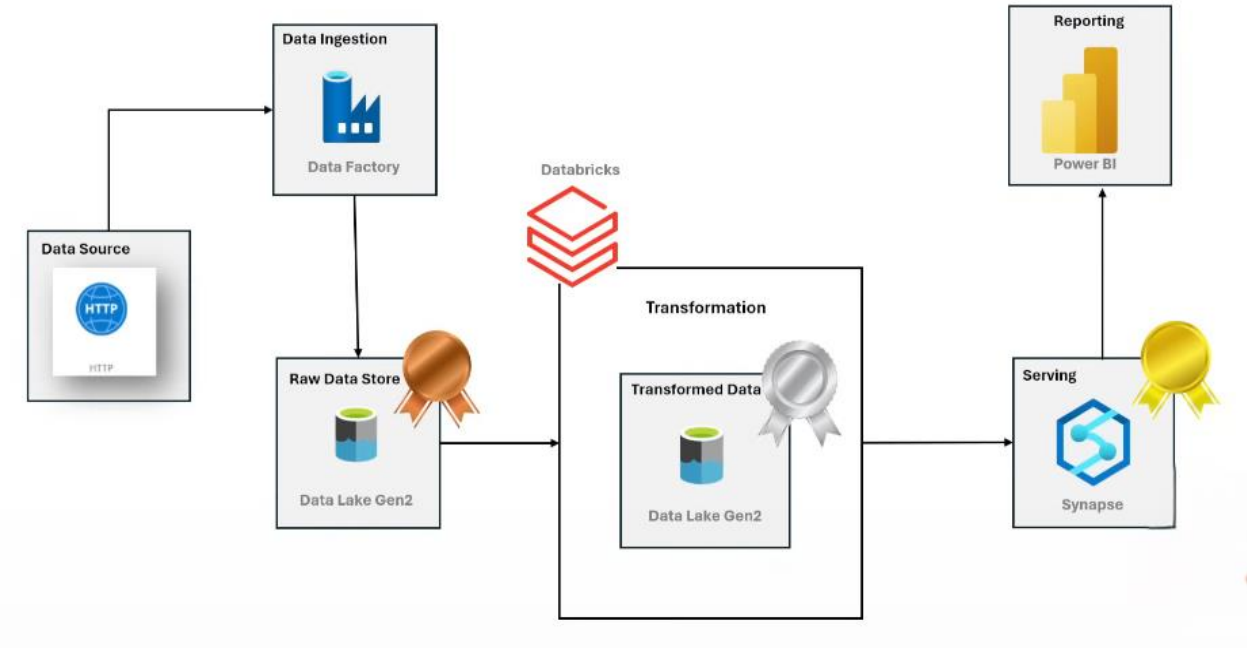
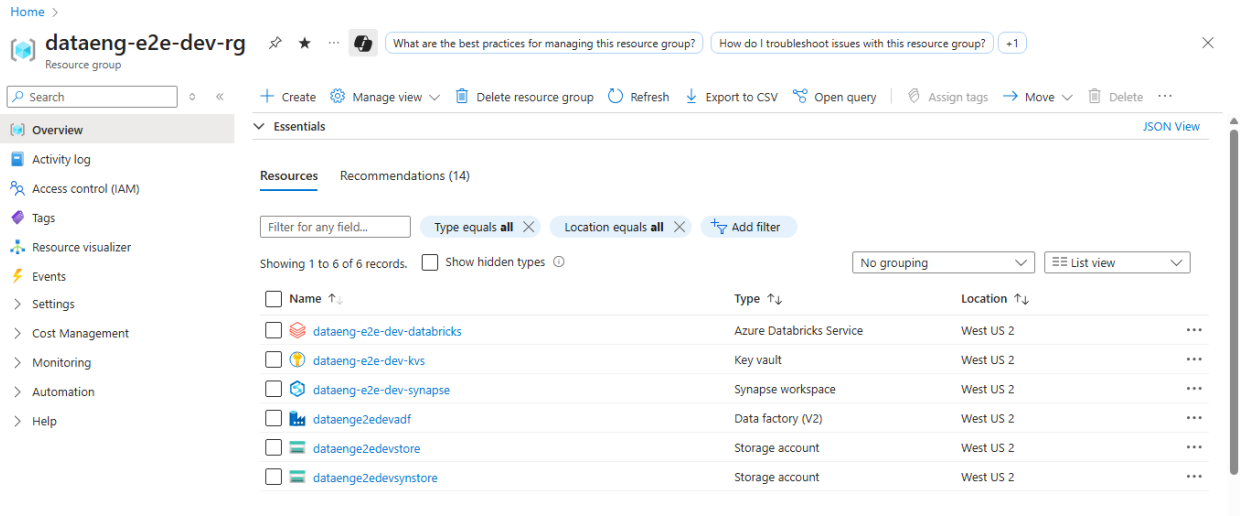
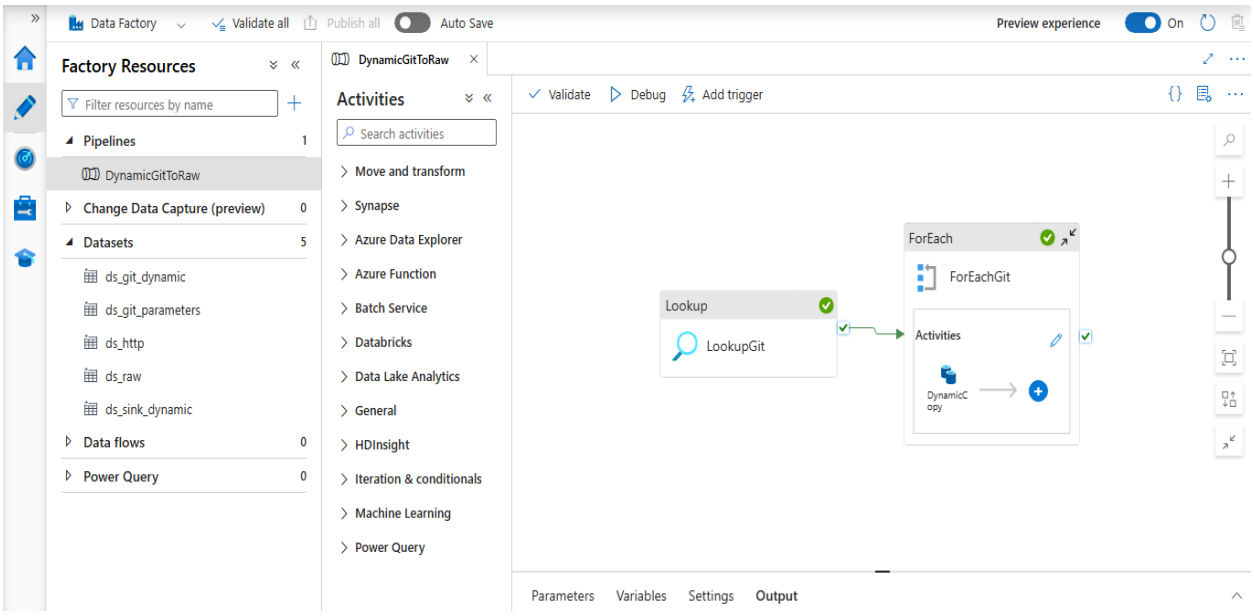
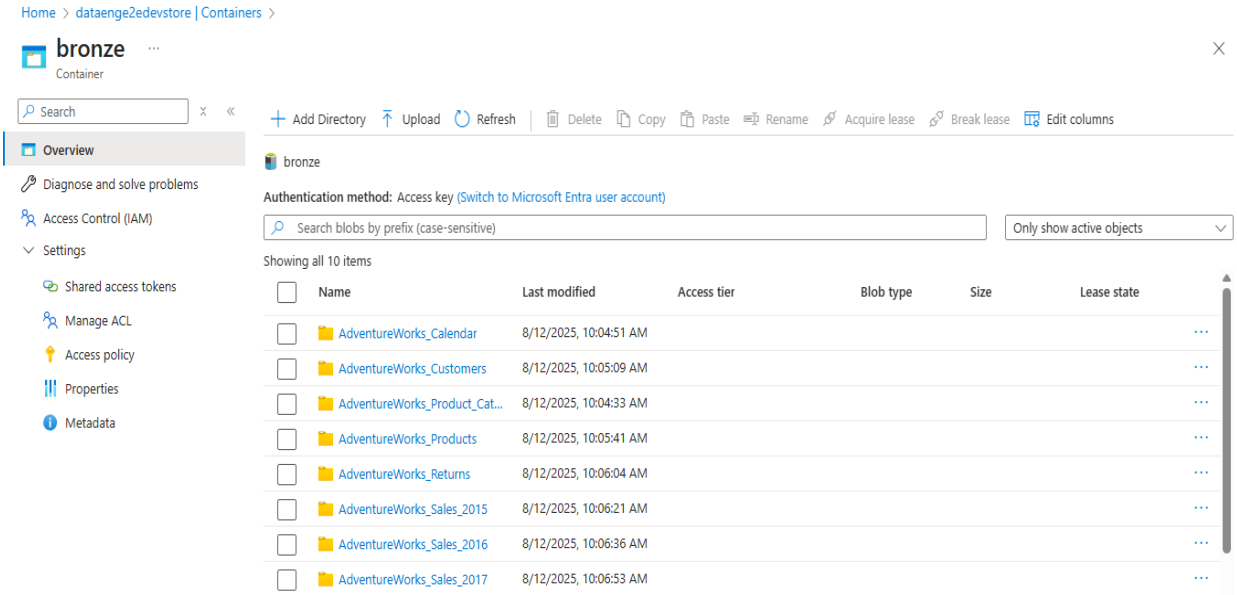
Azure End-To-End Data Engineering Project from Scratch

