IST 615

Cloud Management

Final Project Report

E-COMMERCE ANALYTICS USING AZURE SERVICES FOR OLIST

Group 9:

M K Sowmeya

Dev Jindani

Raaj Mutreja

Sourabh Gavhane

Table of Contents

PROJECT OBJECTIVE

This project is designed to leverage Azure cloud services, specifically Azure Blob Storage, Azure SQL Database, and Azure Data Factory, to create a robust solution for analyzing and predicting customer behavior. The primary goal is to enhance sales strategies for Olist, a prominent Brazilian e-commerce platform.

The core focus revolves around incorporating three key datasets—customer orders, order items, and customer details—into Azure's cloud environment. The datasets will undergo processing and transformation, including cleaning and structuring, to ensure they are well-prepared for analysis.

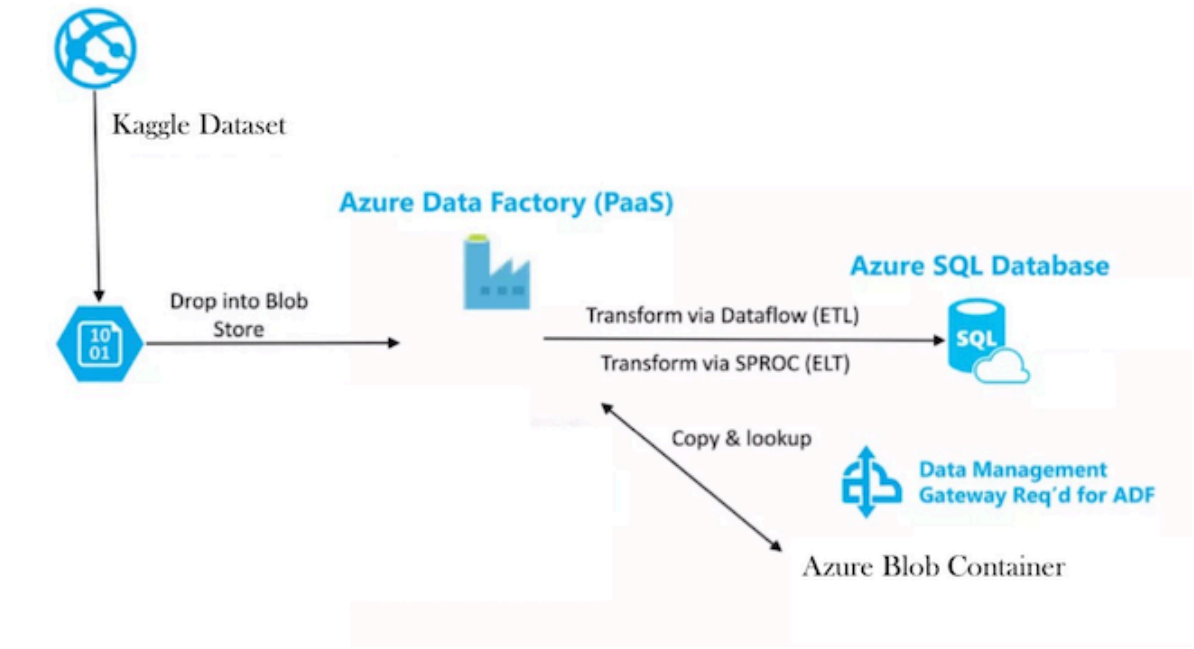
Within Azure's cloud ecosystem, these datasets will be stored and managed efficiently using Azure Blob Storage. Azure SQL Database will then come into play, providing a reliable and scalable relational database service for seamless integration and retrieval of structured data.

The innovative aspect of this project lies in the application of Azure Data Studio. Using this platform, analysis will be developed to identify trends and patterns in customer behavior. This analytical capability aims to provide valuable insights for Olist, enabling the platform to anticipate customer preferences and optimize sales strategies.

In summary, the project aims to harness Azure's cloud services to streamline data processing and develop predictive models for enhanced customer behavior analysis. The ultimate objective is to equip Olist with the tools needed to refine its sales strategies and stay ahead in the competitive e-commerce landscape.

AZURE CLOUD SERVICES

Flow diagram of Azure Service Integration:



INTEGRATION OF CLOUD SERVICES

Our project's integration strategy leverages the capabilities of Azure Blob Storage, Azure SQL Database, and Azure Machine Learning to create a cohesive and powerful data processing ecosystem:

1. **Azure Blob Storage:**

- Serves as the primary storage solution for datasets.
- Enables efficient data ingestion and staging.
- Provides secure and scalable storage for both raw and processed datasets.

2. **Azure SQL Database:**

- Facilitates seamless data integration from diverse sources, including customer orders, order items, and customer details.
- Acts as a fully managed relational database service for structured and organized data storage.
- Ensures optimal query performance, allowing for swift data retrieval during analysis.

