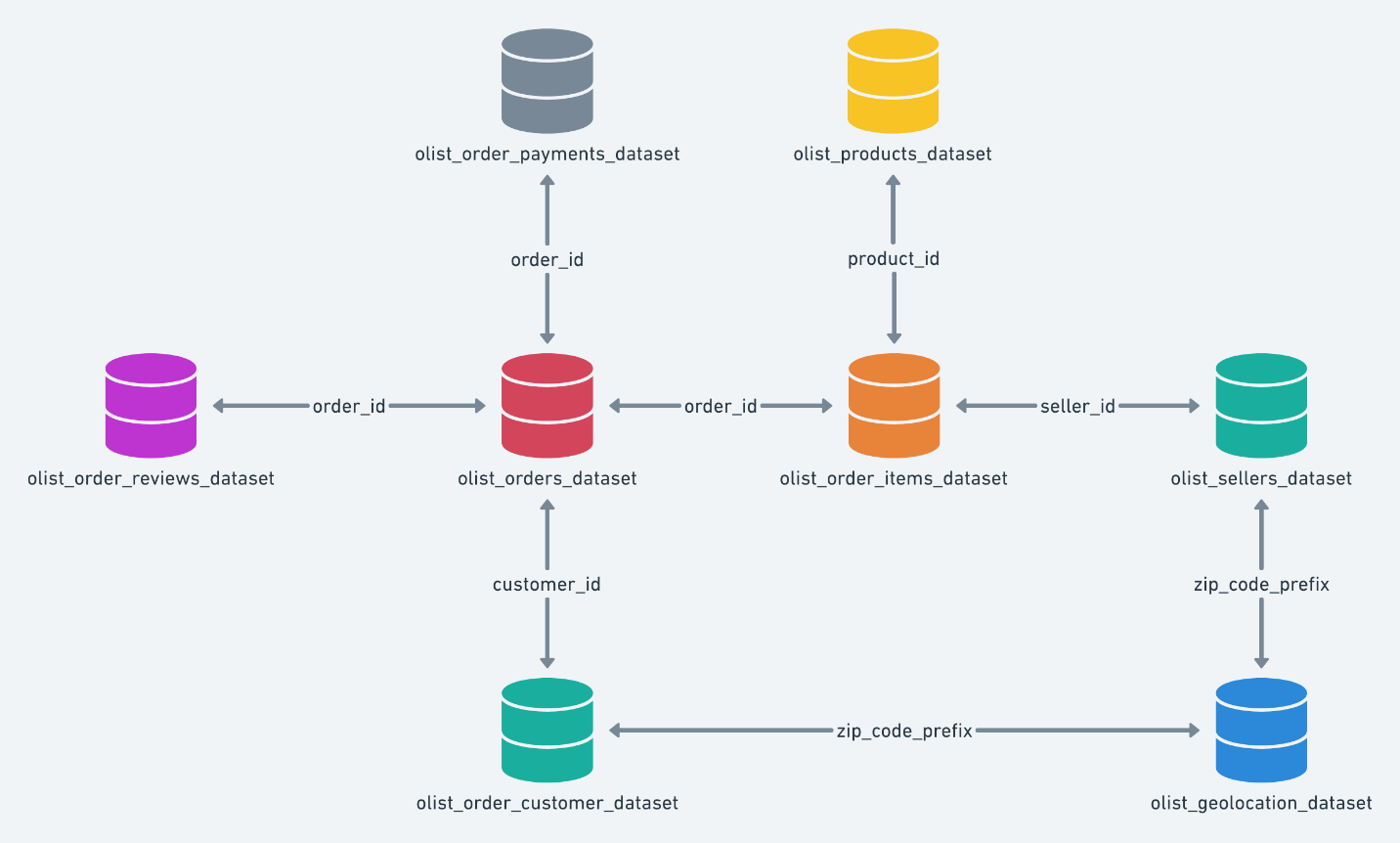
***Giới thiệu về Dataset***

# Bộ dữ liệu công khai thương mại điện tử Brazil của Olist

Đây là tập dữ liệu công khai thương mại điện tử của Brazil về các đơn hàng được thực hiện tại [Olist Store](http://www.olist.com/" \t "_blank) . Tập dữ liệu có thông tin về 100 nghìn đơn hàng từ năm 2016 đến năm 2018 được thực hiện tại nhiều thị trường ở Brazil. Bộ dữ liệu này cho phép xem đơn hàng từ nhiều góc độ : từ trạng thái đơn hàng, giá cả, thanh toán và hiệu suất vận chuyển đến vị trí khách hàng, thuộc tính sản phẩm và cuối cùng là các đánh giá do khách hàng viết. Ngoài ra, đi kèm đó là một tập dữ liệu định vị địa lý liên kết mã bưu chính của Brazil với tọa độ vĩ độ/kinh độ.

