Olist Ecommerce Datawarehouse ETL

Problem Setting

In the past few years, we have seen a rapid rise in e-commerce and digital marketing. Multiple factors are involved in ensuring the smooth running of these business models e.g., sellers, delivery agents, product stock etc. It is important to understand the trends in these factors to ensure maximum profits and minimize losses. Analyzing these trends also helps forecast product demand in the future, which will help with efficient planning.

Problem Definition

Olist is a departmental store that operates in the e-commerce industry with its headquarters in Brazil. The company acts as intermediary between small businesses and customers. The businesses can sell their products through Olist Store after acquiring a contract. They are required to fulfil an order when a customer purchases their products. The Store keeps track of customer satisfaction through a satisfaction survey by email after every purchase. From a market funneling perspective, a seller is contacted by a sales development representative for a consultancy upon signing up at the Olist seller page. If the deal is closed after consultancy, a lead becomes a seller and can then sell products on Olist. It is also helpful for olist to understand the potential customer base they could reach depending on the population of various areas.

Project Goal

- Analyse product sales made over the years and the trend in sales
- Understand Customer behavior when making purchases
- Analyse customer satisfaction through reviews
- Check for trends in purchases and payments over the years
- Analyse closed deals with respect to sellers contacting Olist
- Analyse the time taken for a deal to be closed from the time of first contact solicitation by a seller
- Analyse the effect of population in an area with the total sales

Data Sources

The data sources are as follows:

1) Orders made at Olist: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=product_category_name_translation.csv

2) Marketing funnel of sellers at Olist: https://www.kaggle.com/datasets/olistbr/marketing-funnel-olist?select=olist_marketing_qualified_leads_dataset.csv

3) Population by zipcode in Brazil: https://postal-codes.cybo.com/brazil/?p=4

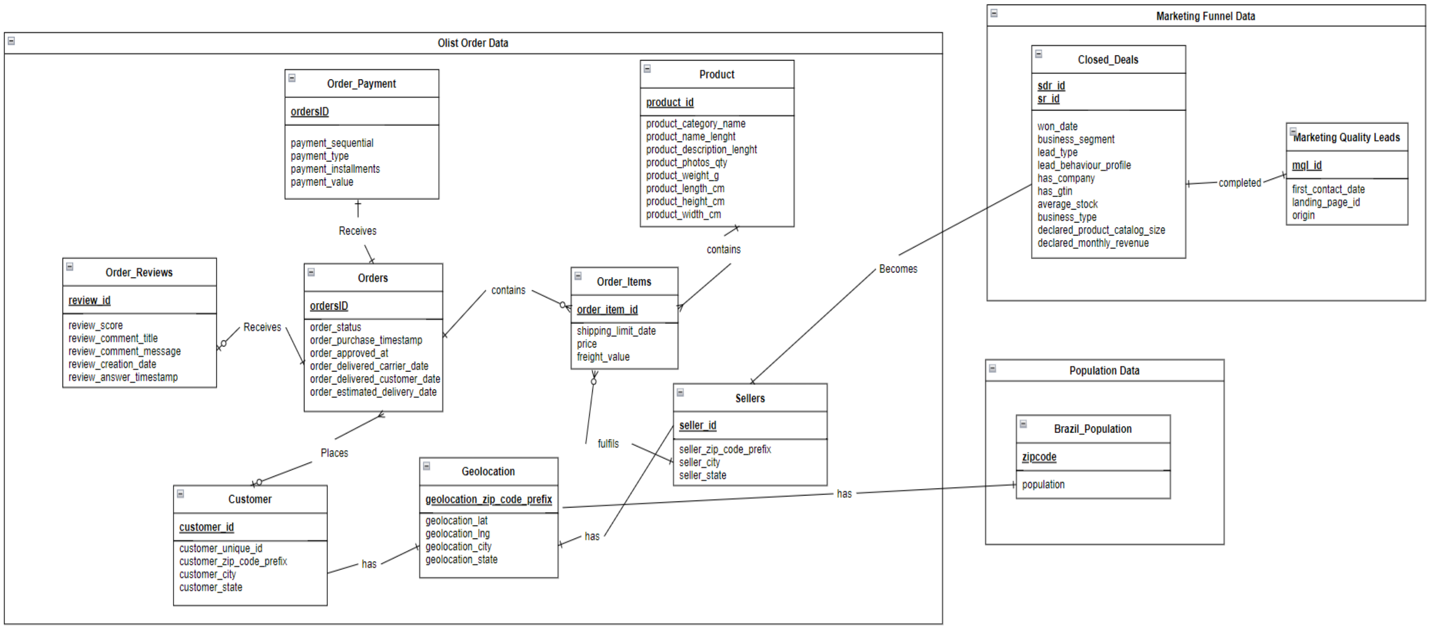
Data Description

The Order data is about orders made at Olist Store. It comprises of 100k orders made from the year 2016 to 2018. The data contains 9 datasets. The datasets are namely: customer, geolocation, products, sellers, orders, order_items, order_payments, order_reviews, product_category_name.

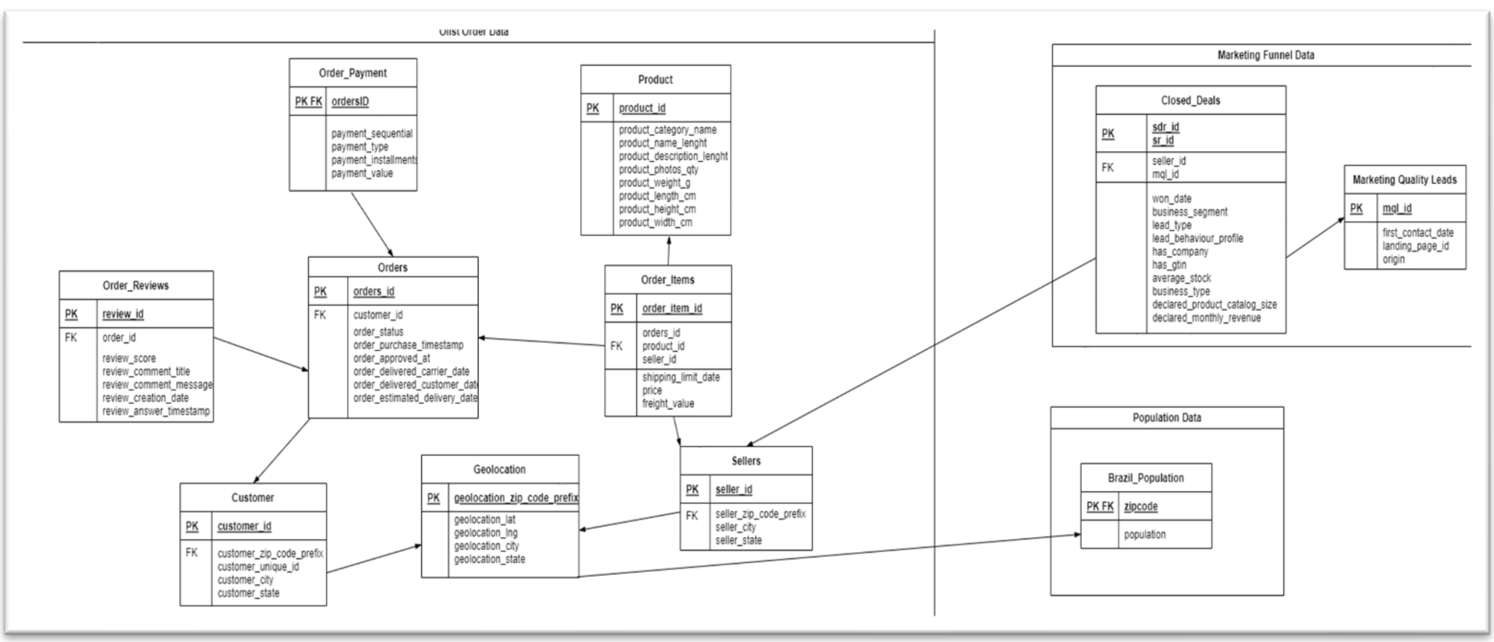
The Marketing Funnel data is about sellers who contacted Olist to sell their products. It comprises of 8k sellers between the time period Jun. 1st 2017 and Jun 1st 2018. It contains two datasets: closed deals and marketing qualified leads. It can be joined to the Order data through seller_id.

The Population dataset contains information about the population in different regions of Brazil based on the zip code. It can be joined to the Order, customer and seller data through zipcode .

ER Diagram



Relational Diagram:



Data Population:

Loaded the data into Postgres Database The following are the tables:

Query   Query History   Scratch Pad ×

```sql
SELECT * FROM olist_customer limit 10;
```

Data Output   Messages   Notifications

	customer_id character varying (80)	customer_unique_id character varying (80)	customer_zip_code integer	customer_city character varying (50)	customer_state character varying (50)
1	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
2	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333...	9790	sao bernardo do campo	SP
3	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2...	1151	sao paulo	SP
4	b2b6027bc5c5109e529d4dc6358b12...	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
5	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d...	13056	campinas	SP
6	879864dab9bc3047522c92c82e1212...	4c93744516667ad3b8f1fb645a3116...	89254	jaragua do sul	SC
7	fd826e7cf63160e536e0908c76c3f441	addec96d2e059c80c30fe6871d30d1...	4534	sao paulo	SP
8	5e274e7a0c3809e14aba7ad5aae0d4...	57b2a98a409812fe9618067b6b8ebe...	35182	timoteo	MG
9	5adf08e34b2e993982a47070956c5c65	1175e95fb47ddff9de6b2b06188f7e0d	81560	curitiba	PR
10	4b7139f34592b3a31687243a302fa75b	9afe194fb833f79e300e37e580171f22	30575	belo horizonte	MG

```sql
SELECT * FROM brazil_population limit 10;
```

Data Output   Messages   Notifications

	zip_code character varying (50)	population integer
1	00000-000	[null]
2	00331-000	[null]
3	00750-000	[null]
4	00790-000	[null]
5	00904-000	[null]
6	00991-000	[null]
7	01000-000	[null]
8	01001-000	3276
9	01002-000	2679
10	01003-000	957

Query Query History

```sql
SELECT * FROM olist_sellers limit 10;
```

Data Output Messages Notifications

	seller_id character varying (50)	seller_zipcode_prefix integer	seller_city character varying (50)	seller_state character varying (50)
1	3442f8959a84dea7ee197c632cb2df15	13023	campinas	SP
2	d1b65fc7debc3361ea86b5f14c68d2e2	13844	mogi guacu	SP
3	ce3ad9de960102d0677a81f5d0bb7b...	20031	rio de janeiro	RJ
4	c0f3eea2e14555b6faeea3dd58c1b1c3	4195	sao paulo	SP
5	51a04a8a6bdcb23deccc82b0b80742cf	12914	braganca paulista	SP
6	c240c4061717ac1806ae6ee72be353...	20920	rio de janeiro	RJ
7	e49c26c3edfa46d227d5121a6b6e4d37	55325	brejao	PE
8	1b938a7ec6ac5061a66a3766e0e75f90	16304	penapolis	SP
9	768a86e36ad6aae3d03ee3c6433d61df	1529	sao paulo	SP
10	ccc4bbb5f32a6ab2b7066a4130f114e3	80310	curitiba	PR

Query Query History Scratch Pad ×

```sql
SELECT * FROM olist_products limit 10;
```

Data Output Messages Notifications

	product_id character varying (50)	product_category character varying (50)	product_name_length integer	product_description_length integer	product_photos_qty integer	product_weight_g integer	product_length_cm integer	produ integ
1	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40	287	1	225	16	
2	3aa071139cb16b67ca9e5dea641aaa2f	artes	44	276	1	1000	30	
3	96bd76ec8810374ed1b65e291975717f	esporte_lazer	46	250	1	154	18	
4	cef67bcfe19066a932b7673e239eb23d	bebes	27	261	1	371	26	
5	9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	37	402	4	625	20	
6	41d3672d4792049fa1779bb35283ed13	instrumentos_musicais	60	745	1	200	38	
7	732bd381ad09e530fe0a5f457d81becb	cool_stuff	56	1272	4	18350	70	
8	2548af3e6e77a690cf3eb6368e9ab61e	moveis_decoracao	56	184	2	900	40	
9	37cc742be07708b53a98702e77a21a02	eletrodomesticos	57	163	1	400	27	
10	8c92109888e8cdf9d66dc7e463025574	brinquedos	36	1156	1	600	17	

Query Query History Scratch Pad ×

```sql
SELECT * FROM olist_orders limit 10;
```

Data Output Messages Notifications

	order_id character varying (50)	customer_id character varying (50)	order_status character varying (50)	order_purchase_timestamp timestamp without time zone	order_approved_at timestamp without time zone	order_delivered_carrier_dat timestamp without time zo
1	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00
2	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00
3	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00
4	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	2017-11-18 19:28:06	2017-11-18 19:45:59	2017-11-22 13:39:59
5	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39	2018-02-13 22:20:29	2018-02-14 19:46:34
6	a4591c265e18cb1dcee52889e2d8ac...	503740e9ca751ccdda7ba28e9ab8f608	delivered	2017-07-09 21:57:05	2017-07-09 22:10:13	2017-07-11 14:58:04
7	136cce7faa42fdb2cefd53fdc79a6098	ed0271e0b7da060a393796590e7b737a	invoiced	2017-04-11 12:22:08	2017-04-13 13:25:17	[null]
8	6514b8ad8028c9f2cc2374ded245783f	9bdf08b4b3b52b5526ff42d37d47f222	delivered	2017-05-16 13:10:30	2017-05-16 13:22:11	2017-05-22 10:07:46
9	76c6e866289321a7c93b82b54852dc...	f54a9f0e6b351c431402b8461ea51999	delivered	2017-01-23 18:29:09	2017-01-25 02:50:47	2017-01-26 14:16:31
10	e69bfb5eb88e0ed6a785585b27e16dbf	31ad1d1b63eb9962463f764d4e6e0c9d	delivered	2017-07-29 11:55:02	2017-07-29 12:05:32	2017-08-10 19:45:24