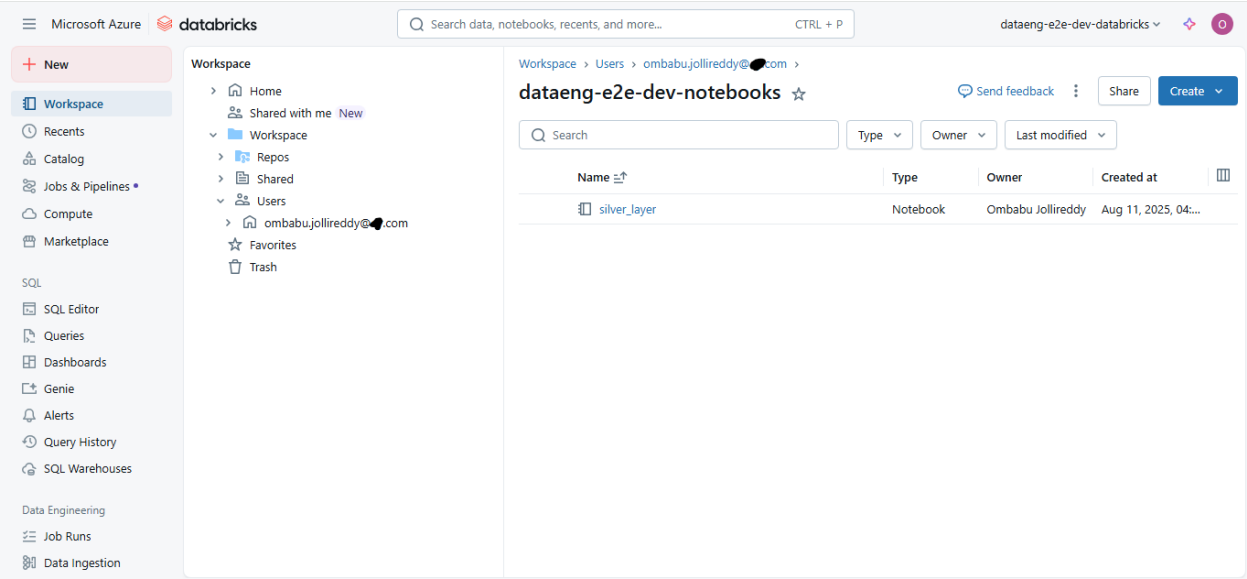
ARCHITECTURE






Data Factory
Validate all
Publish all
Auto Save
Preview experience
On
Factory Resources
Filter resources by name
Pipelines 1
DynamicGitToRaw
Change Data Capture (preview) 0
Datasets 5
ds_git_dynamic
ds_git_parameters
ds_http
ds_raw
ds_sink_dynamic
Data flows 0
Power Query 0
DynamicGitToRaw
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
ForEach
ForEachGit
Lookup
LookupGit
Activities
DynamicCopy
Parameters
Variables
Settings
Output


Home > dataenge2edevstore | Containers >
bronze
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Add Directory
Upload
Refresh
Delete
Copy
Paste
Rename
Acquire lease
Break lease
Edit columns
bronze
Authentication method: Access key (Switch to Microsoft Entra user account)
Search blobs by prefix (case-sensitive)
Only show active objects
Showing all 10 items
Name
Last modified
Access tier
Blob type
Size
Lease state
AdventureWorks_Calendar 8/12/2025, 10:04:51 AM
AdventureWorks_Customers 8/12/2025, 10:05:09 AM
AdventureWorks_Product_Cat... 8/12/2025, 10:04:33 AM
AdventureWorks_Products 8/12/2025, 10:05:41 AM
AdventureWorks_Returns 8/12/2025, 10:06:04 AM
AdventureWorks_Sales_2015 8/12/2025, 10:06:21 AM
AdventureWorks_Sales_2016 8/12/2025, 10:06:36 AM
AdventureWorks_Sales_2017 8/12/2025, 10:06:53 AM



