

# Laboratoire de Business Intelligence : Olist

## 1) Données publiques du e-commerce brésilien Olist :

Voici un ensemble de données publiques issu des commandes passées sur Olist Store. L'ensemble de données contient des informations sur 100 000 commandes de 2016 à 2018. Le jeu de données renseigne, entre autres : l'état de la commande, le prix, les moyens de paiement et de transport, l'emplacement du client, les attributs du produit et enfin les avis rédigés par les clients. Est également fourni, un ensemble de données de géolocalisation qui relie les codes postaux brésiliens à des coordonnées lat / lng.

Ce sont de vraies données commerciales, elles ont été anonymisées en ce qui concerne les clients et les produits. En outre, les références aux entreprises et partenaires dans les textes des évaluations ont été remplacées par les noms des grandes maisons de Game of Thrones.

## 2) Contexte :

Cet ensemble de données a été fourni par Olist, le plus grand des e-marketplaces brésiliens. Olist permet de connecter les petites entreprises de tout le Brésil à son réseau de distribution. Les marchands peuvent vendre leurs produits via la boutique Olist et les expédier directement aux clients en utilisant les partenaires logistiques Olist.

Une fois qu'un client a acheté un produit sur le Olist Store, le vendeur est notifié pour exécuter cette commande. Dès que le client a reçu le produit ou que la date de livraison prévue de livraison est expirée, le client reçoit une enquête de satisfaction par e-mail où il peut donner une note relative à son expérience d'achat et écrire quelques commentaires.

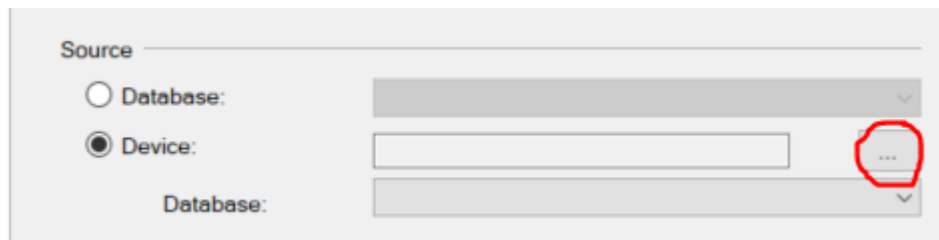
Plus d'infos: <https://olist.com/>

## 3) Ressources :

- Les données que vous allez utiliser sont issues du lien suivant : <https://www.kaggle.com/olistbr/brazilian-ecommerce>. Il s'agit de 9 fichiers csv qui suivent le schéma présenté au point suivant (cf. schéma des données). Cependant, pour vous faciliter la tâche, ces fichiers csv ont déjà été transformés en tables de base de données. **Ces tables constitueront votre *staging* à partir duquel vous construirez votre *datawarehouse*.** Ces tables suivent exactement le même schéma que ceux des fichiers csv. Le *staging* vous aura été fourni par le formateur. Voici la procédure à suivre pour restaurer son backup :
  - Téléchargez le fichier « Olist\_Staging.bak »
  - Déplacez ce fichier à l'emplacement qui suit (ou qui y ressemble) : C:\Program Files\Microsoft SQLServer\MSSQL15.MSSQLSERVER\MSSQL\Backup

## Laboratoire de Business Intelligence : Olist

- Ouvrez SSMS, connectez-vous à votre serveur et cliquez droit sur : Databases > Restore Database...
- Une fenêtre s'ouvre, sélectionnez « Device » (Support) et cliquez sur les ... :



- Une autre fenêtre s'ouvre. Cliquez sur « Add »
- Une nouvelle fenêtre apparaît. Sélectionnez le fichier « Olist\_Staging.bak » à l'emplacement où vous venez de déplacer votre fichier :

