Group Assignment - Group Number 50

Business Topic: Tracking order value of products in Olist E-Commerce site

Contributors:

Carol John - 2141478

Chi Nguyen - 2134771

JuYeong Lee - 2142916

Matthijs de Wildt - 2126952

Tra Mi Nguyen - 2057817

Tilburg School of Humanities and Digital Sciences, Tilburg University

Course: Business Intelligence for Data Science (880682-M-6)

Semester: Fall 2024/Block 2

Dr. Nevena Rankovic

December 13, 2024

1. **Motivation and Goals:**

The motivation behind creating a DW for the Olist data is to create a streamlined version of all the sales order data. This is done in order for us to be able to draw meaningful insights, store it adeptly and to standardize the version of the data. Through this we achieve a unified repository of data and we now have an analytical platform for us to evaluate on customer buying patterns, sales trends, product reviews and so on. The DW plays a major role by ensuring the data is accurate, consistent and helps to avoid redundancy. It also helps us to plan the business effectively through data driven decision making and to analytically plan strategies and forecast prediction to enhance the business. Conclusively DW is efficient in acting as a significant infrastructure in fostering inventive ideas and solutions, bringing growth and positioning the business strategically.

2. **Source OLTP schema:**

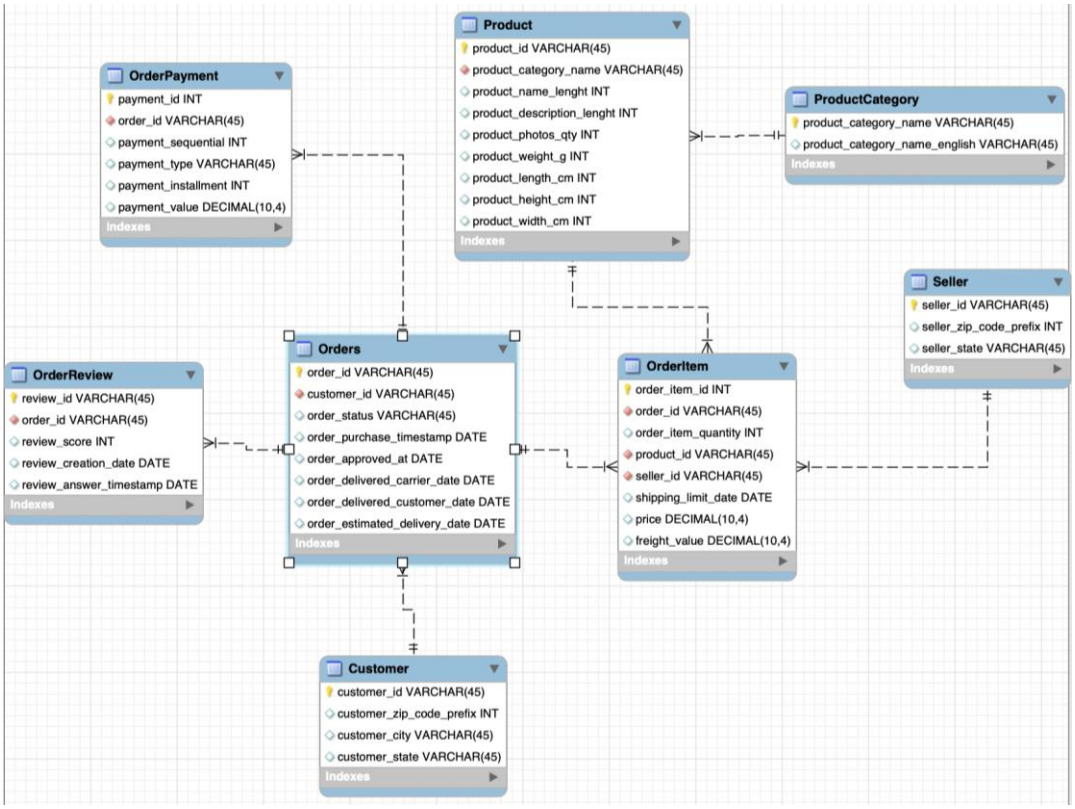


Figure 1: OLTP schema Olist

Describe the relationship between entities

- **OrderPayment** needs at most one order and uses the order_id as the foreign key, however an **Order** can have one or more payments.
- The **OrderReview** must be written for at most one order and use the order_id as the foreign key, and one **Order** can receive zero or many reviews
- A **Customer** can make zero or many orders, however an **Order** needs at most one customer and uses the customer_id as the foreign key.

