

# Brazilian E-Commerce Public Dataset by OLIST

## Information About The Dataset

Data Period ⇒ 2016 to 2018

Welcome! This is a Brazilian e-commerce public dataset of orders made at the **OLIST** Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow 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 anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

### Context

This dataset was generously provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without the hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using **OLIST** logistics partners. See more on our website: [www.olist.com](http://www.olist.com)

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note about the purchase experience and write down some comments.

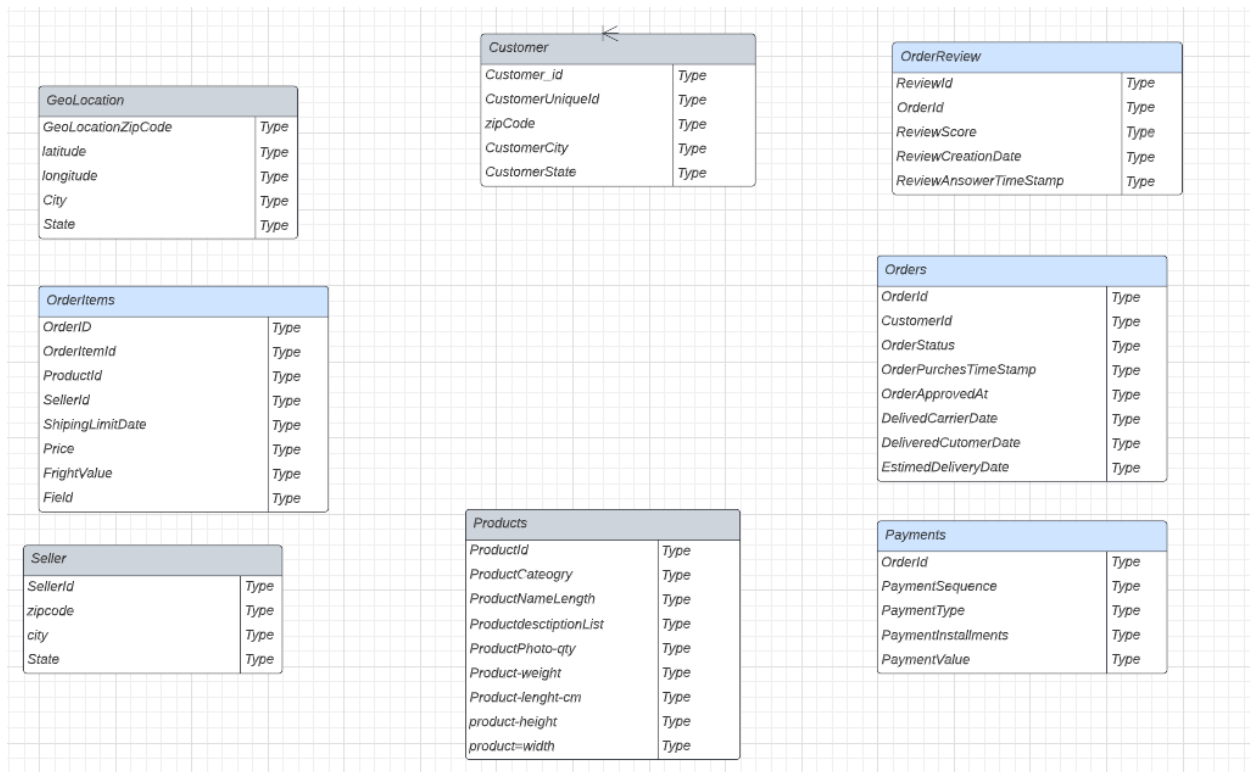
## Attention

An order might have multiple items.

Each item might be fulfilled by a distinct seller.

All text identifying stores and partners where replaced by the names of Game of Thrones great houses.

## Data Origin Design



[https://lucid.app/lucidchart/f7912a3a-2dc9-4df2-8599-8c68e5aaa1e6/edit?beaconFlowId=86E6D6D927078E9A&invitationId=inv\\_b9a4a841-db24-4397-b596-7eb0fe55aba4&page=0\\_0#](https://lucid.app/lucidchart/f7912a3a-2dc9-4df2-8599-8c68e5aaa1e6/edit?beaconFlowId=86E6D6D927078E9A&invitationId=inv_b9a4a841-db24-4397-b596-7eb0fe55aba4&page=0_0#)

