))
```

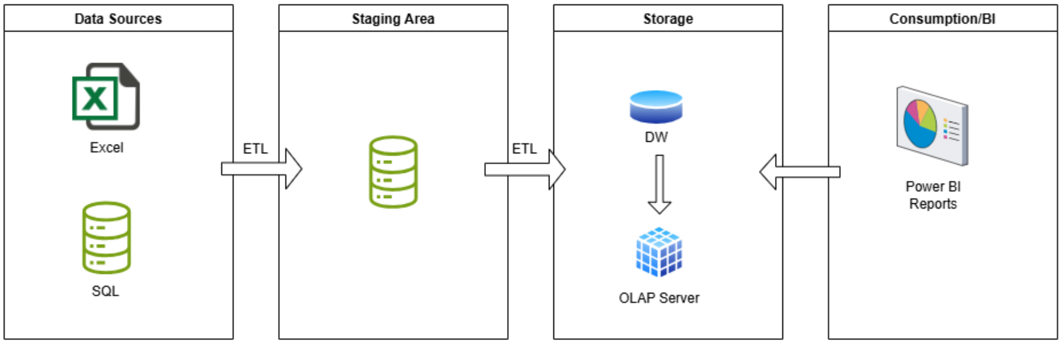
```python
products_df['product_category_name'] = products_df['product_category_name'].map(translation_dict).fillna(products_df['product_category_name'])
```

```python
products_df.to_excel('products_dataset_translated.xlsx', index=False)
```

The core datasets, including customers, products, sellers, orders, and order items were imported into the Olist database to enable structured querying and analysis.

Order payment details were retained in their original format as a CSV file.

Solution architecture



Data Sources:

- Excel Files: Contains raw data such as financials, operations, or customer data; manually maintained or exported from systems.
- SQL Database: Holds structured data from applications, systems, or transactional processes; typically, the main authoritative data source.

Staging Area:

- Cleansed and lightly transformed data ready for further processing into the final warehouse.

ETL (Extract, Transform, Load):

- Automated processes that extract data from Excel and SQL, clean and standardize it, and load it into storage layers.

Storage:

- Data Warehouse (DW): Centralized repository where fully transformed, integrated, and historical data is stored for analysis and reporting.
- OLAP Server (Online Analytical Processing) Server: Organizes data into multidimensional cubes, enabling fast, complex querying and drill-down capabilities.

BI (Business Intelligence Layer):

- Power BI Reports: Dashboards and reports built using Power BI, directly connected to the Data Warehouse for live or scheduled updates.

Data warehouse design & development

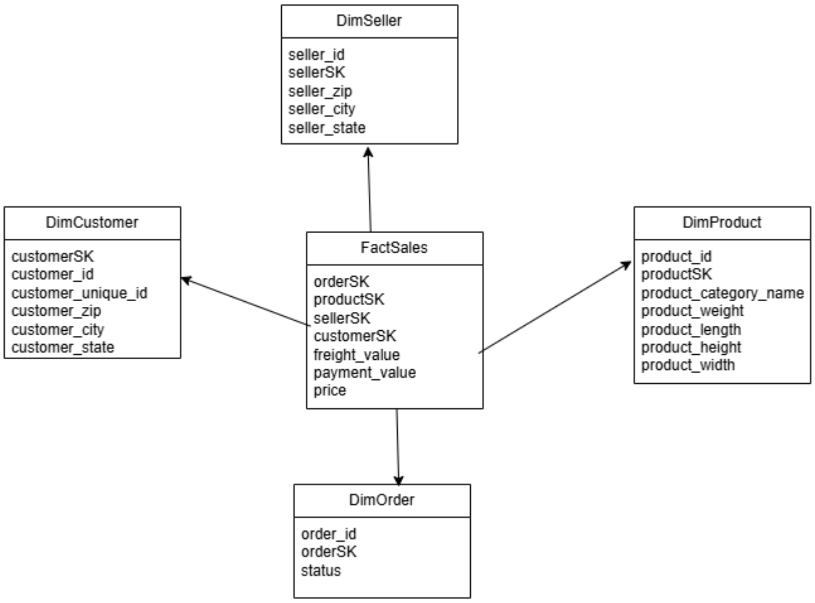
A star schema was designed and implemented for the selected dataset. The schema centers around a Sales fact table and is connected to 5 dimension tables. Sales Fact Table captures transactional data related to each order item. It contains measurable metrics such as payment value and freight value, along with foreign keys referencing the associated dimension tables.

