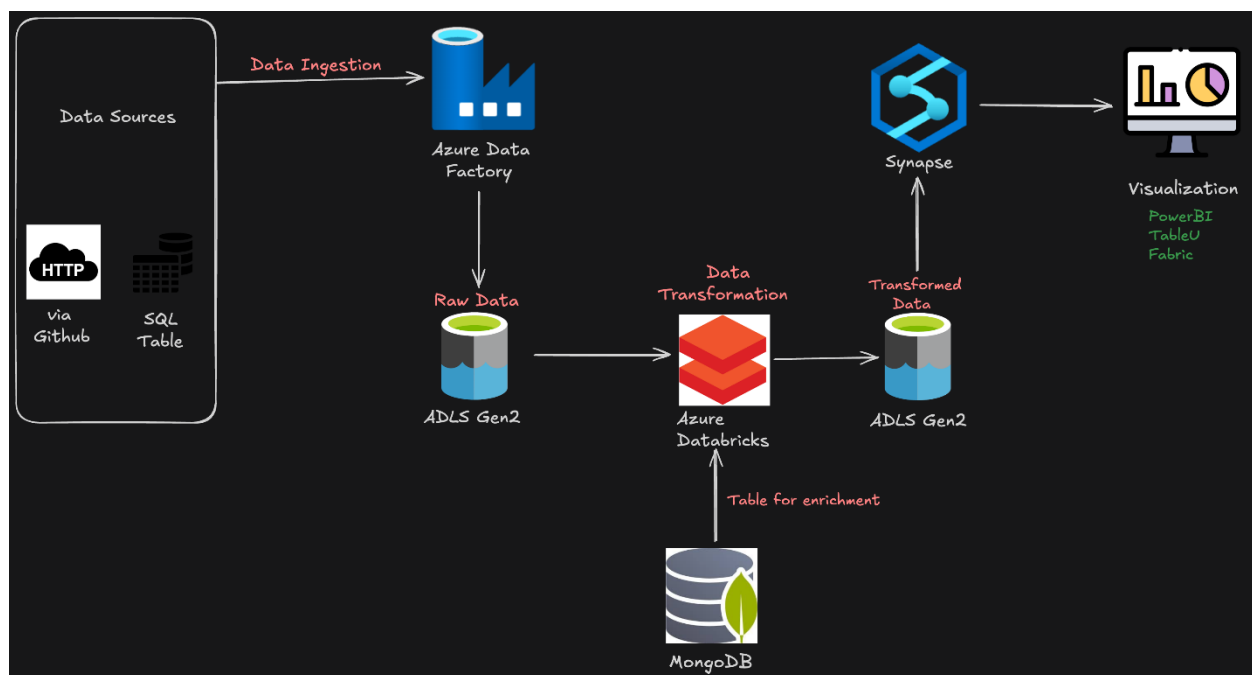


End-to-End Data Engineering Project

Project Architecture Overview:

- Complete E-commerce data pipeline implementation
- Azure-powered cloud architecture
- Real-world data handling techniques
- Industry-standard best practices

Architecture:














Dataset:

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

Multiple Data Sources:

1. HTTP (git hub):

<https://github.com/NisanthTumu/End-to-End-Data-Engineering-Project--olist-/tree/main/Data>

 NisanthTumu Final Push 3b1d149 · 34 minutes	
Name	Last commit message
 ..	
 olist_customers_dataset.csv	Final Push
 olist_geolocation_dataset.csv	Final Push
 olist_order_items_dataset.csv	Final Push
 olist_order_payments_dataset.csv	Final Push
 olist_order_reviews_dataset.csv	Final Push
 olist_orders_dataset.csv	Final Push
 olist_products_dataset.csv	Final Push
 olist_sellers_dataset.csv	Final Push
 product_category_name_translation.csv	Final Push

2. SQL and NoSQL Databases:

From Files.io

Databases







+ New Database

✓ Standard Databases

★ Dedicated Databases New

i Standard VS Dedicated

Q Search DB name...

