HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Phân tích và trực quan hóa dữ liệu thương mại điện tử của Olist

GV: Nguyễn Hữu Đức

Đào Duy Anh - 20235000
Lê Hoài Nam - 20235173
Hoàng Tấn Phúc - 20235189
Vũ Hải Minh - 20235166

ONE LOVE. ONE FUTURE.

Mục lục

I. Giới thiệu về dataset

II. Kiến trúc hệ thống

III. Kết luận

I. Giới thiệu về dataset

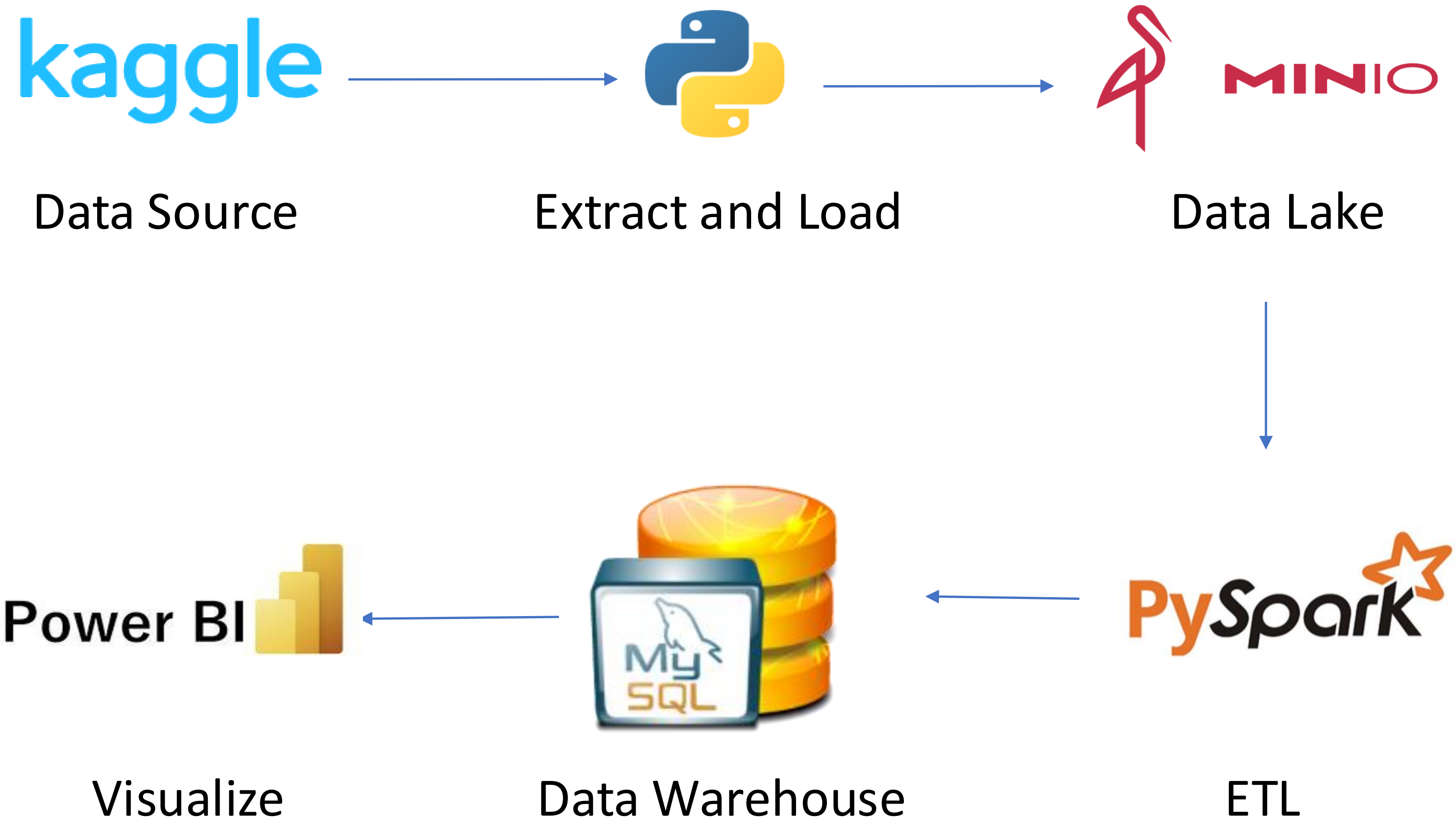
1. Thông tin chung:

- Nguồn: Dữ liệu từ các giao dịch trên trang Olist Store.
- Thời gian: Từ năm 2016 – 2018.
- Quy mô: Khoảng 100.000 đơn hàng.