2 minutes ago (<1s) 1

```
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

SILVER LAYER SCRIPT

DATA ACCESS USING MOUNT

3 minutes ago (1s) 4

```
spark.conf.set(
  "fs.azure.account.key.dataenge2edevstore.dfs.core.windows.net",
  dbutils.secrets.get(scope="dataeng-e2e-dev-scope", key="dataeng-storage-secret")
)
```

✓ 2 minutes ago (<1s) 5

```
mount_point = "/mnt/dataenge2edevstore/bronze"

if not any(mount.mountPoint == mount_point for mount in dbutils.fs.mounts()):
    dbutils.fs.mount(
        source = "wasbs://bronze@dataenge2edevstore.blob.core.windows.net",
        mount_point = mount_point,
        extra_configs = {
            "fs.azure.account.key.dataenge2edevstore.blob.core.windows.net": dbutils.secrets.get
            (scope="dataeng-e2e-dev-scope", key="dataeng-storage-secret")
        }
    )
else:
    print(f"Already mounted: {mount_point}")
```

```
Already mounted: /mnt/dataenge2edevstore/bronze
```

DATA LOADING

Read the Data

✓ 3 minutes ago (1s) 8

```
df_cal = spark.read.format("csv")\
            .option("header", "true")\
            .option("inferSchema", "true")\
            .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Calendar")
```

▸ (2) Spark Jobs

▸ df_cal: pyspark.sql.dataframe.DataFrame = [Date: date]

✓ 3 minutes ago (1s) 9

```
df_cus = spark.read.format("csv")\
            .option("header", "true")\
            .option("inferSchema", "true")\
            .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Customers")
```

▸ (2) Spark Jobs

▸ df_cus: pyspark.sql.dataframe.DataFrame = [CustomerKey: integer, Prefix: string ... 11 more fields]

✓ 3 minutes ago (1s) 10

```
df_procat = spark.read.format("csv")\
            .option("header", "true")\
            .option("inferSchema", "true")\
            .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Product_Categories")
```

▸ (2) Spark Jobs

▸ df_procat: pyspark.sql.dataframe.DataFrame = [ProductCategoryKey: integer, CategoryName: string]

✓ 4 minutes ago (1s) 11

```
df_pro = spark.read.format("csv")\
        .option("header", "true")\
        .option("inferSchema", "true")\
        .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Products")
```

▸ (2) Spark Jobs

▸ df_pro: pyspark.sql.dataframe.DataFrame = [ProductKey: integer, ProductSubcategoryKey: integer ... 9 more fields]

✓ 4 minutes ago (1s) 12

```
df_ret = spark.read.format("csv")\
        .option("header", "true")\
        .option("inferSchema", "true")\
        .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Returns")
```

▸ (2) Spark Jobs

▸ df_ret: pyspark.sql.dataframe.DataFrame = [ReturnDate: date, TerritoryKey: integer ... 2 more fields]

✓ 4 minutes ago (1s) 13

```
df_sales = spark.read.format("csv")\
        .option("header", "true")\
        .option("inferSchema", "true")\
        .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Sales*")
```

▸ (2) Spark Jobs

▸ df_sales: pyspark.sql.dataframe.DataFrame = [OrderDate: date, StockDate: date ... 6 more fields]

✓ 4 minutes ago (1s) 14

```
df_ter = spark.read.format("csv")\
        .option("header", "true")\
        .option("inferSchema", "true")\
        .load("/mnt/dataenge2edevstore/bronze/AdventureWorks_Territories")
```

▸ (2) Spark Jobs

▸ df_ter: pyspark.sql.dataframe.DataFrame = [SalesTerritoryKey: integer, Region: string ... 2 more fields]

✓ 10:18 AM (1s) 15

```
df_subcat = spark.read.format("csv")\
        .option("header", "true")\
        .option("inferSchema", "true")\
        .load("/mnt/dataenge2edevstore/bronze/Product_Subcategories")
```

▸ (2) Spark Jobs

▸ df_subcat: pyspark.sql.dataframe.DataFrame = [ProductSubcategoryKey: integer, SubcategoryName: string ... 1 more field]

TRANSFORMATIONS

Calendar

10:18 AM (<1s) 18

```
df_cal.display()
```

(1) Spark Jobs

Table +

	Date
1	2015-01-01
2	2015-01-02
3	2015-01-03
4	2015-01-04
5	2015-01-05

10:18 AM (<1s) 19 Python

```
df_cal = df_cal.withColumn('Month', month(col('Date')))\
               .withColumn('Year', year(col('Date')))
df_cal.display()
```

(1) Spark Jobs

df_cal: pyspark.sql.dataframe.DataFrame = [Date: date, Month: integer ... 1 more field]

Table +

	Date	Month	Year
1	2015-01-01	1	2015
2	2015-01-02	1	2015
3	2015-01-03	1	2015
4	2015-01-04	1	2015
5	2015-01-05	1	2015

10:18 AM (<1s) 20

```
mount_point = "/mnt/dataenge2edevstore/silver"

if not any(mount.mountPoint == mount_point for mount in dbutils.fs.mounts()):
    dbutils.fs.mount(
        source = "wasbs://silver@dataenge2edevstore.blob.core.windows.net",
        mount_point = mount_point,
        extra_configs = {
            "fs.azure.account.key.dataenge2edevstore.blob.core.windows.net": dbutils.secrets.get
            (scope="dataeng-e2e-dev-scope", key="dataeng-storage-secret")
        }
    )
else:
    print(f"Already mounted: {mount_point}")
```

```
Already mounted: /mnt/dataenge2edevstore/silver
```

10:18 AM (2s) 21

```
df_cal.write.format('parquet')\
            .mode('append')\
            .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Calendar')
```

(1) Spark Jobs

Customers

10:18 AM (1s) 23

```
df_cus.display()
```

(1) Spark Jobs

Table +

	CustomerKey	Prefix	FirstName	LastName	BirthDate	MaritalStatus	Gender
1	11000	MR.	JON	YANG	1966-04-08	M	M
2	11001	MR.	EUGENE	HUANG	1965-05-14	S	M
3	11002	MR.	RUBEN	TORRES	1965-08-12	M	M
4	11003	MS.	CHRISTY	ZHU	1968-02-15	S	F
5	11004	MRS.	ELIZABETH	JOHNSON	1968-08-08	S	F

10:18 AM (1s) 24

