**Change of Dataset**

Due to the inconsistency of the data the previously selected data set has been changes and the similar but more consistent dataset has been used.

Dataset Used:

# [**Brazilian E-Commerce Public Dataset by Olist**](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_order_payments_dataset.csv)

Description:

The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

This is real commercial data, it has been anonymised, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

**Progress Till Now:**

**Transform CSV file:**

The dataset is in CSV format originally. The dataset contains following file:

* Customer dataset
* Geolocation dataset
* Order dataset
* Payments dataset
* Reviews dataset
* Order status dataset
* Product dataset
* Seller dataset
* Category name translation dataset

All the above dataset are first transformed into pandas dataframe in python. This helps in analysis the dataset and performing the preliminary analysis on:

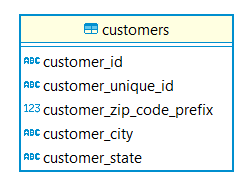
* what are the columns of each dataset
* how many null values are there in each field
* what are the types of each column
* what is the relationship between each datasets?

**Load data into transactional database**

Corresponding MySQL table is created for each of the csv dataset files. MySQL tables are:

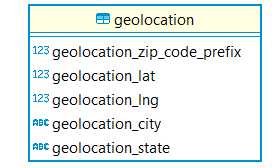
**customers**

* customer\_id
* customer\_unique\_id
* customer\_zip\_code\_prefix
* customer\_city
* customer\_state



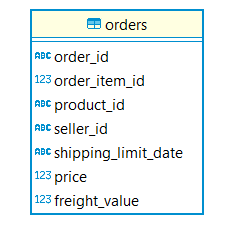
**geolocation**

* geolocation\_zip\_code\_prefix
* geolocation\_lat
* geolocation\_lng
* geolocation\_city
* geolocation\_state



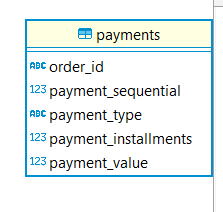
**orders**

* order\_id
* order\_item\_id
* product\_id
* seller\_id
* shipping\_limit\_date
* price
* freight\_value



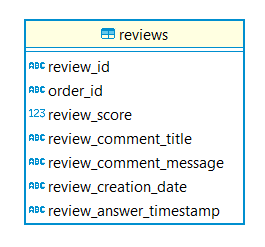
**payments**

* order\_id
* payment\_sequential
* payment\_type
* payment\_installments
* payment\_value



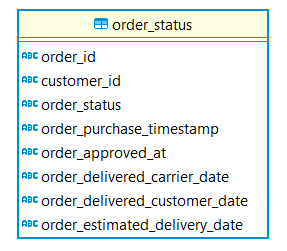
**reviews**

* review\_id
* order\_id
* review\_score
* review\_comment\_title
* review\_comment\_message
* review\_creation\_date
* review\_answer\_timestamp



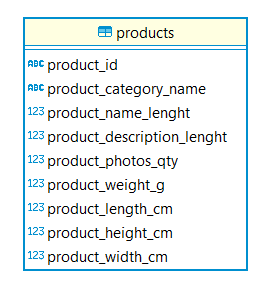
**order\_status**

* order\_id
* customer\_id
* order\_status
* order\_purchase\_timestamp
* order\_approved\_at
* order\_delivered\_carrier\_date
* order\_delivered\_customer\_date
* order\_estimated\_delivery\_date



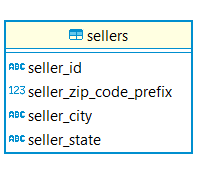
**products**

* product\_id
* product\_category\_name
* product\_name\_lenght
* product\_description\_lenght
* product\_photos\_qty
* product\_weight\_g
* product\_length\_cm
* product\_height\_cm
* product\_width\_cm