- A **Product** can have one or many order items but must belong to one category and use the product_category_name as the foreign key. One **ProductCategory** can have many products.
- An **OrderItem** must need at most one order and uses the order_id as the foreign key, in contrast, an **Order** can include one or many items.
- An **OrderItem** must also belong to one **Product** and use the product_id as the foreign key, in the meanwhile a **Product** can have one or many items.
- A **Seller** can have one or many order items, however, one **OrderItem** needs at most one supplier and uses the seller_id as the foreign key.

3. Transfer the OLTP schema diagram into Tables in MySQL

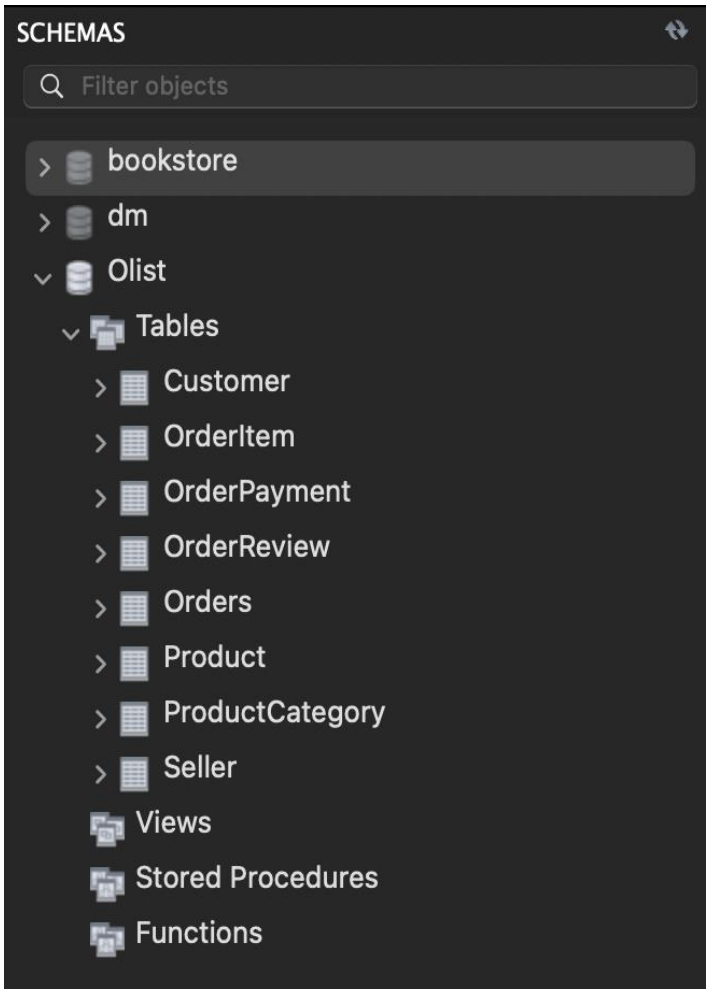


Figure 2: Transfering OLTP into Tables

customer_id	customer_zip_code_prefix	customer_city	customer_state
00012a2ce6f8dcda20d059ce98491703	6273	osasco	SP
000161a058600d5901f007fab4c27140	35550	itapecerica	MG
0001fd6190edaaf884bcaf3d49edf079	29830	nova venecia	ES
0002414f95344307404f0ace7a26f1d5	39664	mendonca	MG
000379cdec625522490c315e70c7a9fb	4841	sao paulo	SP
0004164d20a9e969af783496f3408652	13272	valinhos	SP
000419c5494106c306a97b5635748086	24220	niteroi	RJ
00046a560d407e99b969756e0b10f282	20540	rio de janeiro	RJ
00050bf6e01e69d5c0fd612f1bcfb69c	98700	ijui	RS
000598caf2ef4117407665ac33275130	35540	oliveira	MG
0005aefbb696d34b3424dccd0a0e9fd0	3052	sao paulo	SP
00062b33cb9f6fe976afdcff967ea74d	2306	sao paulo	SP
00066ccbe787a588c52bd5ff404590e3	93525	novo hamburgo	RS
00072d033fe2e59061ae5c3aff1a2be5	45026	vitoria da conquista	BA
0009a69b72033b2d0ec8c69fc70ef768	13106	campinas	SP
000bf8121c3412d3057d32371c5d3395	12335	jacarei	SP
000e943451fc2788ca6ac98a682f2f49	99460	colorado	RS
000f17e290c26b28549908a04cfe36c1	98400	frederico westphalen	RS
000fd45d6fedae68fc6676036610f879	12970	piracaia	SP
0010068a73b7c56da5758c3f9e5c7375	63680	parambu	CE
001028b78fd413e19704b3867c369d3a	5387	sao paulo	SP
00104a47c29da701ce41ee52077587d9	38400	uberlandia	MG