```
df_cus = df_cus.withColumn('fullname', concat_ws(' ', col('Prefix'), col('FirstName'), col('LastName')))
df_cus.display()
```

(1) Spark Jobs

df_cus: pyspark.sql.dataframe.DataFrame = [CustomerKey: integer, Prefix: string ... 12 more fields]

Table +

	CustomerKey	Prefix	FirstName	LastName	BirthDate	MaritalStatus	Gender
1	11000	MR.	JON	YANG	1966-04-08	M	M
2	11001	MR.	EUGENE	HUANG	1965-05-14	S	M
3	11002	MR.	RUBEN	TORRES	1965-08-12	M	M
4	11003	MS.	CHRISTY	ZHU	1968-02-15	S	F
5	11004	MRS.	ELIZABETH	JOHNSON	1968-08-08	S	F

10:18 AM (2s) 25

```
df_cus.write.format('parquet')\
        .mode('append')\
        .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Customers')
```

(1) Spark Jobs

Sub Categories

10:18 AM (<1s) 27

```
df_subcat.display()
```

(1) Spark Jobs

Table +

	ProductSubcategoryKey	SubcategoryName	ProductCategoryKey
1	1	Mountain Bikes	1
2	2	Road Bikes	1
3	3	Touring Bikes	1
4	4	Handlebars	2
5	5	Bottom Brackets	2

10:18 AM (1s) 28

```
df_subcat.write.format('parquet')\
        .mode('append')\
        .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Subcategories')
```

(1) Spark Jobs

Products

10:18 AM (<1s) 30

```
df_pro.display()
```

(1) Spark Jobs

Table +

	ProductKey	ProductSubcategoryKey	ProductSKU	ProductName	ModelName	Prod
1	214	31	HL-U509-R	Sport-100 Helmet, Red	Sport-100	Universa
2	215	31	HL-U509	Sport-100 Helmet, Black	Sport-100	Universa
3	218	23	SO-B909-M	Mountain Bike Socks, M	Mountain Bike Socks	Combina
4	219	23	SO-B909-L	Mountain Bike Socks, L	Mountain Bike Socks	Combina
5	220	31	HL-U509-B	Sport-100 Helmet, Blue	Sport-100	Universa

10:18 AM (<1s) 31

```
df_pro = df_pro.withColumn('ProductSKU', split(col('ProductSKU'), '-')[0])\
               .withColumn('ProductName', split(col('ProductName'), ' ')[0])
df_pro.display()
```

(1) Spark Jobs

df_pro: pyspark.sql.dataframe.DataFrame = [ProductKey: integer, ProductSubcategoryKey: integer ... 9 more fields]

Table +

	ProductKey	ProductSubcategoryKey	ProductSKU	ProductName	ModelName	ProductDe
1	214	31	HL	Sport-100	Sport-100	Universal fit, w
2	215	31	HL	Sport-100	Sport-100	Universal fit, w
3	218	23	SO	Mountain	Mountain Bike Socks	Combination o
4	219	23	SO	Mountain	Mountain Bike Socks	Combination o
5	220	31	HL	Sport-100	Sport-100	Universal fit, w

10:18 AM (1s) 32

```
df_pro.write.format('parquet')\
            .mode('append')\
            .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Products')
```

(1) Spark Jobs

10:18 AM (<1s) 33

```
df_ret.display()
```

(1) Spark Jobs

Table +

	ReturnDate	TerritoryKey	ProductKey	ReturnQuantity
1	2015-01-18	9	312	1
2	2015-01-18	10	310	1
3	2015-01-21	8	346	1
4	2015-01-22	4	311	1
5	2015-02-02	6	312	1

10:18 AM (1s) 34

```
df_ret.write.format('parquet')\
        .mode('append')\
        .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Returns')
```

(1) Spark Jobs

Territories

10:18 AM (<1s) 36

```
df_ter.display()
```

(1) Spark Jobs

Table +

	SalesTerritoryKey	Region	Country	Continent
1	1	Northwest	United States	North America
2	2	Northeast	United States	North America
3	3	Central	United States	North America
4	4	Southwest	United States	North America
5	5	Southeast	United States	North America

10:18 AM (1s) 37

```
df_ter.write.format('parquet')\
        .mode('append')\
        .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Territories')
```

(1) Spark Jobs

Sales

10:18 AM (<1s) 39

```
df_sales.display()
```

(1) Spark Jobs

Table +

	OrderDate	StockDate	OrderNumber	ProductKey	CustomerKey	TerritoryKey	O
1	2017-01-01	2003-12-13	SO61285	529	23791	1	
2	2017-01-01	2003-09-24	SO61285	214	23791	1	
3	2017-01-01	2003-09-04	SO61285	540	23791	1	
4	2017-01-01	2003-09-28	SO61301	529	16747	1	
5	2017-01-01	2003-10-21	SO61301	377	16747	1	

10:18 AM (<1s) 40

```
df_sales = df_sales.withColumn('StockDate', to_timestamp('StockDate'))
```

df_sales: pyspark.sql.dataframe.DataFrame = [OrderDate: date, StockDate: timestamp ... 6 more fields]

10:18 AM (<1s) 41

```
df_sales = df_sales.withColumn('OrderNumber', regexp_replace(col('OrderNumber'), 'S', 'T'))
```

df_sales: pyspark.sql.dataframe.DataFrame = [OrderDate: date, StockDate: timestamp ... 6 more fields]

10:18 AM (<1s) 42

```
df_sales = df_sales.withColumn('multiply', col('OrderLineItem') * col('OrderQuantity'))
```

df_sales: pyspark.sql.dataframe.DataFrame = [OrderDate: date, StockDate: timestamp ... 7 more fields]

10:18 AM (<1s) 43

```
df_sales.display()
```

(1) Spark Jobs

Table +

	OrderDate	StockDate	OrderN...	ProductKey	CustomerKey	TerritoryKey	(
1	2017-01-01	2003-12-13T00:00:00.000+00:...	TO61285	529	23791	1	
2	2017-01-01	2003-09-24T00:00:00.000+00:...	TO61285	214	23791	1	
3	2017-01-01	2003-09-04T00:00:00.000+00:...	TO61285	540	23791	1	
4	2017-01-01	2003-09-28T00:00:00.000+00:...	TO61301	529	16747	1	
5	2017-01-01	2003-10-21T00:00:00.000+00:...	TO61301	377	16747	1	

10:18 AM (2s) 44