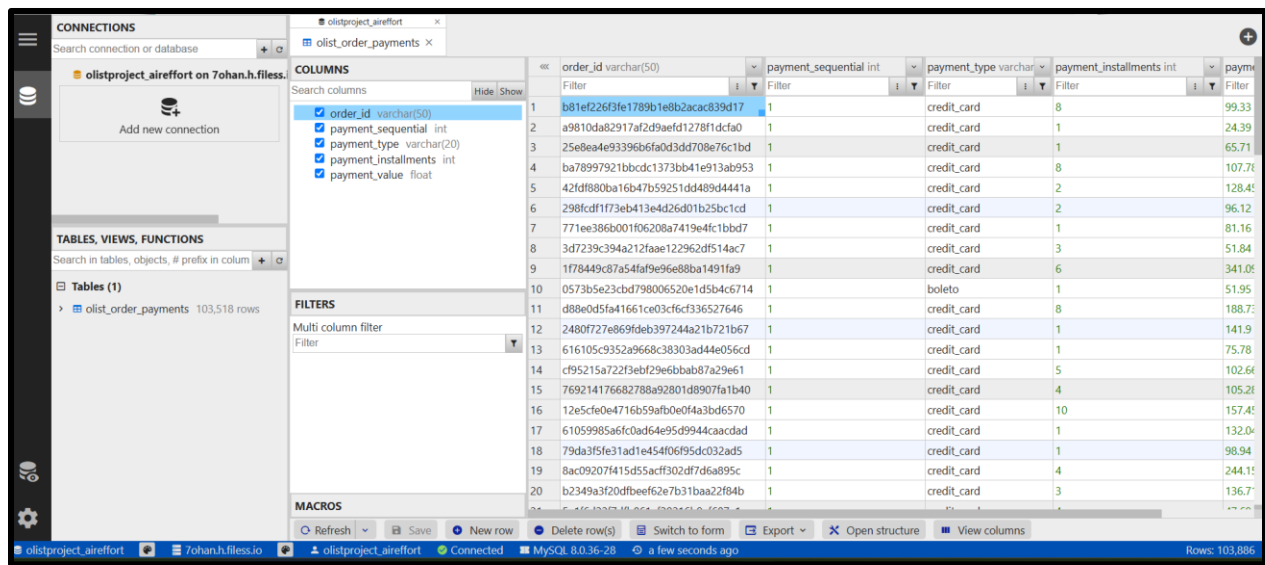
	Motor	Identifier	Available	Location	
	 v8.0.29	olistproject	Yes		⋮
	 v7.0.2	olistprojectNoSQL	Yes		⋮

Rows per page: 5

1-2 of 2

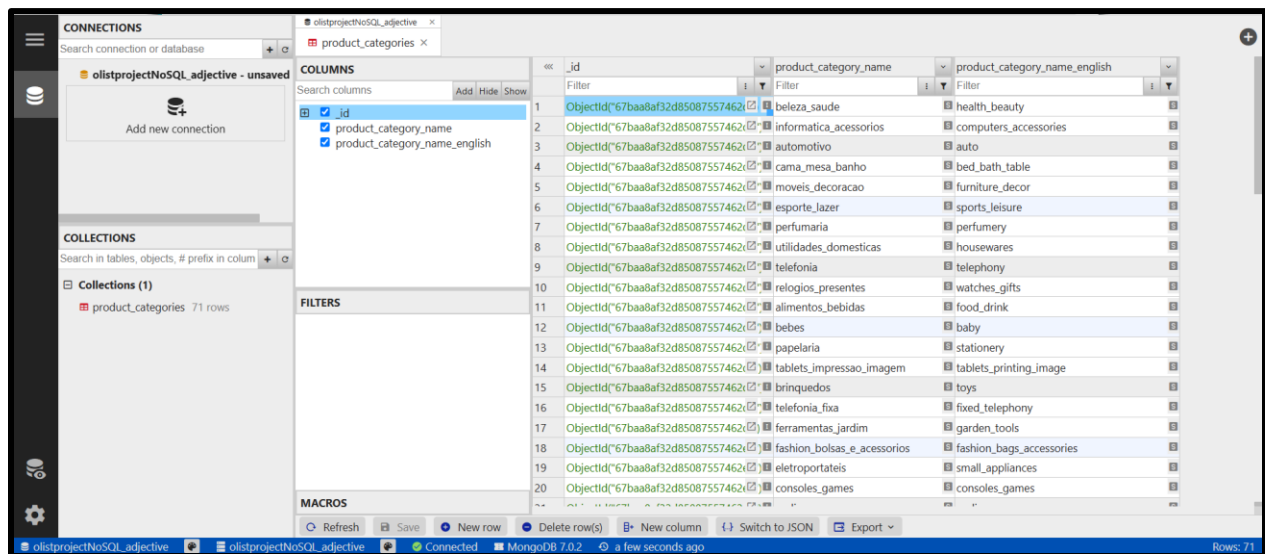
< >

SQL Database:



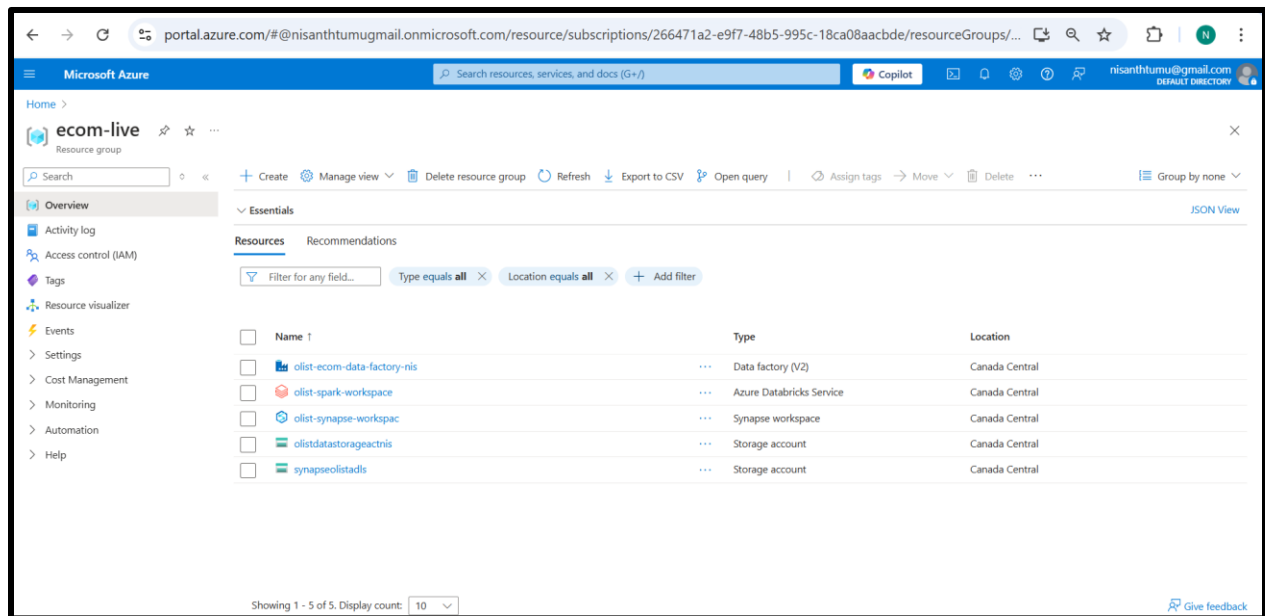
order_id	payment_sequential	payment_type	payment_installments	payment_value
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
a9810da82917af2d9aef1278f1dcfa0	1	credit_card	1	24.39
25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71
ba78997921bbcd1373bb41e913ab953	1	credit_card	8	107.78
42fd880ba16b47b59251dd489d4441a	1	credit_card	2	128.45
298fcd1f73eb413e4d26d01b25bc1cd	2	credit_card	2	96.12
771ee386b001f06208a7419e4fc1bbd7	1	credit_card	1	81.16
3d7239c394a212faae122962df514ac7	1	credit_card	3	51.84
1f78449c87a54fa9e9e688ba1491fa9	1	credit_card	6	341.05
0573b5e23cbd798006520e1d5b4c6714	1	boleto	1	51.95
d88e0d5fa41661ce03cf6d336527646	1	credit_card	8	188.73
2480f727e869deb397244a21b721b67	1	credit_card	1	141.9
616105c9352a9668c38303ad44e05ecd	1	credit_card	1	75.78
cf95215a72f3ebf29e6bbab87a29e61	1	credit_card	5	102.66
769214176682788a92801d8907fa1b40	1	credit_card	4	105.28
12e5cfe0e4716b59afb0e0f4a3bd6570	1	credit_card	10	157.45
61059985a6fc0ad64e95d9944caacdad	1	credit_card	1	132.04
79da3f5fe31ad1e454f0e695dc032ad5	1	credit_card	1	98.94
8ac09207f415d55acf302df7d6a895c	1	credit_card	4	244.15
b2349a3f20dfbeef62e7b31baa2f84b	1	credit_card	3	136.7