Figure 3: Imported data for the Customer table

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at
00b44ba3d7c4a5e9a9ebafef9150781d	0b5f6687d659478f1747caed607c4ec5	delivered	2018-08-28 00:00:00	2018-08-28 00:00:00
00eead1d5a799277ef3d6b387b0bebef	c713fee6f70301b6bf60701902502703	delivered	2018-08-27 00:00:00	2018-08-27 00:00:00
0091a677651feaf5a08d7bb147681e14	93d82c58a8f2b0d2db80f16867b99edd	delivered	2018-08-24 00:00:00	2018-08-24 00:00:00
0097aafe2a01dd54e39888976160c377	6e6943d8c4822ead782b17d532e6ee63	delivered	2018-08-23 00:00:00	2018-08-25 00:00:00
0128412231f9fe9d7e944eea5392fc6b	a7089cdfb6b910afd377368f3335b1e5	delivered	2018-08-22 00:00:00	2018-08-24 00:00:00
022f80b8a6f58b11018082b505f15c38	e55ddd4fde4608bc77a6494772b5bd3b	delivered	2018-08-22 00:00:00	2018-08-22 00:00:00
00e365f4fc03d1098841af23d05c17a6	491609eeefc72ee67ee5b42e63ba9859	delivered	2018-08-20 00:00:00	2018-08-20 00:00:00
01ccfbb0ddf88340b6eed3f1370bcf0c	0b81ac66ced97049984bd7faa500cdda	delivered	2018-08-21 00:00:00	2018-08-22 00:00:00
02036bc298ab3c1efc98b116f420d646	1b0555269127c13e727a1f3cb545faf2	delivered	2018-08-23 00:00:00	2018-08-23 00:00:00
0216b8a71f44f126ac083fa5d639fe47	6721f296aee89a6caef0bf53c7547103	delivered	2018-08-19 00:00:00	2018-08-20 00:00:00
003a7f59d7e08a9c61d9e2881fe6459c	f7838e5eeb3271df42376952e651e403	delivered	2018-08-19 00:00:00	2018-08-20 00:00:00
003d9fc84ad902adf2265248b5ffe1a4	8c66482ec333eae596c93a9131981a9a	delivered	2018-08-19 00:00:00	2018-08-20 00:00:00
00fc308dd7f0937682698becaa9dcc45	f40f31f8bfcc5f74d9d246354b90af09	delivered	2018-08-20 00:00:00	2018-08-20 00:00:00
0191b6a5a70bcbc6cdd39427f216602e	3f62271cc144183c56e7db438c8cba08	delivered	2018-08-22 00:00:00	2018-08-22 00:00:00
01d9a9aa7feccdaf99f7ea832bfa7301	4e2a379354e75a4545a46fe69949844f	delivered	2018-08-21 00:00:00	2018-08-21 00:00:00
01f3ed0a20c6db39216985812a092bea	88cd4760d06074426489e89bb4116184	delivered	2018-08-18 00:00:00	2018-08-21 00:00:00
02014f2495eef0e869616829d481d743	f0d62fcceaafaea74a143d7fce54f1f4	delivered	2018-08-22 00:00:00	2018-08-22 00:00:00
0229e66693fbbd736e5022221fc8148b	de7123a14192ec3066359926049b0f28	delivered	2018-08-18 00:00:00	2018-08-18 00:00:00
028aa70283170bf3cbfdce2c1b751cfa	f146be55cefde6a5c21dd2c13d699bc8	delivered	2018-08-22 00:00:00	2018-08-22 00:00:00
00345f338696283410b7977d2e3efc89	3f9d223f86d2f243dd5a85fdc286c62a	delivered	2018-08-20 00:00:00	2018-08-20 00:00:00
018bb1508d9156b81990407b91fb35a8	ab65db409c590bb188616224a7e76cf7	delivered	2018-08-15 00:00:00	2018-08-15 00:00:00
00f86b368251d739f1896d41469b2b7a	14c152bfbd60c86aa6bc4d4e68a9378e	delivered	2018-08-17 00:00:00	2018-08-17 00:00:00
01bdc09119b132e176e5c8f2cf68b0ac	9b5f0ef5de7b703b77e1131e68263de9	delivered	2018-08-16 00:00:00	2018-08-17 00:00:00