```
df_sales.write.format('parquet')\
        .mode('append')\
        .save('/mnt/dataenge2edevstore/silver/AdventureWorks_Sales')
```

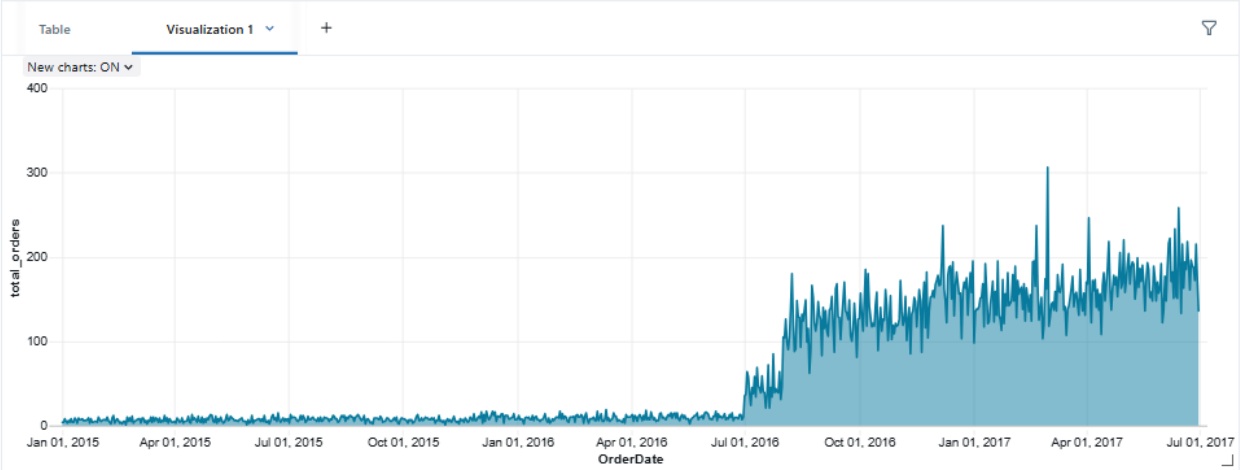
(1) Spark Jobs

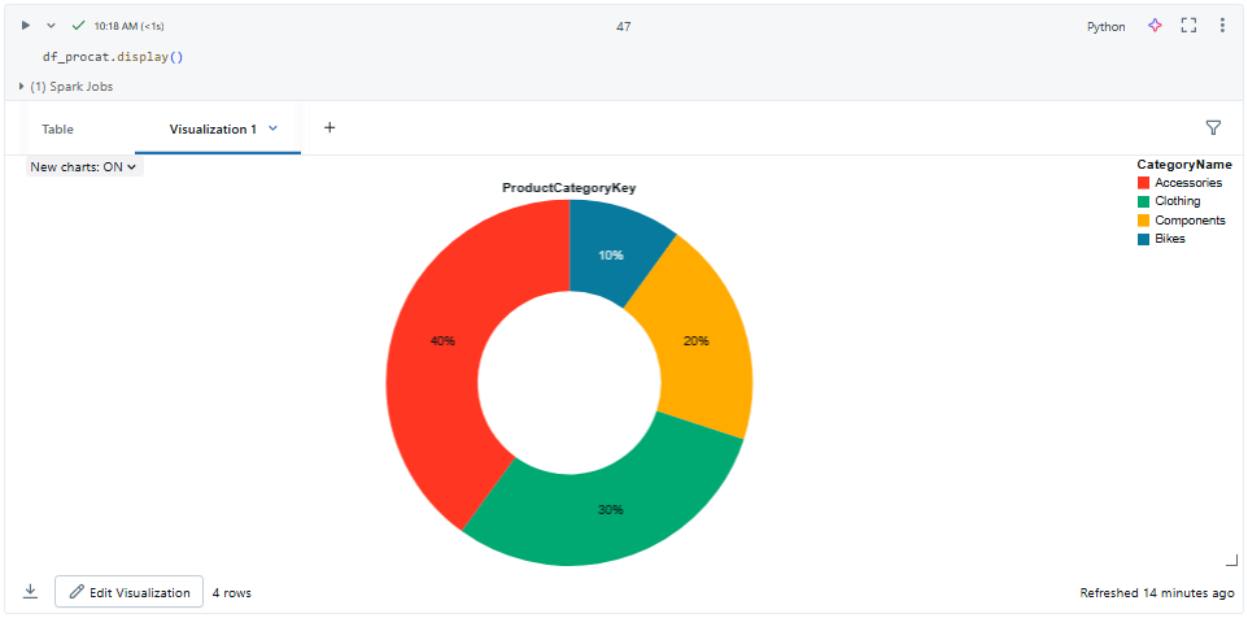
Sales Analysis

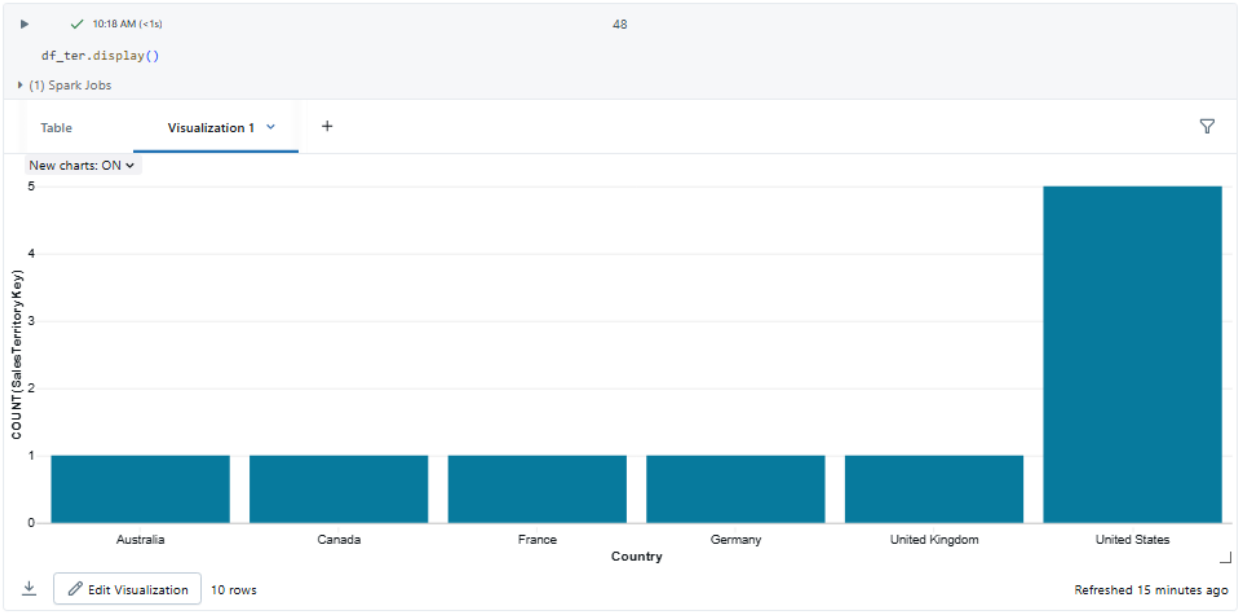
10:18 AM (1s) 46

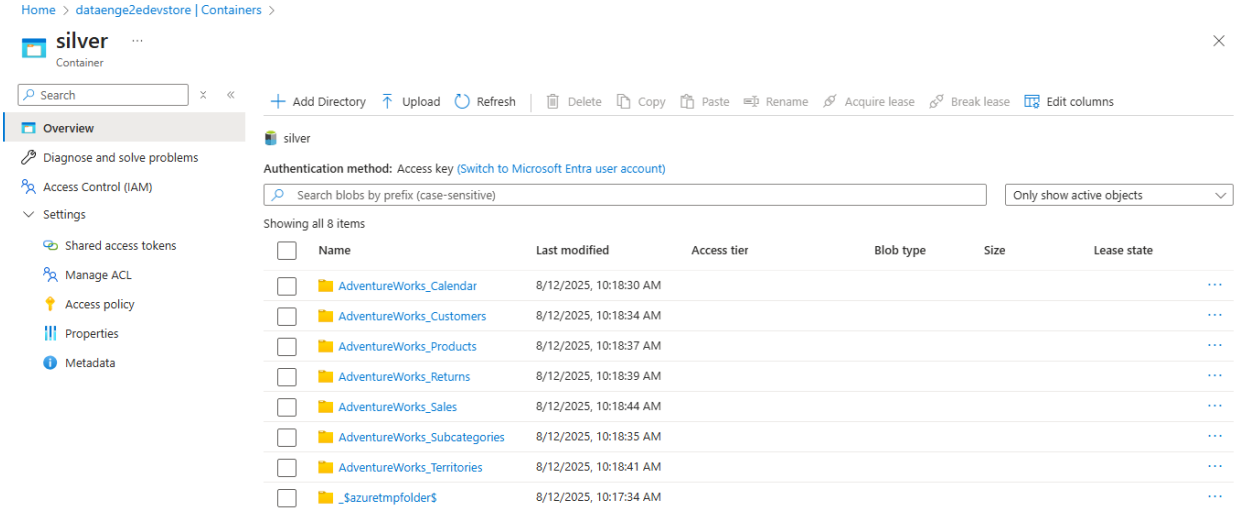
```
df_sales.groupBy('OrderDate').agg(count('OrderNumber').alias('total_orders')).display()
```

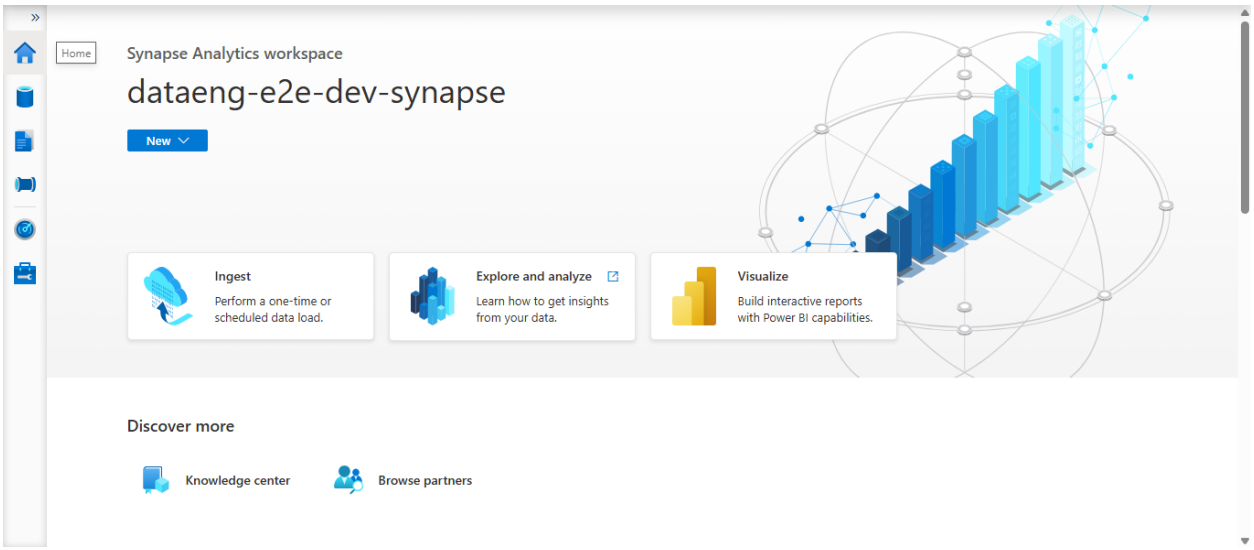
(2) Spark Jobs

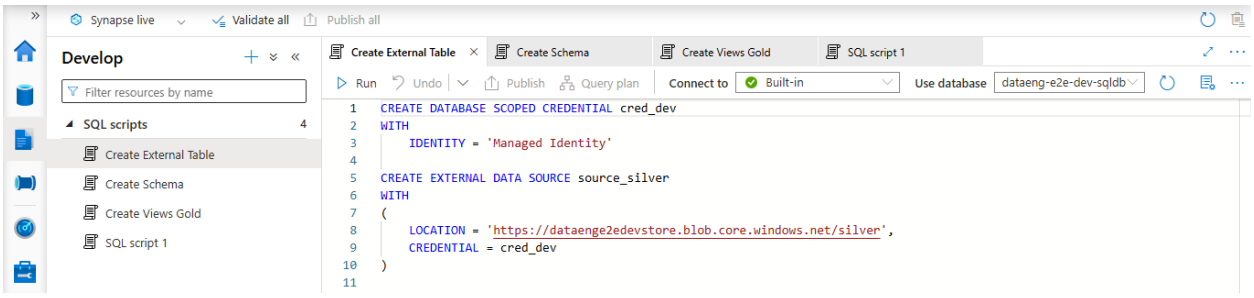


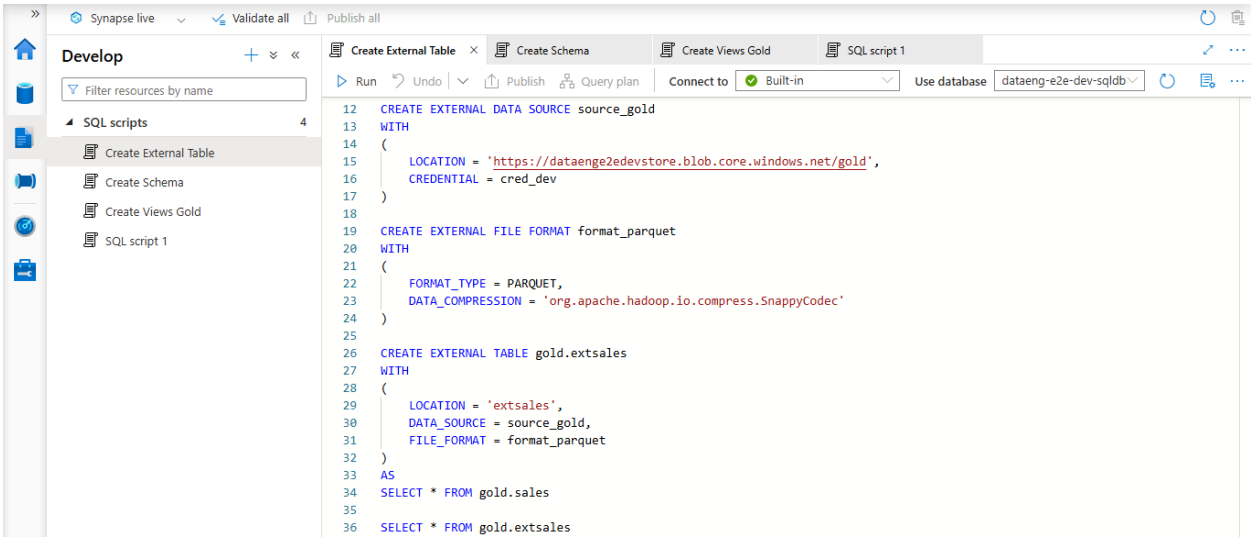

10:18 AM (<1s)
47
Python
df_procat.display()
(1) Spark Jobs
Table
Visualization 1
New charts: ON
ProductCategoryKey
CategoryName
Accessories
Clothing
Components
Bikes
10%
20%
30%
40%
Edit Visualization
4 rows
Refreshed 14 minutes ago

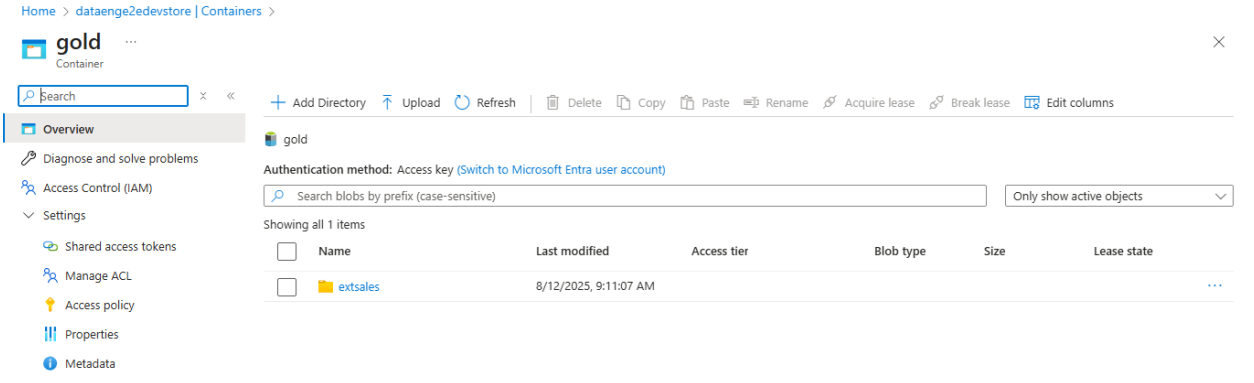

10:18 AM (<1s)
48
df_ter.display()
(1) Spark Jobs
Table
Visualization 1
New charts: ON
COUNT(SalesTerritoryKey)
0
1
2
3
4
5
Australia
Canada
France
Germany
United Kingdom
United States
Country
Edit Visualization
10 rows
Refreshed 15 minutes ago

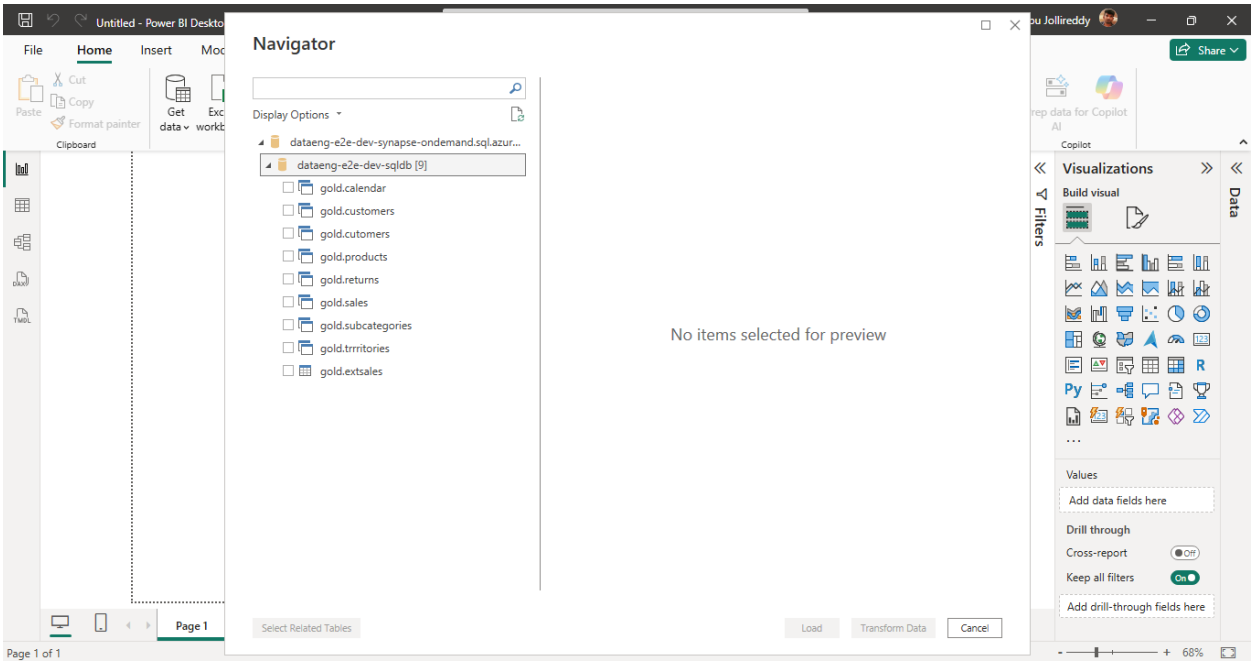

Home > dataenge2edevstore | Containers >
silver