Query Query History Scratch Pad ×

```sql
SELECT * FROM olist_order_reviews limit 10;
```

Data Output Messages Notifications

	review_id character varying (50)	order_id character varying (50)	review_score integer	review_comment_title character varying (300)	review_comment_message character varying (300)
1	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	[null]	[null]
2	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	[null]	[null]
3	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	[null]	[null]
4	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d351...	5	[null]	Recebi bem antes do prazo estipulado.
5	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	[null]	Parabéns lojas lannister adorei comprar pela Internet seguro e prático
6	15197aa66ff4d0650b5434f1b46cda19	b18dcdf73be66366873cd26c5724d1dc	1	[null]	[null]
7	07f9bee5d1b850860defd761afa7ff16	e48aa0d2dcec3a2e87348811bcfdf22b	5	[null]	[null]
8	7c6400515c67679fbee952a7525281ef	c31a859e34e3adac22f376954e19b39d	5	[null]	[null]
9	a3f6f7f6f433de0aefbb97da197c554c	9c214ac970e84273583ab523dfafd09b	5	[null]	[null]
10	8670d52e15e00043ae7de4c01cc2fe06	b9bf720beb4ab3728760088589c62129	4	recomendo	aparelho eficiente. no site a marca do aparelho esta impresso como 3

Query Query History

```sql
SELECT * FROM olist_order_payments limit 10;
```

Data Output Messages Notifications

	order_id character varying (50)	payment_sequential integer	payment_type character varying (50)	payment_installments integer	payment_value numeric
1	b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
2	a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
3	25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71
4	ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78
5	42fdf880ba16b47b59251dd489d4441a	1	credit_card	2	128.45
6	298fcdf1f73eb413e4d26d01b25bc1cd	1	credit_card	2	96.12
7	771ee386b001f06208a7419e4fc1bbd7	1	credit_card	1	81.16
8	3d7239c394a212faae122962df514ac7	1	credit_card	3	51.84
9	1f78449c87a54faf9e96e88ba1491fa9	1	credit_card	6	341.09
10	0573b5e23cbd798006520e1d5b4c6714	1	boleto	1	51.95

```sql
SELECT * FROM olist_order_items limit 10;
```

Data Output Messages Notifications

	order_id character varying (50)	order_item_id integer	product_id character varying (50)	seller_id character varying (50)	shipping_limit_date timestamp without time zone	price numer
1	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e...	48436dade18ac8b2bce089ec2a0412...	2017-09-19 09:45:35	
2	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	2
3	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	1
4	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2...	2018-08-15 10:10:18	
5	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	1
6	00048cc3ae777c65dbb7d2a0634bc1...	1	ef92defde845ab8450f9d70c526ef70f	6426d21aca402a131fc0a5d0960a3c90	2017-05-23 03:55:27	
7	00054e8431b9d7675808bcb819fb4a32	1	8d4f2bb7e93e6710a28f34fa83ee7d28	7040e82f899a04d1b434b795a43b46...	2017-12-14 12:10:31	
8	000576fe39319847cbb9d288c5617fa6	1	557d850972a7d6f792fd18ae1400d9b6	5996cddab893a4652a15592fb58ab8...	2018-07-10 12:30:45	8
9	0005a1a1728c9d785b8e2b08b90457...	1	310ae3c140ff94b03219ad0adc3c778f	a416b6a846a11724393025641d4edd...	2018-03-26 18:31:29	1
10	0005f50442cb953dcd1d21e1fb923495	1	4535b0e1091c278dfd193e5a1d63b39f	ba143b05f0110f0dc71ad71b4466ce92	2018-07-06 14:10:56	

```sql
SELECT * FROM olist_marketing_qualified_leads limit 10;
```

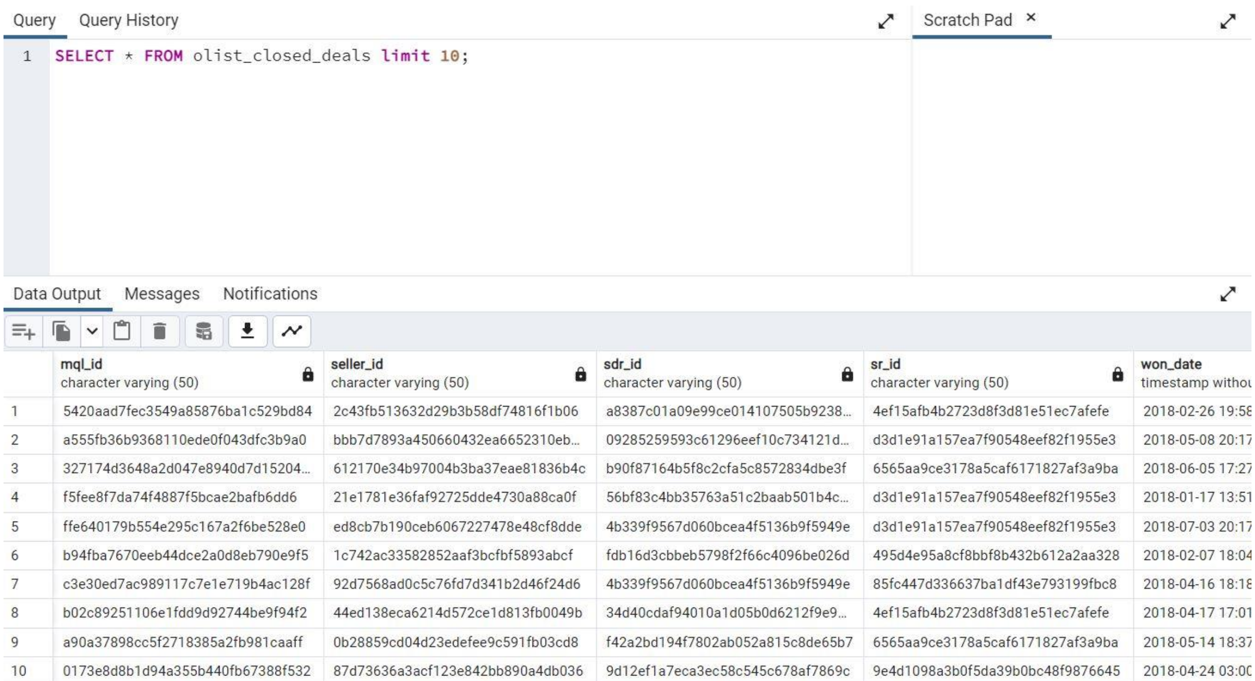
Data Output | Messages | Notifications

	mql_id character varying (50)	first_contact_date timestamp without time zone	landing_page character varying (50)	origin character varying (50)
1	dac32acd4db4c29c230538b72f8dd87d	2018-02-01 00:00:00	88740e65d5d6b056e0cda098e1ea6313	social
2	8c18d1de7f67e60dbd64e3c07d7e9d5d	2017-10-20 00:00:00	007f9098284a86ee80ddeb25d53e0af8	paid_search
3	b4bc852d233dfefc5131f593b538befa	2018-03-22 00:00:00	a7982125ff7aa3b2054c6e44f9d28522	organic_search
4	6be030b81c75970747525b843c1ef4f8	2018-01-22 00:00:00	d45d558f0daeecf3cccdffe3c59684aa	email
5	5420aad7fec3549a85876ba1c529bd84	2018-02-21 00:00:00	b48ec5f3b04e9068441002a19df93c6c	organic_search
6	28bdfd5f057764b54c38770f95c69f2f	2018-01-14 00:00:00	22c29808c4f815213303f8933030604c	organic_search
7	126a0d10becbaafcb2e72ce6848cf32c	2018-05-15 00:00:00	6a110e795dd487f1cf8d7583671987af	email
8	f76136f54d14a3345951f25b7932366b	2018-05-24 00:00:00	d51b0d02f063ba1d053db6d97226eec3	email
9	2f838cade4a6012a6cb1016d1d8d95ed	2017-11-10 00:00:00	aeac92c0f5ae22a04ed3b746cce3a1b6	organic_search
10	7281942387a1a0c3f72a50a8b0bb0920	2017-12-25 00:00:00	88740e65d5d6b056e0cda098e1ea6313	social

```sql
SELECT * FROM olist_geolocation limit 10;
```

Data Output | Messages | Notifications

	geolocation_zip_code integer	geolocation_lat numeric	geolocation_lng numeric	geolocation_city character varying (50)	geolocation_state character varying (50)
1	1037	-23.54562128115268	-46.63929204800168	sao paulo	SP
2	1046	-23.546081127035535	-46.64482029837157	sao paulo	SP
3	1046	-23.54612896641469	-46.64295148361138	sao paulo	SP
4	1041	-23.5443921648681	-46.63949930627844	sao paulo	SP
5	1035	-23.541577961711493	-46.64160722329613	sao paulo	SP
6	1012	-23.547762303364266	-46.63536053788448	são paulo	SP
7	1047	-23.546273112412678	-46.64122516971552	sao paulo	SP
8	1013	-23.546923208436723	-46.6342636964915	sao paulo	SP
9	1029	-23.543769055769133	-46.63427784085132	sao paulo	SP
10	1011	-23.547639550320632	-46.63603162315495	sao paulo	SP



	mql_id character varying (50)	seller_id character varying (50)	sdr_id character varying (50)	sr_id character varying (50)	won_date timestamp withou
1	5420aad7fec3549a85876ba1c529bd84	2c43fb513632d29b3b58df74816f1b06	a8387c01a09e99ce014107505b9238...	4ef15afb4b2723d8f3d81e51ec7afefe	2018-02-26 19:58
2	a555fb36b9368110ede0f043dfc3b9a0	bbb7d7893a450660432ea6652310eb...	09285259593c61296eef10c734121d...	d3d1e91a157ea7f90548eef82f1955e3	2018-05-08 20:17
3	327174d3648a2d047e8940d7d15204...	612170e34b97004b3ba37eae81836b4c	b90f87164b5f8c2cfa5c8572834dbe3f	6565aa9ce3178a5caf6171827af3a9ba	2018-06-05 17:27
4	f5fee8f7da74f4887f5bcae2bafb6dd6	21e1781e36faf92725dde4730a88ca0f	56bf83c4bb35763a51c2baab501b4c...	d3d1e91a157ea7f90548eef82f1955e3	2018-01-17 13:51
5	ffe640179b554e295c167a2f6be528e0	ed8cb7b190ceb6067227478e48cf8dde	4b339f9567d060bcea4f5136b9f5949e	d3d1e91a157ea7f90548eef82f1955e3	2018-07-03 20:17
6	b94fba7670eeb44dce2a0d8eb790e9f5	1c742ac33582852aaf3bcfbf5893abcf	fdb16d3cbbeb5798f2f66c4096be026d	495d4e95a8cf8bbf8b432b612a2aa328	2018-02-07 18:04
7	c3e30ed7ac989117c7e1e719b4ac128f	92d7568ad0c5c76fd7d341b2d46f24d6	4b339f9567d060bcea4f5136b9f5949e	85fc447d336637ba1df43e793199fbc8	2018-04-16 18:18
8	b02c89251106e1fdd9d92744be9f94f2	44ed138eca6214d572ce1d813fb0049b	34d40cdaf94010a1d05b0d6212f9e9...	4ef15afb4b2723d8f3d81e51ec7afefe	2018-04-17 17:01
9	a90a37898cc5f2718385a2fb981caaff	0b28859cd04d23edefee9c591fb03cd8	f42a2bd194f7802ab052a815c8de65b7	6565aa9ce3178a5caf6171827af3a9ba	2018-05-14 18:37
10	0173e8d8b1d94a355b440fb67388f532	87d73636a3acf123e842bb890a4db036	9d12ef1a7eca3ec58c545c678af7869c	9e4d1098a3b0f5da39b0bc48f9876645	2018-04-24 03:00

Dimensions, Hierarchies and Measures:

Dimensions:

Customer, Orders, Payments, Reviews, Sellers, Location, Product, Leads and Deals

Hierarchy:

Time: Year -> Quarter -> Month

Geography: state-> city -> pincode

Product: Category -> Product

Measures:

Price, Total Sales, Total items, Payments, Total Payments, Average Review Score

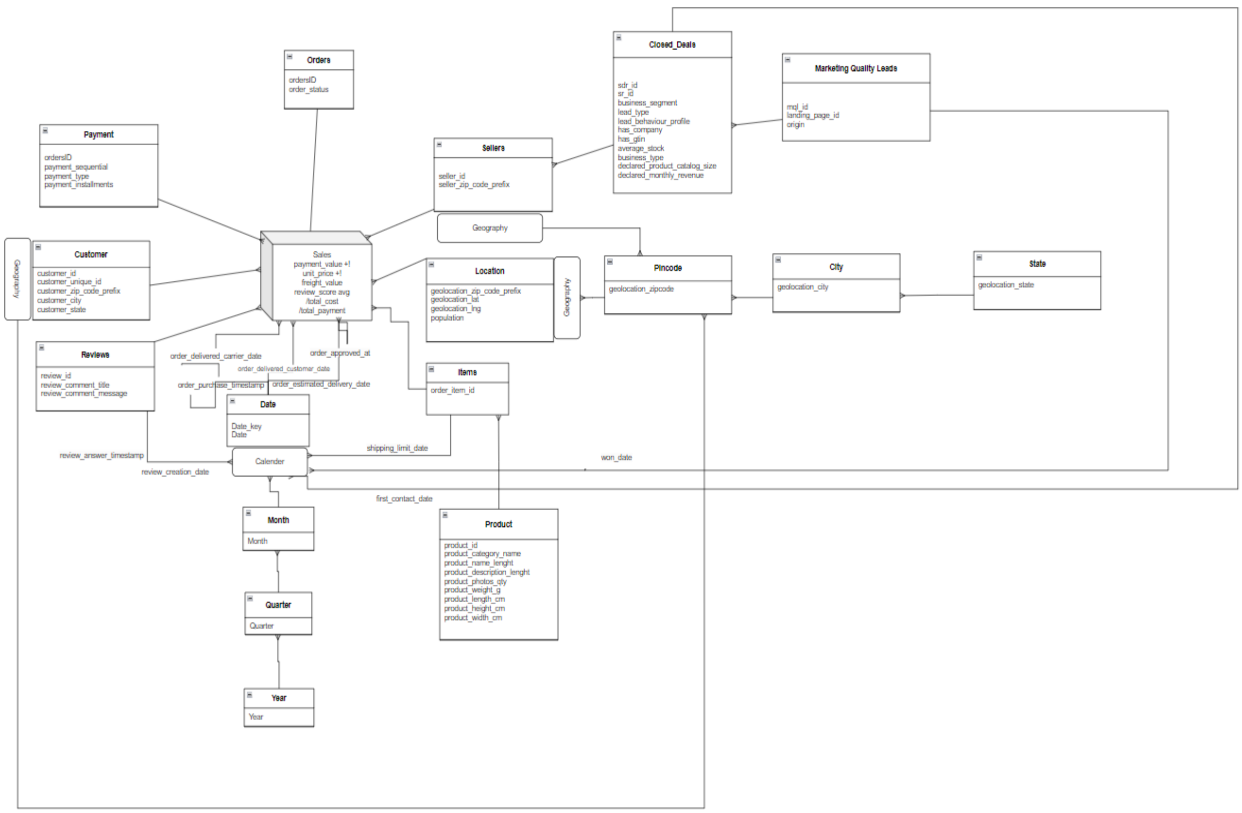
Designing Warehouse:

The data warehouse of this project consists of one fact table i.e. sales and dimension tables: Customer, Orders, Order_item, Payments, Reviews, Sellers, Location, Product, Leads and Deals.

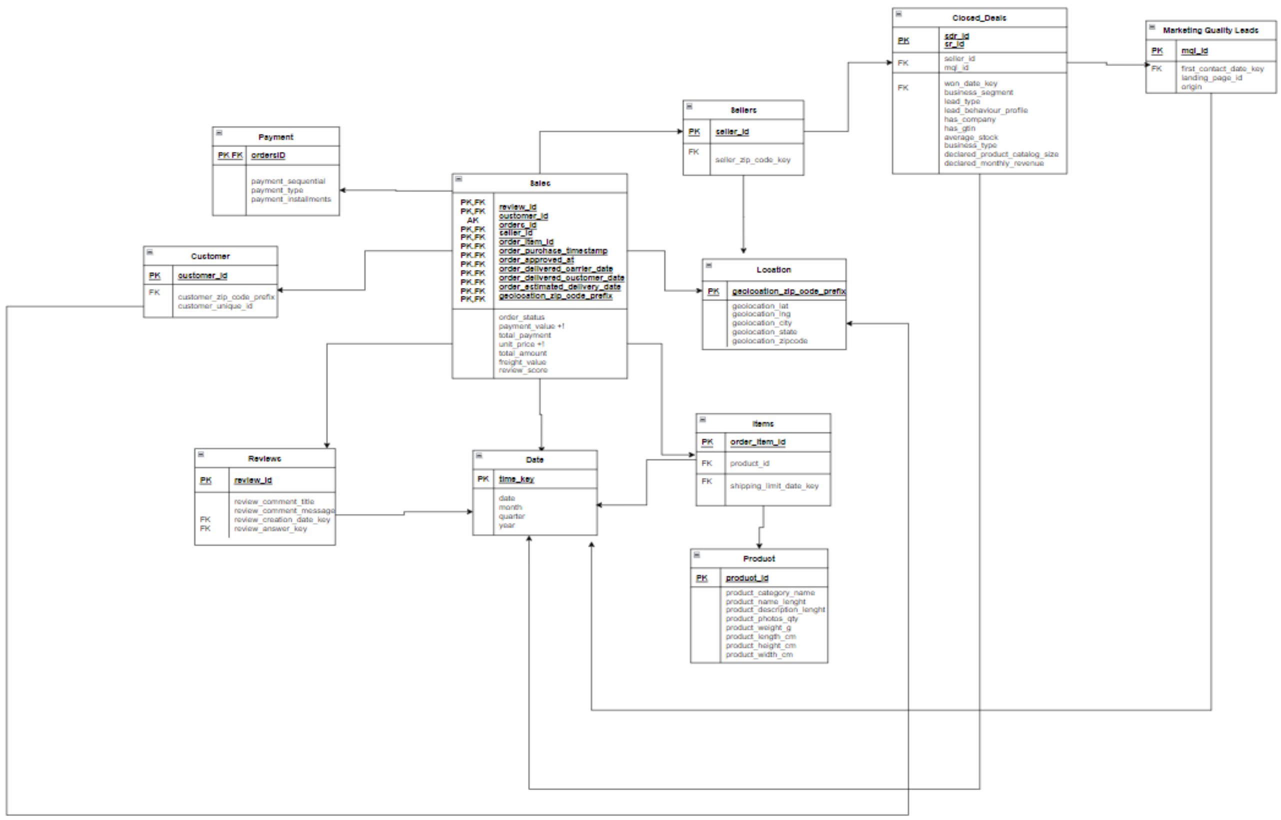
The Fact table is aggregated at order item level, due to this some measures such as unit price, payment are semi-additive. We also have derived values as total price and total items.

We have also separated 3 hierarchies, Date, Geography and product. We will also be removing time from our data as it is not significant to our analysis and will only proceed with date values.

Conceptual Diagram



Logical Diagram:



OLAP Queries:

1) Most popular payment types:
   Sales1 ← ROLLUP*(Sales, Payment → Payment_type, SUM(TotalAmount))
   Result ← MAX(Sales1, TotalAmount, 3)

2) Best selling product in category in a month:
   Sales1 ← ROLLUP*(Sales, Orderitem -> Product, Product → ProductCategory, OrderDate → Month SUM(TotalAmount))
   Result ← MAX(Sales1, TotalAmount) BY ProductCategory and Month

3) Average review score for a product:
   RESULT ← ROLLUP*(Sales, Orderitem -> Product, Avg(Review_score))

4) Top ten percent states giving maxmium sales:
   Sales1 ← ROLLUP*(Sales, Customer → State, SUM(TotalAmount))
   Result ← TOPPERCENT(Sales1, Customer, 10) ORDER BY TotalAmount DESC

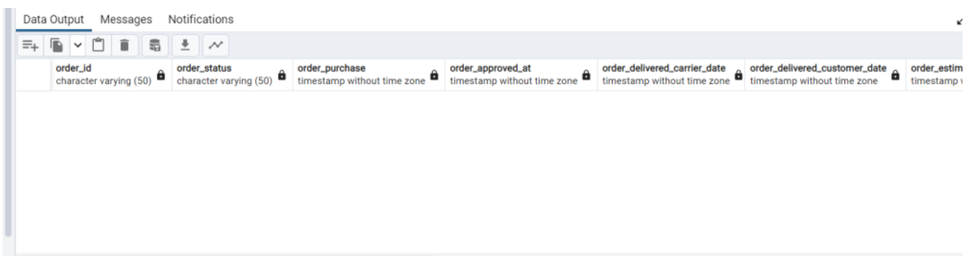
5) Number of cities, states and total sales of each seller:

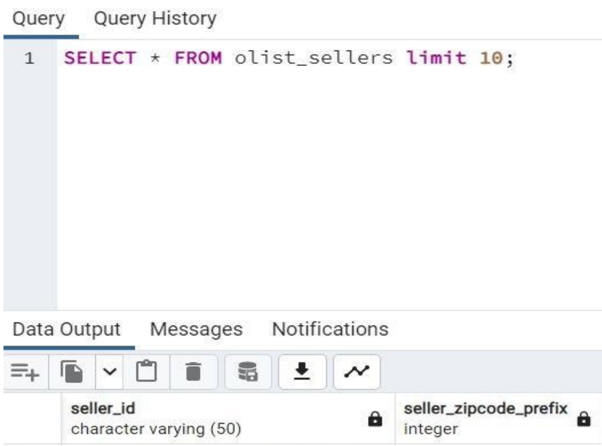
ROLLUP*(Sales, Seller → State, SUM(TotalAmount), COUNT(DISTINCT City) AS No_of_Cities, COUNT(DISTINCT State) AS No_of_States)

6) Total Sale and Average Monthly Sale by seller and year:

   Sales1 ← ROLLUP*(Sales, Seller → Seller, OrderDate → Month, SUM(TotalAmount))

   Result ← ROLLUP*(Sales1, Seller → Seller, OrderDate → Year, SUM(TotalAmount), AVG(TotalAmount))

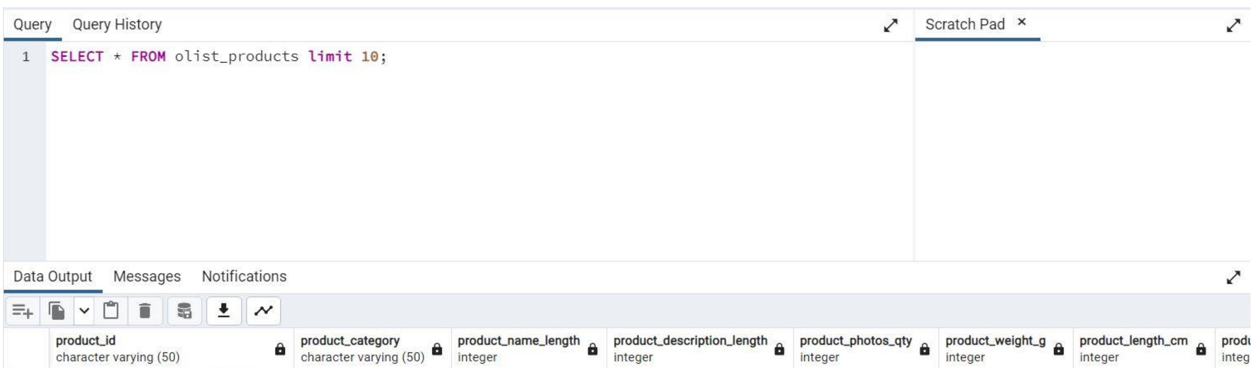



Query
Query History
1 SELECT * FROM olist_sellers limit 10;
Data Output
Messages
Notifications
seller_id
character varying (50)
seller_zipcode_prefix
integer


Query
Query History
Scratch Pad ×
1 SELECT * FROM olist_customer limit 10;
Data Output
Messages
Notifications
customer_id
character varying (80)
customer_unique_id
character varying (80)
customer_zip_code
integer
customer_city
character varying (50)
customer_state
character varying (50)


1 SELECT * FROM brazil_population limit 10;
Data Output
Messages
Notifications
zip_code
character varying (50)
population
integer