# Lược đồ quan hệ dữ liệu

****

\*Tham khảo thêm ở file “ecom.drawio”

# Database “olist\_customers\_dataset.csv”

Bộ dữ liệu này có thông tin về khách hàng và vị trí của họ. Sử dụng nó để xác định những khách hàng duy nhất trong bộ dữ liệu đơn hàng và tìm vị trí giao hàng.

Trong hệ thống của chúng tôi, mỗi đơn hàng được gán cho một customer\_id duy nhất. Điều này có nghĩa là cùng một khách hàng sẽ nhận được các id khác nhau cho các đơn hàng khác nhau. Mục đích của việc có một customer\_unique\_id trên tập dữ liệu là cho phép bạn xác định những khách hàng đã mua lại tại cửa hàng. Nếu không, bạn sẽ thấy rằng mỗi đơn hàng có một khách hàng khác nhau được liên kết.

Gồm 5 cột :

* **customer\_id (khóa chính):** Khóa chính của bộ dữ liệu đơn hàng. Mỗi đơn hàng chứa 1 *customer\_id* duy nhất. Tổng cộng chứa 99441 giá trị.
* **customer\_unique\_id:** định danh của khách hàng khi mua hàng, chứa 96096 giá trị tương ứng với 96096 khách hàng từng mua hàng tại cửa hàng.
* **customer\_zip\_code\_prefix:** 5 giá trị đầu tiên của mã bưu chính nơi mua hàng.
* **customer\_city:** Tên thành phố nơi mua hàng (ví dụ : Rio de Janeiro, Sao Paulo, ...).
* **customer\_state:** kí hiệu viết tắt của *customer\_city*(Rio de Janeiro -> RJ, Sao Paulo -> SP, ...)

# Database “olist\_geolocation\_dataset.csv”

Bộ dữ liệu này có thông tin về mã bưu chính của Brazil và tọa độ vĩ độ/kinh độ của nó. Sử dụng để vẽ bản đồ và tìm khoảng cách giữa người bán và khách hàng.

Gồm 5 thông tin :

* **geolocation\_zip\_code\_prefix:** là 5 giá trị đầu tiên trong mã bưu chính.
* **geolocation\_lat , geolocation\_ing :** vĩ độ, kinh độ.
* **geolocation\_city :** tên thành phố.
* **Geolocation\_state :** viết tắt của *geolocation\_city.*

# Database “olist\_order\_items\_dataset.csv”

Bộ dữ liệu này bao gồm dữ liệu về các mặt hàng được mua trong mỗi đơn hàng.

Ví dụ:

Với order\_id là **“00143d0f86d6fbd9f9b38ab440ac16f5”** có 3 mặt hàng (cùng một sản phẩm). Mỗi mặt hàng có cước phí được tính theo kích thước và trọng lượng của mặt hàng đó. Để có tổng giá trị cước phí cho mỗi đơn hàng, bạn chỉ cần cộng lại.

* Tổng giá trị order\_item là: 21.33 \* 3 = 63.99
* Tổng giá trị vận chuyển là: 15.10 \* 3 = 45.30
* Tổng giá trị đơn hàng (sản phẩm + cước phí) là: 45.30 + 63.99 = 109.29

Gồm 7 trường thông tin :

* **Order\_id** : id tương ứng với mỗi đơn hàng, gồm 98666 giá trị tương ứng với 98666 đơn hàng đã được mua.
* **Order\_item\_id :** số thứ tự xác định số lượng các mục có trong cùng một đơn hàng.
* **Product\_id :** id tương ứng với mỗi sản phẩm, chứa 32951 giá trị tương ứng với mỗi sản phẩm.
* **Seller\_id :** id tương ứng mỗi người bán, chứa 108.632 người.
* **Shipping\_limit\_date:** Hiển thị ngày đến hạn giao hàng của người bán để xử lý đơn hàng cho dịch vụ logistic, trải dài trong giai đoạn từ 19/9/2016 đến 10/4/2020.
* **Price :** giá sản phẩm.
* **Freight\_value :** cước phí vận chuyển.