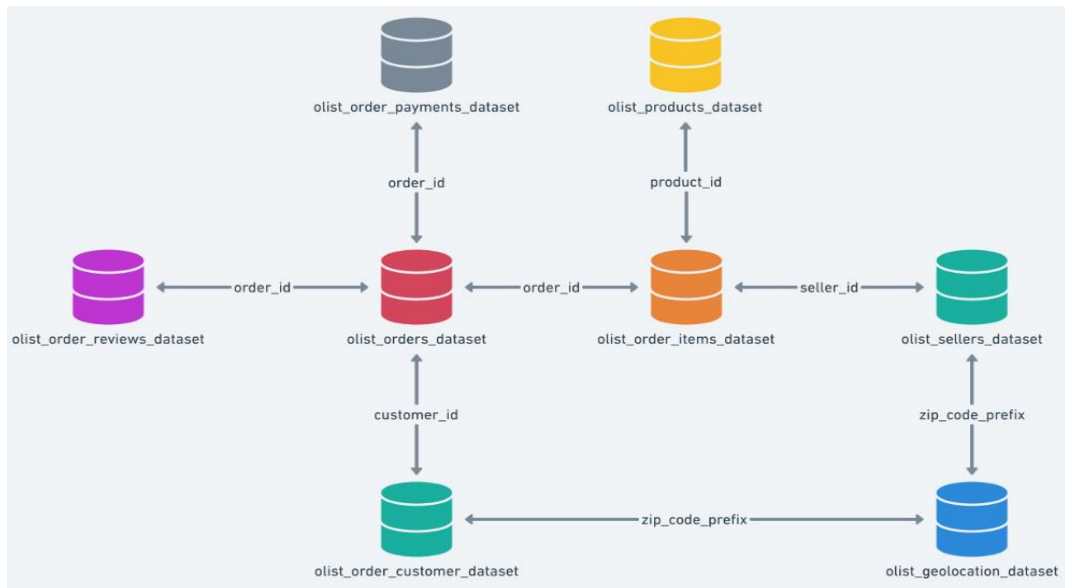


- Cliquez sur OK pour fermer toutes les fenêtres : votre *staging* est à présent restaurer et prêt à être utiliser.
- Comme autre ressource, vous avez également le script Sql de deux tables de dimension : celle de Date (D\_Date) et celle de Time (D\_Time). Vous pouvez retrouver ces fichiers sql sous le nom de « Create\_Populate\_DateDimension » et « Create\_Populate\_TimeDimension » dans le dossier que l'on vous aura fourni. Il ne vous restera qu'à les exécuter.
- Tout autre ressource qu'il vous semble utile d'utiliser : le site d'où sont tirées les données (<https://www.kaggle.com/olistbr/brazilian-ecommerce>), le site web d'Olist (<https://olist.com/>), votre cours de BI et de Datawarehousing, des recherches Internet (ex : site Microsoft détaillant les types de données SSIS (<https://docs.microsoft.com/en-us/sql/integration-services/data-flow/integration-services-data-types?view=sql-server-ver15>), le dossier que vous être en train de lire,... et bien sûr votre tête :-)

### 4) Schéma des données :

Les données sont divisées en plusieurs ensembles de données pour une meilleure compréhension et organisation (NB : le dataset Category\_name\_translation ne figure pas sur le schéma) :

# Laboratoire de Business Intelligence : Olist



Voici à quoi ressemble les tables du staging avec lesquelles vous devez travailler :

Order payments
order_id
payment_sequential
payment_type
payment_installments
payment_value

Products
product_id
product_category_name
product_name_length
product_description_length
product_photos_qty
product_weight_g
product_length_cm
product_height_cm
product_width_cm

Category name translation
product_category_name
product_category_name_english

Order reviews
review_id
order_id
review_score
review_comment_title
review_comment_message
review_creation_date
review_answer_timestamp

Orders
order_id
customer_id
order_status
order_purchase_timestamp
order_approved_at
order_delivered_carrier_date
order_delivered_customer_date
order_estimated_delivery_date

Order item
order_id
order_item_id
product_id
seller_id
shipping_limit_date
price
freight_value

Sellers
seller_id
seller_zip_code_prefix
seller_city
seller_state

Customer
customer_id
customer_unique_id
customer_zip_code_prefix
customer_city
customer_state

Geolocalisation
geolocation_zip_code_prefix
geolocation_lat
geolocation_lng
geolocation_city
geolocation_state

### 5) Métadonnées des fichiers :

Toutes les informations (description des datasets, des colonnes, ...) se trouvent sur : <https://www.kaggle.com/olistbr/brazilian-ecommerce>. La plupart d'entre-elles ont néanmoins été rassemblées dans les pages qui suivent :

- **Customer :**

Cet ensemble de données contient des informations sur les clients et leur localisation. Utilisez-le pour identifier des clients particuliers dans l'ensemble de données des commandes et pour trouver le lieu de livraison des commandes.