NoSql Database:

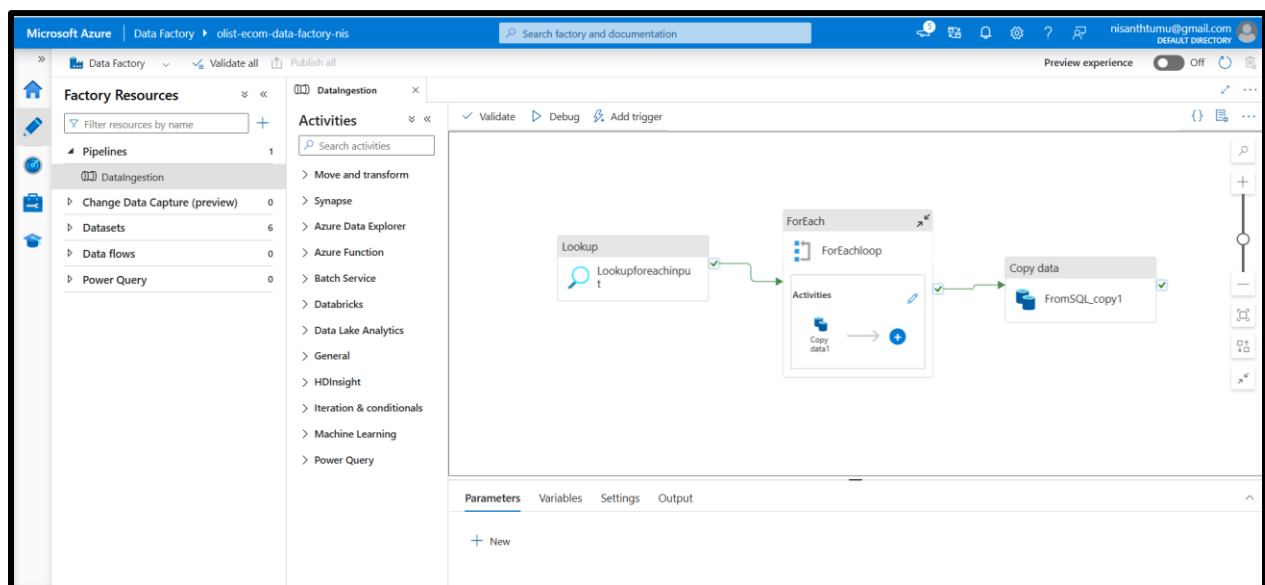


_id	product_category_name	product_category_name_english
Objectid("67baa8af32d85087557462")	beleza_saude	health_beauty
Objectid("67baa8af32d85087557462")	informatica_acessorios	computers_accessories
Objectid("67baa8af32d85087557462")	automotivo	auto
Objectid("67baa8af32d85087557462")	cama_mesa_banho	bed_bath_table
Objectid("67baa8af32d85087557462")	moveis_decoracao	furniture_decor
Objectid("67baa8af32d85087557462")	esporte_lazer	sports_leisure
Objectid("67baa8af32d85087557462")	perfumaria	perfumery
Objectid("67baa8af32d85087557462")	utilidades_domesticas	housewares
Objectid("67baa8af32d85087557462")	telefonia	telephony
Objectid("67baa8af32d85087557462")	relorios_presentes	watches_gifts
Objectid("67baa8af32d85087557462")	alimentos_bebidas	food_drink
Objectid("67baa8af32d85087557462")	bebes	baby
Objectid("67baa8af32d85087557462")	papelaria	stationery
Objectid("67baa8af32d85087557462")	tablets_impresao_imagem	tablets_printing_image
Objectid("67baa8af32d85087557462")	brinquedos	toys
Objectid("67baa8af32d85087557462")	telefonos_fixa	fixed_telephony
Objectid("67baa8af32d85087557462")	ferramentas_jardim	garden_tools
Objectid("67baa8af32d85087557462")	fashion_bolsas_e_acessorios	fashion_bags_accessories
Objectid("67baa8af32d85087557462")	eletroportateis	small_appliances
Objectid("67baa8af32d85087557462")	console_games	console_games