# Database “olist\_order\_payments\_dataset.csv”

Bộ dữ liệu này bao gồm dữ liệu về các tùy chọn thanh toán đơn hang, gồm 5 cột thông tin :

* **Order\_id** : chứa 99440 id đơn hàng tương ứng các đơn hàng đã thanh toán.
* **Payment\_sequential** : ví dụ, khách hàng có thể thanh toán một đơn hàng bằng nhiều phương thức thanh toán. Nếu khách hàng làm như vậy, một chuỗi sẽ được tạo để đáp ứng tất cả các khoản thanh toán.
* **Payment\_type** : các phương thức thanh toán.
* **Payment\_installments** : số kỳ trả góp do khách hàng lựa chọn.
* **Payment\_value** : giá trị thanh toán của đơn hàng.

# Database “olist\_order\_reviews\_dataset.csv”

Bộ dữ liệu này bao gồm dữ liệu về các đánh giá của khách hàng.

Sau khi khách hàng mua sản phẩm từ Olist Store, người bán sẽ được thông báo để thực hiện đơn hàng đó. Khi khách hàng nhận được sản phẩm hoặc ngày giao hàng dự kiến ​​đến, khách hàng sẽ nhận được khảo sát mức độ hài lòng qua email, trong đó khách hàng có thể ghi chú về trải nghiệm mua hàng và viết một số nhận xét.

Gồm 7 trường thông tin :

* **Reviews\_id :** id tương ứng mỗi đánh giá của khách hàng về đơn hàng, gồm 98410 đánh giá.
* **Order\_id :** id của đơn hàng được đánh giá.
* **Review\_score :** mức độ đánh giá của khách hàng tính theo thang từ 1 đến 5.
* **Review\_comment\_title :** tiêu đề bình luận của đánh giá từ khách hàng, viết bằng ngôn ngữ là tiếng Bồ Đào Nha.
* **review\_comment\_message :** đánh giá của khách hàng.
* **Review\_creation\_date :** ngày viết đánh giá.
* **Review\_answer\_timestamp :** ngày phản hồi đánh giá từ người bán.

# Database “olist\_orders\_dataset.csv”

Đây là tập dữ liệu cốt lõi. Từ mỗi đơn hàng, bạn có thể tìm thấy tất cả thông tin khác.

Gồm 8 thông tin :

* Order\_id : mã đơn hàng, gồm 99441 đơn hàng.
* Customer\_id : mã khách hàng tương ứng với đơn hàng.
* Order\_status : trạng thái đơn hàng ( đã vận chuyển, đã nhận hàng,...).
* Order\_purchase\_timestamp : ngày mua hàng của đơn hàng.
* order\_approved\_at : ngày người bán duyệt đơn hàng.
* **order\_delivered\_carrier\_date:** ngày gửi hàng cho đơn vị vận chuyển.
* **order\_delivered\_customer\_date:** ngày nhận hàng của khách hàng.
* **order\_delivered\_carrier\_date:** Hiển thị ngày giao hàng dự kiến ​​(đã được thông báo cho khách hàng tại thời điểm mua hàng).

# Database “olist\_products\_dataset.csv”

Bộ dữ liệu này bao gồm dữ liệu về các sản phẩm được bán bởi Olist, gồm :

* product\_id: 32951 mã sản phẩm.
* Product\_category\_name: tên sản phẩm bằng tiếng Bồ Đào Nha.
* **product\_name\_lenght : độ dài tên sản phẩm.**
* **product\_description\_lenght : đọ dài trích xuất từ giới thiệu của sản phẩm.**
* **product\_photos\_qty :** số lượng ảnh sản phẩm.
* **product\_weight\_g: khối lượng của sản phẩm (đơn vị : gram).**
* **product\_length\_cm: chiều dài sản phẩm (đơn vị : cm).**
* **product\_height\_cm: chiều cao (đơn vị : cm).**
* **product\_width\_cm: chiều rộng (đơn vị : cm).**

# Database “olist\_sellers\_dataset.csv”

Bộ dữ liệu này bao gồm dữ liệu về người bán đã bán hàng tại Olist. Sử dụng để tìm vị trí của người bán và xác định người bán nào đã từng bán.

Gồm 4 thông tin :

* Seller\_id : id người bán
* **seller\_zip\_code\_prefix :** mã bưu chính nơi người bán.
* **Seller\_city:** thành phố
* **Seller\_state:** viết tắt của *seller\_city*

# Database “product\_category\_name\_translation.csv”

Bản dịch sang tiếng Anh của tên các sản phẩm, gồm 2 thông tin là “**product\_category\_name” (tên gốc)** và **“product\_category\_name\_english” (tên sau khi dịch)** chứa 71 giá trị.