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Add Directory
Upload
Refresh
Delete
Copy
Paste
Rename
Acquire lease
Break lease
Edit columns
silver
Authentication method: Access key (Switch to Microsoft Entra user account)
Search blobs by prefix (case-sensitive)
Only show active objects
Showing all 8 items
Name
Last modified
Access tier
Blob type
Size
Lease state
AdventureWorks_Calendar 8/12/2025, 10:18:30 AM
AdventureWorks_Customers 8/12/2025, 10:18:34 AM
AdventureWorks_Products 8/12/2025, 10:18:37 AM
AdventureWorks_Returns 8/12/2025, 10:18:39 AM
AdventureWorks_Sales 8/12/2025, 10:18:44 AM
AdventureWorks_Subcategories 8/12/2025, 10:18:35 AM
AdventureWorks_Territories 8/12/2025, 10:18:41 AM
_$azuretmpfolder$ 8/12/2025, 10:17:34 AM

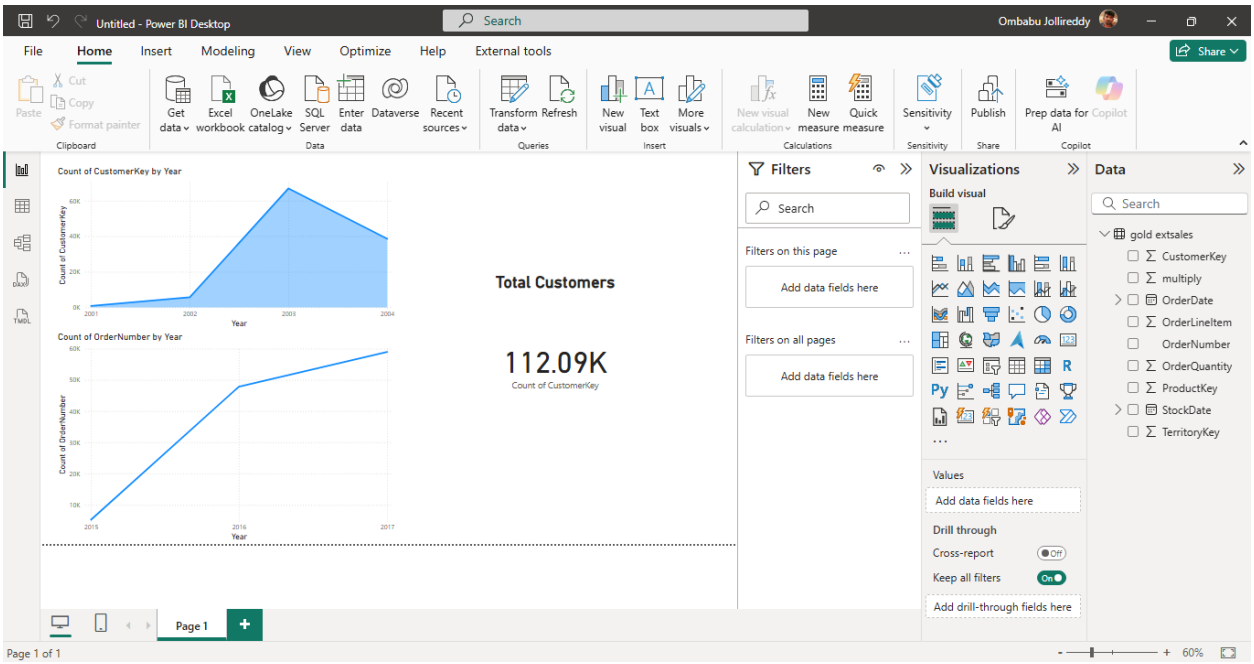

Home
Synapse Analytics workspace
dataeng-e2e-dev-synapse
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Discover more
Knowledge center
Browse partners


Synapse live
Validate all
Publish all
Develop
Filter resources by name
SQL scripts 4
Create External Table
Create Schema
Create Views Gold
SQL script 1
Create External Table
Create Schema
Create Views Gold
SQL script 1
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