The following dimension tables are included:

1. **Customer Dimension**: Stores customer-related information such as customer ID and location (city, state).
2. **Seller Dimension (Slowly Changing Dimension - Type 2)**: Contains seller details including seller ID and location. As seller information may change over time (e.g., change in address), this dimension is modeled as a Slowly Changing Dimension (SCD Type 2) to preserve historical data.
3. **Order Dimension**: Captures order-level attributes such as order ID and order status.
4. **Product Dimension**: Stores product details, including product ID, category, name, and specifications.

Assumptions: Each sales transaction is uniquely identified at the order item level.

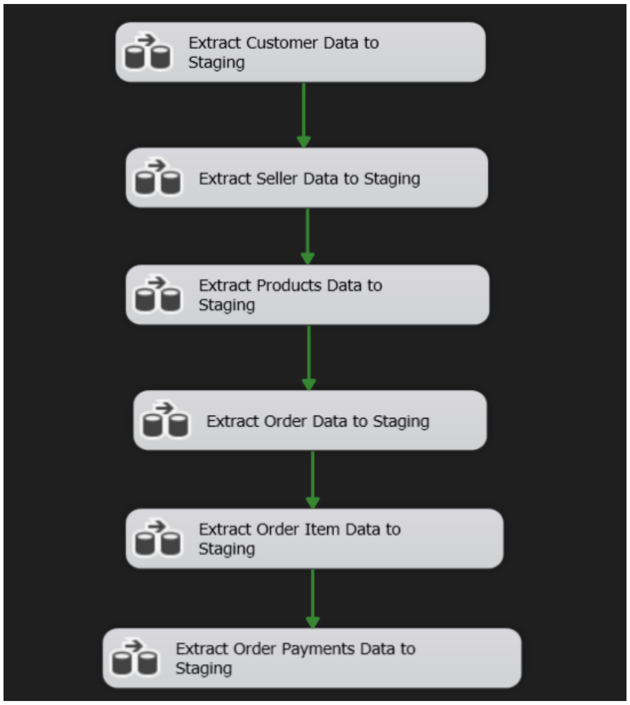
Seller information may change over time.



ETL development

Extract:

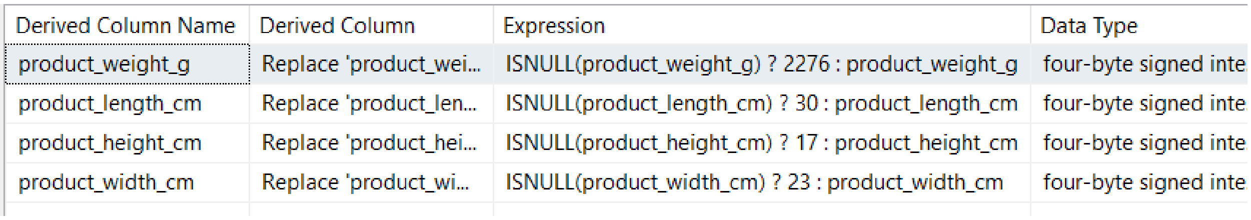
- Extracted the relevant tables - orders, sellers, customers, products, order items – from Olist database and loaded them into the Olist_Staging database.
- Extracted and loaded raw data from order payments CSV file into the Olist_Staging database.


Extract Customer Data to Staging
Extract Seller Data to Staging
Extract Products Data to Staging
Extract Order Data to Staging
Extract Order Item Data to Staging
Extract Order Payments Data to Staging

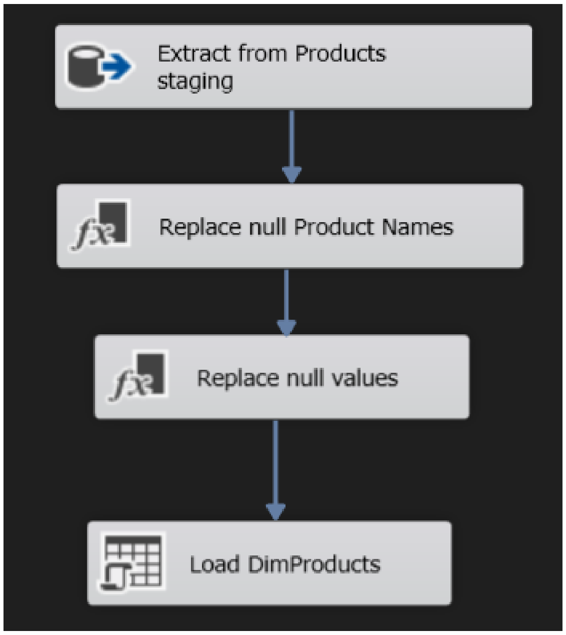
- Created the dimension tables and the fact table in Olist_DW database.