Figure 4: Imported data for the Order table

4. DW schema

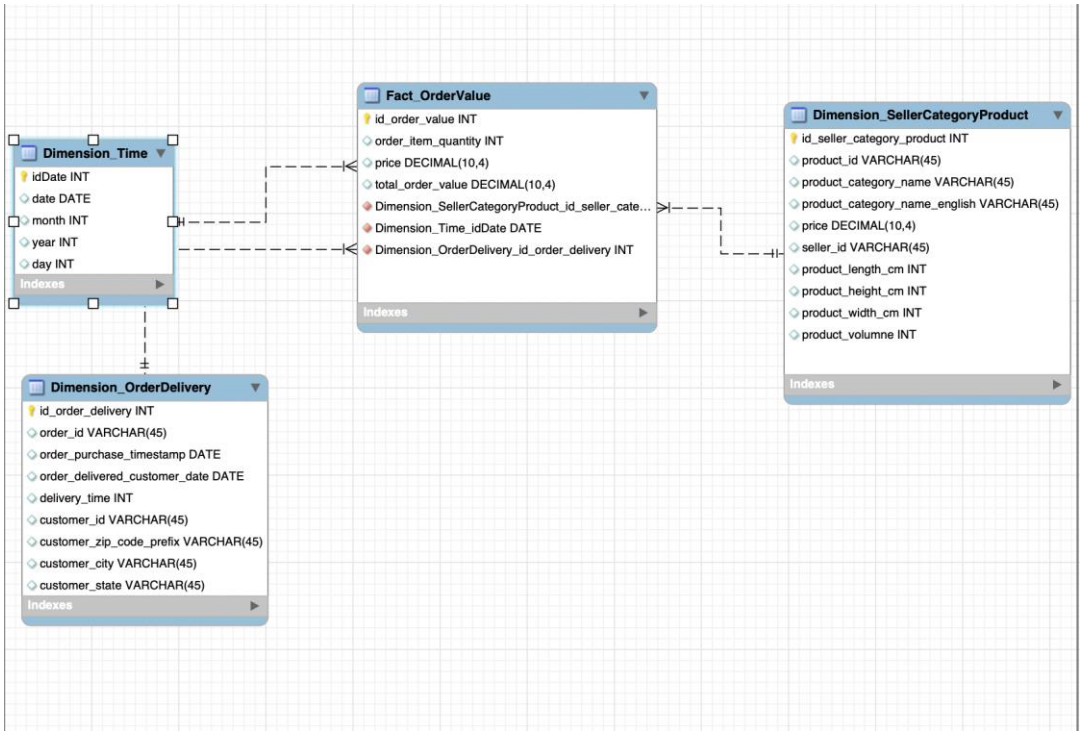


Figure 5: DW schema for Olist

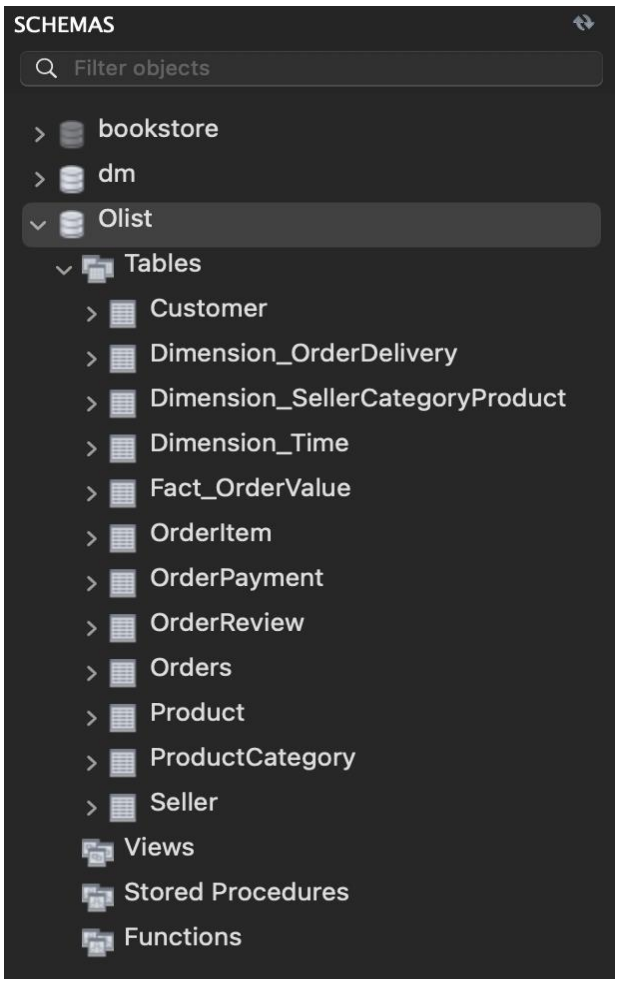


Figure 6: Transferring the DW into Tables

5. Dimensions and Facts of DW

The schema variant chosen is the star variant. The particular variant is chosen for its simplicity and also helps in improving query performance. The structure consists of a single fact table which is Fact_OrderValue which is directly linked to dimension tables which are Dimension_Time, Dimension_OrderDelivery, and Dimension_ SellerCategoryProduct. Therefore the star schema is helpful in efficiency of query and business alignment.

The Dimension_Time represents time and assesses the performance over time. The idDate is generated as a unique identifier for each date row. Other attributes are date, year, month, day.

The Dimension_SellerCategoryProduct describes the product and supplier information. The dimension is the merge of two tables, Seller and Product, and uses its unique key which is id_seller_category_product. It supports analyzing product characteristics, understanding the market demand and identifying suppliers with their diverse product portfolio through product details and sellers' product portfolios.

The Dimension_OrderDelivery gives details about order delivery and customer. The dimension is created by the Orders and Customer tables. The dimension has delivery_time attribute derived from calculation between purchase date delivery date (only working days counted) to track the efficiency of order delivery. By including customer data like customer_city, the dimension helps to analyze regional delivery performance and customer demographics based on the regions.

