|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  |  |  | |  |
| Brazilian E-Commerce  Final Report Project 2: ETL  UCSD Data Science and Visualization Bootcamp | | |
| Team:  Amir Vazquez Narjes Taqvaei Annette Broeren | | February 24, 2020 |

# Introduction

The objective of this report is to describe the approach and process of transforming a dataset into a data warehouse for analytical purposes and business reporting. For this project, Brazilian E-Commerce and marketing funnel datasets released by Olist Stores were used.

The E-Commerce dataset contains data for approximately 100,000 transactions from 2016 to 2018 made at multiple marketplaces in Brazil.

The marketing funnel dataset contains data from sellers that filled-in requests for contact to sell their products on Olist Store. The dataset has information of approximately 8,000 Marketing Qualified Leads (MQLs) collected between June 1, 2017 and June 1, 2018. They were randomly sampled from the total MQLs.

The following tasks were explored:

***Extract*** original data sources;

***Transform*** data into clean repositories

***Load*** the final repositories into a data warehouse

# Sources

1. Brazilian E-Commerce Public Dataset by Olist (<https://www.kaggle.com/olistbr/brazilian-ecommerce>)

9 .CSV files

52 columns

13 string

13 Integer

12 Uuid

14 Other

2. Marketing Funnel by Olist (<https://www.kaggle.com/olistbr/marketing-funnel-olist>)

2 .CSV files

18 columns

6 string

6 Uuid

2 DateTime

4 Other

# Approach

The approach consists of a two-tier stage environment where the raw data is imported into the database, then prepared to populate final tables for the data warehouse.

* Download the source files to a Github repository with branches for all team members
* Create staging tables from the source files in PGAdmin
* Add 4 columns: Create-date, Created\_by, Changed\_date, Changed\_by
* Create a date dimension table to facilitate analysis based on dates and periods
* Use Python to
  + Inspect the tables
  + Convert date/time strings to date dimensions
  + Add triggers to auto-populate date created plus username and date updated plus username
* Load reformatted tables into PGAdmin



# Execution

## Raw data to Staging

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Source file** | To | **Staging table** |
| 1 | olist\_customer\_dataset.csv |  | stg\_olist\_customer\_dataset |
| 2 | olist\_geolocation\_dataset.csv |  | stg\_olist\_geolocation\_dataset |
| 3 | olist\_order\_items\_dataset.csv |  | stg\_olist\_order\_items\_dataset |
| 4 | olist\_order\_payments\_dataset.csv |  | stg\_olist\_order\_payments\_dataset |
| 5 | olist\_order\_reviews\_dataset.csv |  | stg\_olist\_order\_reviews\_dataset |
| 6 | olist\_orders\_dataset.csv |  | stg\_olist\_orders\_dataset |
| 7 | olist\_products\_dataset.csv |  | stg\_olist\_products\_dataset |
| 8 | olist\_sellers\_dataset.csv |  | stg\_olist\_sellers\_dataset |
| 9 | product\_category\_name\_translation.csv |  | stg\_product\_category\_name\_translation |
| 10 | olist\_marketing\_qualified\_leads\_dataset.csv |  | stg\_olist\_marketing\_qualified\_leads\_dataset |
| 11 | olist\_closed\_deals\_dataset.csv |  | stg\_olist\_closed\_deals\_dataset |

## INsert Data

/\*Insert data from olist\_customer\_dataset.csv\*/

COPY stg\_olist\_customer\_dataset (

    customer\_id,

    customer\_unique\_id,

    customer\_zip\_code\_prefix,

    customer\_city,

    customer\_state)

/\*Update your location of the files here\*/

FROM ‘(PATH)\Resources/olist\_customers\_dataset.csv' DELIMITER ',' CSV HEADER;

## staging to final

|  |  |  |  |
| --- | --- | --- | --- |
|  | Staging Table | To | Final Table |
| 1 | stg\_olist\_customer\_dataset |  | customer |
| 2 | stg\_olist\_geolocation\_dataset |  | geolocation |
| 3 | stg\_olist\_order\_items\_dataset |  | order\_items |
| 4 | stg\_olist\_order\_payments\_dataset |  | order\_payments |
| 5 | stg\_olist\_order\_reviews\_dataset |  | order\_reviews |
| 6 | stg\_olist\_orders\_dataset |  | orders |
| 7 | stg\_olist\_products\_dataset |  | products |
| 8 | stg\_olist\_sellers\_dataset |  | sellers |
| 9 | stg\_product\_category\_name\_translation |  | prod\_category |
| 10 | stg\_olist\_marketing\_qualified\_leads\_dataset |  | mkt\_leads |
| 11 | stg\_olist\_closed\_deals\_dataset |  | mkt\_deals |

-- Table: public.products

DROP TABLE IF EXISTS public.products;