dataeng-e2e-dev-sqldb
CREATE DATABASE SCOPED CREDENTIAL cred_dev
WITH
IDENTITY = 'Managed Identity'
CREATE EXTERNAL DATA SOURCE source_silver
WITH
(
LOCATION = 'https://dataenge2edevstore.blob.core.windows.net/silver',
CREDENTIAL = cred_dev
)


Synapse live
Validate all
Publish all
Develop
Filter resources by name
SQL scripts 4
Create External Table
Create Schema
Create Views Gold
SQL script 1
Create External Table
Create Schema
Create Views Gold
SQL script 1
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
dataeng-e2e-dev-sqldb
CREATE EXTERNAL DATA SOURCE source_gold
WITH
(
LOCATION = 'https://dataenge2edevstore.blob.core.windows.net/gold',
CREDENTIAL = cred_dev
)
CREATE EXTERNAL FILE FORMAT format_parquet
WITH
(
FORMAT_TYPE = PARQUET,
DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
)
CREATE EXTERNAL TABLE gold.extsales
WITH
(
LOCATION = 'extsales',
DATA_SOURCE = source_gold,
FILE_FORMAT = format_parquet
)
AS
SELECT * FROM gold.sales
SELECT * FROM gold.extsales


Home
dataenge2edevstore | Containers
gold
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Add Directory
Upload
Refresh
Delete
Copy
Paste
Rename
Acquire lease
Break lease
Edit columns
gold
Authentication method: Access key (Switch to Microsoft Entra user account)