The fact dimension, Fact_OrderValue, aggregates data to measure metrics as total revenue earning from the total orders and determine patterns in sales over time. The fact table connects to all other dimensions by using their primary keys as the foreign keys. A business can gain insights to enhance logistics, distribute resources effectively, and exploit low-productive locations with an appropriate strategy.

6. ETL transformations in Pentaho Data Integration

Dimension_Time: Use the 'Generate Rows' to create a data range from September 1, 2016 to October 17, 2018 with a limit of 1000 rows. Then, 'Add sequence' to create Unique ID called idDate. Next, use the 'Calculator' step and add idDate to the base date and apply the format yyyy-MM-dd. After that, use the 'Select Values' to filter and select idDate, Date, Year, Month, and Day. Finally, use the 'Table Output' step to write into table output.

Figure 7. Dimension_Time transformation

idDate	date	month	year	day
1	2016-09-02	9	2016	2
2	2016-09-03	9	2016	3
3	2016-09-04	9	2016	4
4	2016-09-05	9	2016	5
5	2016-09-06	9	2016	6
6	2016-09-07	9	2016	7
7	2016-09-08	9	2016	8
8	2016-09-09	9	2016	9
9	2016-09-10	9	2016	10
10	2016-09-11	9	2016	11

Figure 8. Dimension_Time in MySQL

Dimension_OrderDelivery: Use the 'Table Input' to load data from Order and Customer tables. Then use the 'Stream Lookup' to look up customer_id. Then use the 'Select Values' to select values including customer_id and make a sequence using 'Add Sequence'. After, use 'Calculator' to calculate delivery_time by subtracting in format Date A – Date B (working days). Finally, use 'Select Value' to select data including delivery_time and use the 'Table Output' to write into table output.