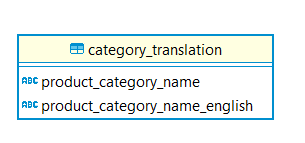
**sellers**

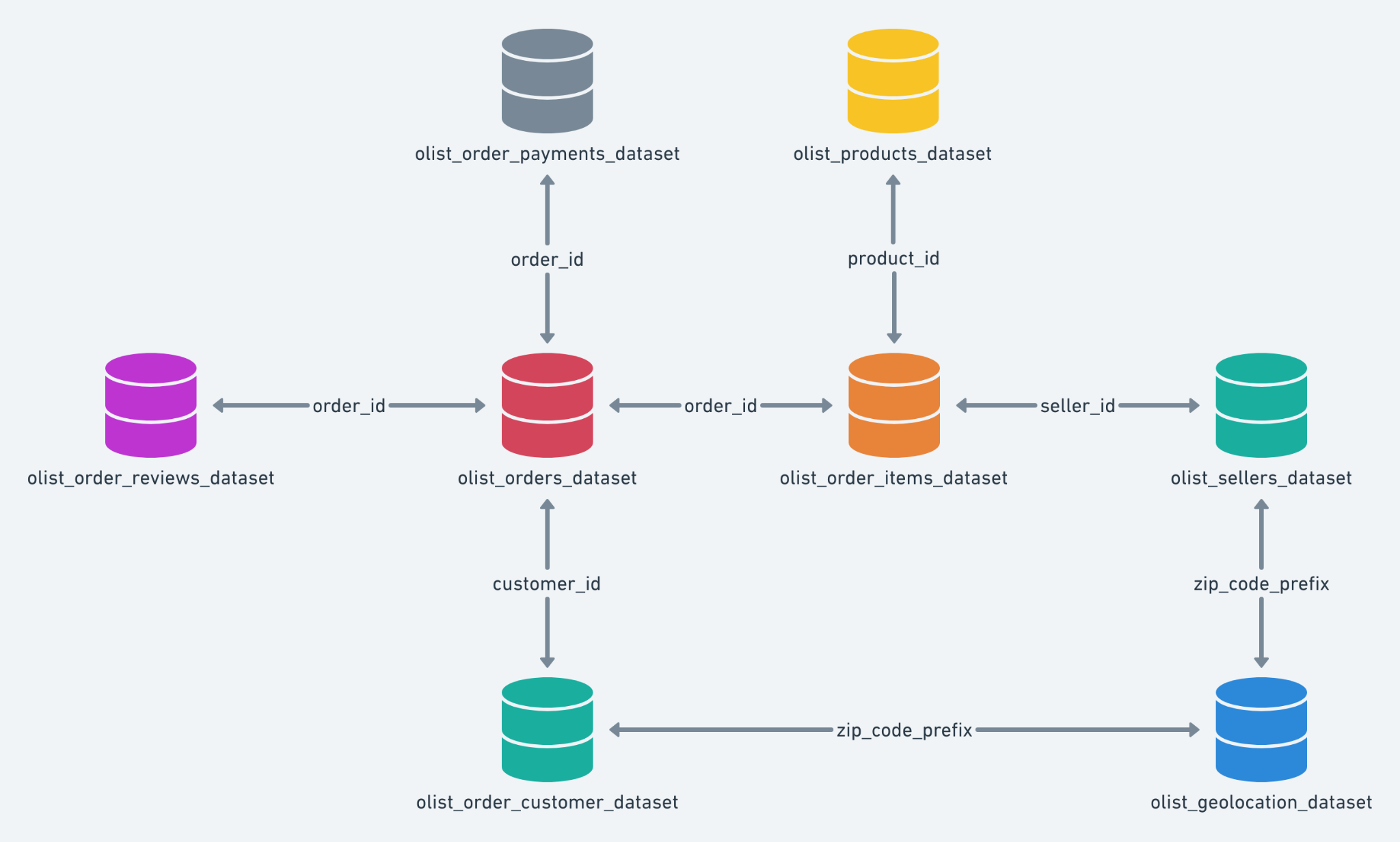
* seller\_id
* seller\_zip\_code\_prefix
* seller\_city
* seller\_state