2. Cấu trúc dữ liệu: 9 bảng

- olist_orders_dataset
- olist_order_items_dataset
- olist_order_payments_dataset
- olist_order_reviews_dataset
- olist_products_dataset
- olist_customers_dataset
- olist_sellers_dataset
- olist_geolocation_dataset
- product_category_name_translation

II. Kiến trúc hệ thống



Data Source

Extract and Load

Data Lake

Visualize

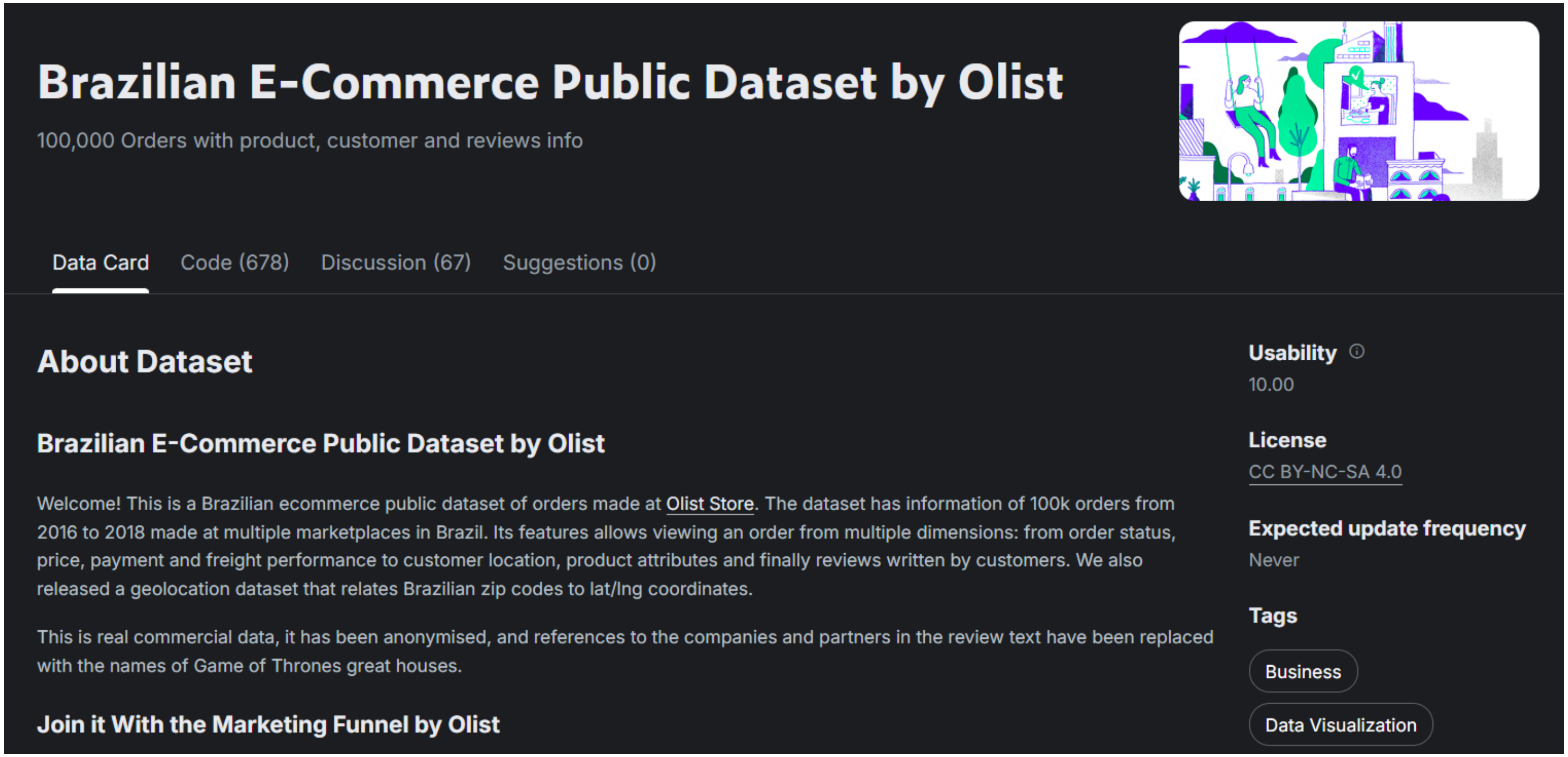
Data Warehouse

ETL

II. Kiến trúc hệ thống

1. Data Source:

- Link bộ dữ liệu: [Brazilian E-Commerce Public Dataset by Olist]



Brazilian E-Commerce Public Dataset by Olist

100,000 Orders with product, customer and reviews info

About Dataset

Brazilian E-Commerce Public Dataset by Olist

Welcome! 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.