- Created surrogate keys for each dimension table to ensure consistency and performance in fact-dimension joins.

Transform:

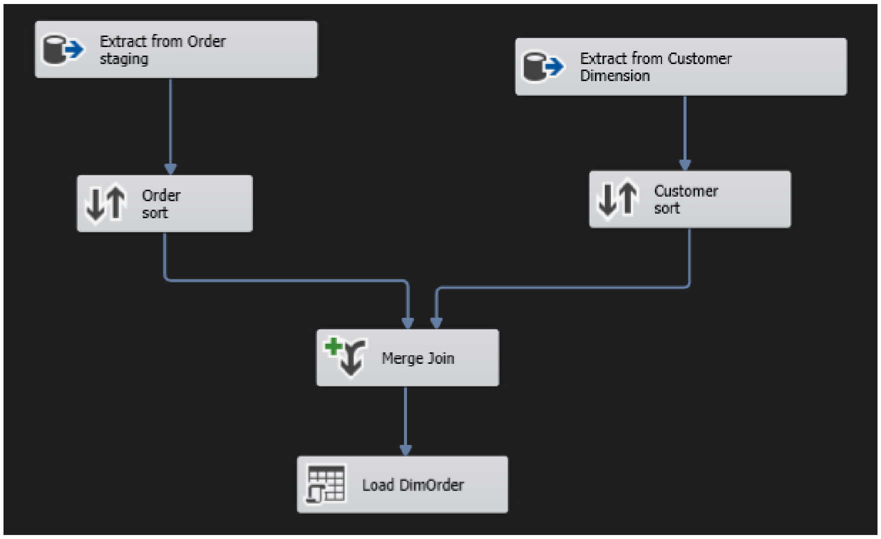
- Handled null values in StgProducts table. There were no null or duplicate values in other tables.



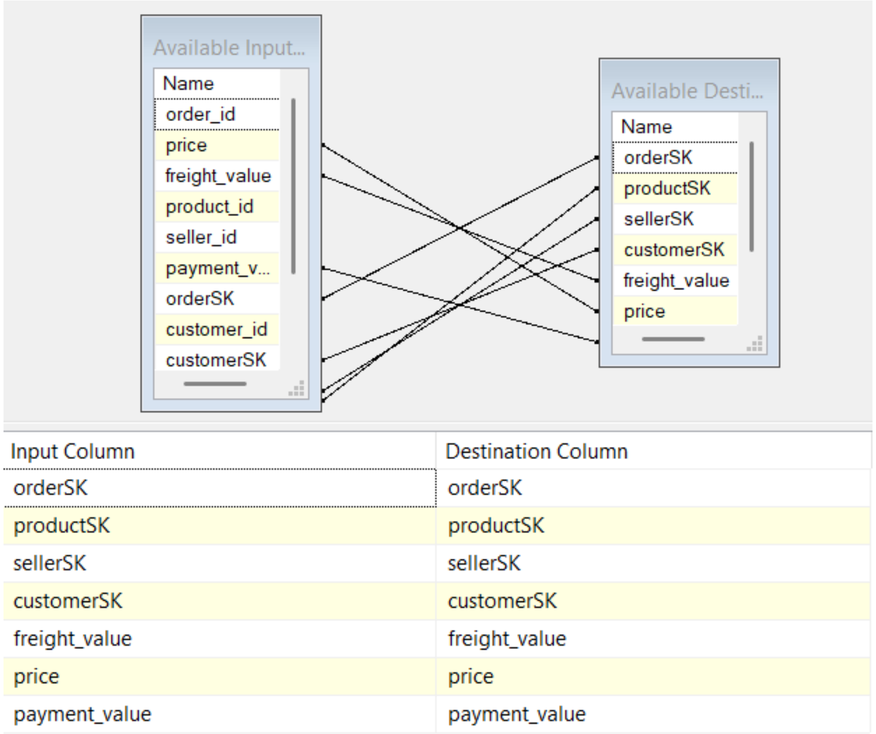
Derived Column Name	Derived Column	Expression	Data Type
product_weight_g	Replace 'product_wei...	ISNULL(product_weight_g) ? 2276 : product_weight_g	four-byte signed inte
product_length_cm	Replace 'product_len...	ISNULL(product_length_cm) ? 30 : product_length_cm	four-byte signed inte
product_height_cm	Replace 'product_hei...	ISNULL(product_height_cm) ? 17 : product_height_cm	four-byte signed inte
product_width_cm	Replace 'product_wi...	ISNULL(product_width_cm) ? 23 : product_width_cm	four-byte signed inte



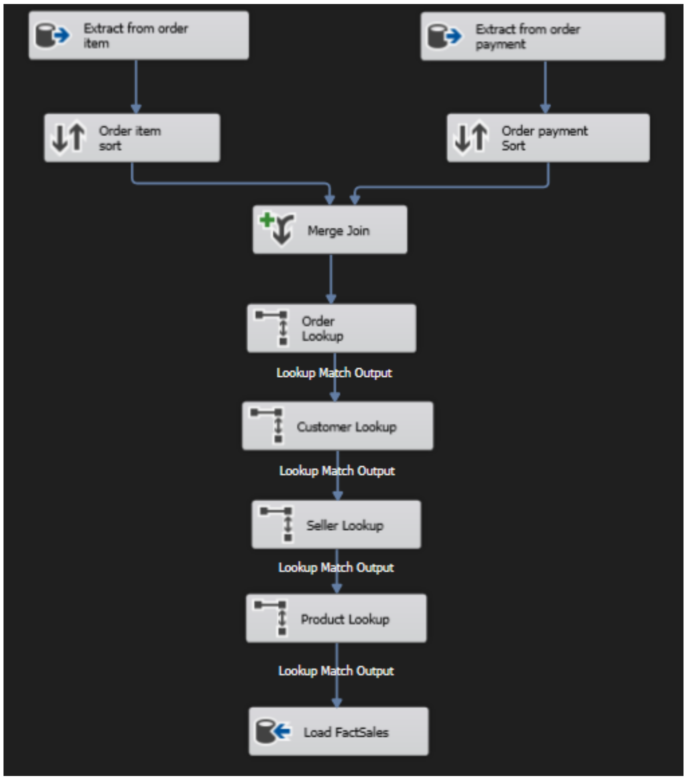
- Joined StgOrder table with DimCustomer to create customer_id foreign key.