## My Work

First

i should make The DataBase in SQL Server to handle anything I want

1. create database in SQL server "**OLIST** "
2. make an integration Project in my Visual Studio "That will add the CSV Files into my OLEDB "

Hint

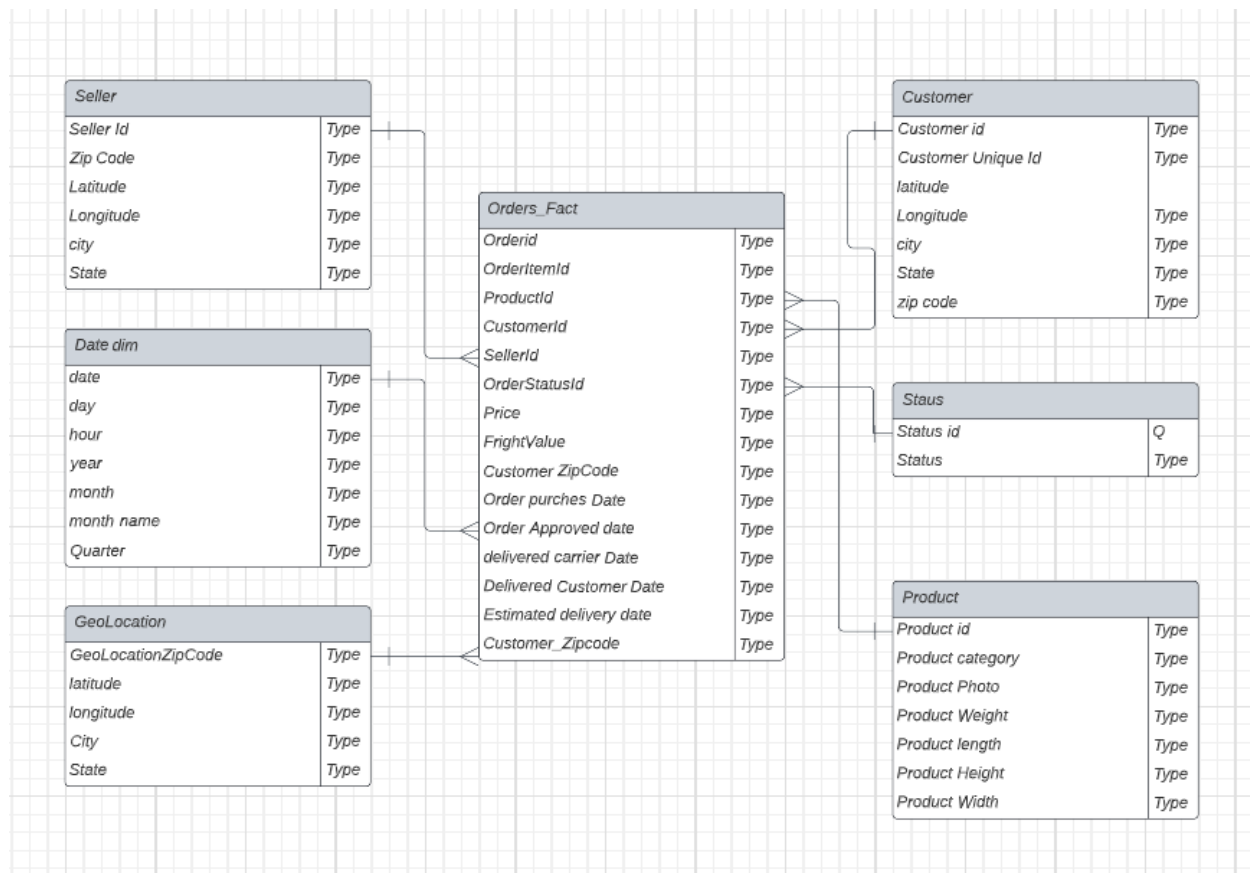
When I see the data I see that The Geo Location have a lot of Rows with The Same "zip code but different in lat and long" this cause because there is a lot of people in the same area that have same zip code

for that I should make a solution for this Problem

The Solution

I will remain one value for one zip code in geo location table then I will add the lat and long in the customer and seller tables to make it easy to make my Star Schema not Snowflek schema "i think it is the best solution for this problem"