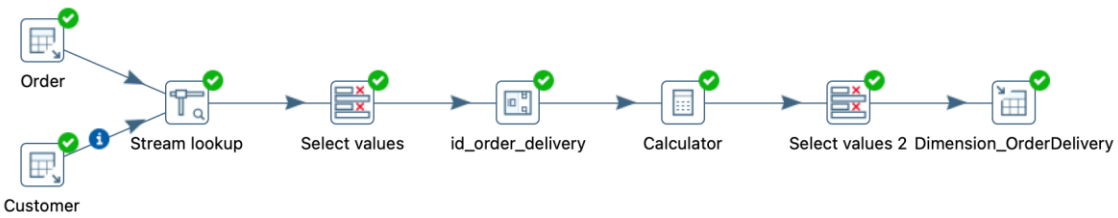


Figure 9. Dimension_OrderDelivery transformation

id_order_delive...	order_id	order_purchase_timesta...	order_delivered_customer_d...	delivery_time	customer_id
1	00010242fe8c5a6d1ba2dd792cb16214	2017-09-13	2017-09-20	6	3ce436f183e68e07877b28
2	00018f77f2f0320c557190d7a144bdd3	2017-04-26	2017-05-12	13	f6dd3ec061db4e3987629f
3	000229ec398224ef6ca0657da4fc703e	2018-01-14	2018-01-22	6	6489ae5e4333f3693df5ad
4	00024acbcdf0a6daa1e931b038114c75	2018-08-08	2018-08-14	5	d4eb9395c8c0431ee92fce
5	00042b26cf59d7ce69dfabb4e55b4fd9	2017-02-04	2017-03-01	18	58dbd0b2d70206bf40e62c
6	00048cc3ae777c65dbb7d2a0634bc1ea	2017-05-15	2017-05-22	6	816cbea969fe5b689b39cf
7	00054e8431b9d7675808bcb819fb4a32	2017-12-10	2017-12-18	6	32e2e6ab09e778d99bf2e0
8	000576fe39319847cbb9d288c5617fa6	2018-07-04	2018-07-09	4	9ed5e522dd9dd85b4af4a0
9	0005a1a1728c9d785b8e2b08b904576c	2018-03-19	2018-03-29	9	16150771dfd4776261284
10	0005f50442cb953dcd1d21e1fb923495	2018-07-02	2018-07-04	3	351d3cb2cee3c7fd0af661

Figure 10. Dimension_OrderDelivery in MySQL

Dimension_SellerCategoryProduct*:* Use the 'Table Input' to load data from Product and Order Item tables. Then use the 'Stream Lookup' to look up product_id. After selecting values, use 'Table Input' for seller tables and use 'Stream Lookup' again to look up 'seller_id'. After, use 'Select Values' to select values including 'product_id' , 'seller_id' and use 'Add Sequence'. Then, we use 'Calculator' twice to calculate Product_volumne in a format A * B. Finally, we use 'Select Values' to select values including product_volumne and use 'Table Output' to write into table output.