*customer\_id*

clé de l'ensemble des données relatives aux commandes. Chaque commande a un numéro de client différent.

*customer\_unique\_id*

l'identifiant unique d'un client. (id propre à chaque client et qui est donc le même s'il passe plusieurs fois commandes)

*customer\_zip\_code\_prefix*

les cinq premiers chiffres du code postal (zip code) du client

*customer\_city*

nom de la ville du client

*customer\_state*

état du client

Dans notre système, chaque commande est attribuée à *customer\_id* différent. Cela signifie qu'un même client obtiendra des identifiants différents pour des commandes différentes.

Exemple.:

Si l'on prend 2 *customer\_id* (1afe8a9c67eec3516c09a8bdcc539090 et 24b0e2bd287e47d54d193e7bbb51103f), on voit bien qu'ils sont assignés à 2 commandes différentes (*order\_id*).

## Laboratoire de Business Intelligence : Olist

```
1 /***** Script for SelectTopNRows command from SSMS *****/
2 SELECT TOP (1000) [order_id]
3     , [customer_id]
4     , [order_status]
5     , [order_purchase_timestamp]
6     , [order_approved_at]
7     , [order_delivered_carrier_date]
8     , [order_delivered_customer_date]
9     , [order_estimated_delivery_date]
10 FROM [Olist_Staging].[dbo].[Orders]
11 WHERE customer_id IN('1afe8a9c67eec3516c09a8bdcc539090', '24b0e2bd287e47d54d193e7bbb51103f')
```

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
1	bb874c45df1a3c97842d52f31efee99a	delivered	2018-07-28 00:23:49	2018-07-28 00:35:19	2018-07-31		
2	c306eca42d32507b970739b5b6a5a33a	anceled	2018-08-13 09:14:07				

L'objectif d'avoir un `customer_unique_id` sur l'ensemble des données est de vous permettre d'identifier les clients qui ont effectué plusieurs commandes. Si ce n'était pas le cas, vous constateriez que chaque commande est associée à un client différent.

### Exemple :

Si on regarde les `customer_unique_id`, on remarque qu'ils sont ici tous les deux égaux ('00172711b30d52eea8b313a7f2cced02'). Cela signifie que les deux commandes ont été passées par le même client.

customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
1afe8a9c67eec3516c09a8bdcc539090	00172711b30d52eea8b313a7f2cced02	45200	jequie	BA
24b0e2bd287e47d54d193e7bbb51103f	00172711b30d52eea8b313a7f2cced02	45200	jequie	BA

- **Geolocation:**

Cet ensemble de données contient des informations sur les codes postaux brésiliens et leurs coordonnées lat/long. Utilisez-le pour tracer des cartes et trouver les distances entre les vendeurs et les clients.

`geolocation_zip_code_prefix`

les cinq premiers chiffres du code postal (zip code) du client

`geolocation_lat`

latitude

`geolocation_lng`

longitude

`geolocation_city`

city name

`geolocation_state`

state

## Laboratoire de Business Intelligence : Olist

- **Order\_items:**

Cet ensemble de données comprend des données sur les articles achetés dans le cadre de chaque commande.

Exemple:

La commande dont l'order\_id = 00143d0f86d6fbd9f9b38ab440ac16f5 contient 3 articles (du même produit). Pour chaque article, le fret (càd le prix du transport des marchandises) est calculé en fonction de ses mesures et de son poids. Pour obtenir la valeur totale du fret pour chaque commande, il suffit de sommer le fret de chaque article (ou de multiplier le fret par le nombre d'articles) :

La valeur totale du fret est de : 15,10 (freight\_value) \* 3 (nb d'articles) = 45,30

La valeur totale de l'article de la commande est : 21,33 (price) \* 3 (nb d'articles) = 63,99

⇒ La valeur totale de la commande (produit + fret) est : 45,30 + 63,99 = 109,29

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
1	00143d0f86d6fbd9f9b38ab440ac16f5	1	e95ee6822b66ac6058e2e4aff656071a	a17f621c590ea0fab3d5d883e1630ec6	2017-10-20 16:07:52.000	21.33	15.10
2	00143d0f86d6fbd9f9b38ab440ac16f5	2	e95ee6822b66ac6058e2e4aff656071a	a17f621c590ea0fab3d5d883e1630ec6	2017-10-20 16:07:52.000	21.33	15.10
3	00143d0f86d6fbd9f9b38ab440ac16f5	3	e95ee6822b66ac6058e2e4aff656071a	a17f621c590ea0fab3d5d883e1630ec6	2017-10-20 16:07:52.000	21.33	15.10