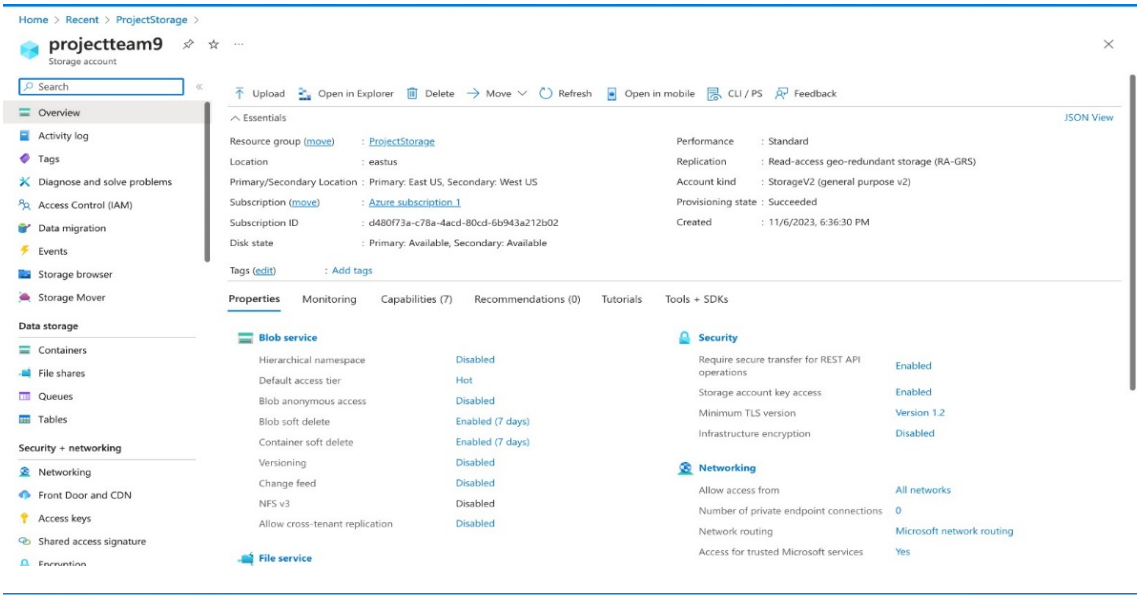
3. **Azure Data Factory:**

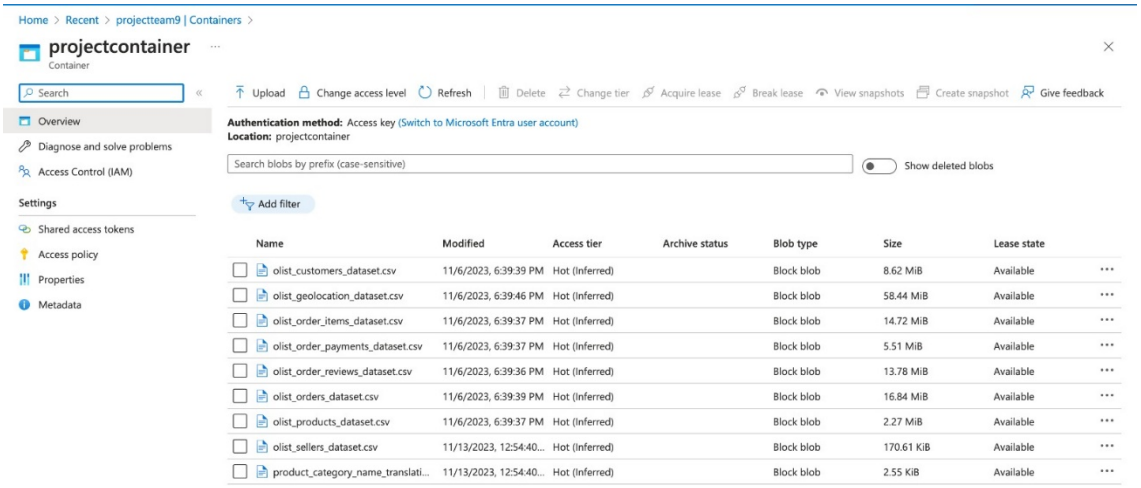
- **Primary Storage Solution:** Serves as the central repository for datasets.
- **Efficient Data Ingestion:** Facilitates streamlined data ingestion and staging processes.
- **Secure and Scalable Storage:** Provides a secure and scalable environment for both raw and processed datasets.

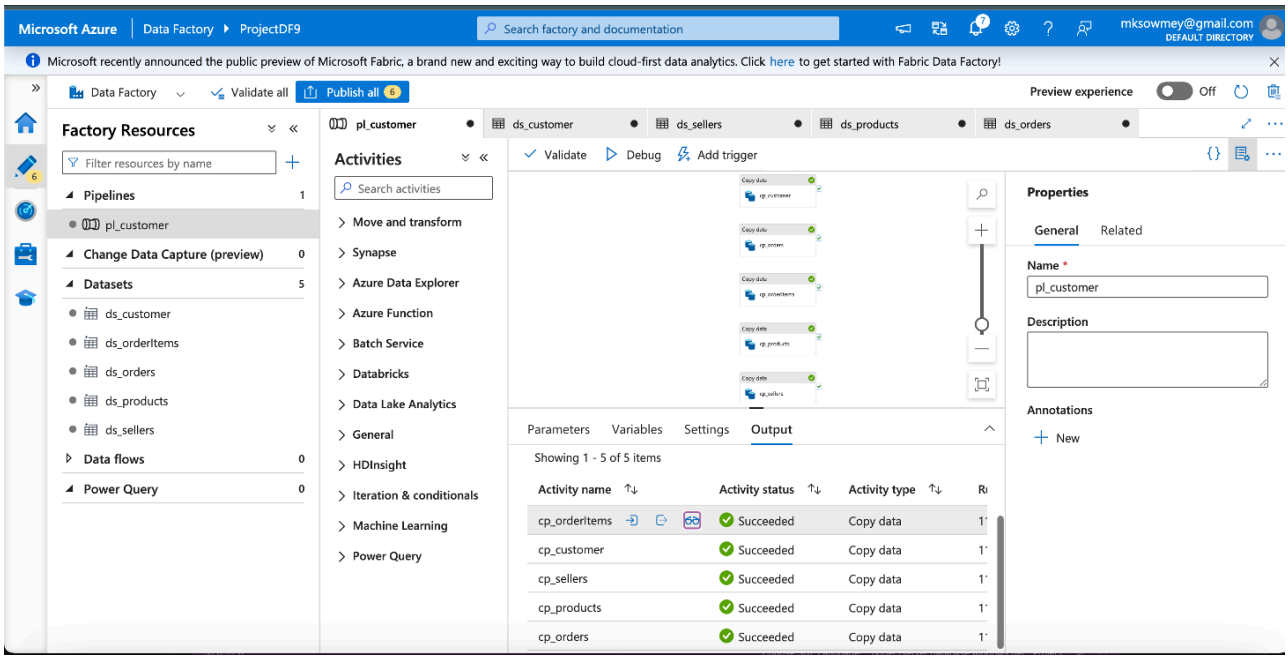
This integrated approach establishes a harmonized cloud environment, where Azure Blob Storage acts as the central repository, Azure SQL Database serves as the relational foundation, and Azure data factory drives advanced analytics. Through this synergy, our project aims to provide Olist with valuable insights, enhancing its sales strategies through a data-driven approach.

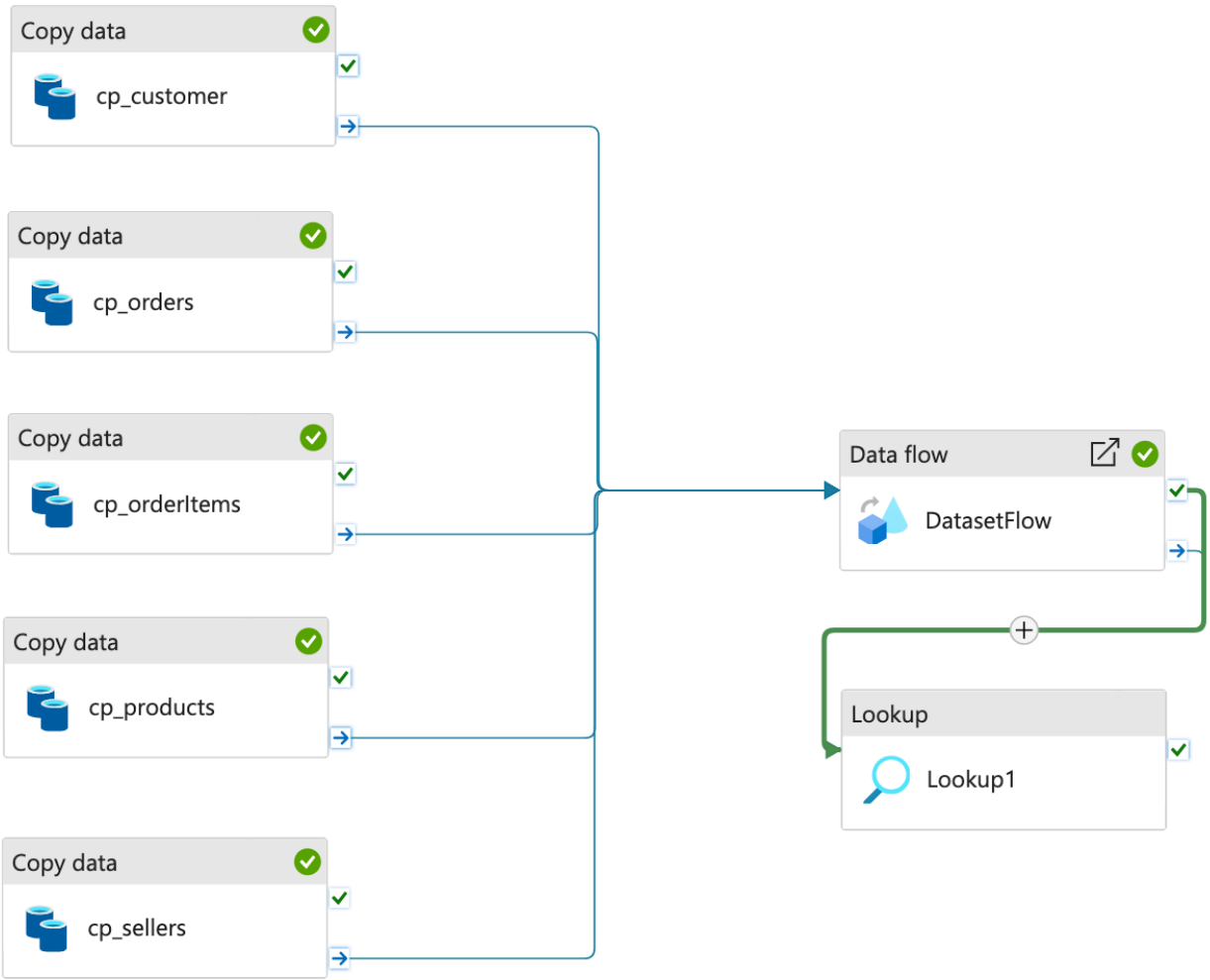
EXECUTION

Azure Blob storage containers:

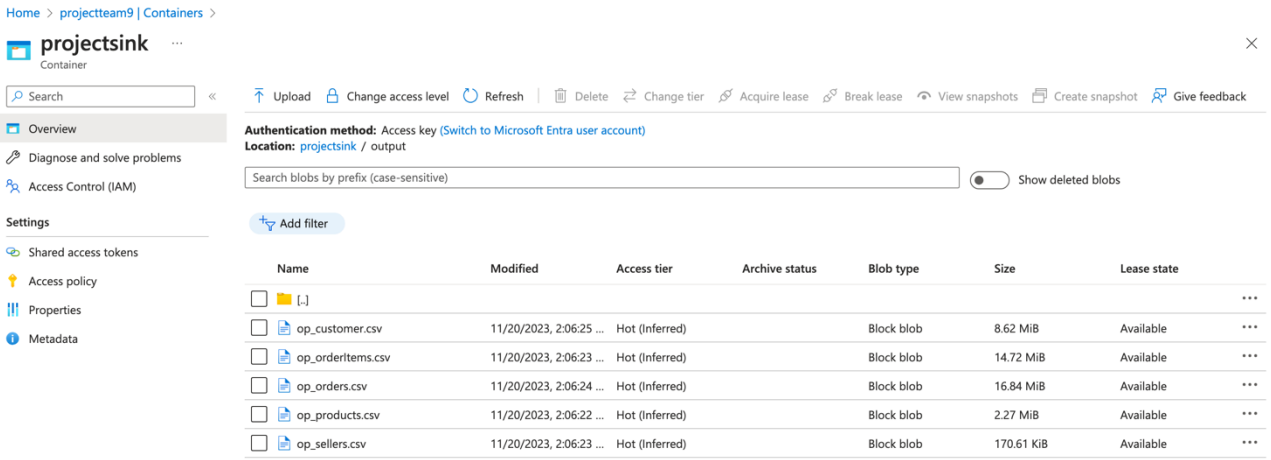

Home > Recent > ProjectStorage >
projectteam9
Storage account
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Storage Mover
Data storage
Containers
File shares
Queues
Tables
Security + networking
Networking
Front Door and CDN
Access keys
Shared access signature
Upload
Open in Explorer
Delete
Move
Refresh
Open in mobile
CLI / PS
Feedback
Essentials
JSON View
Resource group (move) : ProjectStorage
Location : eastus
Primary/Secondary Location : Primary: East US, Secondary: West US
Subscription (move) : Azure subscription 1
Subscription ID : d480f73a-c78a-4acd-80cd-6b943a212b02
Disk state : Primary: Available, Secondary: Available
Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 11/6/2023, 6:36:30 PM
Tags (edit) : Add tags
Properties
Monitoring
Capabilities (7)
Recommendations (0)
Tutorials
Tools + SDKs
Blob service
Hierarchical namespace Disabled
Default access tier Hot
Blob anonymous access Disabled
Blob soft delete Enabled (7 days)
Container soft delete Enabled (7 days)
Versioning Disabled
Change feed Disabled
NFS v3 Disabled
Allow cross-tenant replication Disabled
File service
Security
Require secure transfer for REST API operations Enabled
Storage account key access Enabled
Minimum TLS version Version 1.2
Infrastructure encryption Disabled
Networking
Allow access from All networks
Number of private endpoint connections 0
Network routing Microsoft network routing
Access for trusted Microsoft services Yes