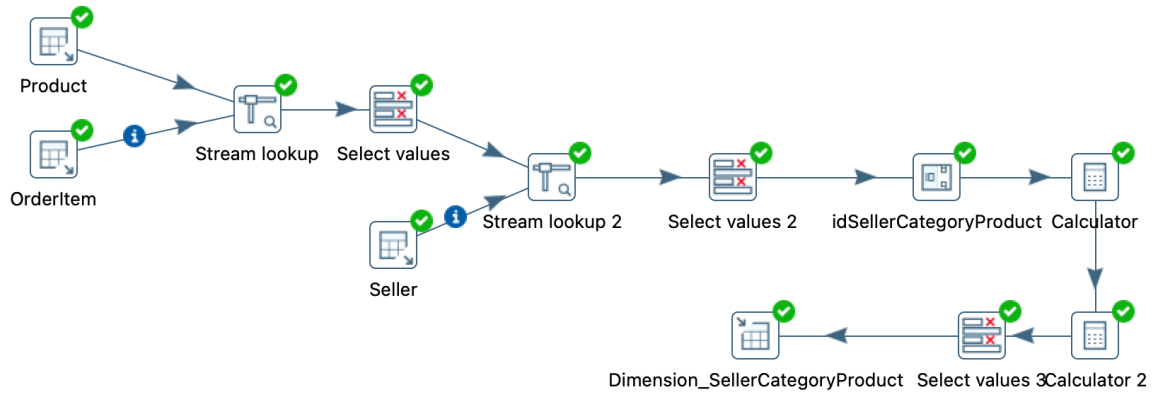


Figure 11. Dimension_SellerCategoryProduct transformation

id_seller_category_prod...	product_id	product_category_na...	price	seller_id	product_lengt
1	00066f42aeeb9f3007548bb9d3f33c38	perfumaria	101.6500	5670f4db5b62c43d542e1b2d56b0cf7c	20
2	00088930e925c41fd95ebfe695fd2655	automotivo	129.9000	7142540dd4c91e2237acb7e911c4eba2	55
3	0009406fd7479715e4bef61dd91f2462	cama_mesa_banho	229.0000	4a3ca9315b744ce9f8e9374361493884	45
4	000b8f95fcb9e0096488278317764d19	utilidades_domesticas	58.9000	40ec8ab6cdafbcc4f544da38c67da39a	19
5	000d9be29b5207b54e86aa1b1ac54872	relogios_presentes	199.0000	8ae520247981aa06bc94abddf5f46d34	22
6	0011c512eb256aa0dbbb544d8dffcf6e	automotivo	52.0000	b4ffb71f0cb1b1c3d63fad021ecf93e1	16
7	00126f27c813603687e6ce486d909d01	cool_stuff	249.0000	cd68562d3f44870c08922d380acae552	25
8	001795ec6f1b187d37335e1c4704762e	consoles_games	38.9000	8b321bb669392f5163d04c59e235e066	30
9	001b237c0e9bb435f2e54071129237e9	cama_mesa_banho	78.9000	d2374cbcbb3ca4ab1086534108cc3ab7	40
10	001b72dfd63e9833e8c02742adf472e3	moveis_decoracao	34.9900	8a32e327fe2c1b3511609d81aaf9f042	26

Figure 12. Dimension_SellerCategoryProduct in MySQL

7. View in MySQL

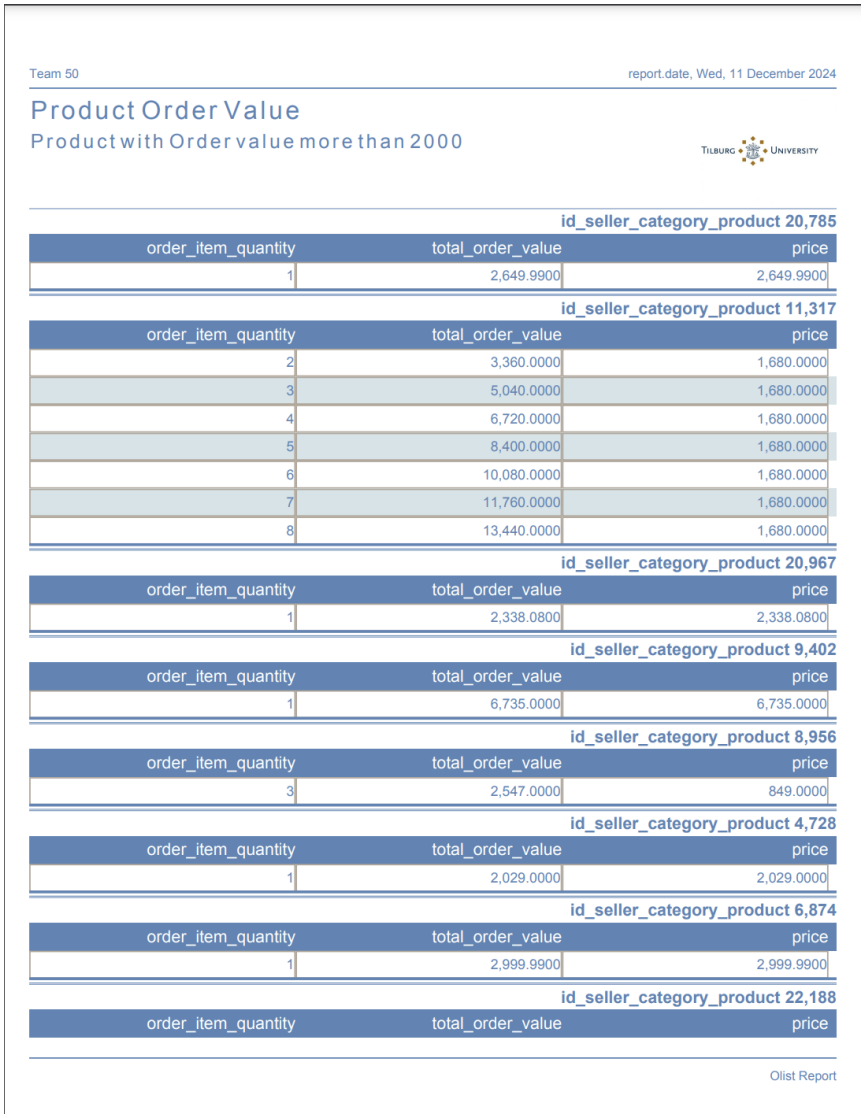
```sql
CREATE VIEW customer_purchases
AS
SELECT id_order_delivery, order_purchase_timestamp, customer_id, customer_city, customer_state
FROM Dimension_OrderDelivery
WHERE (customer_state = 'SP') AND (order_purchase_timestamp BETWEEN '2017-01-01' AND '2017-05-01')
ORDER BY order_purchase_timestamp ASC;

SELECT id_order_delivery, order_purchase_timestamp, customer_id, customer_city, customer_state
FROM customer_purchases;
```

