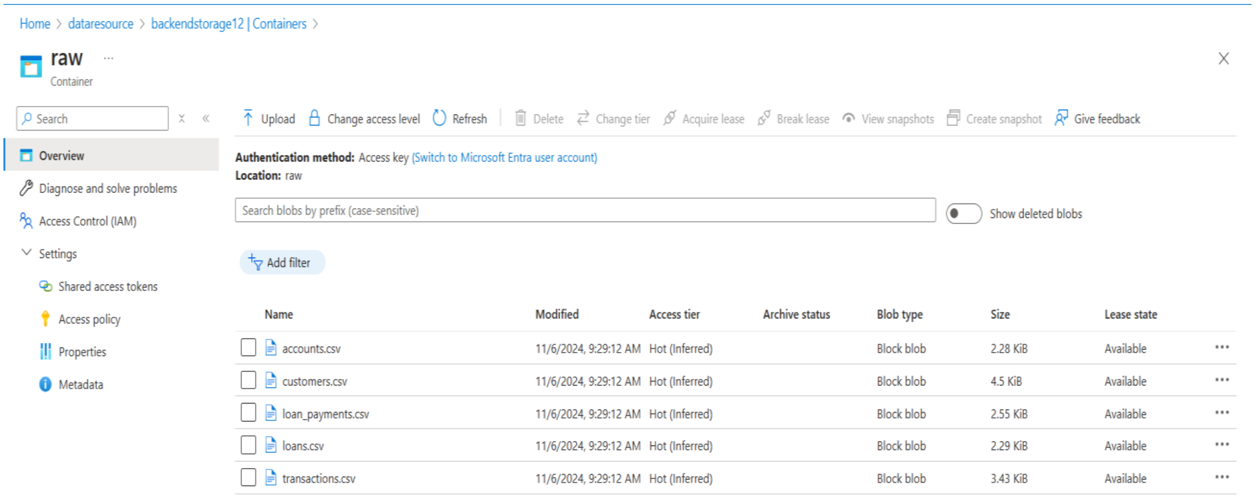
Data Pipeline for Customer

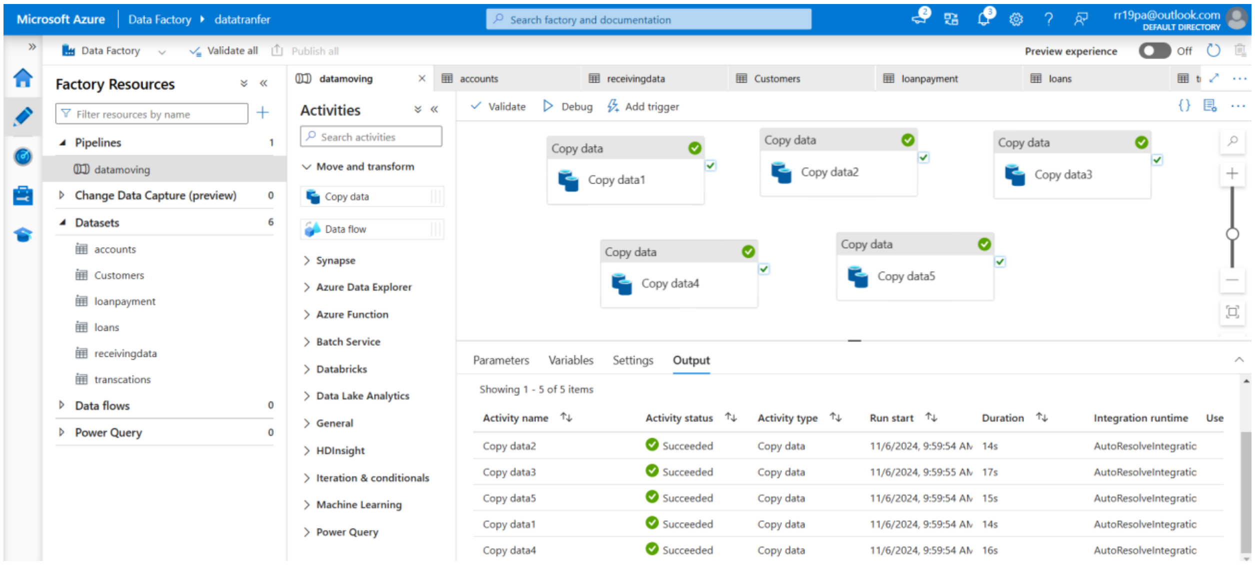
Step 1: Data Ingestion (Backend Storage to Raw(Bronze) Container)