The Design of DWH will be



[https://lucid.app/lucidchart/f2314cb6-0587-4e2b-923e-e4a99f177238/edit?beaconFlowId=D490CD0D22259B07&invitationId=inv\\_6209e3d6-060f-4a19-ab4c-fc9cbd86ee62&page=0\\_0#](https://lucid.app/lucidchart/f2314cb6-0587-4e2b-923e-e4a99f177238/edit?beaconFlowId=D490CD0D22259B07&invitationId=inv_6209e3d6-060f-4a19-ab4c-fc9cbd86ee62&page=0_0#)

## SSIS Work & ETL

i should Get Data From CSV file to my SQL Database "OLIST Data base"

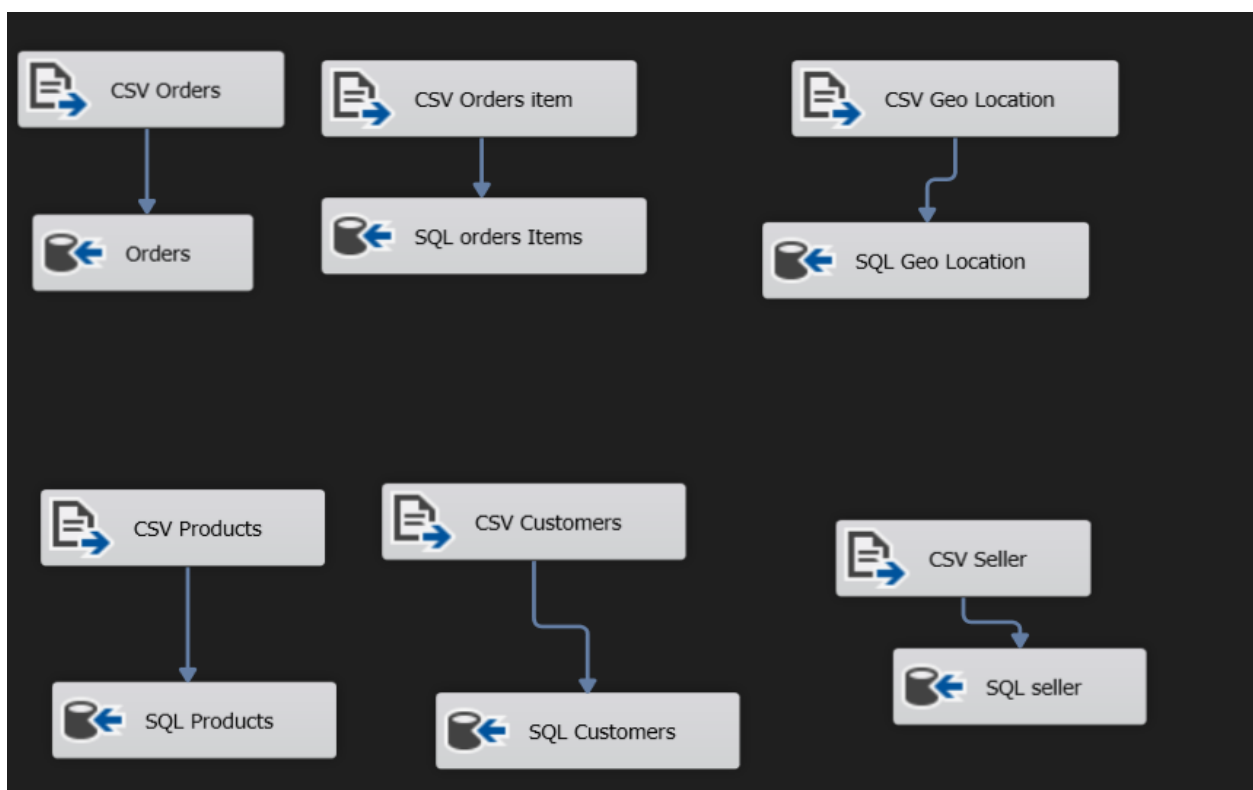
This is the control Flow task :

- here i used SSIS and SQL server for ETL Process
- all things i use are in Stored procedure written in sql and called in SSIS

## Control Flow Task



## Data Flow task :



## SQL Steps

first i should Make The Views Which i should take Data From also  
This Step is a Transform Step

### **Building the Structure of the Tables that i will use in my DWH**

Business Case is "order"

1. Building the Fact table "Orders" Containing (olist orders table + olist order items +Customer\_ZipCode)
2. Date Table as a dim table
3. seller table as a dim table
4. product table as a dim table
5. status dim table
6. customer dim table
7. geolocation dim table

### **Transformation & cleansing**

making a views that will containg the transformed data that i will populate it later in its distination table "using stored procedures"