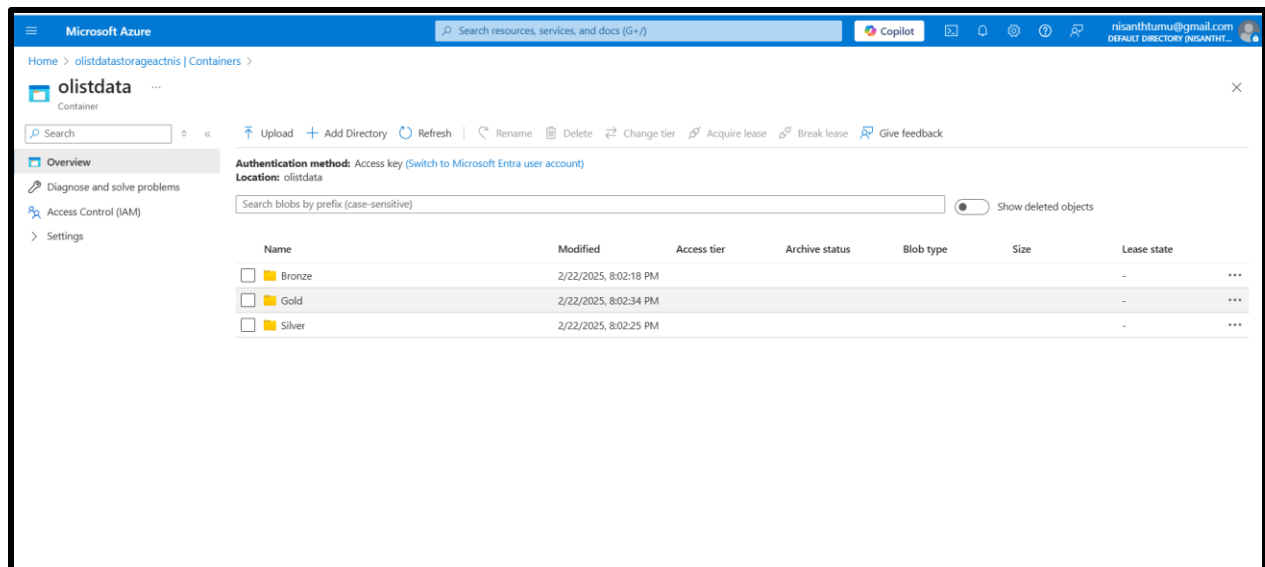
Azure Setup:



DataFactory Pipeline:

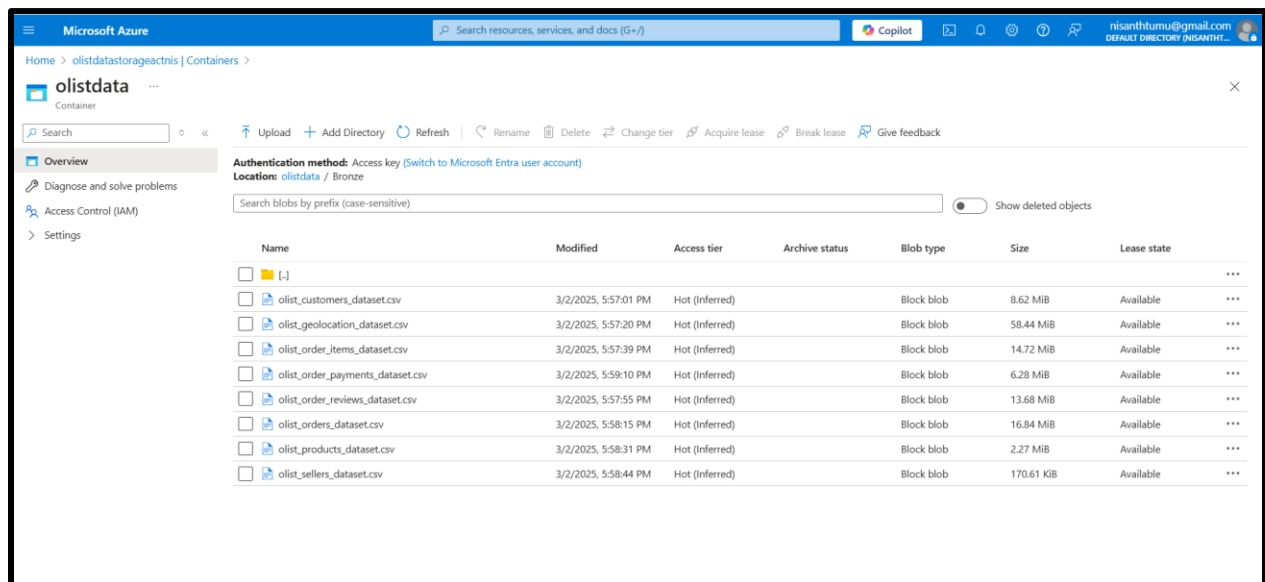


Data Storage account: (Medallion Architecture)