*order\_id*

identifiant unique de la commande

*order\_item\_id*

numéro séquentiel identifiant le nombre d'articles inclus dans une commande. Ces numéros représentent l'ordre de chaque article dans une commande. (ex : 1 = article n°1 de la commande ; 2 = article n°2 de la commande ; etc.)

*product\_id*

identifiant unique du produit

*seller\_id*

identifiant unique du vendeur

*shipping\_limit\_date*

indique la date limite d'expédition (shipping limit date) du vendeur pour le traitement de la commande au partenaire logistique.

*price*

prix de l'article

*freight\_value*

la valeur du fret de l'article (si une commande comporte plus d'un article, la valeur du fret est répartie entre les articles)

## Laboratoire de Business Intelligence : Olist

- **Order\_payments:**

Cet ensemble de données comprend des données sur les options de paiement des commandes. Un paiement est identifié à la fois par son `order_id` et sa séquence de paiement (*payment\_sequential*).

*order\_id*  
identifiant unique de la commande

*payment\_sequential*  
un client peut payer une commande avec plus d'un mode de paiement. S'il le fait, une séquence sera créée pour permettre tous les paiements.

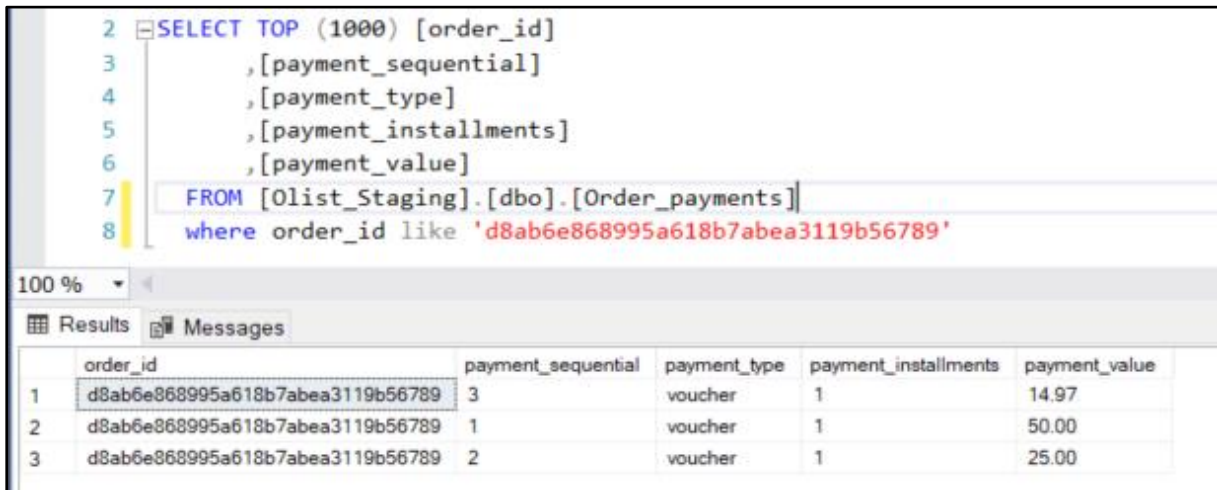
*payment\_type*  
mode de paiement choisi par le client.

*payment\_installments*  
nombre de versements choisis par le client.