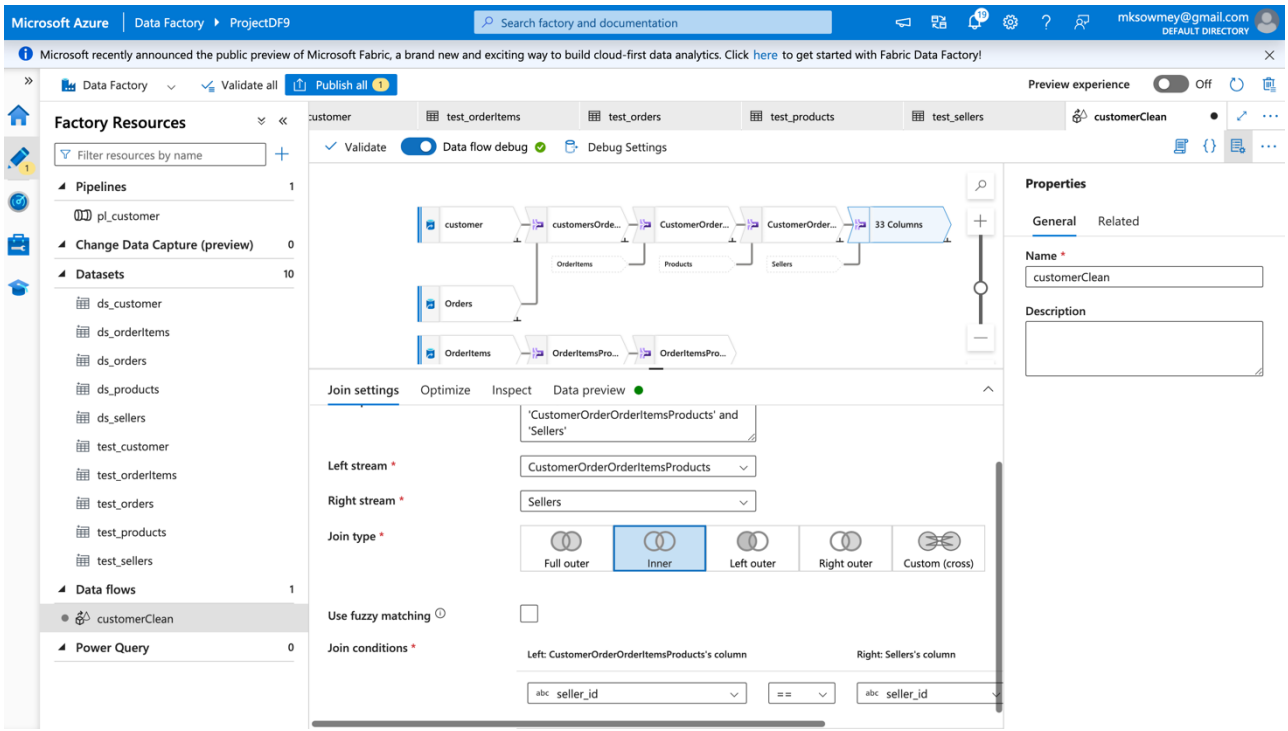

Home > Recent > projectteam9 | Containers >
projectcontainer
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Upload
Change access level
Refresh
Delete
Change tier
Acquire lease
Break lease
View snapshots
Create snapshot
Give feedback
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: projectcontainer
Search blobs by prefix (case-sensitive)
Show deleted blobs
Add filter
Name Modified Access tier Archive status Blob type Size Lease state
olist_customers_dataset.csv 11/6/2023, 6:39:39 PM Hot (Inferred) Block blob 8.62 MiB Available
olist_geolocation_dataset.csv 11/6/2023, 6:39:46 PM Hot (Inferred) Block blob 58.44 MiB Available
olist_order_items_dataset.csv 11/6/2023, 6:39:37 PM Hot (Inferred) Block blob 14.72 MiB Available
olist_order_payments_dataset.csv 11/6/2023, 6:39:37 PM Hot (Inferred) Block blob 5.51 MiB Available
olist_order_reviews_dataset.csv 11/6/2023, 6:39:36 PM Hot (Inferred) Block blob 13.78 MiB Available
olist_orders_dataset.csv 11/6/2023, 6:39:39 PM Hot (Inferred) Block blob 16.84 MiB Available
olist_products_dataset.csv 11/6/2023, 6:39:37 PM Hot (Inferred) Block blob 2.27 MiB Available
olist_sellers_dataset.csv 11/13/2023, 12:54:40... Hot (Inferred) Block blob 170.61 KiB Available
product_category_name_translati... 11/13/2023, 12:54:40... Hot (Inferred) Block blob 2.55 KiB Available

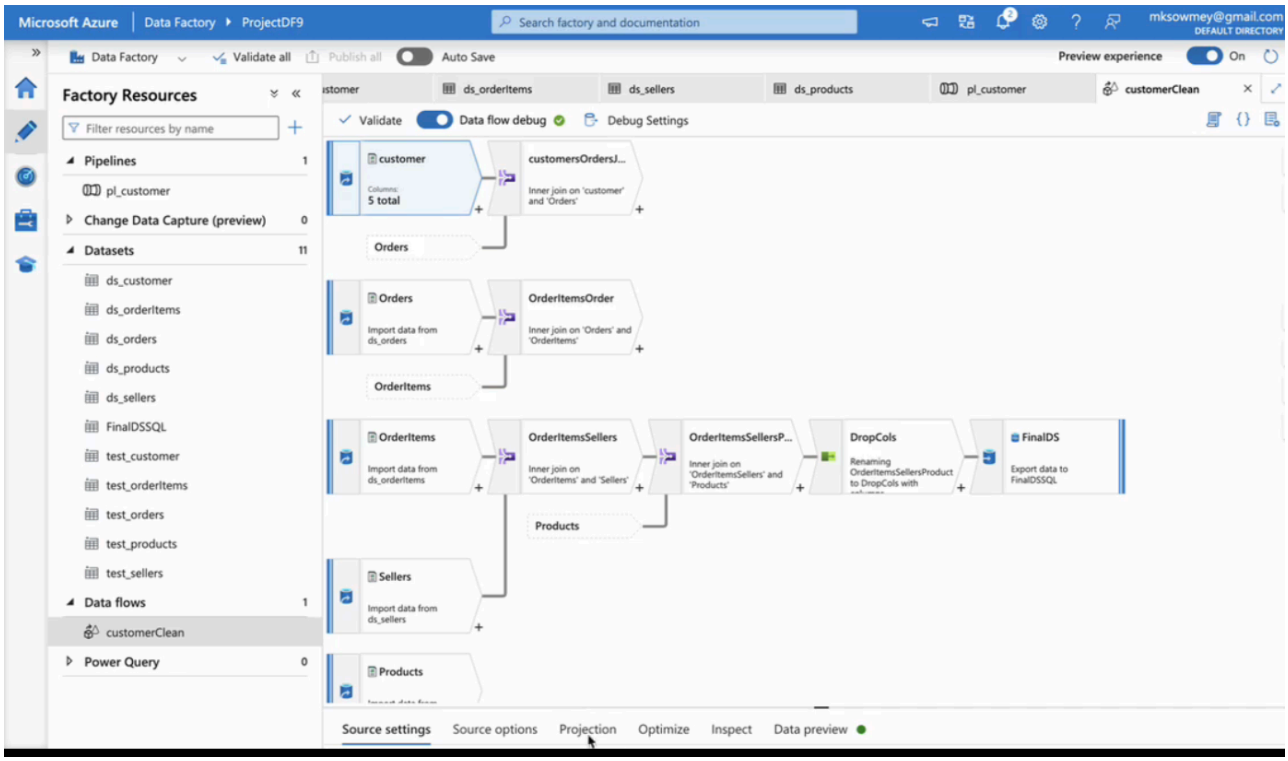

Microsoft Azure
Data Factory
ProjectDF9
Search factory and documentation
mksowmey@gmail.com
DEFAULT DIRECTORY
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all 6
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 1
pl_customer
Change Data Capture (preview) 0
Datasets 5
ds_customer
ds_orderItems
ds_orders
ds_products
ds_sellers
Data flows 0
Power Query 0
pl_customer
ds_customer
ds_sellers
ds_products
ds_orders
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Properties
General
Related
Name *
pl_customer
Description
Annotations
New
Parameters
Variables
Settings
Output
Showing 1 - 5 of 5 items
Activity name
Activity status
Activity type
cp_orderItems
Succeeded
Copy data
cp_customer
Succeeded
Copy data
cp_sellers
Succeeded
Copy data
cp_products
Succeeded
Copy data
cp_orders
Succeeded
Copy data


Copy data
cp_customer
Copy data
cp_orders
Copy data
cp_orderItems
Copy data
cp_products
Copy data
cp_sellers
Data flow
DatasetFlow
Lookup
Lookup1

- In ADF, a pipeline is a logical grouping of activities that together perform a task. The activities in a pipeline define actions to perform on your data.
- The pipeline shown in the image used for copying data from various sources, likely to aggregate them into a common data store or to move them into a transformation process (Dataset Flow).
- The "Copy Data" activities indicate that data is being moved from the source datasets to a destination, which could be a database, a data lake, or some other storage service.
- The "Data flow" activity would then apply transformations to the data that has been copied, such as joining, cleaning, enriching, or aggregating the data before it is loaded into its destination.
- To put it all together, this ADF pipeline is being used to automate the extraction and transformation of data from multiple sources.
- Once the data is prepared through the "Data flow", it can then be utilized for analytics, reporting, or further data processing tasks.
- To perform SQL operations as part of your pipeline, we added a "Lookup" activity if we need to retrieve data.