CREATE TABLE public.products

(

    product\_id character varying(40) COLLATE pg\_catalog."default" NOT NULL,

    product\_category\_name character varying(50) COLLATE pg\_catalog."default",

    product\_name\_lenght integer,

    product\_description\_lenght integer,

    product\_photos\_qty integer,

    product\_weight\_g integer,

    product\_length\_cm integer,

    product\_height\_cm integer,

    product\_width\_cm integer,

    "CREATE\_DATE" date,

    "CREATED\_BY " character varying(60)[] COLLATE pg\_catalog."default",

    "CHANGED\_DATE" date,

    "CHANGED\_BY" character varying(60) COLLATE pg\_catalog."default",

    CONSTRAINT products\_pkey PRIMARY KEY (product\_id)

)

TABLESPACE pg\_default;

ALTER TABLE public.products

    OWNER to postgres;

## DAte Conversions

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Pellentesque id nibh tortor id aliquet. Purus in mollis nunc sed. Elit duis tristique sollicitudin nibh sit amet commodo nulla. At tellus at urna condimentum. Nunc non blandit massa enim nec dui nunc. Massa id neque aliquam vestibulum morbi blandit.

Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida. Nunc non blandit massa enim nec dui nunc mattis. Nulla malesuada pellentesque elit eget gravida cum. Sit amet nulla facilisi morbi

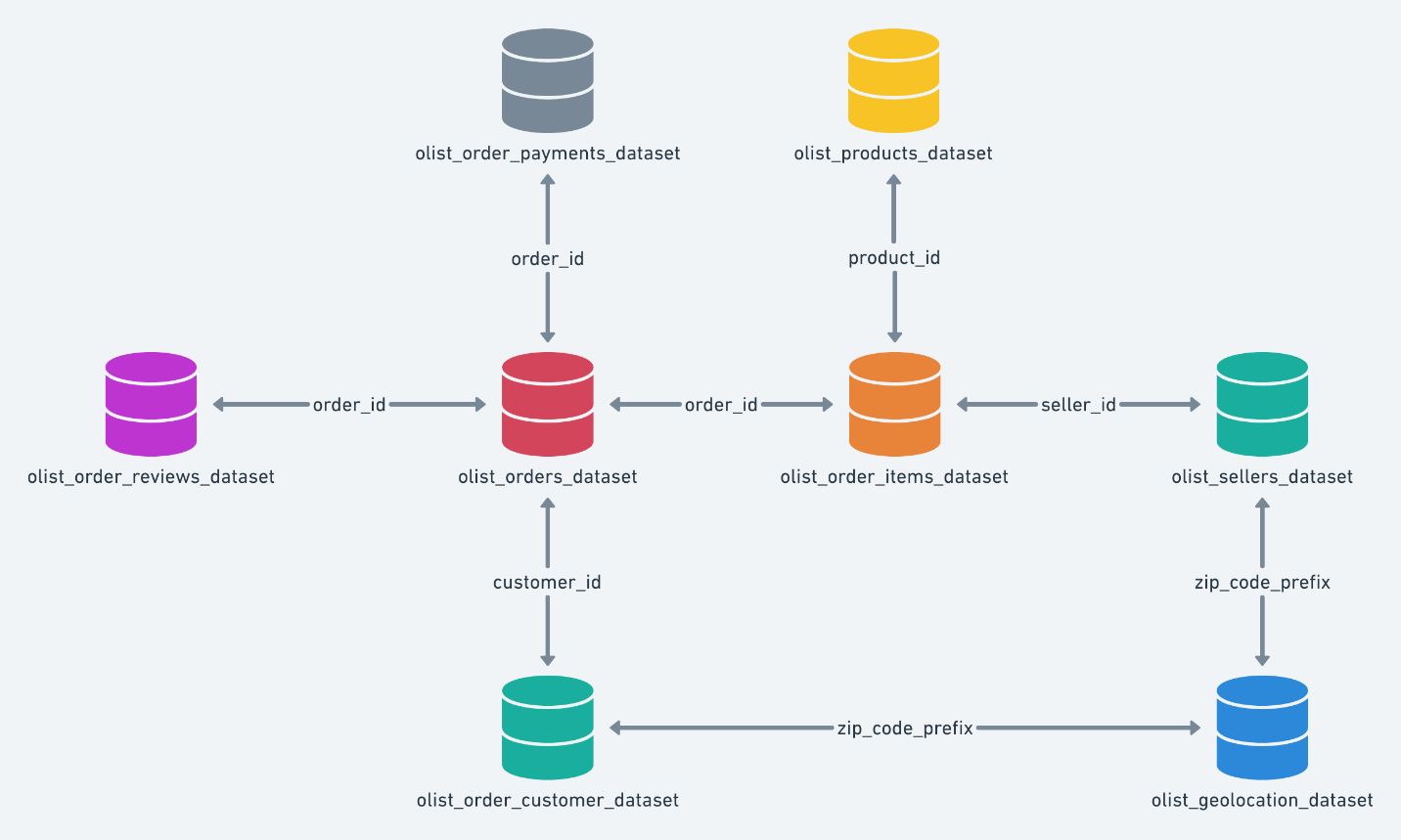
## Triggers

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Pellentesque id nibh tortor id aliquet. Purus in mollis nunc sed. Elit duis tristique sollicitudin nibh sit amet commodo nulla. At tellus at urna condimentum. Nunc non blandit massa enim nec dui nunc. Massa id neque aliquam vestibulum morbi blandit.