Join it With the Marketing Funnel by Olist

Usability ⓘ
10.00

License
CC BY-NC-SA 4.0

Expected update frequency
Never

Tags

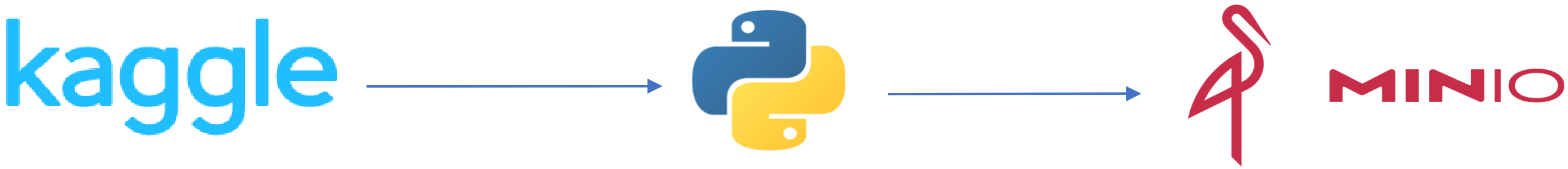
Business

Data Visualization

II. Kiến trúc hệ thống

2. Extract and Load:

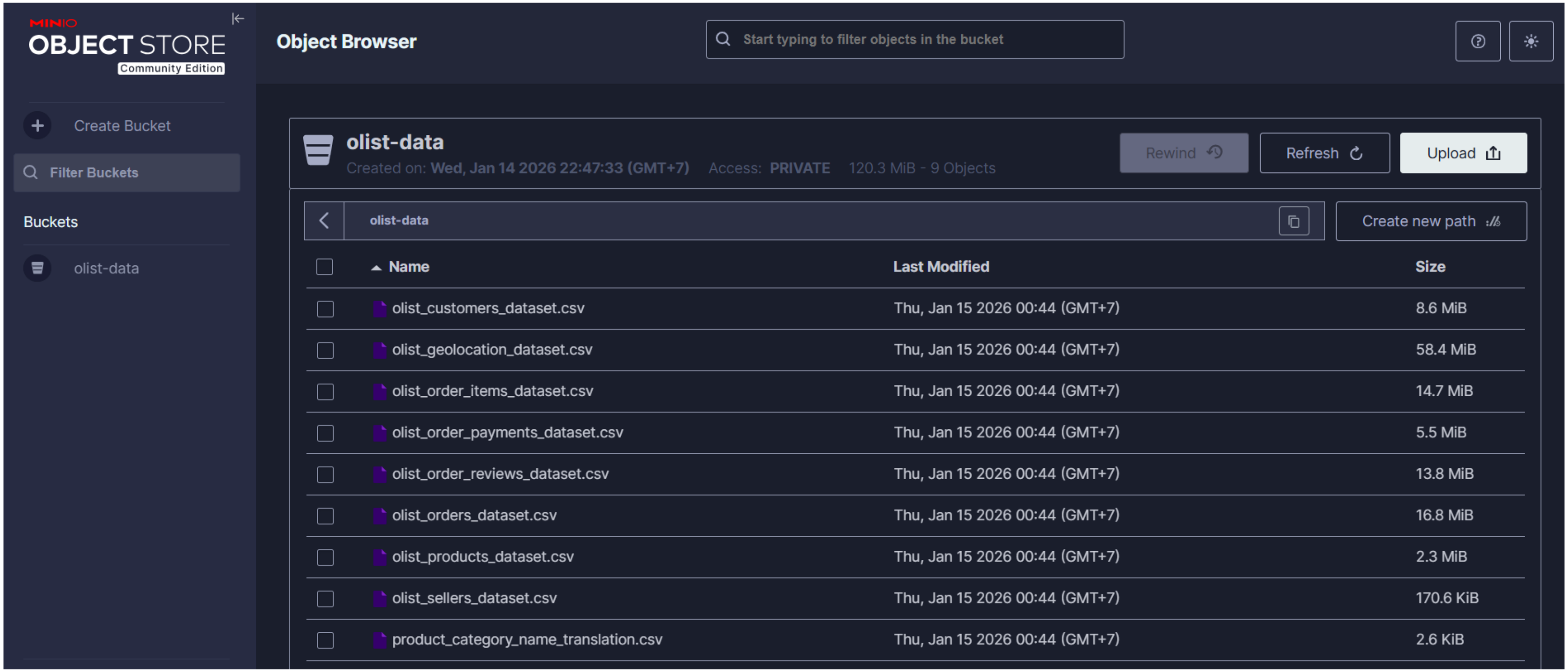
- Sử dụng thư viện Kaggle API để tải dữ liệu về máy local.
- Sử dụng thư viện boto3 để upload các file csv vào bucket.



II. Kiến trúc hệ thống

3. Data Lake:

- Lưu trữ các file csv đã được upload.



olist-data

Created on: Wed, Jan 14 2026 22:47:33 (GMT+7) Access: PRIVATE 120.3 MiB - 9 Objects

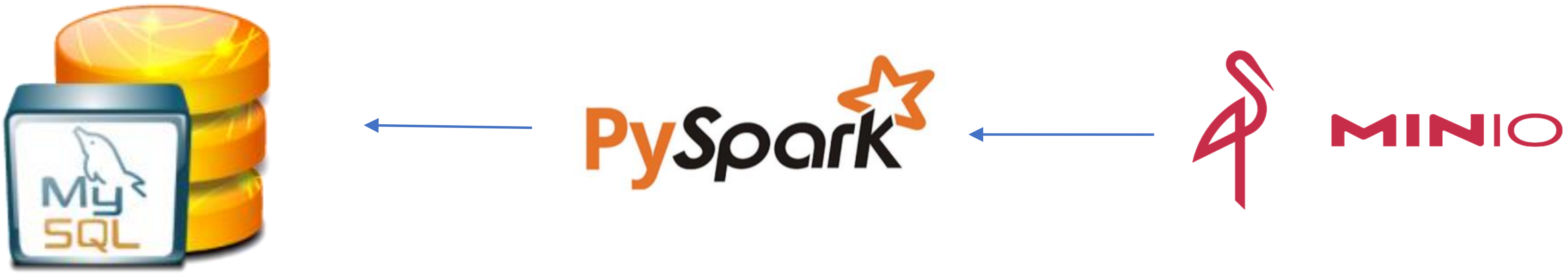
olist-data

☐	Name	Last Modified	Size
☐	olist_customers_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	8.6 MiB
☐	olist_geolocation_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	58.4 MiB
☐	olist_order_items_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	14.7 MiB
☐	olist_order_payments_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	5.5 MiB
☐	olist_order_reviews_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	13.8 MiB
☐	olist_orders_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	16.8 MiB
☐	olist_products_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	2.3 MiB
☐	olist_sellers_dataset.csv	Thu, Jan 15 2026 00:44 (GMT+7)	170.6 KiB
☐	product_category_name_translation.csv	Thu, Jan 15 2026 00:44 (GMT+7)	2.6 KiB