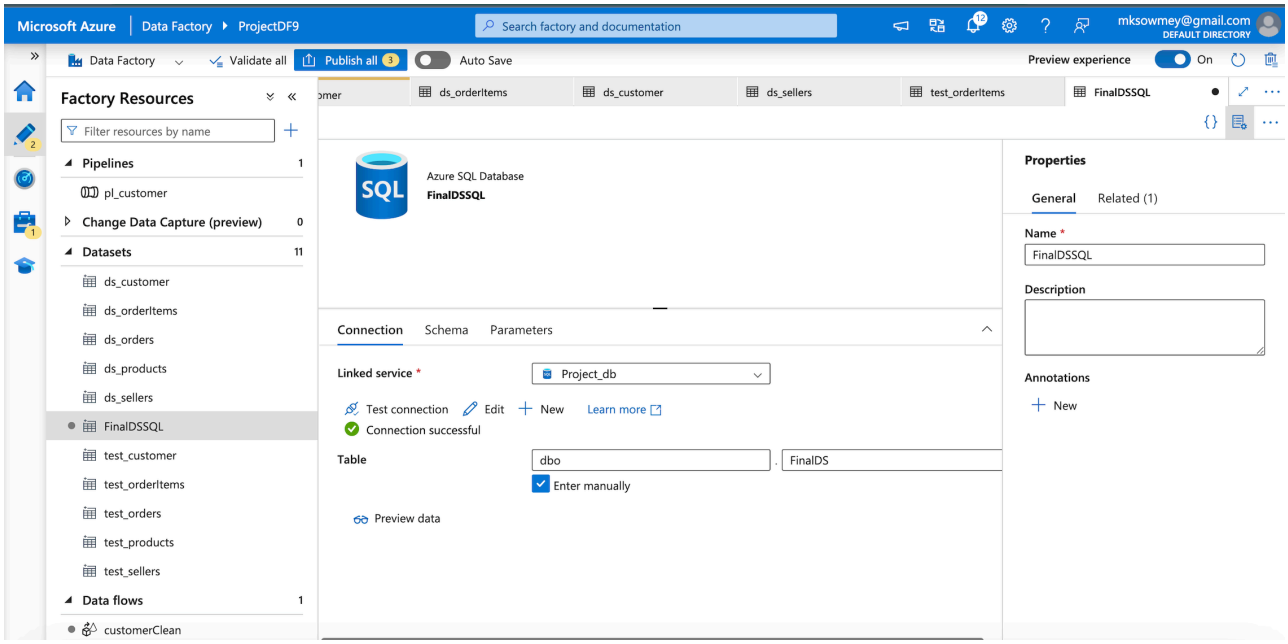
- In Azure Data Factory, a data flow represents a series of data transformation steps that are visually designed using a drag-and-drop interface. The use of data flow within Azure Data Factory is multifaceted and includes the following key aspects as per our scenario.

- Data Transformation: Data flows allow users to perform data transformation processes such as sorting, filtering, aggregating, and merging data without writing code. Transformations

are defined visually and executed as part of the Azure Data Factory pipeline.

- ETL Processes: They are essential for creating ETL (Extract, Transform, Load) processes, which enable businesses to consolidate data from various sources, transform it into a usable format, and load it into a destination data store for analytics or other purposes.

The above data flow is likely part of a larger data pipeline that includes several steps like joining different datasets, transforming datasets through various operations, and eventually preparing a final dataset for reporting, analytics, & operational use.

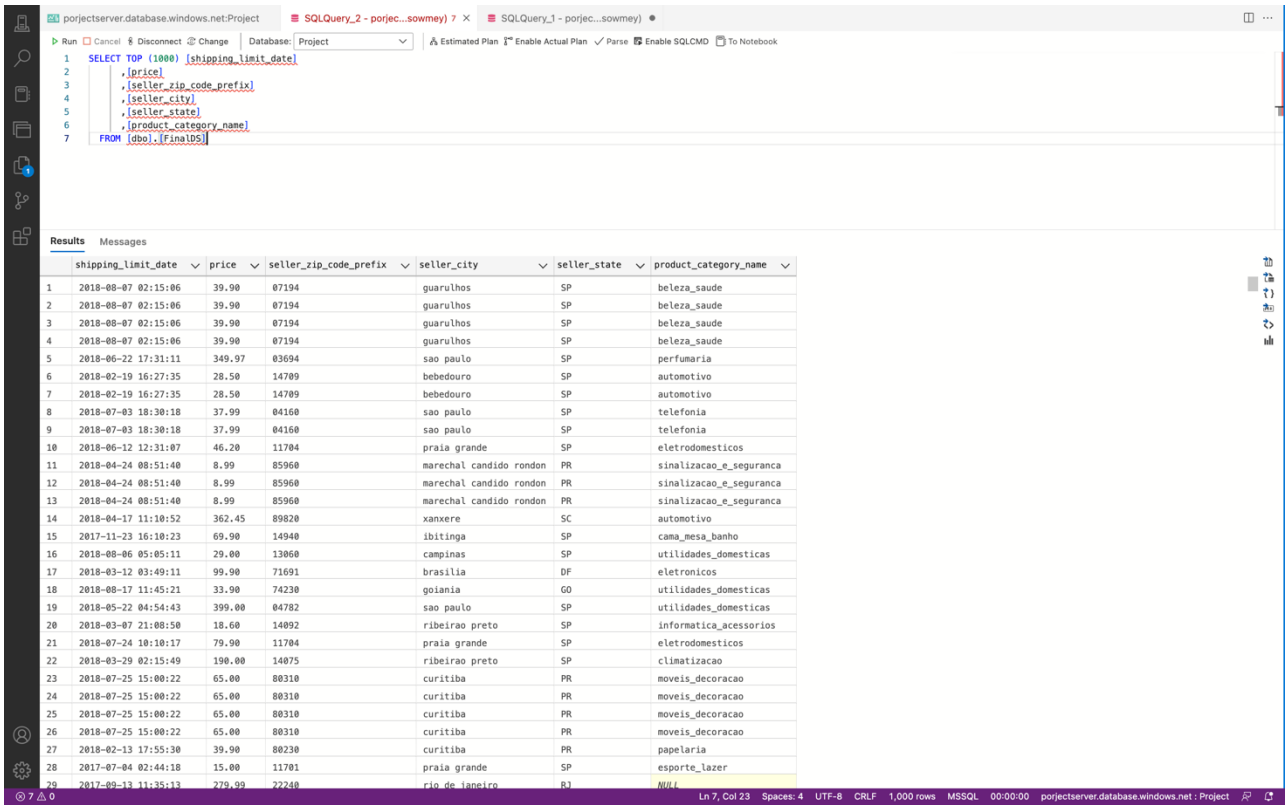


- Sink Configuration: Once the data is processed within the data flow, the final output, known as the "sink," is configured to be stored or loaded into a SQL Server database, commonly referred to as SQL Studio.

- SQL Server Integration: The data is sent to the SQL Server database, where it can be stored, analyzed, or used for various purposes like reporting, querying, or serving as a source for other applications.

- Utilization in SQL Studio: Within SQL Studio, the received data can be further analyzed, joined with other datasets, queried using SQL queries, or used in various business intelligence and reporting tools to derive insights or support decision-making processes.