Figure 13. View that contains only orders made by customers from Sao Paulo state (SP) and from January to May 2017, ordered by purchase date.

id_order_delive...	order_purchase_timesta...	customer_id	customer_city	customer_sta...
85344	2017-01-05	758b633d88b82063db189810084f4ea9	ribeirao preto	SP
72057	2017-01-06	96054b94409f7712eeffe8edfa6959a6	valinhos	SP
51644	2017-01-07	d2b141e8cefd8acb97baee4b25b01ea5	sao paulo	SP
44584	2017-01-08	70210c917bc804a1269942a71308a4f7	indaiatuba	SP
66732	2017-01-09	0b8cca6cd998b579f4a0556a31f388f4	bariri	SP
70773	2017-01-10	7cde25498a31a63d00e82b29b929cc49	cardoso	SP
72182	2017-01-11	cd7fc3b7e625b3b0492197326b461c4c	sao bernardo do campo	SP
60053	2017-01-11	f71daf549aff62ad50e839ea1086d9a1	guarulhos	SP
5212	2017-01-12	5a46fcfb9cb1aa7feef0014acace422c	taquarivai	SP
13081	2017-01-12	a76e74102c5779eaca55a3f9187ea2f2	sao carlos	SP

Figure 14. View result

8. Report in Pentaho Report Designer



Team 50 report.date, Wed, 11 December 2024

Product Order Value

Product with Order value more than 2000

Tilburg University

id_seller_category_product 20,785

order_item_quantity	total_order_value	price
1	2,649.9900	2,649.9900

id_seller_category_product 11,317

order_item_quantity	total_order_value	price
2	3,360.0000	1,680.0000
3	5,040.0000	1,680.0000
4	6,720.0000	1,680.0000
5	8,400.0000	1,680.0000
6	10,080.0000	1,680.0000
7	11,760.0000	1,680.0000
8	13,440.0000	1,680.0000

id_seller_category_product 20,967

order_item_quantity	total_order_value	price
1	2,338.0800	2,338.0800

id_seller_category_product 9,402

order_item_quantity	total_order_value	price
1	6,735.0000	6,735.0000

id_seller_category_product 8,956

order_item_quantity	total_order_value	price
3	2,547.0000	849.0000

id_seller_category_product 4,728

order_item_quantity	total_order_value	price
1	2,029.0000	2,029.0000

id_seller_category_product 6,874

order_item_quantity	total_order_value	price
1	2,999.9900	2,999.9900

id_seller_category_product 22,188

order_item_quantity	total_order_value	price

Olist Report

Figure 15. Report of all products which have a total order value higher than 2000.