II. Kiến trúc hệ thống

4. ETL:

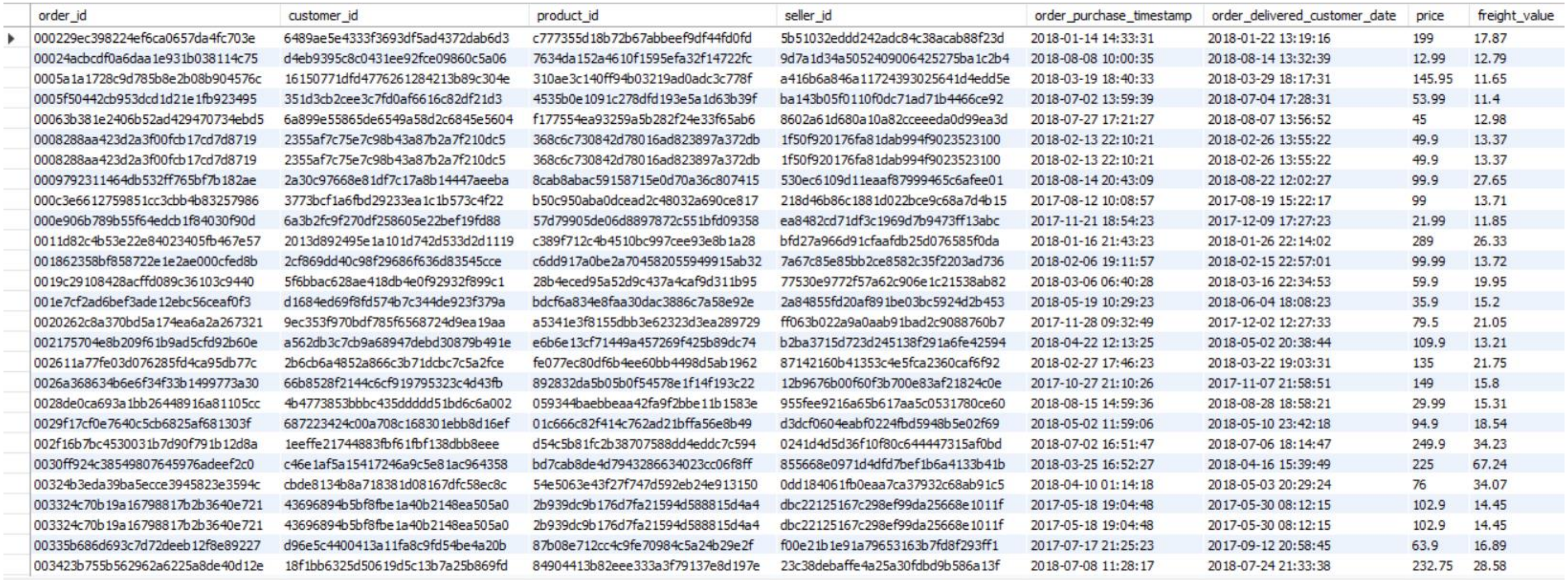
- Làm sạch dữ liệu: Dịch tên sản phẩm, chuẩn hóa tọa độ,…
- Làm giàu dữ liệu: Tính toán thời gian giao hàng, tổng giá trị đơn hàng,…
- Mô hình hóa dữ liệu.