*payment\_value*  
valeur de la transaction (pour une séquence de paiement)

### Exemple :

Les 3 *payment\_sequential* pour la commande 'd8ab6e868995a618b7abea3119b56789' indiquent que le client a payé avec plusieurs méthodes de paiement. Attention, plusieurs méthodes de paiement ne signifient pas forcément des méthodes de paiement différentes (voir exemple ci-dessous)



```
2 SELECT TOP (1000) [order_id]
3     ,[payment_sequential]
4     ,[payment_type]
5     ,[payment_installments]
6     ,[payment_value]
7 FROM [Olist_Staging].[dbo].[Order_payments]
8 where order_id like 'd8ab6e868995a618b7abea3119b56789'
```

	order_id	payment_sequential	payment_type	payment_installments	payment_value
1	d8ab6e868995a618b7abea3119b56789	3	voucher	1	14.97
2	d8ab6e868995a618b7abea3119b56789	1	voucher	1	50.00
3	d8ab6e868995a618b7abea3119b56789	2	voucher	1	25.00

On voit que le total des 3 séquences de paiement est égal à  $14,97 + 50 + 25 = 89,97$  ce qui correspond bien à  $71,90 + 18,07 = 89,97$  c'est-à-dire au total de la commande (ou au total de tous les items de la commande).

## Laboratoire de Business Intelligence : Olist

```
2 SELECT TOP (1000) [order_id]
3     ,[order_item_id]
4     ,[product_id]
5     ,[seller_id]
6     ,[shipping_limit_date]
7     ,[price]
8     ,[freight_value]
9 FROM [Olist Staging].[dbo].[Order_item]
10 where order_id like 'd8ab6e868995a618b7abea3119b56789'
```

100 %

Results Messages

order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value
d8ab6e868995a618b7abea3119b56789	1	58f88ebb71c90b2d46a5b297ae6c3455	01fdafa7697d26ad920e9e0346d4bd1b	2017-07-17 10:10:13	71.90	18.07

- **Order\_reviews:**

Ce dataset comprend des données sur les avis des clients.

Lorsqu'un client achète un produit chez Olist Store, un vendeur est informé de l'exécution de cette commande. Une fois que le client reçoit le produit, ou que la date de livraison prévue est due, il reçoit une enquête de satisfaction par courrier électronique où il peut donner une note pour l'expérience d'achat et noter quelques commentaires.

*review\_id*

identifiant unique de la review (avis)

*order\_id*

identifiant unique de la commande

*review\_score*

note allant de 1 à 5 et qui est donnée par le client lors d'une enquête de satisfaction.

*review\_comment\_title*

Titre du commentaire de l'avis laissé par le client, en portugais.

*review\_comment\_message*

Message de commentaires de l'avis laissé par le client, en portugais.

*review\_creation\_date*

Indique la date à laquelle l'enquête de satisfaction a été envoyée au client.

*review\_answer\_timestamp*

Affiche l'horodatage (timestamp) des réponses aux enquêtes de satisfaction.

- **Orders:**

Il s'agit du dataset de base. A partir de chaque commande, vous pouvez trouver toutes les autres informations.

*order\_id*

identifiant unique de la commande.



## Laboratoire de Business Intelligence : Olist

*customer\_id*

clé de l'ensemble des données sur les clients. Chaque commande a un numéro de client différent.

*order\_status*

Référence à l'état de la commande (livrée, expédiée, etc.).

*order\_purchase\_timestamp*

Indique l'horodatage de l'achat.

*order\_approved\_at*

Indique l'horodatage de l'approbation de l'achat.

*order\_delivered\_carrier\_date*

Affiche l'horodatage de la commande qui indique quand elle a été transmise au partenaire logistique.

*order\_delivered\_customer\_date*

Indique au client la date réelle de livraison de la commande.

*order\_estimated\_delivery\_date*

Indique la date de livraison estimée qui a été communiquée au client au moment de l'achat.

- **Products :**

Ce dataset comprend des données sur les produits vendus par Olist.

*product\_id*

identifiant unique du produit.

*product\_category\_name*

catégorie principale du produit, en portugais.

*product\_name\_lenght*

nombre de caractères extraits du nom du produit.

*product\_description\_lenght*

nombre de caractères extraits de la description du produit.

*product\_photos\_qty*

nombre de photos du produit publiées

*product\_weight\_g*

le poids du produit mesuré en grammes.

*product\_length\_cm*

la longueur du produit mesurée en centimètres.

*product\_height\_cm*

## Laboratoire de Business Intelligence : Olist

la hauteur du produit mesurée en centimètres.

*product\_width\_cm*

product width measured in centimeters.

- **Sellers :**

Ce dataset comprend des données sur les vendeurs qui ont exécuté les commandes passées à Olist. Utilisez-les pour trouver l'emplacement du vendeur et pour identifier quel vendeur a exécuté chaque produit.