Lorem ipsum dolor sit amet consectetur adipiscing. Nisi lacus sed viverra tellus. Orci eu lobortis elementum nibh tellus molestie nunc non. Laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Pharetra massa massa ultricies mi quis hendrerit dolor. Non tellus orci ac auctor augue mauris augue neque gravida. Nunc non blandit massa enim nec dui nunc mattis. Nulla malesuada pellentesque elit eget gravida cum. Sit amet nulla facilisi morbi

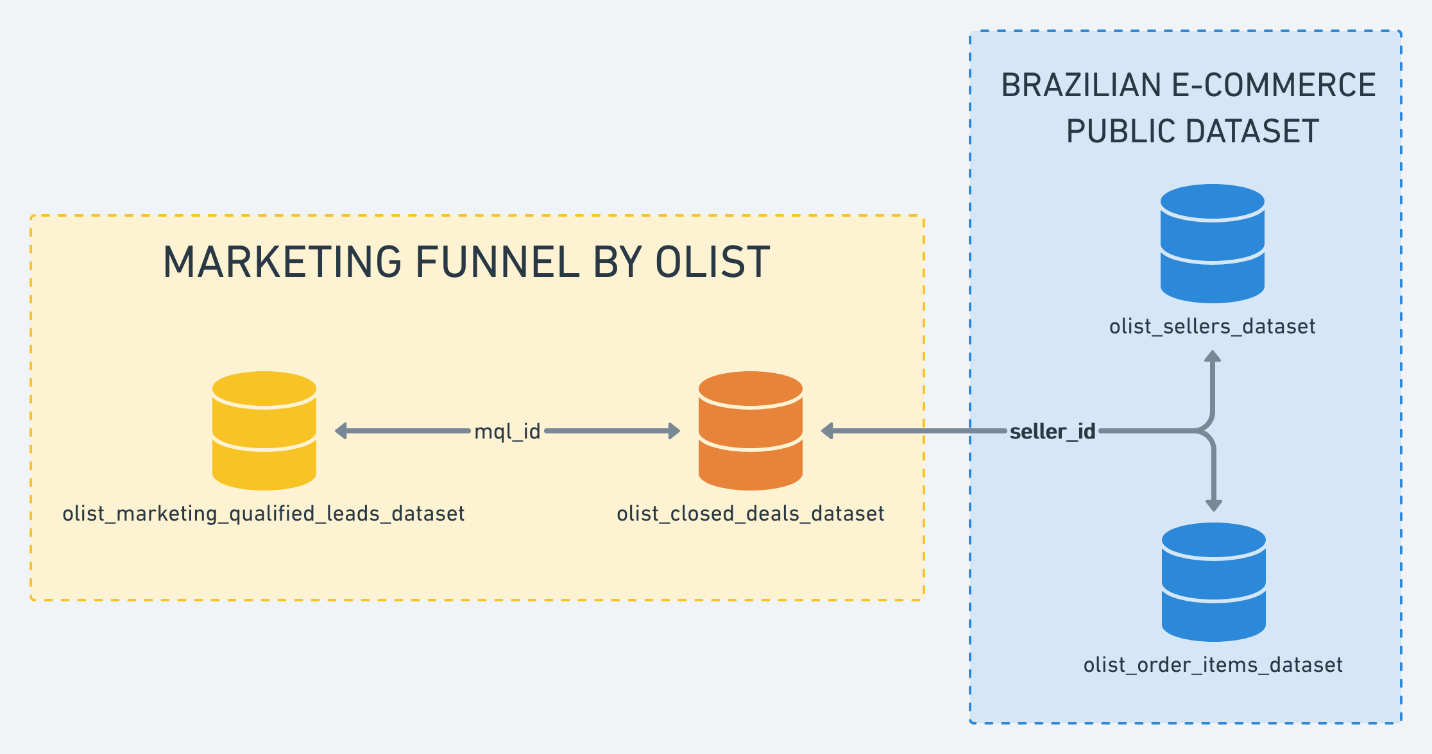
# Data Warehouse

Figure 1: Data Schema for Brazilian E-Commerce Public Dataset



Source: <https://www.kaggle.com/olistbr/brazilian-ecommerce>

Figure 2: Data Schema for Marketing Funnel Dataset



Source: https://www.kaggle.com/olistbr/marketing-funnel-olist/home

# Issues Encountered

## Source Files

Source files were read only, so the SQL server could not read the data.

### Solution:

1. Right click on file name
2. Go to security tab
3. Add Everyone to usernames
4. Change permissions to allow for all

Note: Pushing and pulling from Github may reverse these property-settings.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt.

## String to Date conversions

(Narjes and/or Amir to describe)

## Github

An empty .git file caused fatal push and pull errors for a team member. This file was located in the root of the C-drive and had to be deleted in order for git to resume normal functionality for that team member.