II. Kiến trúc hệ thống

5. Data Warehouse:

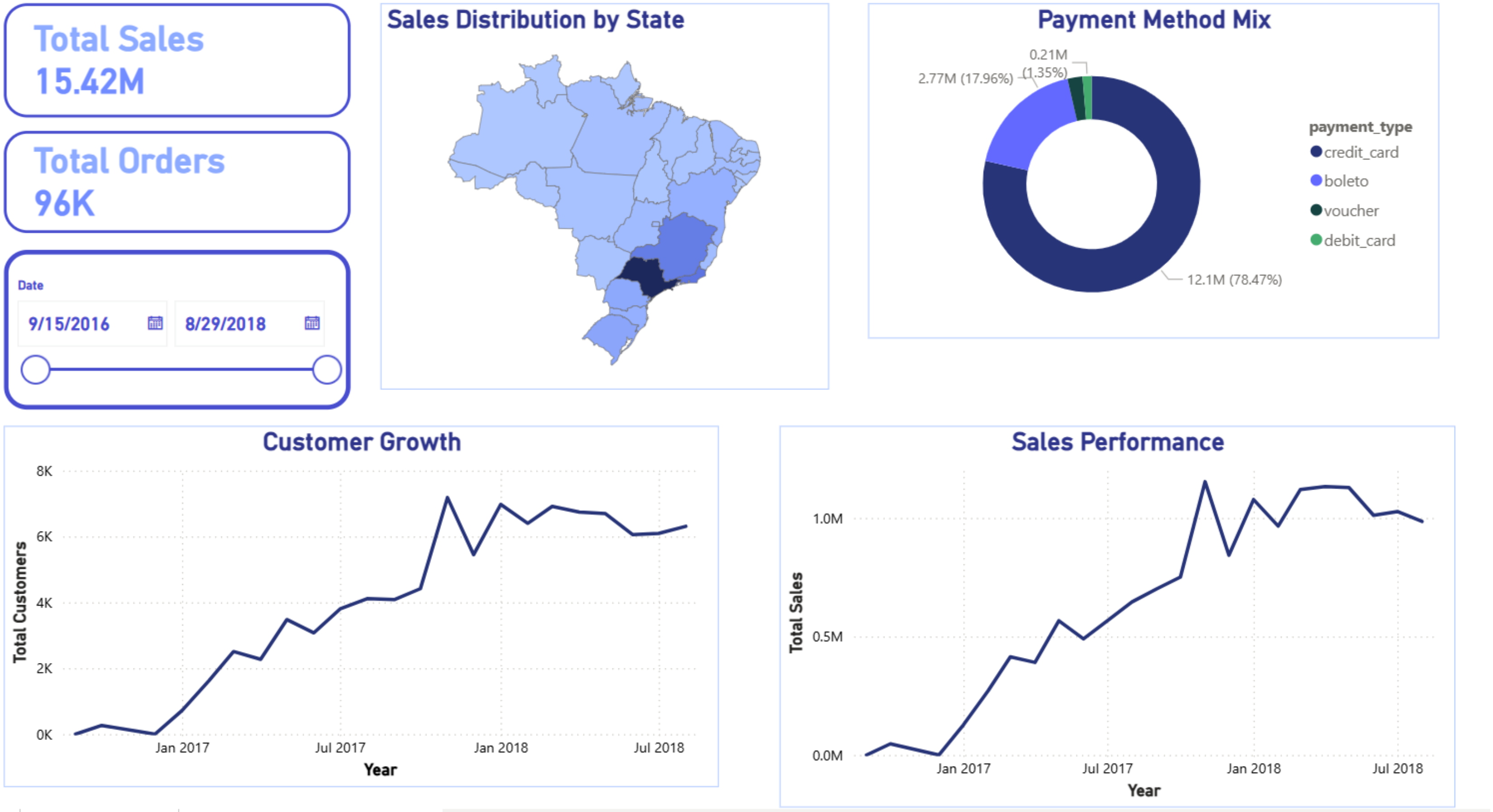
- Cloud database được cung cấp bởi Aiven.