- This seamless flow from data transformation in the data flow to storing the finalized data in SQL Studio ensures the availability of processed and structured data in a database environment for efficient storage, analysis, and utilization within an organization's data ecosystem.




```
10
11  SELECT AVG(price) AS average_price FROM [dbo].[FinalDS];
12
```

Results Messages

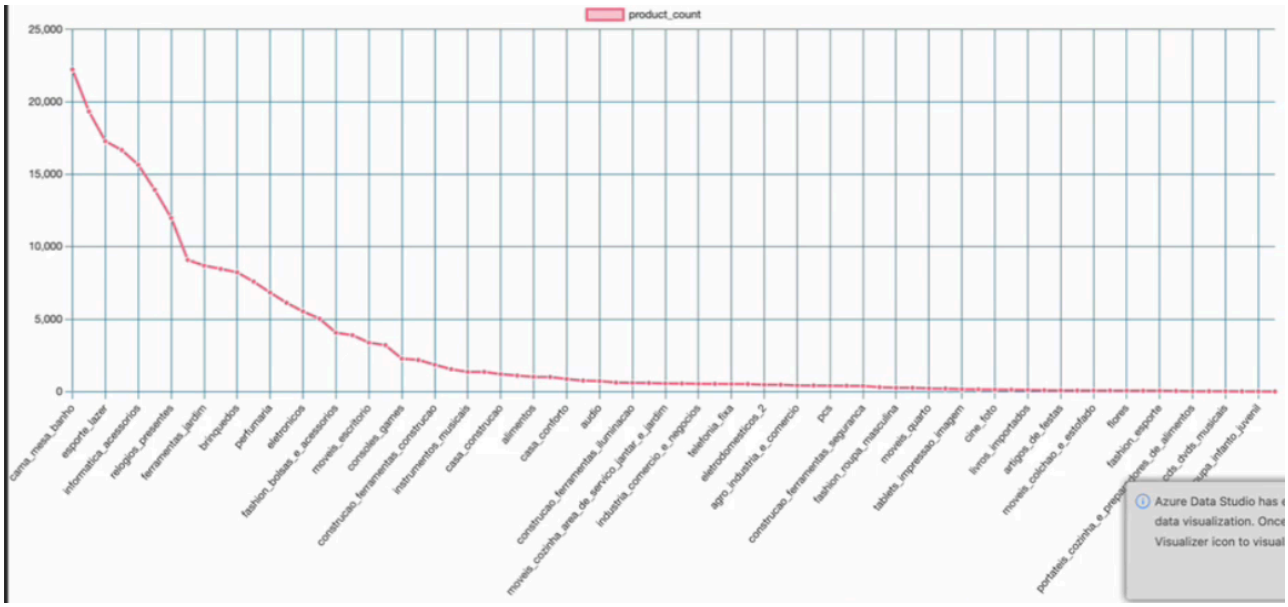
	average_price
1	120.653739

```sql
-- SELECT DISTINCT product_category_name FROM [dbo].[FinalDS];

SELECT product_category_name, COUNT(*) AS product_count
FROM [dbo].[FinalDS]
GROUP BY product_category_name
ORDER BY product_count DESC;
```

Results Messages

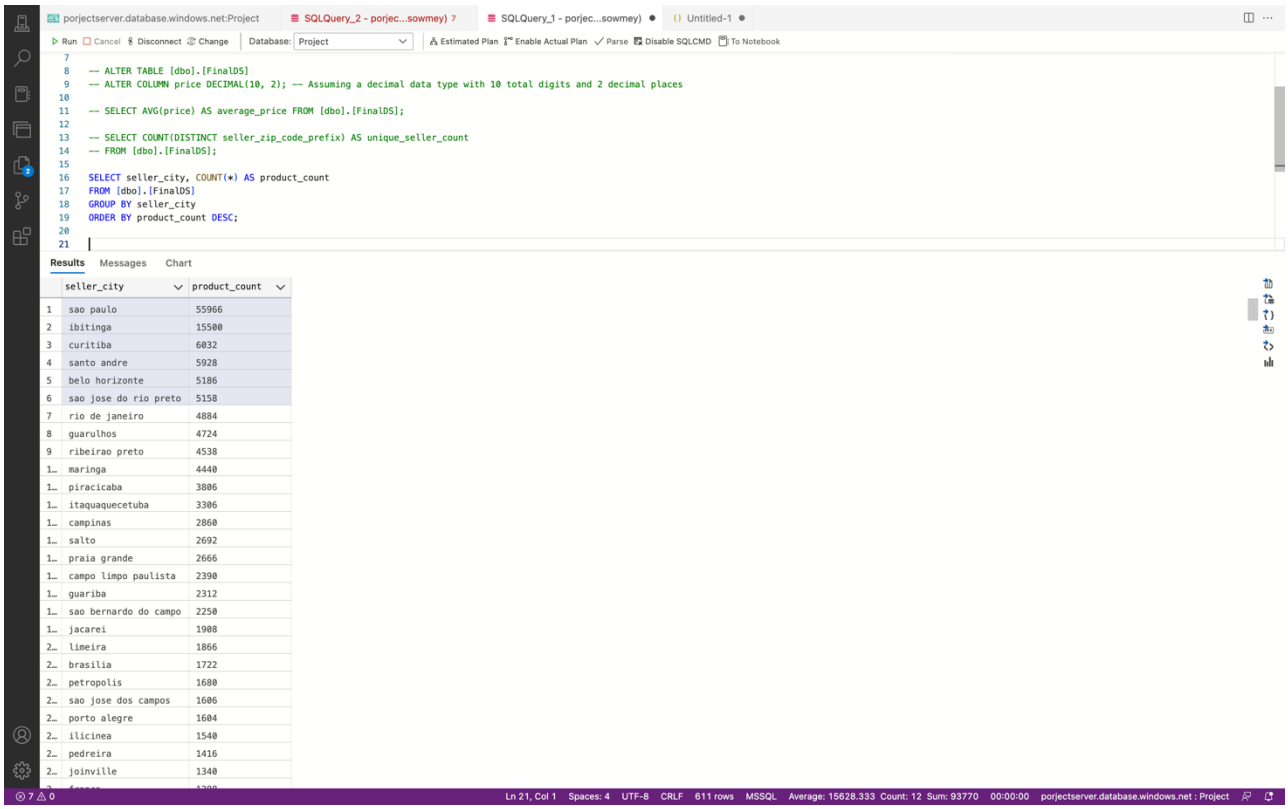
	product_category_name	product_count
1	cama_mesa_banho	22230
2	beleza_saude	19340
3	esporte_lazer	17282
4	moveis_decoracao	16668
5	informatica_acessorios	15654
6	utilidades_domesticas	13928
7	relogios_presentes	11982
8	telefonia	9090
9	ferramentas_jardim	8694
10	automotivo	8470
11	brinquedos	8234
12	cool_stuff	7592
13	perfumaria	6838
14	bebes	6130
15	eletronicos	5534
16	papelaria	5034
17	fashion_bolsas_e_acessorios	4062
18	pet_shop	3894
19	moveis_escritorio	3382
20	*NULL*	3206
21	consoles_games	2274
22	malas_acessorios	2184
23	construcao_ferramentas_construcao	1858
24	eletrodomesticos	1542
25	instrumentos_musicais	1360
26	eletroportateis	1358
27	casa_construcao	1208
28	livros_interesse_geral	1106
29	alimentos	1020
30	moveis_sala	1006
31	casa_conforto	868
32	bebidas	758
33	audio	728