1. Making the View Of Getting the geo location data after deleting the repeated items with same zip code
2. making view of getting fact table data "orders / orders items +customer\_ZipCode"
3. Making a view of Status data Get\_Status\_Data
4. making a view of Products data get\_product\_data
5. making a view of seller data get\_seller\_data
6. making a view of Customer data get\_customer\_data

## Populting Data

Population of data is done with a stored procedure that pull the data from a view and put this data into its location "table"

SPs

1. Creating a Stored procedure of populating product data  
"populate\_product\_data"
2. Creating a Stored procedure of populating seller data  
"populate\_seller\_data"
3. Creating a Stored procedure of populating Customer data "  
populate\_customer\_data"
4. Creating a Stored procedure of populating geo location data  
"populate\_geo\_location\_data"
5. Creating a Stored procedure of populating Status data  
"populate\_status\_data"
6. Creating a Stored procedure of populating date data  
"populate\_Date" (it takes 2 input parameters start\_date , and  
end\_date then it generate dates between this two dates )
7. then it is time to populate the fact table with data  
"populate\_order\_data"

## Cleaning

my data was need some cleaning like removing duplicates and clean the date to make it related to hours only

1. clean with a view "view of geo location data" removing all  
duplicated of zip code
2. cleaning dates to hold hours data only to connect clearly to  
the dim date "cleaning\_dates"

3. cleaning geolocation data cause of the city of it has incorrect typing "cleaning\_geo\_location"
4. also in orders fact table i found that "delivey\_carrier\_Date , order\_approved\_time" have some data that is not valid where the 2 dates are equal to a date in 1999 so i make a Stored procedure that will update those values to go the default value of date that it '2012-01-01 00:00:00.000' those anominal values its executino will be :

```
Messages  
  
(2454 rows affected)  
  
(1 row affected)  
  
Completion time: 2023-03-06T13:45:30.0776391+02:00
```

5. also in my data i found that there is some users that there zipcode and city and state are not stored in geolocation table so that i will add an difault value in geolocation table that will hold the errors like this error "clean\_customer\_zipcode"

```
insert into geolocation (zipcode,latitude,longitude,city,state) values (0,0,0,'N/A','N/A');
```

## Execution of Stored Procedure

to make data go to its distination i should execute all of the stored procedure that i have

- in execution of populate date stored procedure first i get the minimum vales of data that in my data also i get the max



value of date in my data and passed it as parameters to my stored procedure

```
select min(order_purchases_date) from orders_fact  
  
select max(Estimated_delivery_date) from orders_fact  
  
exec populate_Date '2016-01-01 00:00:00.000' , '2018-12-31 23:00:00.000'
```