*seller\_id*

identifiant unique du vendeur

*seller\_zip\_code\_prefix*

5 premiers chiffres du code postal du vendeur

*seller\_city*

nom de la ville du vendeur

*seller\_state*

état du vendeur

- **Product\_category\_name\_translation :**

Ce dataset traduit le nom de la catégorie de produit en anglais

*product\_category\_name*

nom de la catégorie de produit

*product\_category\_name\_english*

nom de la catégorie en anglais

## 6) Données supplémentaires (facultatives)

- Un Excel nommé « Infos états brésiliens » qui contient notamment la traduction des abréviations en noms complets (ex : AM = Amazonas)

	A	B	C	D	E	
1	État	Abréviation	Capitale	Superficie (km²)	Population (2014)	
2	Acre	AC	Rio Branco	1525814	795145	
3	Alagoas	AL	Maceió	277677	3327551	
4	Amapá	AP	Macapá	1428146	756500	
5	Amazonas	AM	Manaus	15707457	3893763	
6	Bahia	BA	Salvador	5646927	15150143	
7	Ceará	CE	Fortaleza	1488256	8867448	
8	District fédéré	DF	Brasília	58221	2867869	
9	Espírito Santo	ES	Vitória	460775	3894899	
10	Goias	GO	Goiânia	3408677	6554333	

## Laboratoire de Business Intelligence : Olist

- Un CSV nommé «taux de change\_BRL-EUR\_2016-2018 » qui contient les taux de change du réal brésilien (R\$) en euro (€) entre le 01/01/2016 et le 31/12/2018.

### 7) Tâche :

- Il vous est demandé de **transformer ce modèle relationnel en un modèle dimensionnel**, en vue de répondre aux questions business suivantes (liste non exhaustive).
  - Montants/Quantités par catégorie de produits
  - Répartition des ventes par type de produits
  - Identification des meilleurs clients
  - Identification des fournisseurs les plus importants
  - Identification des produits les plus vendus
  - Evolution des ventes selon le temps (Jour, Mois, Année, etc.)
  - Localisation des clients et revendeurs
  - Produits les mieux évalués
  - Revendeurs les mieux cotés
  - Ratio du nombre de commandes livrées à temps/en retard
  - Types de paiement les plus utilisés
  - ...
- Après transformation en modèle dimensionnel, il vous est demandé d'exploiter votre nouveau modèle et ses données afin d'**élaborer des rapports permettant de répondre aux questions qui sont posées** (voire d'autres). Ces rapports peuvent être réalisés à l'aide d'outils tels que Power BI, Qlik Sense et Tableau.

### 8) Recommandations :

- Afin de vous aider à vous situer dans votre avancement, voici la durée que devrait prendre approximativement chacune des étapes du labo (6-7 jours au total) :
  - 1 jour : lancement + appréhender le labo et les données. Attention à ne pas vouloir se lancer trop vite dans les étapes suivantes. Veuillez tout d'abord à bien comprendre le cas d'étude.
  - 1 jour : réflexion autour de la conception de votre DW + schéma du modèle dimensionnel (<https://app.diagrams.net/>)
  - 2 à 3 jours : réalisation du datawarehouse avec l'outil d'ETL de votre choix

## Laboratoire de Business Intelligence : Olist

- 1 à 2 jours : reporting avec l'outil de votre choix (attention à ne pas sous-estimer cette étape)

NB : il s'agit là d'une indication, libre à vous de gérer ça comme vous le souhaitez.

- Dans un premier temps, il est sans doute préférable de voir « petit » et de vous concentrer sur un fait d'analyse bien précis et ensuite, de réfléchir à comment vous voudriez l'analyser. Par exemple, **commencez par vous concentrer sur les ventes**.
- Il est possible que vous n'ayez pas tous exactement les mêmes schémas dimensionnels, tout dépend de ce que vous voulez analyser (certains préféreront se concentrer sur certains aspects plutôt que d'autres). Néanmoins, vous devriez tous avoir un schéma dimensionnel qui présente une structure assez similaire.
- Si vous vous sentez perdu... pas de panique ! Demandez à votre formateur quelques indications afin de vous remettre sur la bonne voie 😊
- Ce laboratoire est un travail individuel. Néanmoins, vous pouvez vous mettre en groupe pour discuter et vous entraider (plusieurs têtes valent mieux qu'une).

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