	order_id	customer_id	product_id	seller_id	order_purchase_timestamp	order_delivered_customer_date	price	freight_value
▶	000229ec398224ef6ca0657da4fc703e	6489ae5e4333f3693df5ad4372dab6d3	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-14 14:33:31	2018-01-22 13:19:16	199	17.87
	00024acbcdf0a6daa1e931b038114c75	d4eb9395c8c0431ee92fce09860c5a06	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	2018-08-08 10:00:35	2018-08-14 13:32:39	12.99	12.79
	0005a1a1728c9d785b8e2b08b904576c	16150771dfd4776261284213b89c304e	310ae3c140ff94b03219ad0adc3c778f	a416b6a846a11724393025641d4edd5e	2018-03-19 18:40:33	2018-03-29 18:17:31	145.95	11.65
	0005f50442cb953dcd1d21e1fb923495	351d3cb2cee3c7fd0af6616c82df21d3	4535b0e1091c278dfd193e5a1d63b39f	ba143b05f0110f0dc71ad71b4466ce92	2018-07-02 13:59:39	2018-07-04 17:28:31	53.99	11.4
	00063b381e2406b52ad429470734ebd5	6a899e55865de6549a58d2c6845e5604	f177554ea93259a5b282f24e33f65ab6	8602a61d680a10a82cceeeda0d99ea3d	2018-07-27 17:21:27	2018-08-07 13:56:52	45	12.98
	0008288aa423d2a3f00fcb17cd7d8719	2355af7c75e7c98b43a87b2a7f210dc5	368c6c730842d78016ad823897a372db	1f50f920176fa81dab994f9023523100	2018-02-13 22:10:21	2018-02-26 13:55:22	49.9	13.37
	0008288aa423d2a3f00fcb17cd7d8719	2355af7c75e7c98b43a87b2a7f210dc5	368c6c730842d78016ad823897a372db	1f50f920176fa81dab994f9023523100	2018-02-13 22:10:21	2018-02-26 13:55:22	49.9	13.37
	0009792311464db532ff765bf7b182ae	2a30c97668e81df7c17a8b14447aeeba	8cab8abac59158715e0d70a36c807415	530ec6109d11eaaf87999465c6afee01	2018-08-14 20:43:09	2018-08-22 12:02:27	99.9	27.65
	000c3e6612759851cc3cbb4b83257986	3773bcf1a6fbd29233ea1c1b573c4f22	b50c950aba0dcead2c48032a690ce817	218d46b86c1881d022bce9c68a7d4b15	2017-08-12 10:08:57	2017-08-19 15:22:17	99	13.71
	000e906b789b55f64edcb1f84030f90d	6a3b2fc9f270df258605e22bef19fd88	57d79905de06d8897872c551bfd09358	ea8482cd71df3c1969d7b9473ff13abc	2017-11-21 18:54:23	2017-12-09 17:27:23	21.99	11.85
	0011d82c4b53e22e84023405fb467e57	2013d892495e1a101d742d533d2d1119	c389f712c4b4510bc997cee93e8b1a28	bfd27a966d91cfaafdb25d076585f0da	2018-01-16 21:43:23	2018-01-26 22:14:02	289	26.33
	001862358bf858722e1e2ae000cfed8b	2cf869dd40c98f29686f636d83545cce	c6dd917a0be2a704582055949915ab32	7a67c85e85bb2ce8582c35f2203ad736	2018-02-06 19:11:57	2018-02-15 22:57:01	99.99	13.72
	0019c29108428acffd089c36103c9440	5f6bbac628ae418db4e0f92932f899c1	28b4eced95a52d9c437a4caf9d311b95	77530e9772f57a62c906e1c21538ab82	2018-03-06 06:40:28	2018-03-16 22:34:53	59.9	19.95
	001e7cf2ad6bef3ade12ebc56ceaf0f3	d1684ed69f8fd574b7c344de923f379a	bdcf6a834e8faa30dac3886c7a58e92e	2a84855fd20af891be03bc5924d2b453	2018-05-19 10:29:23	2018-06-04 18:08:23	35.9	15.2