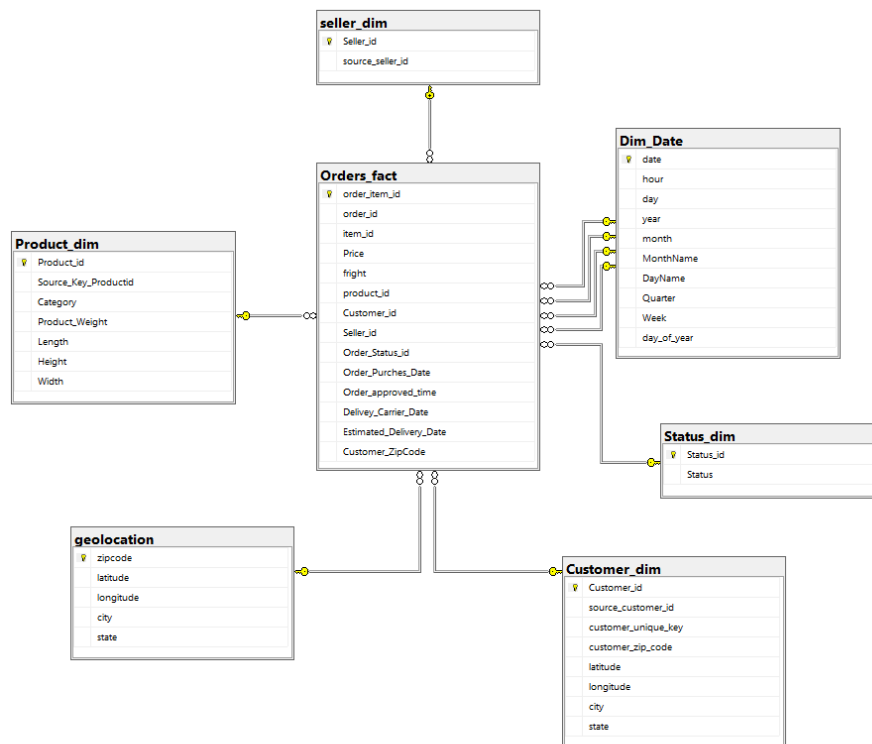
## Relations

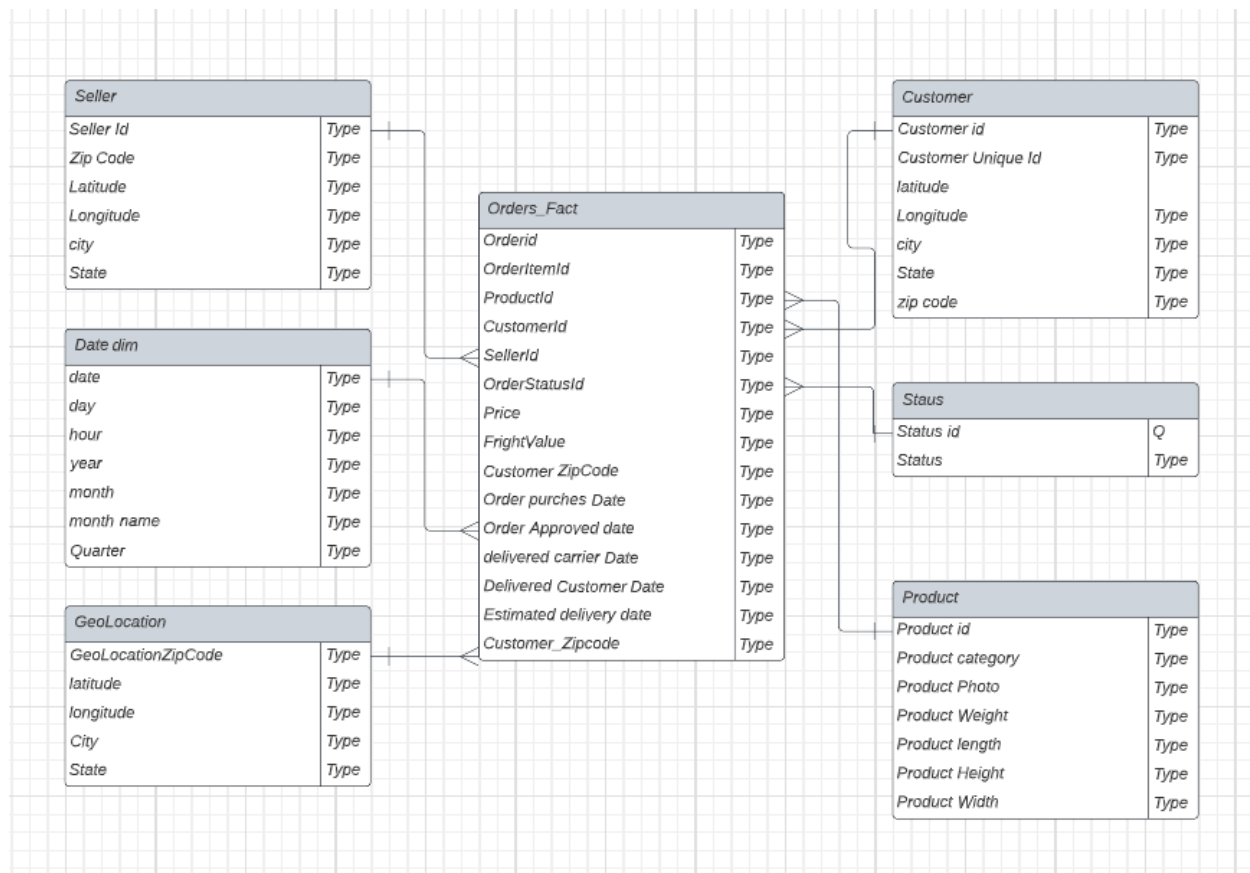
Now it is time to make the relations between tables with also a stored procedure ""

hints:

- in customer data there is customers that have more than 1 id in customer data but the unique customer id will hold one value for all of the repeated
- in the relation between table to make it more easy and more fast to increase the performace of my DWH i change the primary key of each dim table and generate a surgate key that i will use in making relations between dims and fact table
- 

final diagram is : for the OLAP





## ▼ Visualizaion