Bronze Layer:

After Data Ingestion from ETL Pipeline with multiple data sources(HTTP and SQL Database)



Silver Layer:

After data wrangling and enrichment (from NoSQL Database) using Azure Databricks

Microsoft Azure

Search resources, services, and docs (G+J)

Copilot

nisanthtumu@gmail.com
DEFAULT DIRECTORY (NISANTH...

Home >

olist-spark-workspace Azure Databricks Service

Search

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Settings

Monitoring

Automation

Help

Essentials

Status : Active

Resource group : ecom-live

Location : Canada Central

Subscription : Azure subscription 1

Subscription ID : 266471a2-e9f7-48b5-995c-18ca08aacbde

Tags (edit) : Add tags

Managed Resource Group : ecom-databricks-resource-group

URL : https://adb-2442505950395105.5.azuredatabricks.net

Pricing Tier : Premium (+ Role-based access controls) (Click to change)

Launch Workspace

Documentation

Getting Started

Import Data from File

Import Data from Azure Storage

Notebook

Admin Guide

Link Azure ML workspace

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

olist-spark-workspace

New

Workspace

Recents

Catalog

Workflows

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Pipelines

Machine Learning

Playground

DataBricks Code Python

File Edit View Run Help Last edit was 25 days ago

orders.payments_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 19 more fields]

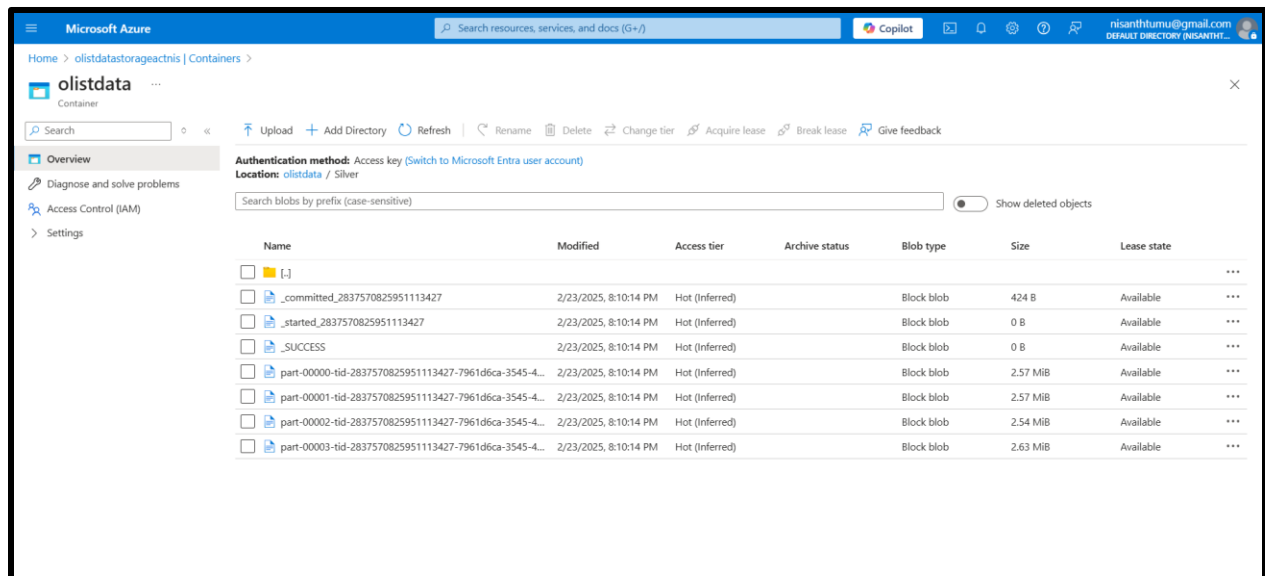
Run all Starting Schedule Share

Feb 23, 2025 (4)

display(final_df)