	0020262c8a370bd5a174ea6a2a267321	9ec353f970bdf785f6568724d9ea19aa	a5341e3f8155dbb3e62323d3ea289729	ff063b022a9a0aab91bad2c9088760b7	2017-11-28 09:32:49	2017-12-02 12:27:33	79.5	21.05
	002175704e8b209f61b9ad5cfd92b60e	a562db3c7cb9a68947debd30879b491e	e6b6e13cf71449a457269f425b89dc74	b2ba3715d723d245138f291a6fe42594	2018-04-22 12:13:25	2018-05-02 20:38:44	109.9	13.21
	002611a77fe03d076285fd4ca95db77c	2b6cb6a4852a866c3b71dcbc7c5a2fce	fe077ec80df6b4ee60bb4498d5ab1962	87142160b41353c4e5fca2360caf6f92	2018-02-27 17:46:23	2018-03-22 19:03:31	135	21.75
	0026a368634b6e6f34f33b1499773a30	66b8528f2144c6cf919795323c4d43fb	892832da5b05b0f54578e1f14f193c22	12b9676b00f60f3b700e83af21824c0e	2017-10-27 21:10:26	2017-11-07 21:58:51	149	15.8
	0028de0ca693a1bb26448916a81105cc	4b4773853bbbc435ddddd51bd6c6a002	059344baebbeaa42fa9f2bbe11b1583e	955fee9216a65b617aa5c0531780ce60	2018-08-15 14:59:36	2018-08-28 18:58:21	29.99	15.31
	0029f17cf0e7640c5cb6825af681303f	687223424c00a708c168301ebb8d16ef	01c666c82f414c762ad21bffa56e8b49	d3dcf0604eabf0224fbd5948b5e02f69	2018-05-02 11:59:06	2018-05-10 23:42:18	94.9	18.54
	002f16b7bc4530031b7d90f791b12d8a	1eeffe21744883fbf61fbf138dbb8eee	d54c5b81fc2b38707588dd4eddc7c594	0241d4d5d36f10f80c644447315af0bd	2018-07-02 16:51:47	2018-07-06 18:14:47	249.9	34.23
	0030ff924c38549807645976adeef2c0	c46e1af5a15417246a9c5e81ac964358	bd7cab8de4d794328663402 3cc06f8ff	855668e0971d4dfd7bef1b6a4133b41b	2018-03-25 16:52:27	2018-04-16 15:39:49	225	67.24
	00324b3eda39ba5ecce3945823e3594c	cbde8134b8a718381d08167dfc58ec8c	54e5063e43f27f747d592eb24e913150	0dd184061fb0eaa7ca37932c68ab91c5	2018-04-10 01:14:18	2018-05-03 20:29:24	76	34.07
	003324c70b19a16798817b2b3640e721	43696894b5bf8fbe1a40b2148ea505a0	2b939dc9b176d7fa21594d588815d4a4	dbc22125167c298ef99da25668e1011f	2017-05-18 19:04:48	2017-05-30 08:12:15	102.9	14.45
	003324c70b19a16798817b2b3640e721	43696894b5bf8fbe1a40b2148ea505a0	2b939dc9b176d7fa21594d588815d4a4	dbc22125167c298ef99da25668e1011f	2017-05-18 19:04:48	2017-05-30 08:12:15	102.9	14.45
	00335b686d693c7d72deeb12f8e89227	d96e5c4400413a11fa8c9fd54be4a20b	87b08e712cc4c9fe70984c5a24b29e2f	f00e21b1e91a79653163b7fd8f293ff1	2017-07-17 21:25:23	2017-09-12 20:58:45	63.9	16.89
	003423b755b562962a6225a8de40d12e	18f1bb6325d50619d5c13b7a25b869fd	84904413b82eee333a3f79137e8d197e	23c38debaffe4a25a30fdbd9b586a13f	2018-07-08 11:28:17	2018-07-24 21:33:38	232.75	28.58

II. Kiến trúc hệ thống

6. Visualize:



Total Sales
15.42M

Total Orders
96K

Date
9/15/2016 8/29/2018

III. Kết luận

- **Kết quả đạt được:** Hệ thống vận hành trơn tru từ Local lên Cloud.

- **Hướng phát triển:** Tự động hóa hoàn toàn quy trình bằng Airflow hoặc tích hợp thêm các mô hình Machine Learning.

hust.edu.vn fb.com/dhbkhn

THANK YOU !