**Description**

**Brazilian E-Commerce Public Dataset by Olist**

This is a Brazilian ecommerce public dataset of orders made at Olist Store. 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.

Context

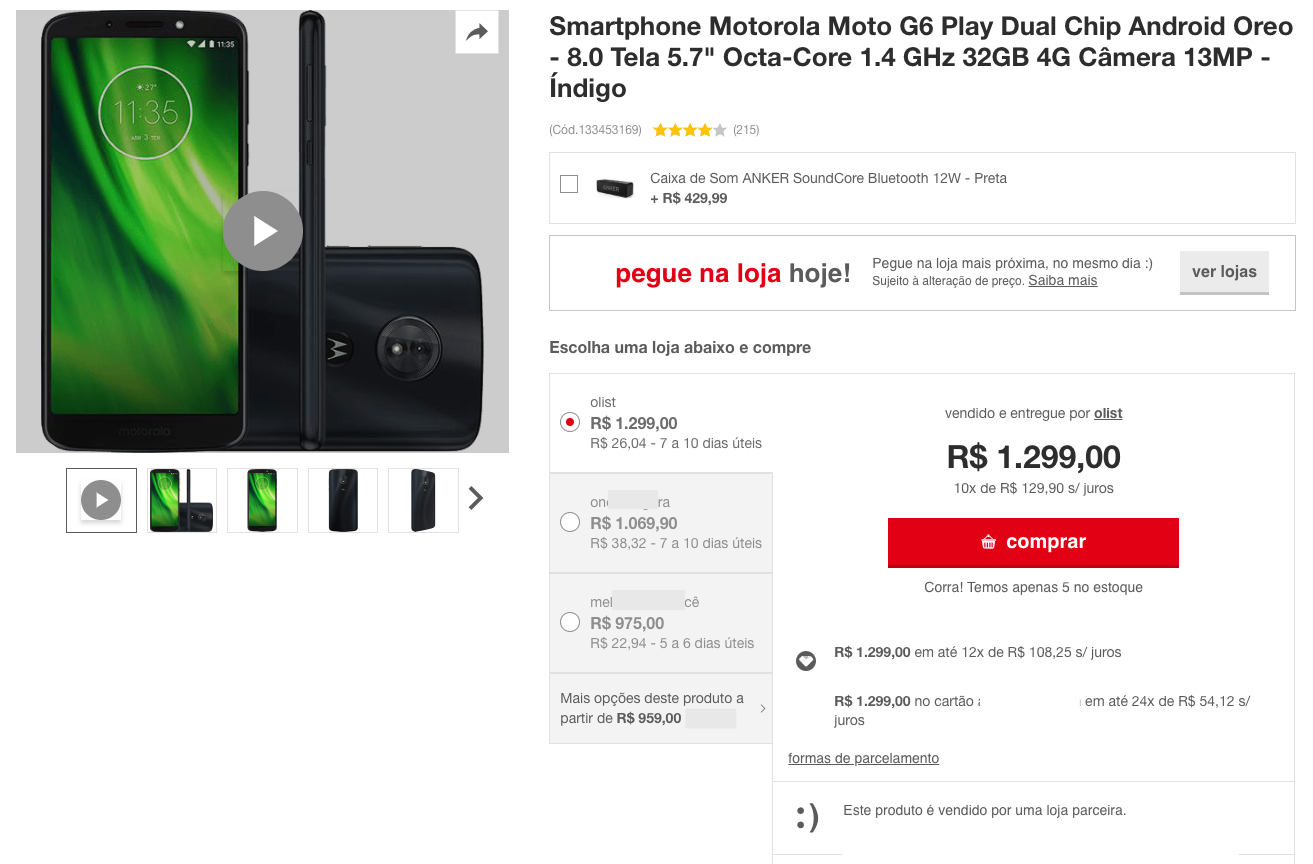
This dataset was generously provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners.

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 for the purchase experience and write down some comments.