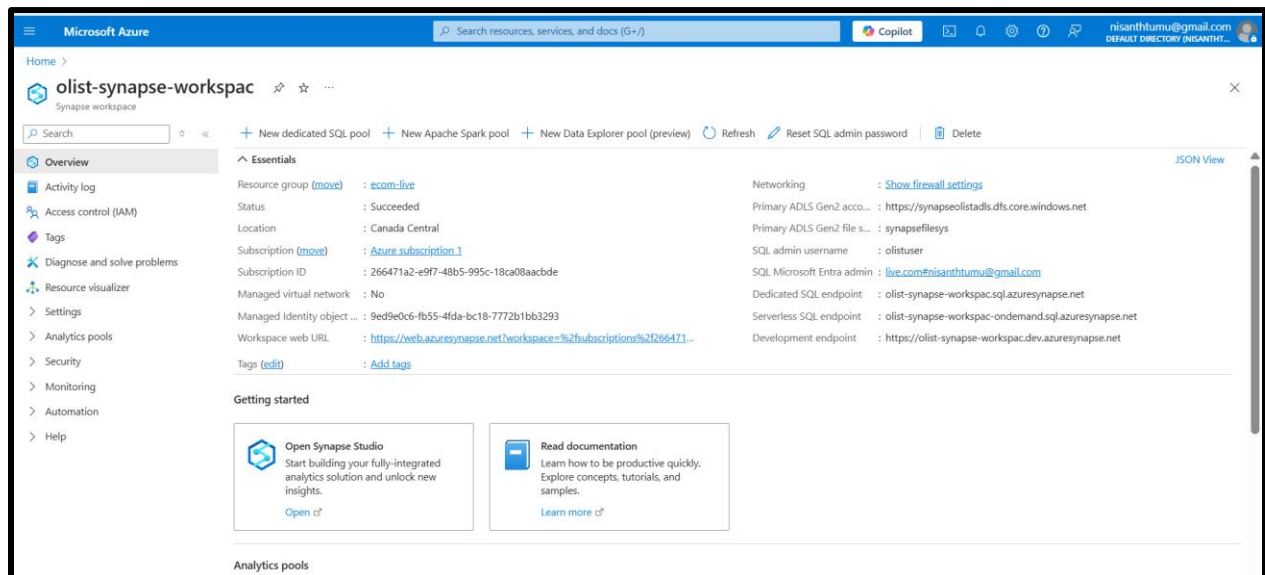
(7) Spark Jobs

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	orde
1	acce194856392f074dbf9dada14db8b2	7e20bf5ca92da58200643bda76c504c6	delivered	2018-06-04	2018-06-05T00:35:10.000+00...	2018-06
2	1d067305b599c1e0dceb3864056ea527	0489975a325480c9e385e9f135bb13c3	delivered	2018-02-14	2018-02-14T13:15:38.000+00...	2018-02
3	6942b8da583c2f9957e99d028607019	52006a9383bf149a4fb24226b173106f	shipped	2018-01-10	2018-01-11T02:32:30.000+00...	2018-01
4	6f841dde94727854eaff3f66432c00ea	a9c9532060c9d245f06526c633a2dfba	delivered	2018-01-02	2018-01-02T19:32:22.000+00...	2018-01
5	ca290a06ee0945b956f79c93b5191633	acd575d738296888941a5a3f37510dd	delivered	2018-04-09	2018-04-09T23:10:27.000+00...	2018-04
6	3735720b6c1cd4212fa0866a8e58a049	94af59d9cac1ae1976312504628ef6f	delivered	2018-05-16	2018-05-16T23:55:11.000+00...	2018-05
7	8b346abc34a6e64bc67fcd3b0eccdc9f	dd1506d329d9b0135855082aae3c79ac	delivered	2018-07-11	2018-07-11T20:26:13.000+00...	2018-07
8	a4d7c8bca45b56444e3c59ddcca7d7c9	fcdb673a0f8d2b84d3862193f17a08f9	delivered	2017-08-29	2017-08-30T02:05:44.000+00...	2017-09
9	b64b1539563f5f15922bd124fd838863	a22203b8cd7210764aa3f442904a076	delivered	2017-05-06	2017-05-06T12:47:06.000+00...	2017-05
10	e05ad3bb40dd7a1f005e828e86493bcc	348da3f77102be30559eb101f7f96	delivered	2018-05-24	2018-05-25T02:55:02.000+00...	2018-05
11	0c00bf3feaced2a198154419c788fc5	a4f809d1f8681f7b4c03a3294528963	delivered	2017-03-20	2017-03-20T17:15:11.000+00...	2017-03
12	69b4682d3ab5ef1f14a8e9e74638f87e	765b56e9a7716dfe3074a58a8aac5974	delivered	2018-08-07	2018-08-08T03:05:48.000+00...	2018-08
13	f222c58035b47dfa1e069a88235d730	b74ca180a63f9ae0443e4e13a2f5bdaf	delivered	2018-01-30	2018-02-04T23:31:47.000+00...	2018-01
14	341dc4ea81fab7de6af564b72fcd3e2	8a8c2e93d6ec33bb908f101211d78227	delivered	2018-07-04	2018-07-05T16:30:57.000+00...	2018-07
15						