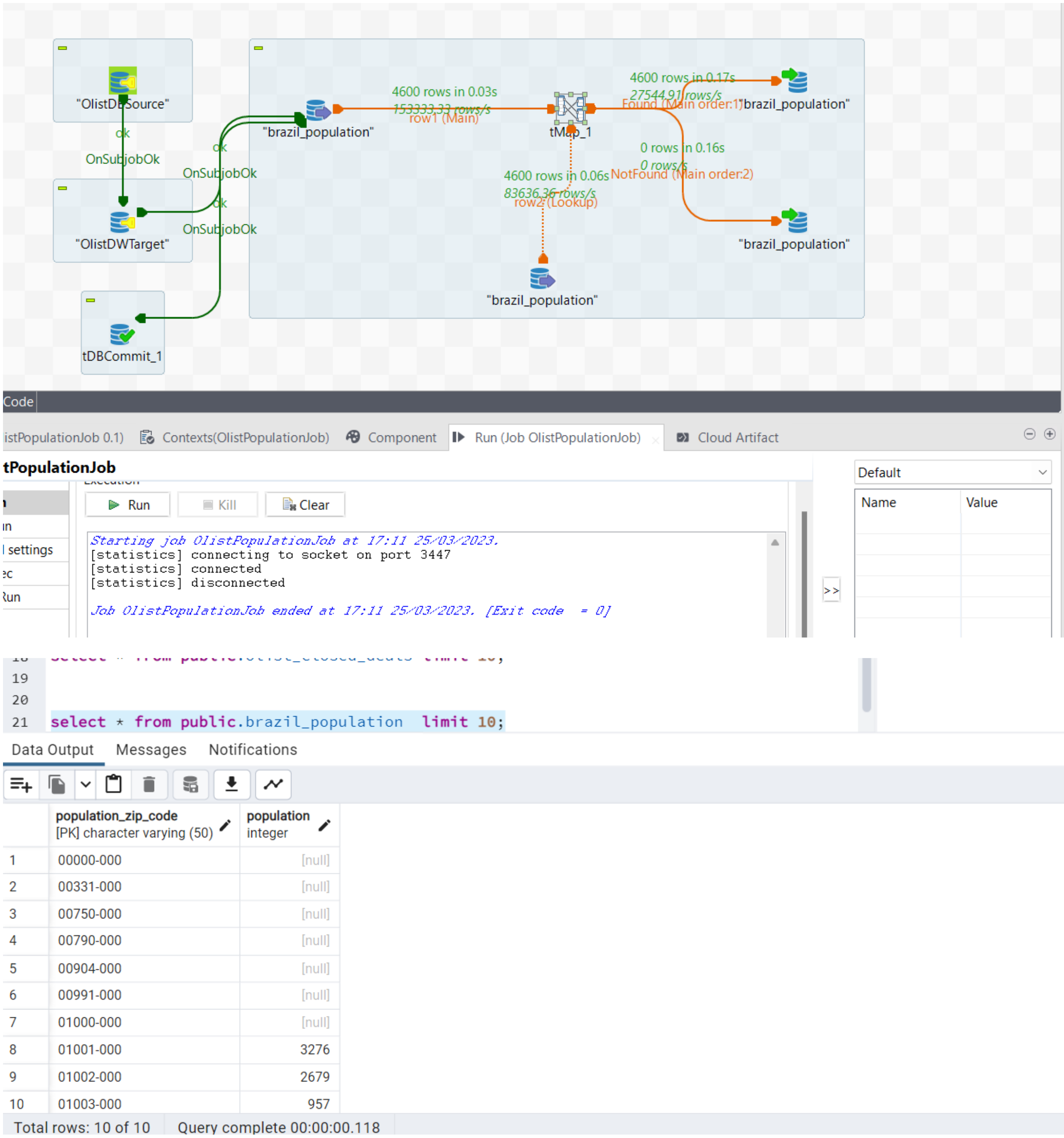

Query
Query History
1 SELECT * FROM olist_sellers limit 10;
Data Output
Messages
Notifications
seller_id
character varying (50)
seller_zipcode_prefix
integer
seller_city
character varying (50)
seller_state
character varying (50)


Query
Query History
Scratch Pad
1 SELECT * FROM olist_products limit 10;
Data Output
Messages
Notifications
product_id
character varying (50)
product_category
character varying (50)
product_name_length
integer
product_description_length
integer
product_photos_qty
integer
product_weight_g
integer
product_length_cm
integer


Query
Query History
Scratch Pad
1 SELECT * FROM olist_orders limit 10;
Data Output
Messages
Notifications
order_id
character varying (50)
customer_id
character varying (50)
order_status
character varying (50)
order_purchase_timestamp
timestamp without time zone
order_approved_at
timestamp without time zone

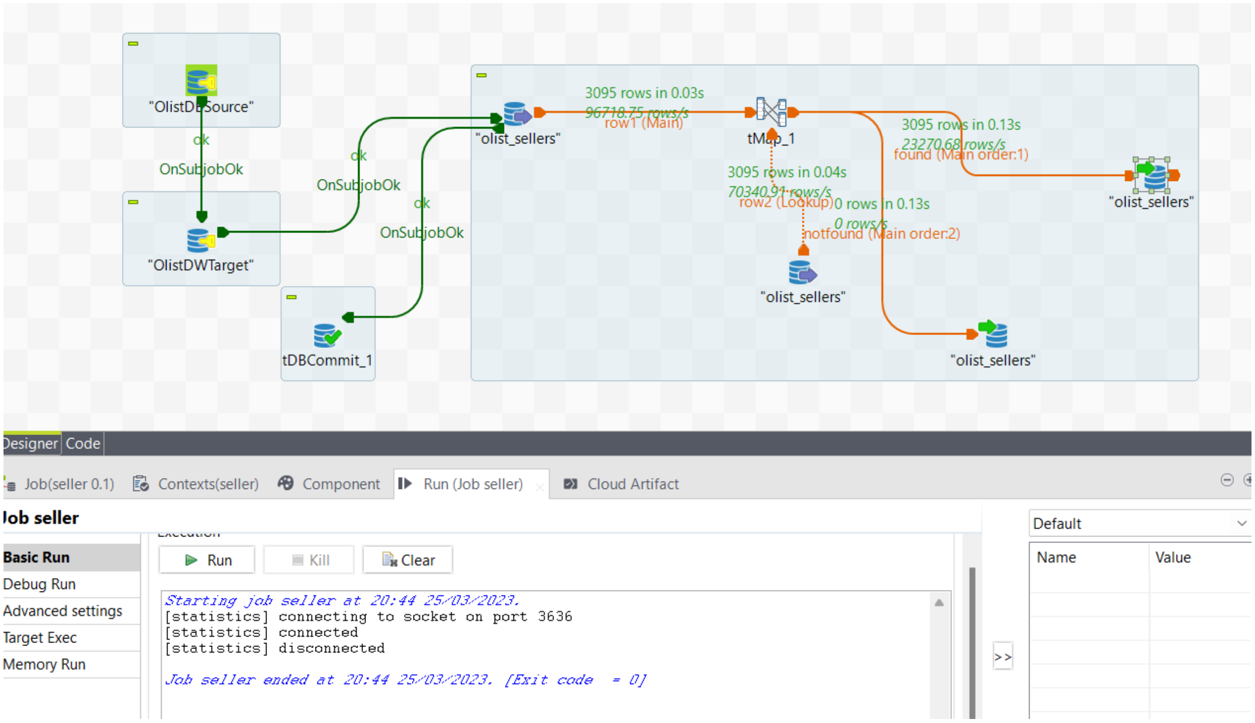
Talend:

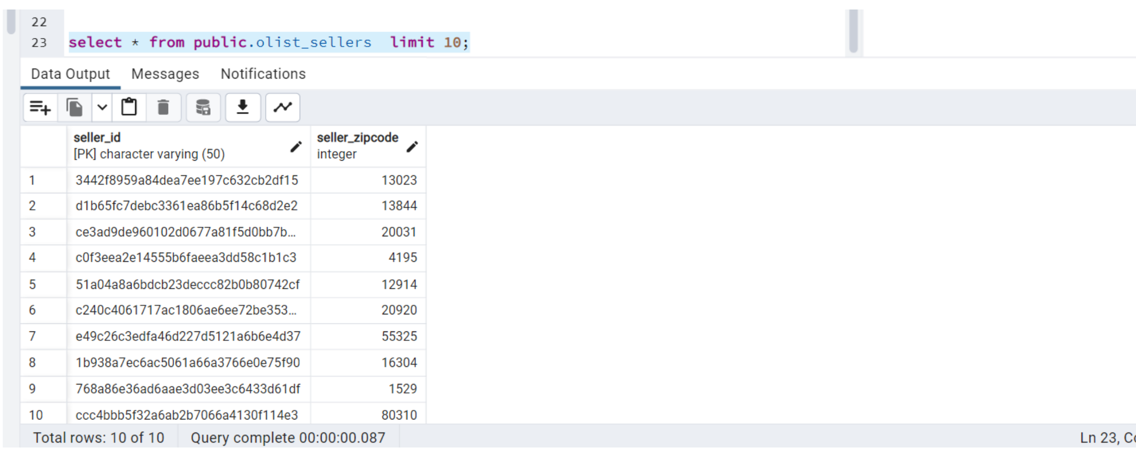
Population:



	population_zip_code [PK] character varying (50)	population integer
1	00000-000	[null]
2	00331-000	[null]
3	00750-000	[null]
4	00790-000	[null]
5	00904-000	[null]
6	00991-000	[null]
7	01000-000	[null]
8	01001-000	3276
9	01002-000	2679
10	01003-000	957

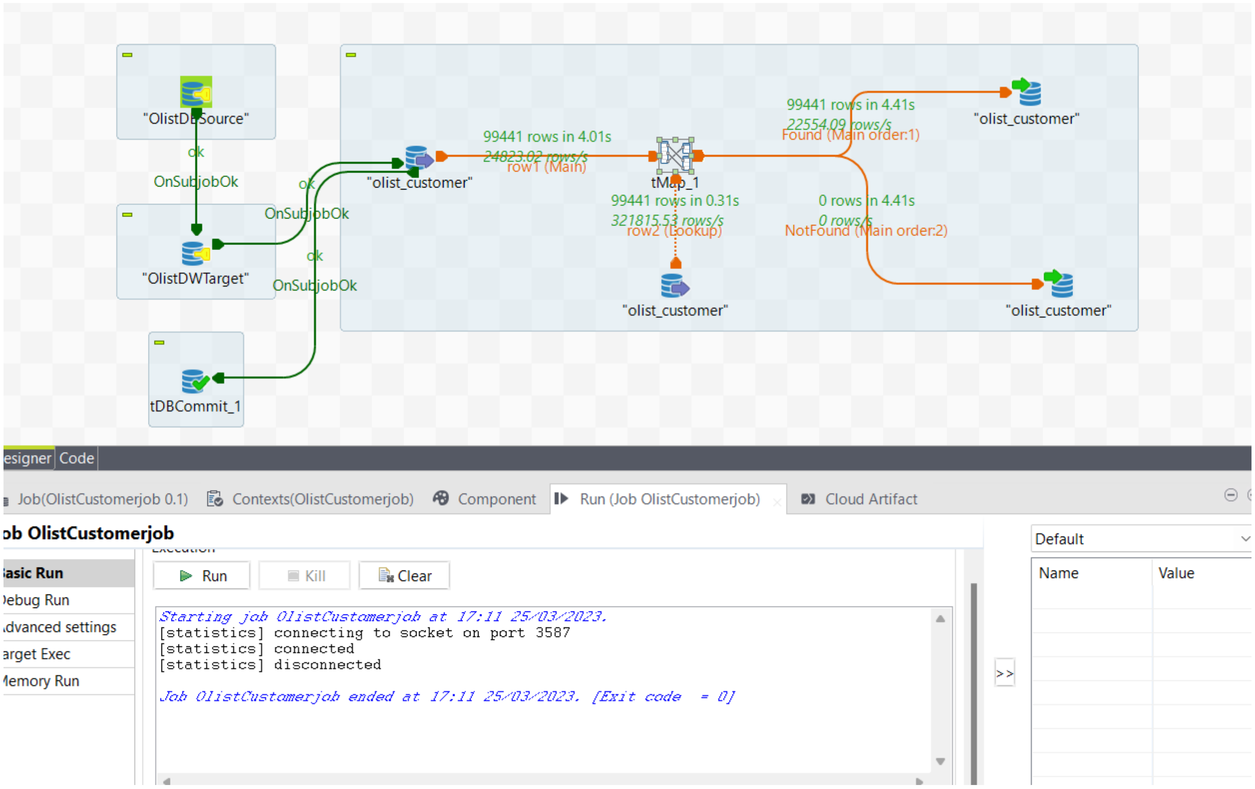
Seller: By creating separate location entity we were able to eliminate excess columns from seller





	seller_id [PK] character varying (50)	seller_zipcode integer
1	3442f8959a84dea7ee197c632cb2df15	13023
2	d1b65fc7debc3361ea86b5f14c68d2e2	13844
3	ce3ad9de960102d0677a81f5d0bb7b...	20031
4	c0f3eea2e14555b6faeea3dd58c1b1c3	4195
5	51a04a8a6bdcb23deccc82b0b80742cf	12914
6	c240c4061717ac1806ae6ee72be353...	20920
7	e49c26c3edfa46d227d5121a6b6e4d37	55325
8	1b938a7ec6ac5061a66a3766e0e75f90	16304
9	768a86e36ad6aae3d03ee3c6433d61df	1529
10	ccc4bbb5f32a6ab2b7066a4130f114e3	80310

Customer: By creating separate location entity we were able to eliminate excess columns from customer

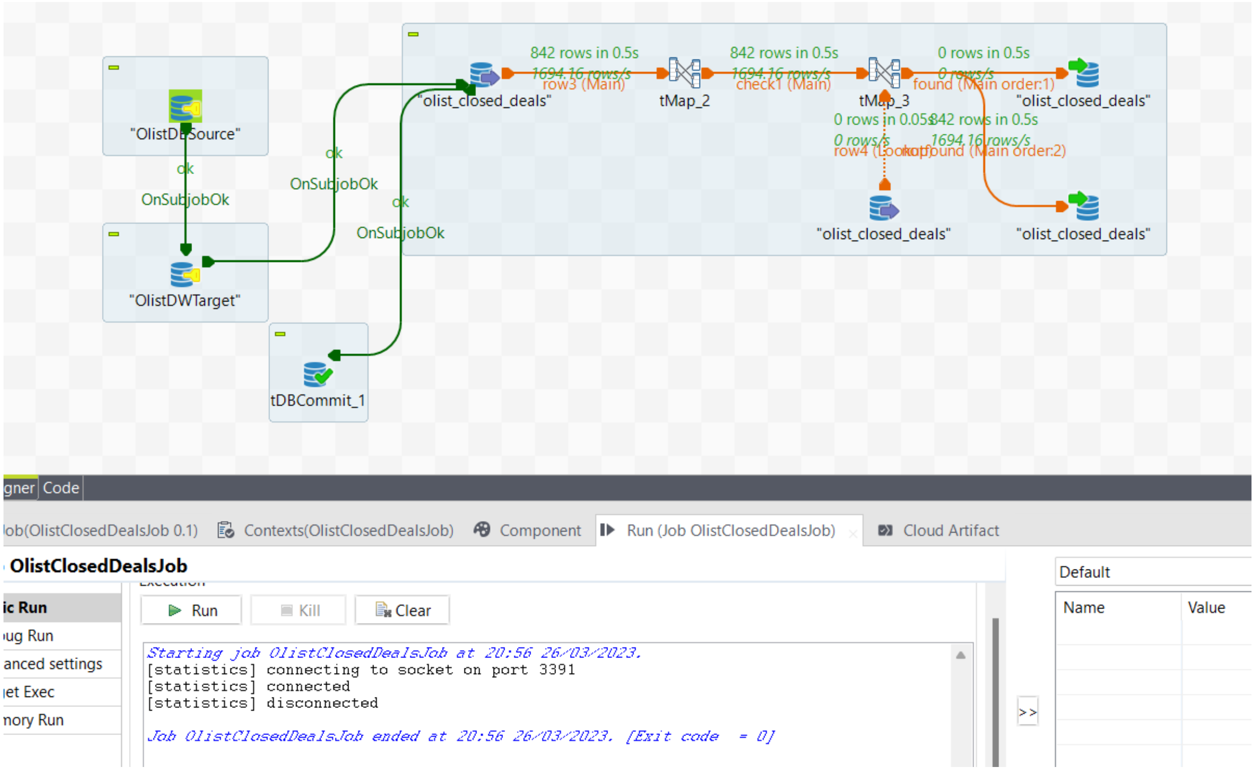



```
select * from public.olist_customer  limit 10;
```