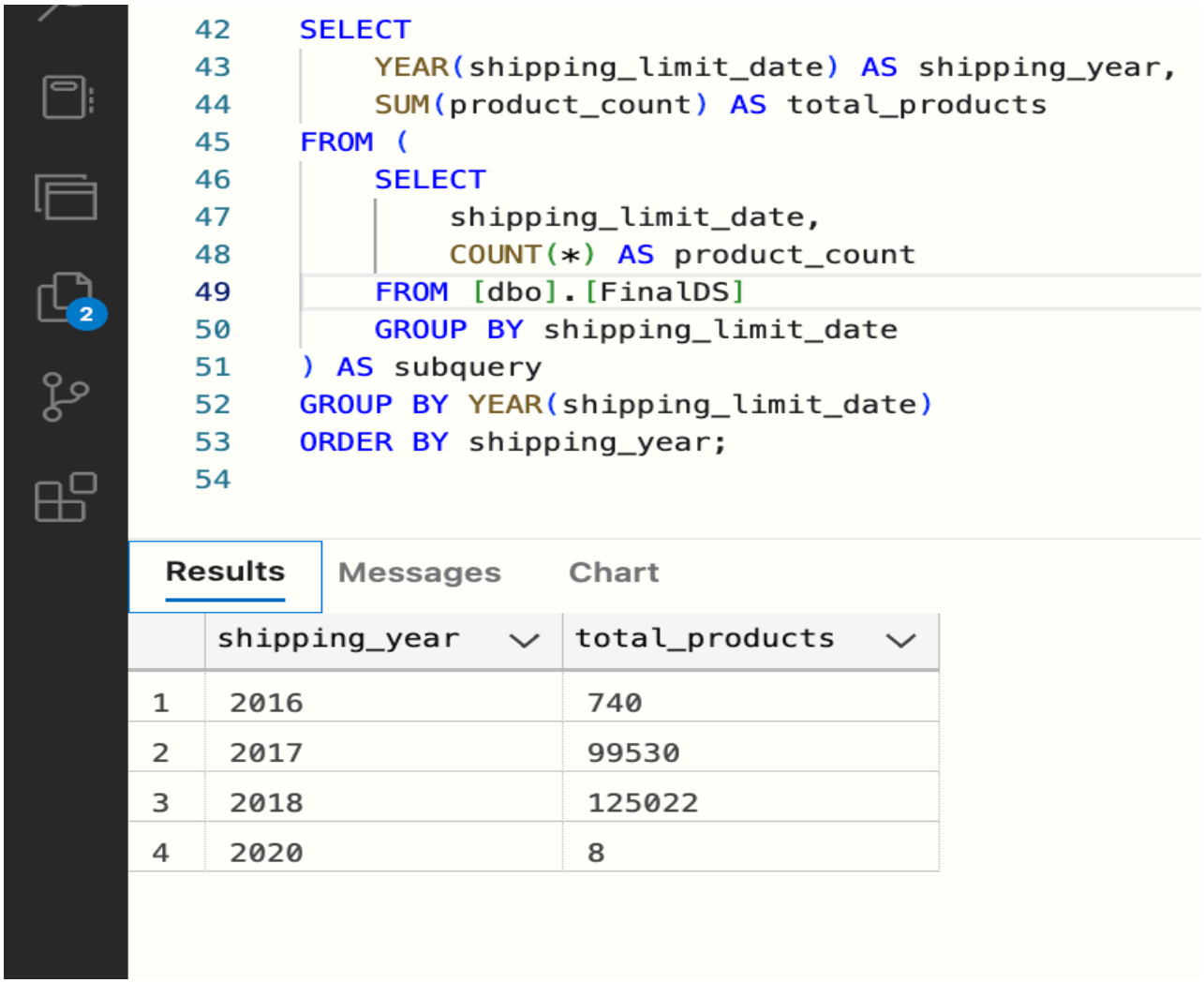



```
12
13  SELECT COUNT(DISTINCT seller_zip_code_prefix) AS unique_seller_count
14  FROM [dbo].[FinalDS];
15
16
```