- Joined the Order Items, and Order Payments tables to aggregate values price, freight value, and payment value for each order item.
- Linked fact entries to appropriate dimension keys using Look ups.
- Ensured that foreign keys in the fact table matched primary keys in corresponding dimension tables.



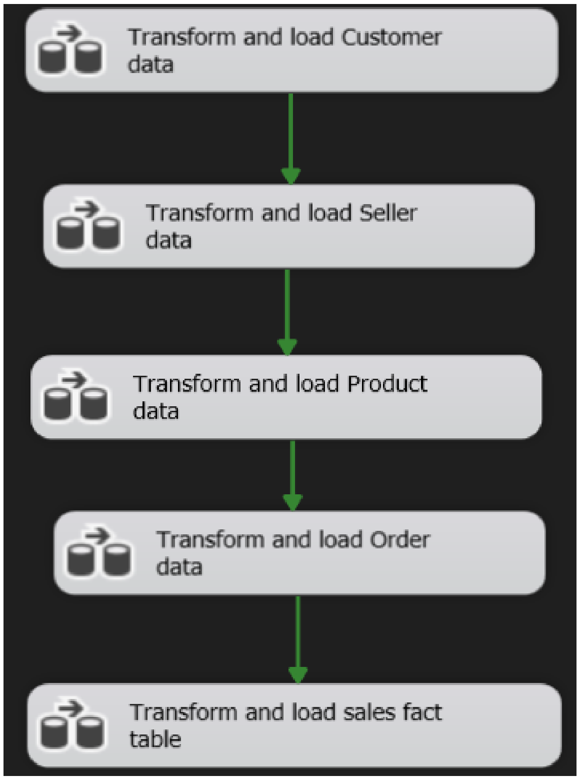
Input Column	Destination Column
orderSK	orderSK
productSK	productSK
sellerSK	sellerSK
customerSK	customerSK
freight_value	freight_value
price	price
payment_value	payment_value



Load:

- Loaded the transformed dimension tables into the data warehouse before loading the fact table to maintain foreign key dependencies.

- Loaded the fact table FactSales with all metrics and foreign keys from dimension tables.


Transform and load Customer data
Transform and load Seller data
Transform and load Product data
Transform and load Order data
Transform and load sales fact table