Data Output | Messages | Notifications

	customer_id [PK] character varying (80)	customer_unique_id character varying (80)	customer_zip_code integer
1	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409
2	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333...	9790
3	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2...	1151
4	b2b6027bc5c5109e529d4dc6358b12...	259dac757896d24d7702b9acbbff3f3c	8775
5	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d...	13056
6	879864dab9bc3047522c92c82e1212...	4c93744516667ad3b8f1fb645a3116...	89254
7	fd826e7cf63160e536e0908c76c3f441	addec96d2e059c80c30fe6871d30d1...	4534
8	5e274e7a0c3809e14aba7ad5aae0d4...	57b2a98a409812fe9618067b6b8ebe...	35182
9	5adf08e34b2e993982a47070956c5c65	1175e95fb47ddff9de6b2b06188f7e0d	81560

ClosedDeals:

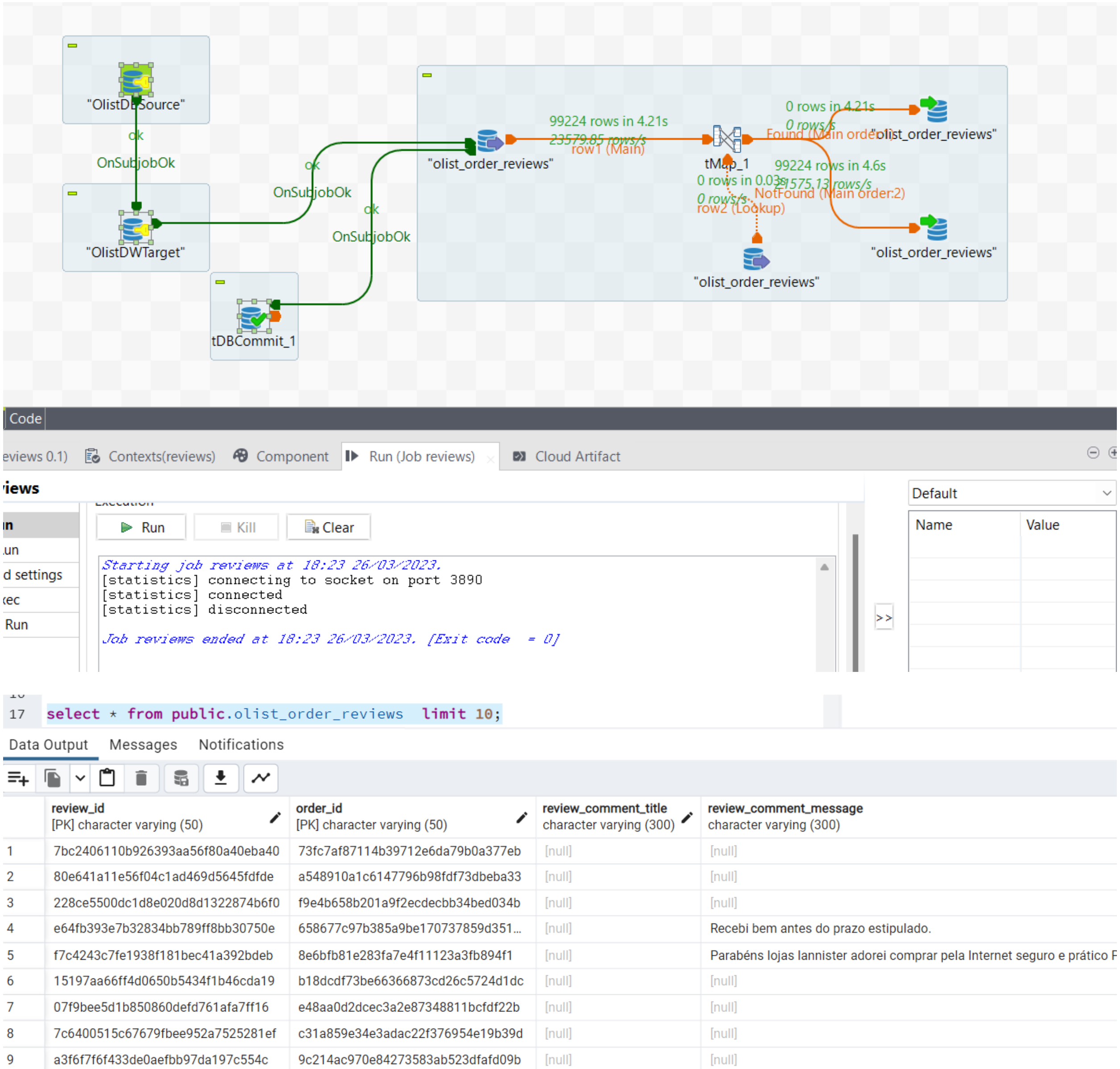



```
24 select * from public.olist_closed_deals limit 10;
25
```

Data Output | Messages | Notifications

	mql_id character varying (50)	seller_id character varying (50)	sdr_id [PK] character varying (50)	sr_id [PK] character varying (50)	won_date_key integer
1	450b97a48568c3118ff397f73abe4e4a	df91910b6a03bb2e3358fa6a35e32f6f	09285259593c61296eef10c734121d5b	060c0a26f19f4d66b42e0d8796688...	2018041
2	a8d2884db49e0769ad4a0b8dce2e1...	d65f74c704563bcb8baa915fbf8cfa1e	6aa3b86a83d784b05f0e37e26b20860d	de63de0d10a6012430098db33c679...	2018092
3	20fcec0873b39b4df3df34140d77d6e7	cee3c487c56d29554a22d112af618301	084b6fbb10729ed4da8c3d3f5a3ae7c9	0a0fb2b07d841f84fb6714e35c7230...	2018111
4	92c7eacf759c266a7663d63876d842...	da3eab58892c04039f4cd5ecb09ae260	9d12ef1a7eca3ec58c545c678af7869c	c638112b43f1d1b86dcabb0da720c...	2018032
5	f0b314f185b80cf35d986e298db53fe3	a663d9c3797e90eac99ff60939416a56	34d40cdaf94010a1d05b0d6212f9e909	060c0a26f19f4d66b42e0d8796688...	2018042
6	902a621ebb8a96c516d34a58334eb...	2aef4c8187ab057d6f1b7e0e636aa481	e7dff61b78bebffa71678e126ce669ad	a8387c01a09e99ce014107505b923...	2018082
7	1124225a47b3f07185073f3df3b449bf	faf3ee2764599d114d7e8f25d4c19845	9749123c950bf8363ace42cb1c2d0815	2695de1affa7750089c0455f8ce270...	2018091
8	ac68bc93ecc07d7daec8ce924ca09b...	6025c79c035c3d772133b8b8238463b2	b90f87164b5f8c2cfa5c8572834dbe3f	56bf83c4bb35763a51c2baab501b4...	2018051
9	d768d7e09a6f9ff4146a1659635cf304	f233b575e585413f12fe2f847d922447	282f16e20d50cb18d05182e80f6b9dd6	d3d1e91a157ea7f90548eef82f1955...	2018103

Review: Transformed review_date to review date_key so that it can be linked to Date dimension



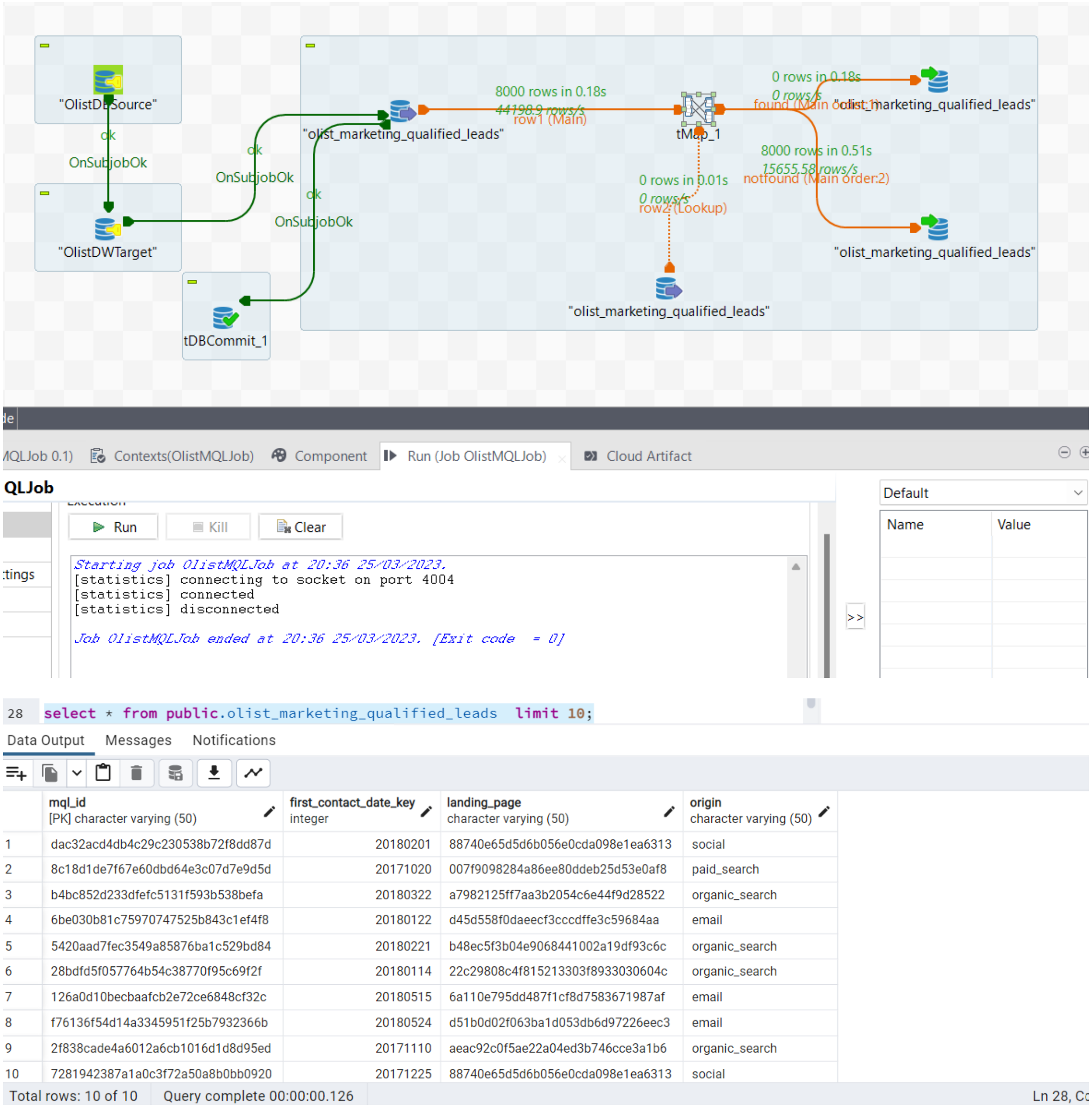

```
Starting job reviews at 18:23 26/03/2023.
[statistics] connecting to socket on port 3890
[statistics] connected
[statistics] disconnected

Job reviews ended at 18:23 26/03/2023. [Exit code  = 0]
```

```sql
select * from public.olist_order_reviews  limit 10;
```