Results | Messages | Chart

	unique_seller_count
1	2246

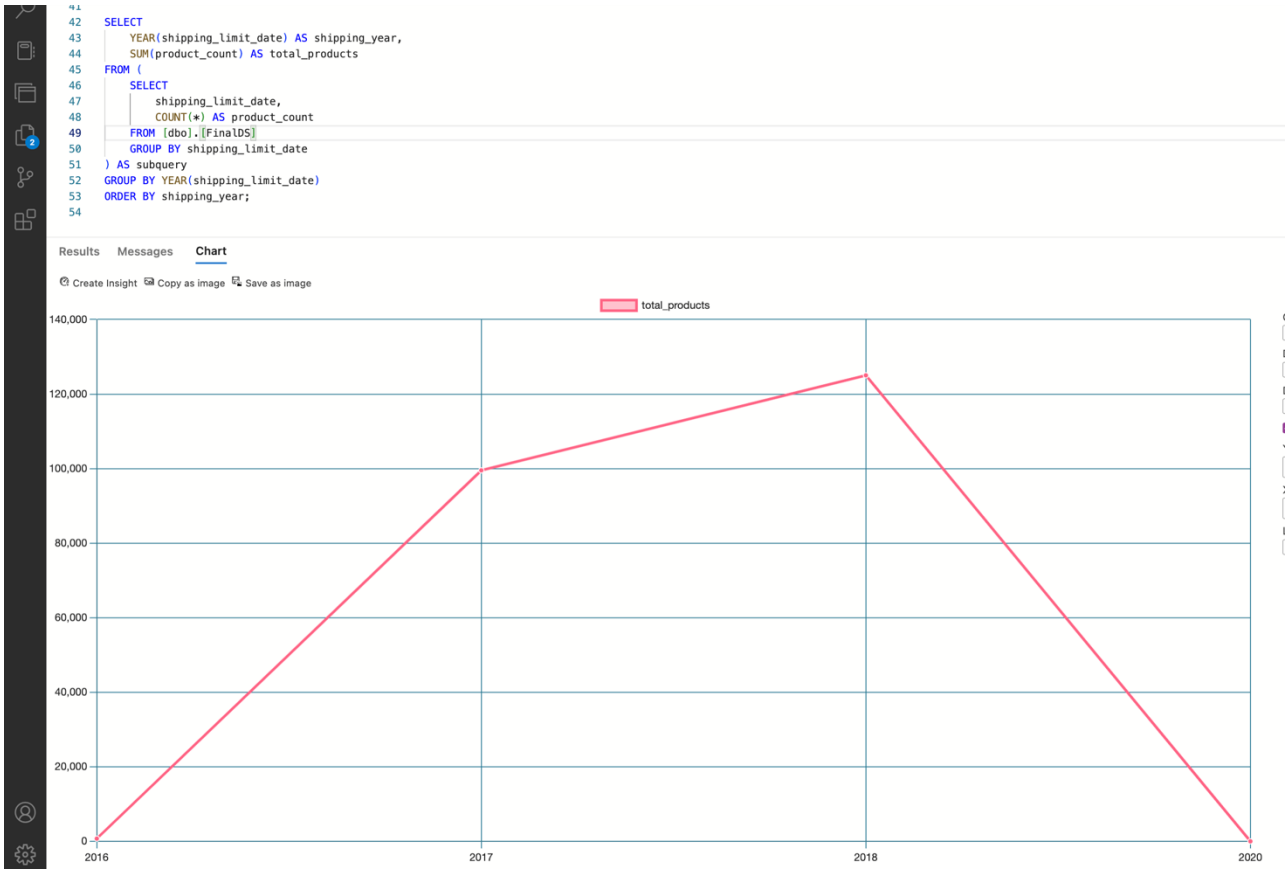

porjectserver.database.windows.net:Project
SQLQuery_2 - porjec...sowmey) 7
SQLQuery_1 - porjec...sowmey)
Untitled-1
Run
Cancel
Disconnect
Change
Database: Project
Estimated Plan
Enable Actual Plan
Parse
Disable SQLCMD
To Notebook
7
8 -- ALTER TABLE [dbo].[FinalDS]
9 -- ALTER COLUMN price DECIMAL(10, 2); -- Assuming a decimal data type with 10 total digits and 2 decimal places
10
11 -- SELECT AVG(price) AS average_price FROM [dbo].[FinalDS];
12
13 -- SELECT COUNT(DISTINCT seller_zip_code_prefix) AS unique_seller_count
14 -- FROM [dbo].[FinalDS];
15
16 SELECT seller_city, COUNT(*) AS product_count
17 FROM [dbo].[FinalDS]
18 GROUP BY seller_city
19 ORDER BY product_count DESC;
20
21
Results
Messages
Chart
seller_city
product_count
1 sao paulo 55966
2 ibitinga 15500
3 curitiba 6032
4 santo andre 5928
5 belo horizonte 5186
6 sao jose do rio preto 5158
7 rio de janeiro 4884
8 guarulhos 4724
9 ribeirao preto 4538
1… maringa 4440
1… piracicaba 3806
1… itaquaquecetuba 3306
1… campinas 2860
1… salto 2692
1… praia grande 2666
1… campo limpo paulista 2390
1… guariba 2312
1… sao bernardo do campo 2250
1… jacarei 1908
2… limeira 1866
2… brasilia 1722
2… petropolis 1680
2… sao jose dos campos 1606
2… porto alegre 1604
2… ilicinea 1540
2… pedreira 1416
2… joinville 1340
⊗ 7 ⚠ 0
Ln 21, Col 1
Spaces: 4
UTF-8
CRLF
611 rows
MSSQL
Average: 15628.333 Count: 12 Sum: 93770
00:00:00
porjectserver.database.windows.net : Project



```sql
SELECT
    YEAR(shipping_limit_date) AS shipping_year,
    SUM(product_count) AS total_products
FROM (
    SELECT
        shipping_limit_date,
        COUNT(*) AS product_count
    FROM [dbo].[FinalDS]
    GROUP BY shipping_limit_date
) AS subquery
GROUP BY YEAR(shipping_limit_date)
ORDER BY shipping_year;
```

Results | Messages | Chart

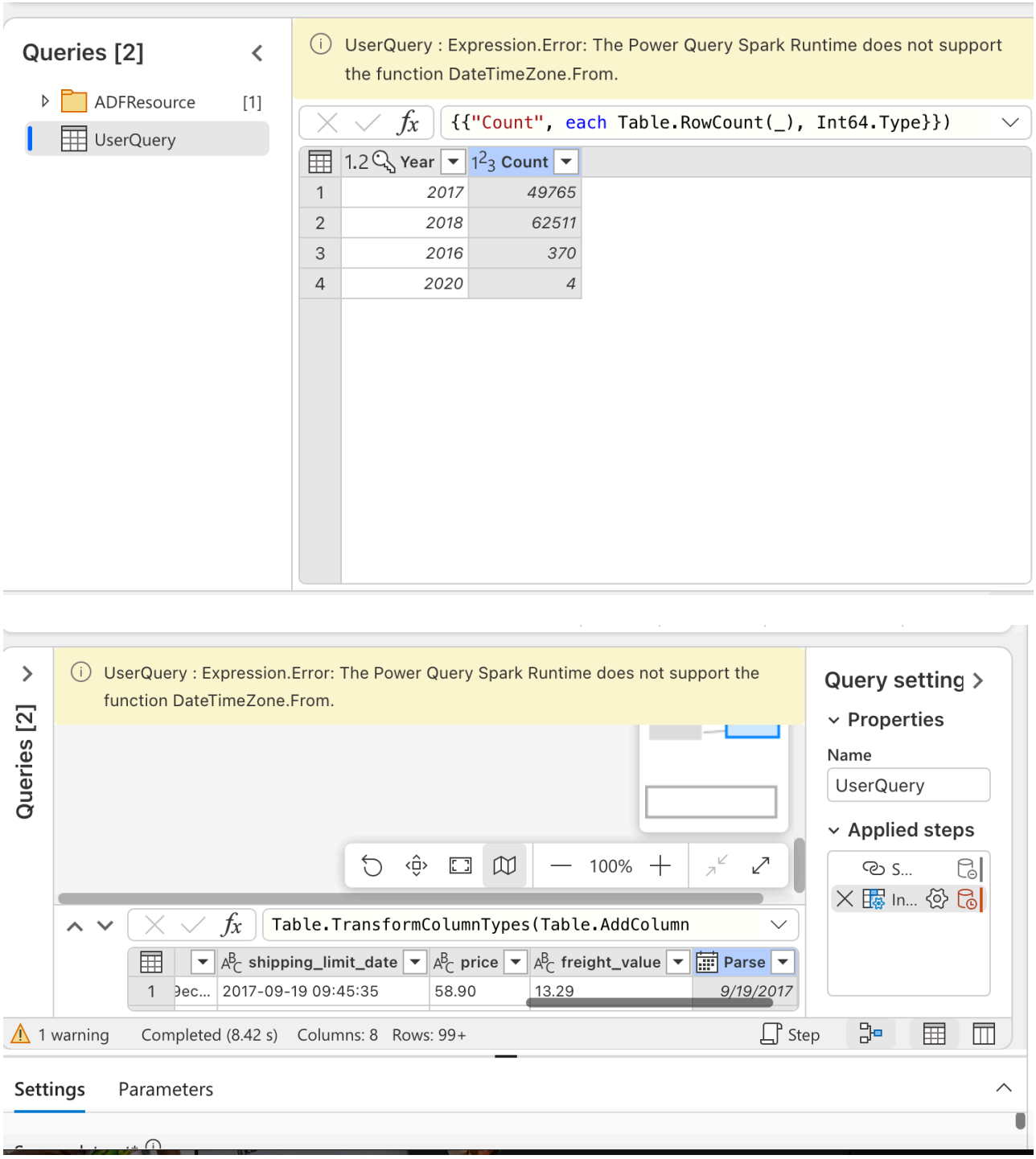
	shipping_year	total_products
1	2016	740
2	2017	99530
3	2018	125022
4	2020	8



KEY CHALLENGES FACED

1. The challenge we faced was with datetime zone in Power Query within Azure Data Factory (ADF) which is a common hurdle when integrating various cloud services. Handling datetime zones, especially in data transformation processes, can be complex due to different time zone formats or inconsistencies across data sources.

2. In cases where the dataset lacks certain qualities required for predictive modeling, such as insufficient features, data quality issues, or inadequate sample size, predictive modeling might not yield reliable results. When faced with a dataset that isn't suitable for predictive modeling, leveraging SQL queries for analytics purposes becomes a valuable alternative enabling the extraction of insights critical for informed decision-making and strategy formulation.



CONCLUSION

In conclusion, this project marks a strategic initiative to harness the power of Azure cloud services—Azure Blob Storage, Azure SQL Database, and Azure Data Factory—to revolutionize the analysis and prediction of customer behavior for Olist, a leading Brazilian e-commerce platform. By seamlessly integrating and transforming key datasets within Azure's cloud

environment, we have laid the foundation for a robust and scalable solution.

Azure Blob Storage facilitates efficient data storage and management, while Azure SQL Database ensures reliable integration and retrieval of structured data. The innovative use of Azure Data Studio adds a layer of sophistication, enabling us to discover trends and patterns in customer behavior, thereby providing Olist with actionable insights to refine sales strategies.

This project not only demonstrates the prowess of Azure's cloud ecosystem but also underscores its practical applications in enhancing e-commerce analytics. By empowering Olist with advanced predictive models and analytical capabilities, the project aligns with the objective of staying ahead in the competitive e-commerce landscape. In essence, the successful implementation of this project equips Olist with the tools needed to optimize customer engagement and elevate sales strategies to new height.

REFERENCES

- https://learn.microsoft.com/en-us/azure/?product=popular
- https://olist.com/
- https://azure.microsoft.com/en-us/blog/
- https://azure.microsoft.com/en-us/products/azure-sql/database