Search blobs by prefix (case-sensitive)
Only show active objects
Showing all 1 items
Name
Last modified
Access tier
Blob type
Size
Lease state
extsales
8/12/2025, 9:11:07 AM


Untitled - Power BI Deskto
File
Home
Insert
Navigator
Display Options
dataeng-e2e-dev-synapse-ondemand.sql.azur...
dataeng-e2e-dev-sqldb [9]
gold.calendar
gold.customers
gold.cutomers
gold.products
gold.returns
gold.sales
gold.subcategories
gold.trrritories
gold.extsales
No items selected for preview
Select Related Tables
Load
Transform Data
Cancel
Visualizations
Build visual
Filters
Data
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Page 1
Page 1 of 1
68%


Untitled - Power BI Desktop
Search
Ombabu Jollireddy
File
Home
Insert
Modeling
View
Optimize
Help
External tools
Share
Count of CustomerKey by Year
Count of OrderNumber by Year
Year
Total Customers
112.09K
Count of CustomerKey
Filters
Filters on this page
Add data fields here
Filters on all pages
Visualizations
Build visual
Data
gold extsales
CustomerKey
multiply
OrderDate
OrderLineItem
OrderNumber
OrderQuantity
ProductKey
StockDate
TerritoryKey
Values
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Page 1
Page 1 of 1
60%