Data Output | Messages | Notifications

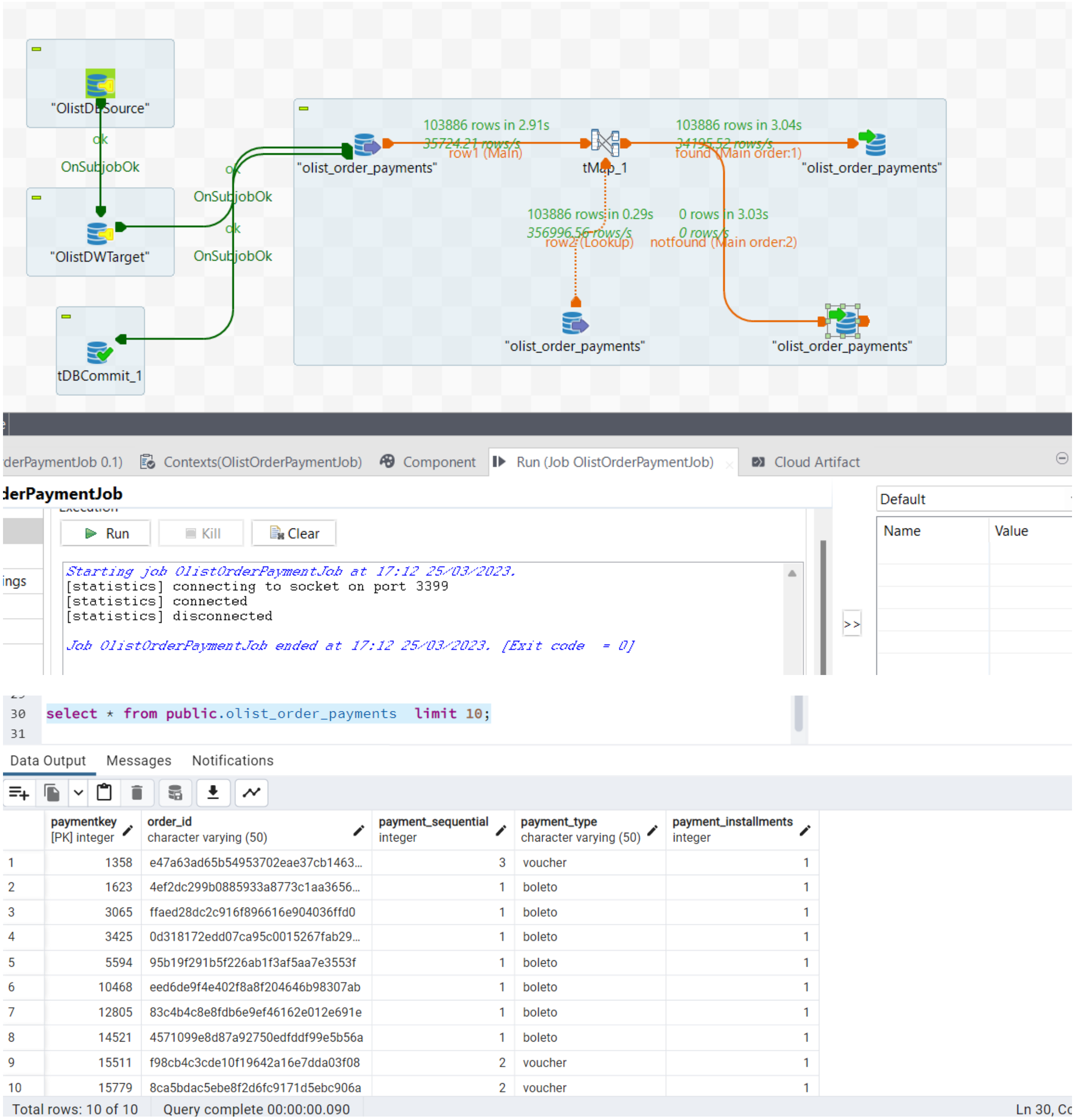
	review_id [PK] character varying (50)	order_id [PK] character varying (50)	review_comment_title character varying (300)	review_comment_message character varying (300)
1	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	[null]	[null]
2	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	[null]	[null]
3	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	[null]	[null]
4	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d351…	[null]	Recebi bem antes do prazo estipulado.
5	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	[null]	Parabéns lojas lannister adorei comprar pela Internet seguro e prático F
6	15197aa66ff4d0650b5434f1b46cda19	b18dcdf73be66366873cd26c5724d1dc	[null]	[null]
7	07f9bee5d1b850860defd761afa7ff16	e48aa0d2dcec3a2e87348811bcfdf22b	[null]	[null]
8	7c6400515c67679fbee952a7525281ef	c31a859e34e3adac22f376954e19b39d	[null]	[null]
9	a3f6f7f6f433de0aefbb97da197c554c	9c214ac970e84273583ab523dfafd09b	[null]	[null]

Marketing Qualified Leads:



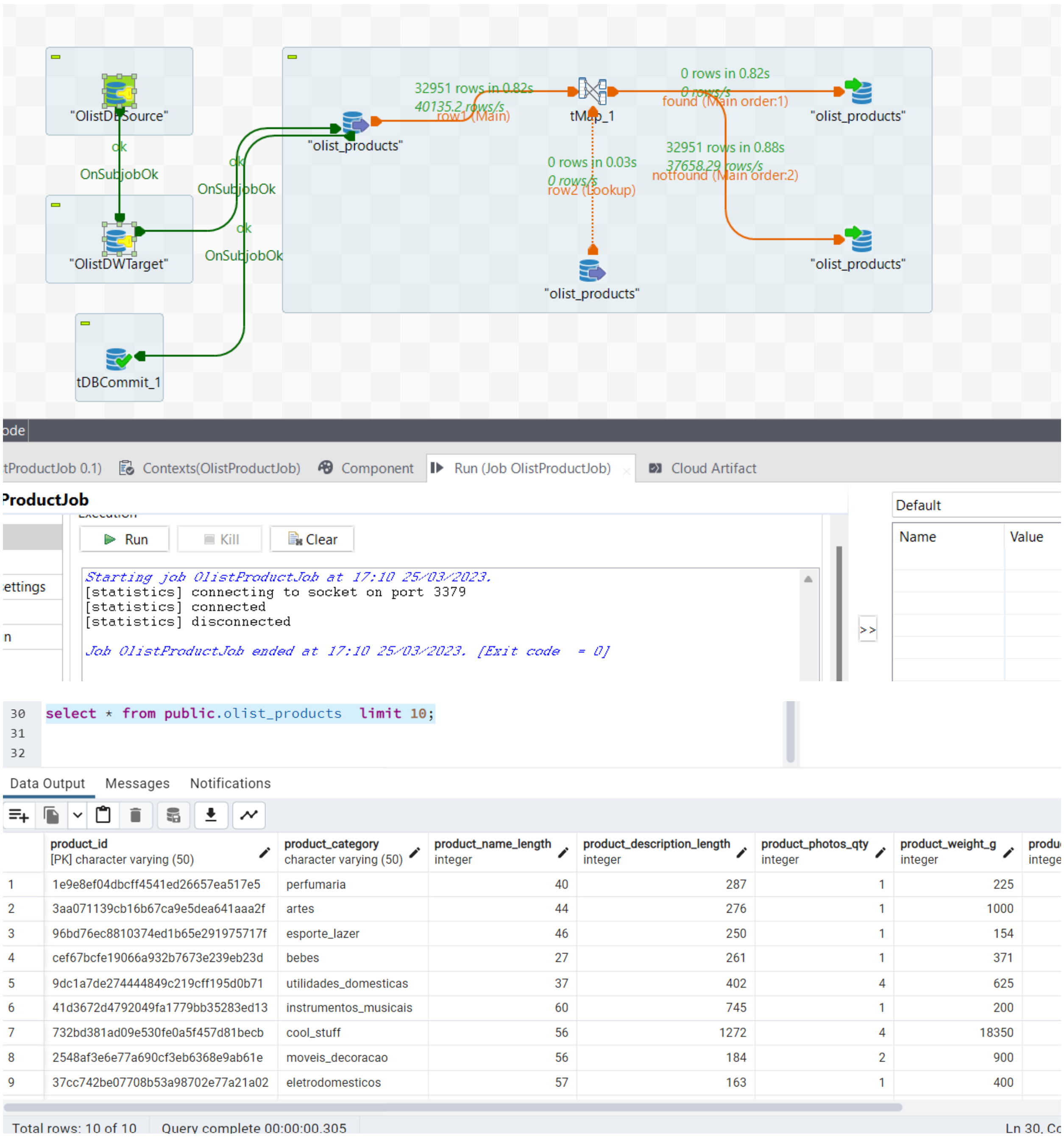
	mql_id [PK] character varying (50)	first_contact_date_key integer	landing_page character varying (50)	origin character varying (50)
1	dac32acd4db4c29c230538b72f8dd87d	20180201	88740e65d5d6b056e0cda098e1ea6313	social
2	8c18d1de7f67e60dbd64e3c07d7e9d5d	20171020	007f9098284a86ee80ddeb25d53e0af8	paid_search
3	b4bc852d233dfefc5131f593b538befa	20180322	a7982125ff7aa3b2054c6e44f9d28522	organic_search
4	6be030b81c75970747525b843c1ef4f8	20180122	d45d558f0daeecf3cccdffe3c59684aa	email
5	5420aad7fec3549a85876ba1c529bd84	20180221	b48ec5f3b04e9068441002a19df93c6c	organic_search
6	28bdfd5f057764b54c38770f95c69f2f	20180114	22c29808c4f815213303f8933030604c	organic_search
7	126a0d10becbaafcb2e72ce6848cf32c	20180515	6a110e795dd487f1cf8d7583671987af	email
8	f76136f54d14a3345951f25b7932366b	20180524	d51b0d02f063ba1d053db6d97226eec3	email
9	2f838cade4a6012a6cb1016d1d8d95ed	20171110	aeac92c0f5ae22a04ed3b746cce3a1b6	organic_search
10	7281942387a1a0c3f72a50a8b0bb0920	20171225	88740e65d5d6b056e0cda098e1ea6313	social

Payment:

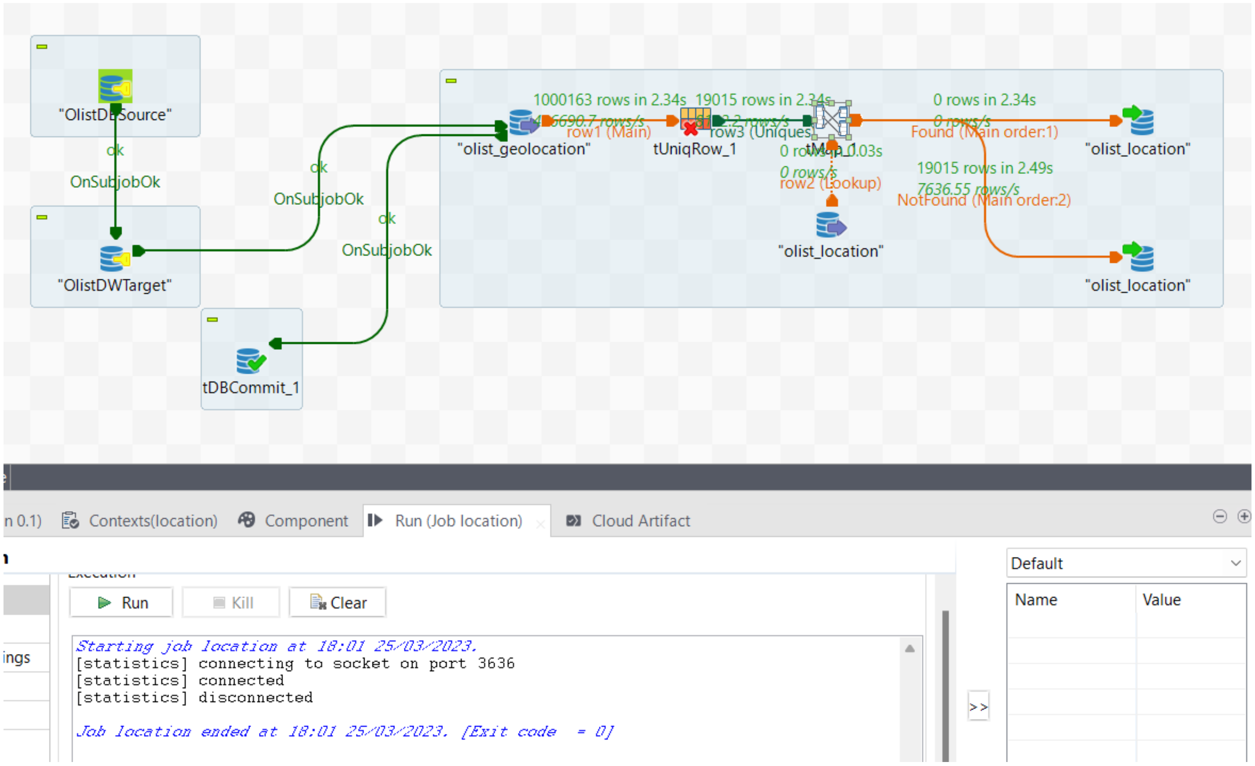


	paymentkey [PK] integer	order_id character varying (50)	payment_sequential integer	payment_type character varying (50)	payment_installments integer
1	1358	e47a63ad65b54953702eae37cb1463...	3	voucher	1
2	1623	4ef2dc299b0885933a8773c1aa3656...	1	boleto	1
3	3065	ffaed28dc2c916f896616e904036ffd0	1	boleto	1
4	3425	0d318172edd07ca95c0015267fab29...	1	boleto	1
5	5594	95b19f291b5f226ab1f3af5aa7e3553f	1	boleto	1
6	10468	eed6de9f4e402f8a8f204646b98307ab	1	boleto	1
7	12805	83c4b4c8e8fdb6e9ef46162e012e691e	1	boleto	1
8	14521	4571099e8d87a92750edfddf99e5b56a	1	boleto	1
9	15511	f98cb4c3cde10f19642a16e7dda03f08	2	voucher	1
10	15779	8ca5bdac5ebe8f2d6fc9171d5ebc906a	2	voucher	1

Product:



Location: Used tUniqRow to get rid of dirty data in the form of duplicates




```
select * from public.olist_location  limit 10;
```