Gold Layer:

After Data Transformation from Azure Synapse Workspace



Microsoft Azure | Synapse Analytics | olist-synapse-workspace

Synapse live | Validate all | Publish all

Develop

Filter resources by name

SQL scripts 4

- create View
- Create CETAS
- Create Username
- SQL script Create Username

create View | Create CETAS | Create Username | SQL script 1

Run | Undo | Publish | Query plan | Connect to Built-in | Use database olist

```
1 CREATE EXTERNAL FILE FORMAT extfileformat WITH (
2   FORMAT_TYPE = PARQUET,
3   DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
4 );
5
6 CREATE EXTERNAL DATA SOURCE goldlayer WITH (
7   LOCATION = 'https://olistdatastorageactnis.dfs.core.windows.net/olistdata/Gold/',
8   CREDENTIAL = nisanthadmin
9 );
10
11
12 CREATE EXTERNAL TABLE gold.finaltable WITH (
13   LOCATION = 'Serving',
14   DATA_SOURCE = goldlayer,
15   FILE_FORMAT = extfileformat
16 ) AS
```

Results | Messages

View Table | Chart | Export results

Search

product_cat...	order_status	order_purchas...	order_approve...	order_delivere...	order_delivere...	order_estimate...	actual_delivery...	estimated_deli...	Delay Time	customer_uni...
bebes	delivered	2018-06-04	2018-06-05T00:...	2018-06-05T13:...	2018-06-16	2018-07-18	12	44	-32	576ea0cab426
cama_mesa_ba...	delivered	2018-02-14	2018-02-14T13:...	2018-02-20T20:...	2018-03-09	2018-03-09	23	23	0	b577af9a54b0
beleza_saude	delivered	2018-04-09	2018-04-09T23:...	2018-04-10T22:...	2018-04-25	2018-05-10	16	31	-15	3234e9d7817

00:00:03 Query executed successfully.

Microsoft Azure

Search resources, services, and docs (G+I)

Copilot

nisanthtumu@gmail.com
DEFAULT DIRECTORY NISANTHT...

Home > olistdatastorageactnis | Containers >

olistdata
Container

Search

Upload | Add Directory | Refresh | Rename | Delete | Change tier | Acquire lease | Break lease | Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: olistdata / Gold / Serving

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> [.]						...
<input type="checkbox"/> ..	2/27/2025, 8:58:27 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> 8C33AE76-2091-44EA-856D-D4FE24D37620_6_0-1.parquet	2/27/2025, 8:58:29 PM	Hot (Inferred)		Block blob	2.5 MiB	Available
<input type="checkbox"/> 8C33AE76-2091-44EA-856D-D4FE24D37620_6_0-12.parquet	2/27/2025, 8:58:29 PM	Hot (Inferred)		Block blob	2.47 MiB	Available
<input type="checkbox"/> 8C33AE76-2091-44EA-856D-D4FE24D37620_6_0-13.parquet	2/27/2025, 8:58:29 PM	Hot (Inferred)		Block blob	2.51 MiB	Available
<input type="checkbox"/> 8C33AE76-2091-44EA-856D-D4FE24D37620_6_0-14.parquet	2/27/2025, 8:58:29 PM	Hot (Inferred)		Block blob	2.57 MiB	Available