Attention

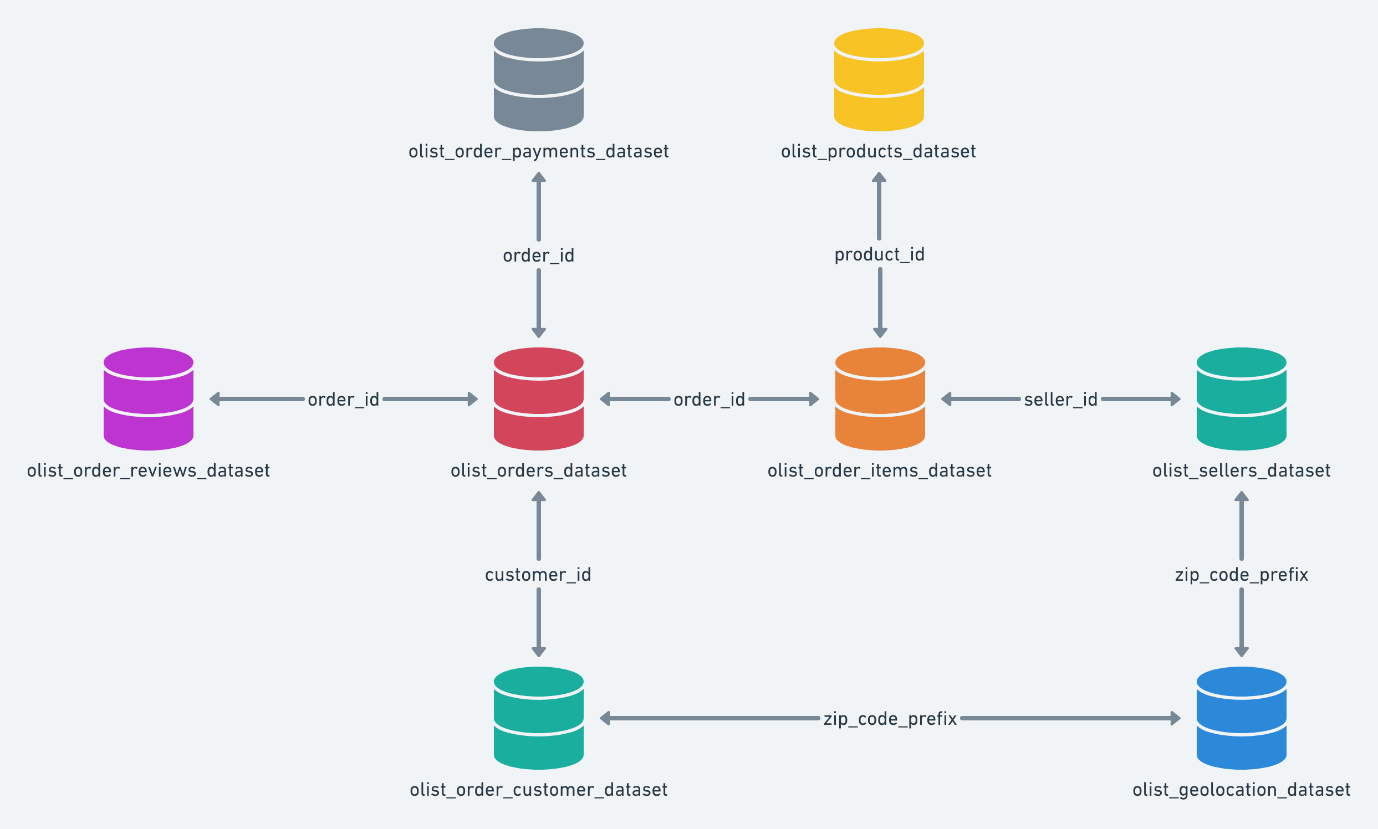
1. An order might have multiple items.
2. Each item might be fulfilled by a distinct seller.
3. All text identifying stores and partners where replaced by the names of Game of Thrones great houses.

Example of a product listing on a marketplace



Data Schema

The data is divided in multiple datasets for better understanding and organization. Please refer to the following data schema when working with it:



# Customers Dataset

This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the orders delivery location.

At our system each order is assigned to a unique customerid. This means that the same customer will get different ids for different orders. The purpose of having a customerunique\_id on the dataset is to allow you to identify customers that made repurchases at the store. Otherwise you would find that each order had a different customer associated with.

# Geolocation Dataset

This dataset has information Brazilian zip codes and its lat/lng coordinates. Use it to plot maps and find distances between sellers and customers.

# Order Items Dataset

This dataset includes data about the items purchased within each order.  
  
**Example:**

The order\_id = 00143d0f86d6fbd9f9b38ab440ac16f5 has 3 items (same product). Each item has the freight calculated accordingly to its measures and weight. To get the total freight value for each order you just have to sum.

**The total order\_item value is:** 21.33 \* 3 = 63.99

**The total freight value is:** 15.10 \* 3 = 45.30

**The total order value (product + freight) is:** 45.30 + 63.99 = 109.29

# Payments Dataset

This dataset includes data about the orders payment options.

# Order Reviews Dataset

This dataset includes data about the reviews made by the customers.

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 for the purchase experience and write down some comments.

# Order Dataset

This is the core dataset. From each order you might find all other information.

# Products Dataset

This dataset includes data about the products sold by Olist.

# Sellers Dataset

This dataset includes data about the sellers that fulfilled orders made at Olist. Use it to find the seller location and to identify which seller fulfilled each product.

# Category Name Translation

Translates the productcategoryname to english.

**Data Preparation**

1. Prepare a service in java that reads these input files and prepares the insert statement dynamically and store in MySQL database. Each of these files should be maintained in separate tables.

**Note:** Work on designing the tables with proper constraints and data types first.

1. These MySQL tables should be used as the source for further processing.
2. Write Sqoop scripts to bring the data from MySQL tables in HDFS.

**Schema Design**

1. Design hive tables for each of the input source.

**Data Processing**

**(Each of the below processing should yield a o/p file)**

1. List customer-wise product orders which is delivered.
2. Generate the list of orders which performed badly. Review score should be considered as the performance indicator of a product.
3. Generate list of three customers for each state who have ordered between 01-01-2017 to 31-05-2018, order status is delivered and order value is greater than 100.
4. Analytics team wants to analyse the contribution of different types of payments made across the platform. Generate the data that explains this contribution.
5. Analyse product wise top 5 total order value (please refer to above description for calculating the total order value)
6. Analyse the top-rated sellers for all the different states (A top rated seller would be the one who has fulfilled max order withing a given region).
7. Give insights on the most demanding products in different states.

**Preparation for Reporting**

1. These output files must be maintained in hbase tables for further reporting.

**Additional Info**

1. All these tasks should be part of a single oozie pipeline.