Data Output | Messages | Notifications

	geolocation_zip_code [PK] character varying (50)	geolocation_city character varying (50)	geolocation_state character varying (50)
1	1037	sao paulo	SP
2	1046	sao paulo	SP
3	1041	sao paulo	SP
4	1035	sao paulo	SP
5	1012	são paulo	SP
6	1047	sao paulo	SP
7	1013	sao paulo	SP
8	1029	sao paulo	SP
9	1011	sao paulo	SP
10	1032	sao paulo	SP

Date: This field is prepopulated field

```sql
CREATE TABLE olist_date
(
  date_key              INT NOT NULL,
  date_actual              DATE NOT NULL,
  epoch                    BIGINT NOT NULL,
  day_suffix               VARCHAR(4) NOT NULL,
  day_name                 VARCHAR(9) NOT NULL,
  day_of_week              INT NOT NULL,
  day_of_month             INT NOT NULL,
  day_of_quarter           INT NOT NULL,
  day_of_year              INT NOT NULL,
  week_of_month            INT NOT NULL,
  week_of_year             INT NOT NULL,
  week_of_year_iso         CHAR(10) NOT NULL,
  month_actual             INT NOT NULL,
  month_name               VARCHAR(9) NOT NULL,
  month_name_abbreviated   CHAR(3) NOT NULL,
  quarter_actual           INT NOT NULL,
  quarter_name             VARCHAR(9) NOT NULL,
  year_actual              INT NOT NULL,
  first_day_of_week        DATE NOT NULL,
  last_day_of_week         DATE NOT NULL,
  first_day_of_month       DATE NOT NULL,
```

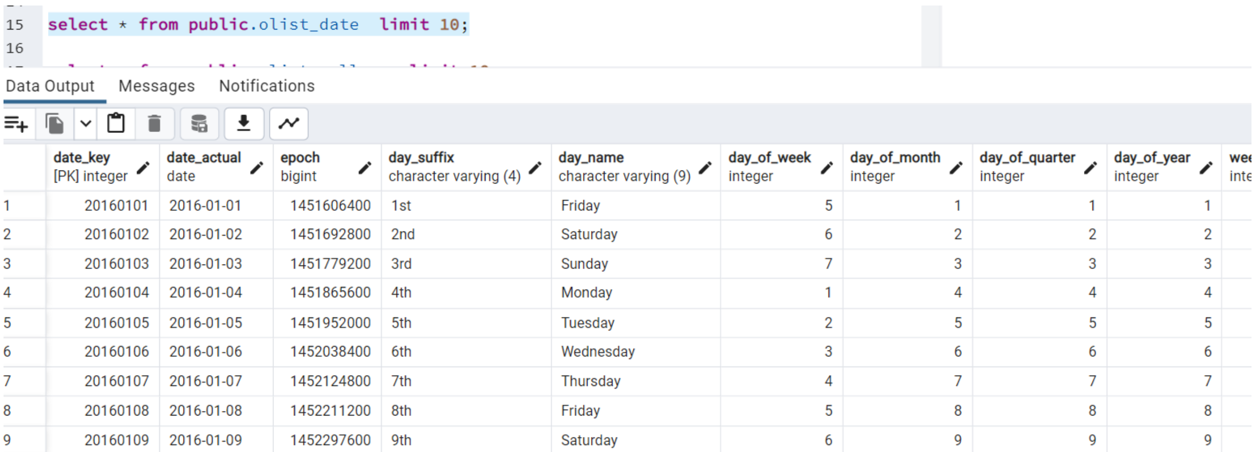
```sql
  last_day_of_month        DATE NOT NULL,
  first_day_of_quarter     DATE NOT NULL,
  last_day_of_quarter      DATE NOT NULL,
  first_day_of_year        DATE NOT NULL,
  last_day_of_year         DATE NOT NULL,
  mmyyyy                   CHAR(6) NOT NULL,
  mmddyyyy                 CHAR(10) NOT NULL,
  weekend_indr             BOOLEAN NOT NULL
);

ALTER TABLE public.olist_date ADD CONSTRAINT d_date_date_dim_id_pk PRIMARY KEY (date_key);

CREATE INDEX d_date_date_actual_idx
  ON olist_date(date_actual);

COMMIT;

INSERT INTO olist_date
SELECT TO_CHAR(datum, 'yyyymmdd')::INT AS date_dim_id,
       datum AS date_actual,
       EXTRACT(EPOCH FROM datum) AS epoch,
       TO_CHAR(datum, 'fmDDth') AS day_suffix,
       TO_CHAR(datum, 'TMDay') AS day_name,
       EXTRACT(ISODOW FROM datum) AS day_of_week,
```

```sql
        EXTRACT(DAY FROM datum) AS day_of_month,
        datum - DATE_TRUNC('quarter', datum)::DATE + 1 AS day_of_quarter,
        EXTRACT(DOY FROM datum) AS day_of_year,
        TO_CHAR(datum, 'W')::INT AS week_of_month,
        EXTRACT(WEEK FROM datum) AS week_of_year,
        EXTRACT(ISOYEAR FROM datum) || TO_CHAR(datum, '"-W"IW-') || EXTRACT(ISODOW FROM datum) AS week_of_year_iso,
        EXTRACT(MONTH FROM datum) AS month_actual,
        TO_CHAR(datum, 'TMMonth') AS month_name,
        TO_CHAR(datum, 'Mon') AS month_name_abbreviated,
        EXTRACT(QUARTER FROM datum) AS quarter_actual,
        CASE
            WHEN EXTRACT(QUARTER FROM datum) = 1 THEN 'First'
            WHEN EXTRACT(QUARTER FROM datum) = 2 THEN 'Second'
            WHEN EXTRACT(QUARTER FROM datum) = 3 THEN 'Third'
            WHEN EXTRACT(QUARTER FROM datum) = 4 THEN 'Fourth'
            END AS quarter_name,
        EXTRACT(YEAR FROM datum) AS year_actual,
        datum + (1 - EXTRACT(ISODOW FROM datum))::INT AS first_day_of_week,
        datum + (7 - EXTRACT(ISODOW FROM datum))::INT AS last_day_of_week,
        datum + (1 - EXTRACT(DAY FROM datum))::INT AS first_day_of_month,
        (DATE_TRUNC('MONTH', datum) + INTERVAL '1 MONTH - 1 day')::DATE AS last_day_of_month,
        DATE_TRUNC('quarter', datum)::DATE AS first_day_of_quarter,
        (DATE_TRUNC('quarter', datum) + INTERVAL '3 MONTH - 1 day')::DATE AS last_day_of_quarter,
        TO_DATE(EXTRACT(YEAR FROM datum) || '-01-01', 'YYYY-MM-DD') AS first_day_of_year,
```

```sql
            WHEN EXTRACT(QUARTER FROM datum) = 1 THEN 'First'
            WHEN EXTRACT(QUARTER FROM datum) = 2 THEN 'Second'
            WHEN EXTRACT(QUARTER FROM datum) = 3 THEN 'Third'
            WHEN EXTRACT(QUARTER FROM datum) = 4 THEN 'Fourth'
            END AS quarter_name,
        EXTRACT(YEAR FROM datum) AS year_actual,
        datum + (1 - EXTRACT(ISODOW FROM datum))::INT AS first_day_of_week,
        datum + (7 - EXTRACT(ISODOW FROM datum))::INT AS last_day_of_week,
        datum + (1 - EXTRACT(DAY FROM datum))::INT AS first_day_of_month,
        (DATE_TRUNC('MONTH', datum) + INTERVAL '1 MONTH - 1 day')::DATE AS last_day_of_month,
        DATE_TRUNC('quarter', datum)::DATE AS first_day_of_quarter,
        (DATE_TRUNC('quarter', datum) + INTERVAL '3 MONTH - 1 day')::DATE AS last_day_of_quarter,
        TO_DATE(EXTRACT(YEAR FROM datum) || '-01-01', 'YYYY-MM-DD') AS first_day_of_year,
        TO_DATE(EXTRACT(YEAR FROM datum) || '-12-31', 'YYYY-MM-DD') AS last_day_of_year,
        TO_CHAR(datum, 'mmyyyy') AS mmyyyy,
        TO_CHAR(datum, 'mmddyyyy') AS mmddyyyy,
        CASE
            WHEN EXTRACT(ISODOW FROM datum) IN (6, 7) THEN TRUE
            ELSE FALSE
            END AS weekend_indr
FROM (SELECT '2016-01-01'::DATE + SEQUENCE.DAY AS datum
      FROM GENERATE_SERIES(0, 29219) AS SEQUENCE (DAY)
      GROUP BY SEQUENCE.DAY) DQ
ORDER BY 1;
```

Data Output | Messages | Notifications

```
CREATE TABLE

Query returned successfully in 72 msec.
```

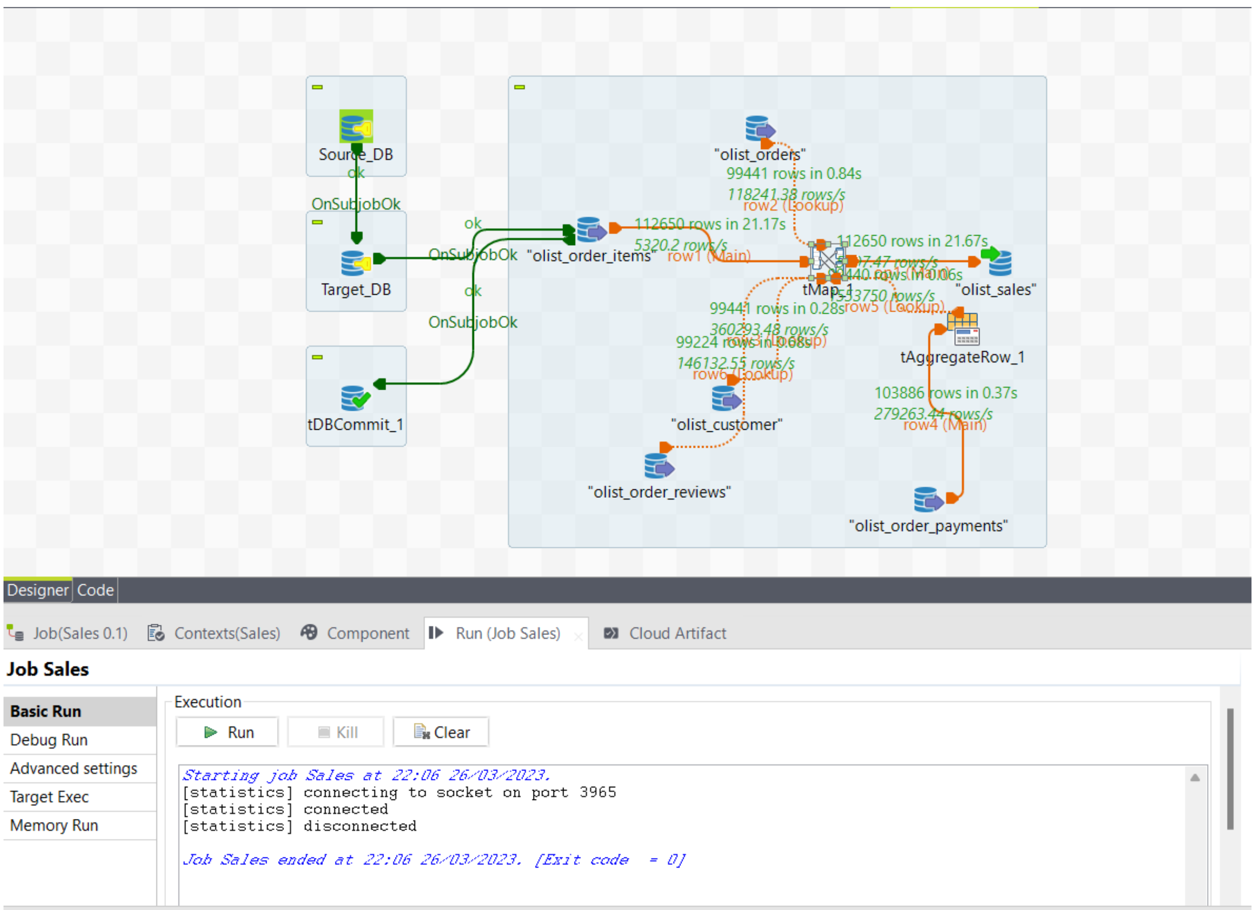


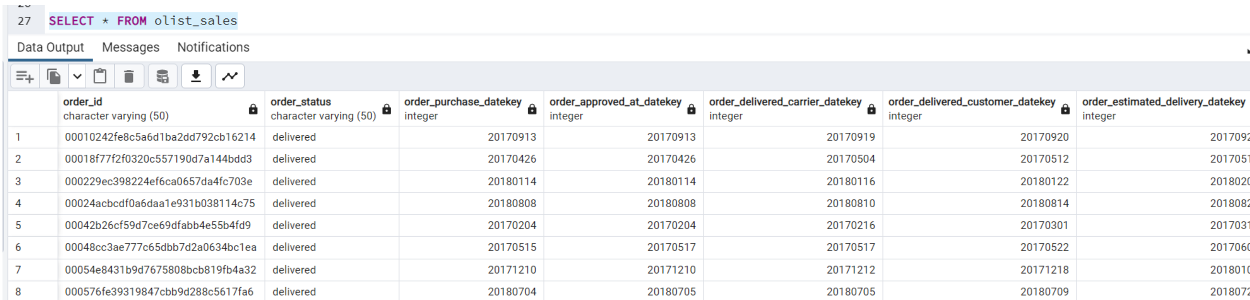
```
15  select * from public.olist_date  limit 10;
16
```

Data Output | Messages | Notifications

	date_key [PK] integer	date_actual date	epoch bigint	day_suffix character varying (4)	day_name character varying (9)	day_of_week integer	day_of_month integer	day_of_quarter integer	day_of_year integer	wee inte
1	20160101	2016-01-01	1451606400	1st	Friday	5	1	1	1	
2	20160102	2016-01-02	1451692800	2nd	Saturday	6	2	2	2	
3	20160103	2016-01-03	1451779200	3rd	Sunday	7	3	3	3	
4	20160104	2016-01-04	1451865600	4th	Monday	1	4	4	4	
5	20160105	2016-01-05	1451952000	5th	Tuesday	2	5	5	5	
6	20160106	2016-01-06	1452038400	6th	Wednesday	3	6	6	6	
7	20160107	2016-01-07	1452124800	7th	Thursday	4	7	7	7	
8	20160108	2016-01-08	1452211200	8th	Friday	5	8	8	8	
9	20160109	2016-01-09	1452297600	9th	Saturday	6	9	9	9	

Sales (Fact): This is our fact table it is maintained at the lowest level of granularity I.e. order_item, we also use aggregation to aggregate payment value as that is at installment level in the warehouse.