**category\_transaltion**

* product\_category\_name
* product\_category\_name\_english



****

**In Progress**

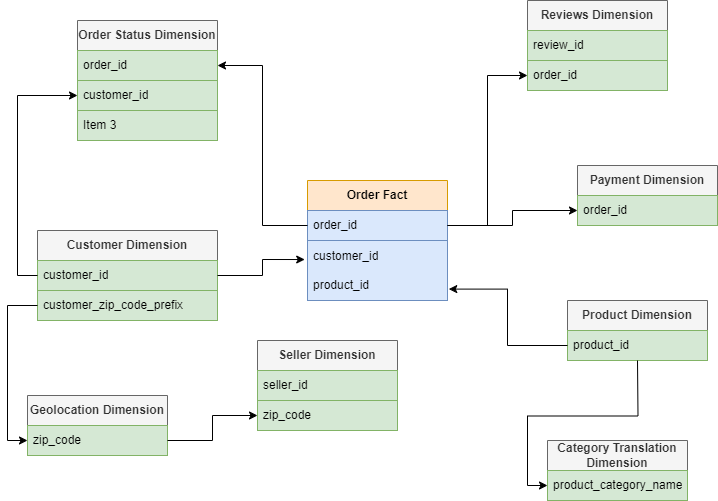
**Maintain Data Warehouse:** The central repository where data from various sources is stored in a structured format optimised for querying and analysis. It typically consists of fact tables and dimension tables.

OLAP cubes are multi-dimensional structures that allow for fast analysis of data along multiple dimensions. The cube can be pre-aggregated to various levels of granularity to support different types of analyses.

The data from the above transactional database is extracted, transformed and loaded in the Data warehouse.

**Multidimensional data warehouse architecture**

The architecture of a data warehouse schema is best represented by a star schema which is characterised by a central fact table that contains measurable, quantitative data about transactions, and a series of surrounding dimension tables that contain descriptive attributes related to the fact table's measurements.

****

**Central Fact Table:**

**orders\_fact**

* order\_id (Primary Key)
* customer\_id (Foreign Key)
* product\_id (Foreign Key)
* seller\_id (Foreign Key)
* price freight\_value
* payment\_value (sum of all payments from payments table) review\_score (average score from reviews table)
* order\_item\_count (could be a count of items per order if available)

**Dimension Tables:**

**customers\_dim**

* customer\_id (Primary Key)
* customer\_unique\_id
* customer\_zip\_code\_prefix (Foreign Key)
* customer\_city
* customer\_state

**products\_dim**

* product\_id (Primary Key)
* product\_category\_name\_english (from category\_translation table) product\_name\_length
* product\_description\_length
* product\_photos\_qty
* product\_weight\_g
* product\_length\_cm
* product\_height\_cm
* product\_width\_cm

**payments\_dim**

* order\_id (Foreign Key)
* payment\_sequential
* payment\_type
* payment\_installments

**reviews\_dim**

* order\_id (Foreign Key)
* review\_id (Foreign Key)
* review\_comment\_title
* review\_comment\_message
* review\_creation\_date
* review\_answer\_timestamp

**order\_status\_dim**

* order\_id (Foreign Key)
* order\_status
* order\_purchase\_timestamp
* order\_approved\_at
* order\_delivered\_carrier\_date
* order\_delivered\_customer\_date
* order\_estimated\_delivery\_date

**sellers\_dim**

* seller\_id (Primary Key)
* seller\_zip\_code\_prefix
* seller\_city
* seller\_state

**geolocation\_dim**

* geolocation\_zip\_code\_prefix (Primary Key)
* geolocation\_lat
* geolocation\_lng
* geolocation\_city
* geolocation\_state

**category\_translation\_dim**

* product\_category\_name (Primary Key)
* Product\_category\_name\_english (Foreign Key)

**Tools and Technologies Used:**

**MySQL:** Creating facts and dimension tables

**Programming language:** Python with libraries such as pandas for data manipulation and transformation

**To Do**

**Schedule:**

* Schedule the ETL process to run periodically (e.g., daily, hourly) to ensure the data warehouse reflects the latest data from the source systems.

**Analysis and Report:**

Customer Segmentation

Analytics:

* Sales figures
* Customer Demographic
* Product vs Revenue
* Payment Analysis
* customer satisfaction

Prediction:

* Churn Rate
* Sales Prediction