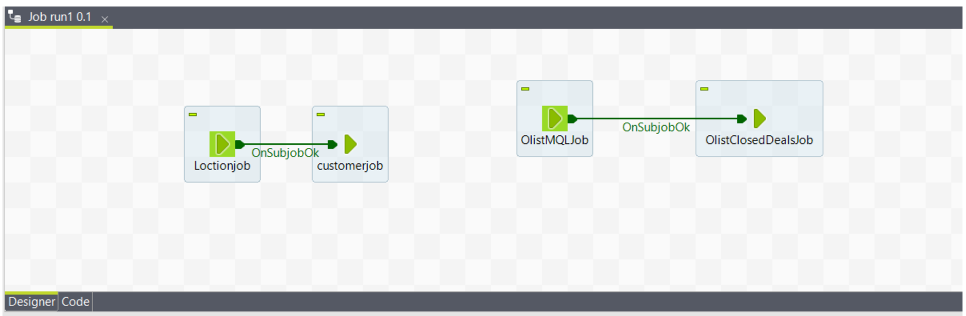


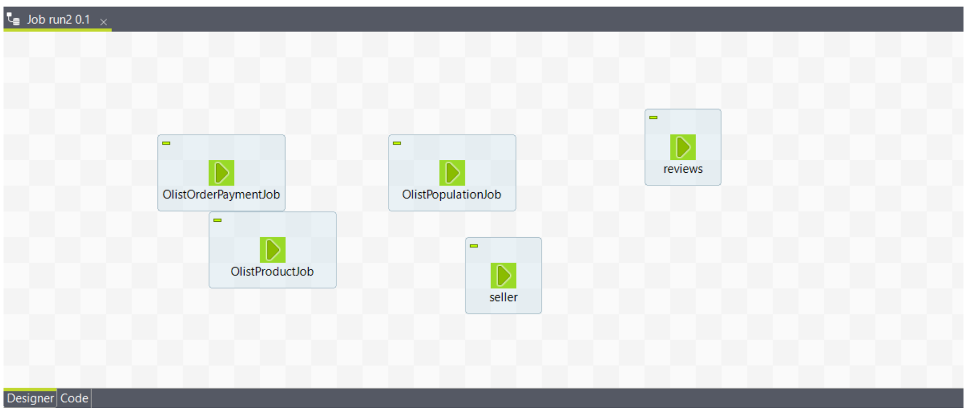
```sql
SELECT * FROM olist_sales
```

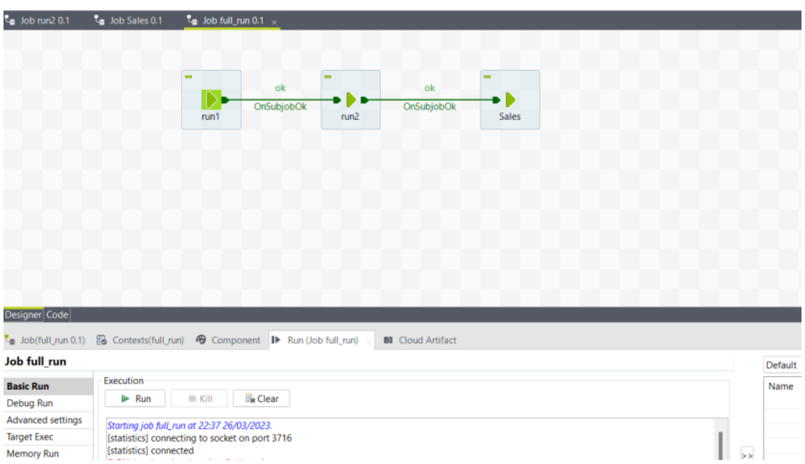
Data Output | Messages | Notifications

	order_id character varying (50)	order_status character varying (50)	order_purchase_datekey integer	order_approved_at_datekey integer	order_delivered_carrier_datekey integer	order_delivered_customer_datekey integer	order_estimated_delivery_datekey integer
1	00010242fe8c5a6d1ba2dd792cb16214	delivered	20170913	20170913	20170919	20170920	2017092
2	00018f77f2f0320c557190d7a144bdd3	delivered	20170426	20170426	20170504	20170512	2017051
3	000229ec398224ef6ca0657da4fc703e	delivered	20180114	20180114	20180116	20180122	2018020
4	00024acbcdf0a6daa1e931b038114c75	delivered	20180808	20180808	20180810	20180814	2018082
5	00042b26cf59d7ce69dfabb4e55b4fd9	delivered	20170204	20170204	20170216	20170301	2017031
6	00048cc3ae777c65dbb7d2a0634bc1ea	delivered	20170515	20170517	20170517	20170522	2017060
7	00054e8431b9d7675808bcb819fb4a32	delivered	20171210	20171210	20171212	20171218	2018010
8	000576fe39319847cbb9d288c5617fa6	delivered	20180704	20180705	20180705	20180709	2018072

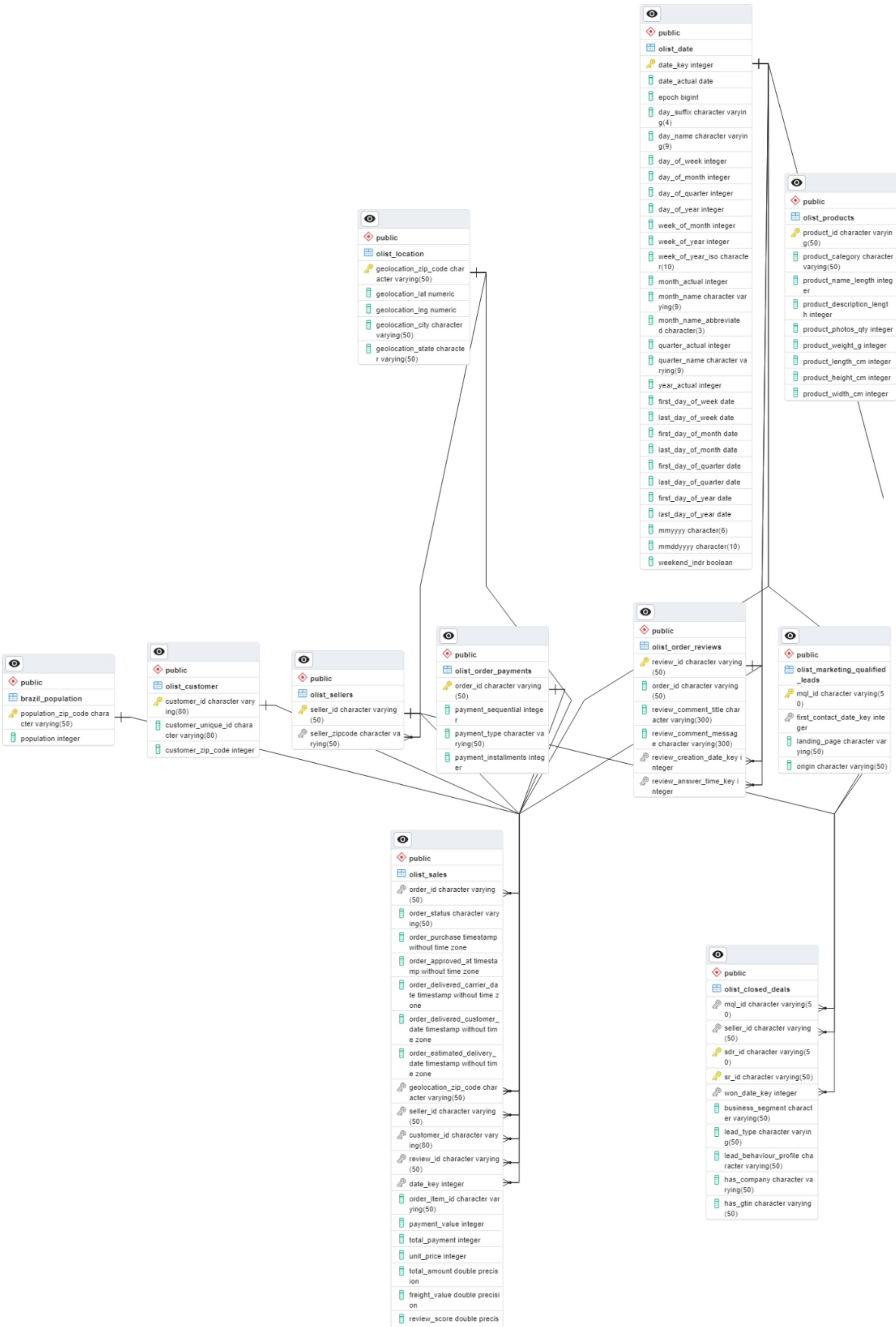
Finally we merge all these jobs to run together in parallel; we divide them in to two files and the final file has the fact table job:







ERD Diagram of Datawarehouse:



SAMPLE QUERIES ON WAREHOUSE:

Query Query History

```sql
select product_category,avg(payment_value)
from olist_sales A
inner join olist_products B
ON A.product_id = B.product_id
group by product_category
```

Data Output Messages Notifications

	product_category character varying (50)	avg numeric
1	agro_industria_e_comercio	560.0500471698113208
2	alimentos	91.0366078431372549
3	alimentos_bebidas	98.1644964028776978
4	artes	148.2915311004784689
5	artes_e_artesanato	96.9237500000000000
6	artigos_de_festas	138.7513953488372093
7	artigos_de_natal	124.1488235294117647
8	audio	165.7269780219780220

Query Query History

```sql
select day_name,count(distinct order_id)
from olist_sales A
inner join olist_date B
ON A.order_purchase_datekey = B.date_key
WHERE B.month_actual = 1
group by day_name
```

Data Output Messages Notifications

	day_name character varying (9)	count bigint
1	Friday	1012
2	Monday	1340
3	Saturday	843
4	Sunday	861
5	Thursday	1148
6	Tuesday	1419
7	Wednesday	1386

Query | Query History

```sql
select A.product_id,avg(review_score)
from olist_sales A
inner join olist_products B
ON A.product_id = B.product_id
group by A.product_id
```

Data Output | Messages | Notifications

	product_id character varying (50)	avg numeric
1	027293c3b6d9e221268d9d6a5ffe5d0b	5.0000000000000000
2	9ad4fb75641c9a71f995a06d0386a6f9	3.0000000000000000
3	5ed2407e517ceac8615bb4d1ea238fe6	4.0000000000000000
4	8e630a6101b33d99671d79b67b8705...	3.5000000000000000
5	cf41f723df264ec6f1c74d1caca1642e	5.0000000000000000
6	f838e96e79703e5e94c13c86609530c5	1.00000000000000000000
7	3d5837f86205fe83f03fb5f7e4d5b9cf	4.6363636363636364
8	46d92d8c6c48c2ac51bdd6be618d05...	4.0000000000000000
9	c13753983dc6a777c5e17ebf6db9fceb	5.0000000000000000

Query | Query History

```sql
select seller_id,avg(payment_value)
from olist_sales
group by seller_id
```

Data Output | Messages | Notifications

	seller_id character varying (50)	avg numeric
1	a61cc04793308395a84080710436...	38.3700000000000000
2	6d2f2e3b539480db1e0842b3a4e3...	163.1833333333333333
3	c53bcd3be457a342a97e39e5a9f0b...	163.1380000000000000
4	9591fc341b1bfb7ef561e2968ec6e...	94.2100000000000000
5	f9ec7093df3a7b346b7bcf7864069...	86.5577777777777778
6	fc38b5dceee1a730fad8853453437...	35.6020000000000000
7	5def4c3732941a971cba8fdee992e...	41.4366666666666667
8	b39d7fe263ef469605dbb32608aee...	161.7932653061224490
9	ea65d8b58316a6f2362f2a9e4b3e8	273.5425000000000000

Query Query History

```sql
select geolocation_city,avg(payment_value)
from olist_sales A
INNER JOIN olist_location B
ON A.geolocation_zip_code = B.geolocation_zip_code
group by geolocation_city
```

Data Output Messages Notifications

	geolocation_city character varying (50)	avg numeric
1	bom jardim de minas	130.1609090909090909
2	alto rio doce	69.6375000000000000
3	são tomás de aquino	127.8000000000000000
4	batatais	227.3583870967741935
5	capao da porteira	107.5650000000000000
6	icatu	892.9000000000000000
7	osorio	315.9700000000000000
8	populina	71.3000000000000000
9	novo mundo	77.1300000000000000