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

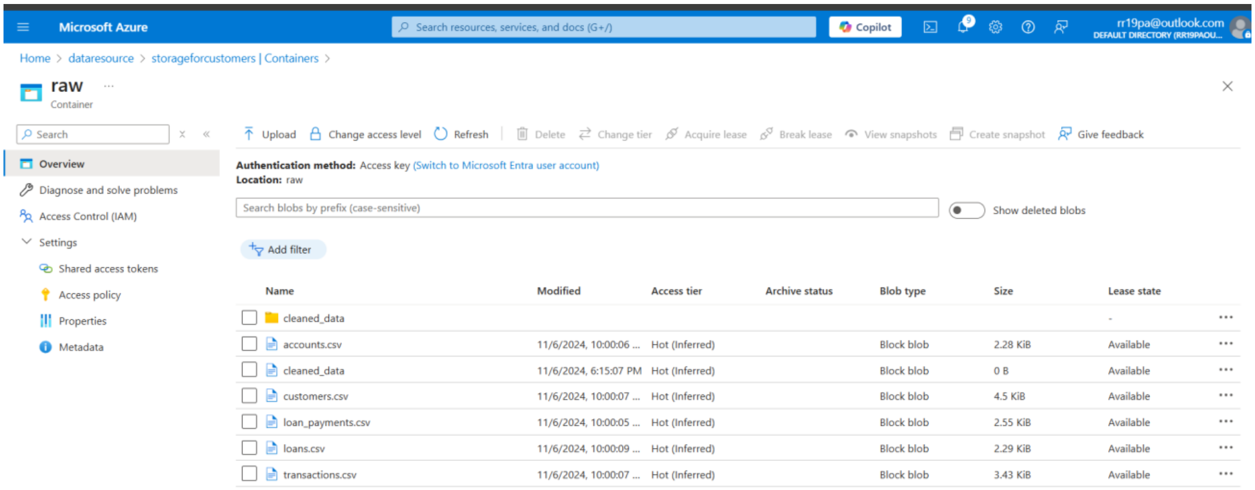
	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
□	accounts.csv	11/6/2024, 9:29:12 AM	Hot (Inferred)		Block blob	2.28 KiB	Available	...
□	customers.csv	11/6/2024, 9:29:12 AM	Hot (Inferred)		Block blob	4.5 KiB	Available	...
□	loan_payments.csv	11/6/2024, 9:29:12 AM	Hot (Inferred)		Block blob	2.55 KiB	Available	...
□	loans.csv	11/6/2024, 9:29:12 AM	Hot (Inferred)		Block blob	2.29 KiB	Available	...
□	transactions.csv	11/6/2024, 9:29:12 AM	Hot (Inferred)		Block blob	3.43 KiB	Available	...

We moved files from backendstorage to new storage using Data Factory



Showing 1 - 5 of 5 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	Use
Copy data2	✓ Succeeded	Copy data	11/6/2024, 9:59:54 AN	14s	AutoResolveIntegratic	
Copy data3	✓ Succeeded	Copy data	11/6/2024, 9:59:55 AN	17s	AutoResolveIntegratic	
Copy data5	✓ Succeeded	Copy data	11/6/2024, 9:59:54 AN	15s	AutoResolveIntegratic	
Copy data1	✓ Succeeded	Copy data	11/6/2024, 9:59:54 AN	14s	AutoResolveIntegratic	
Copy data4	✓ Succeeded	Copy data	11/6/2024, 9:59:54 AN	16s	AutoResolveIntegratic	

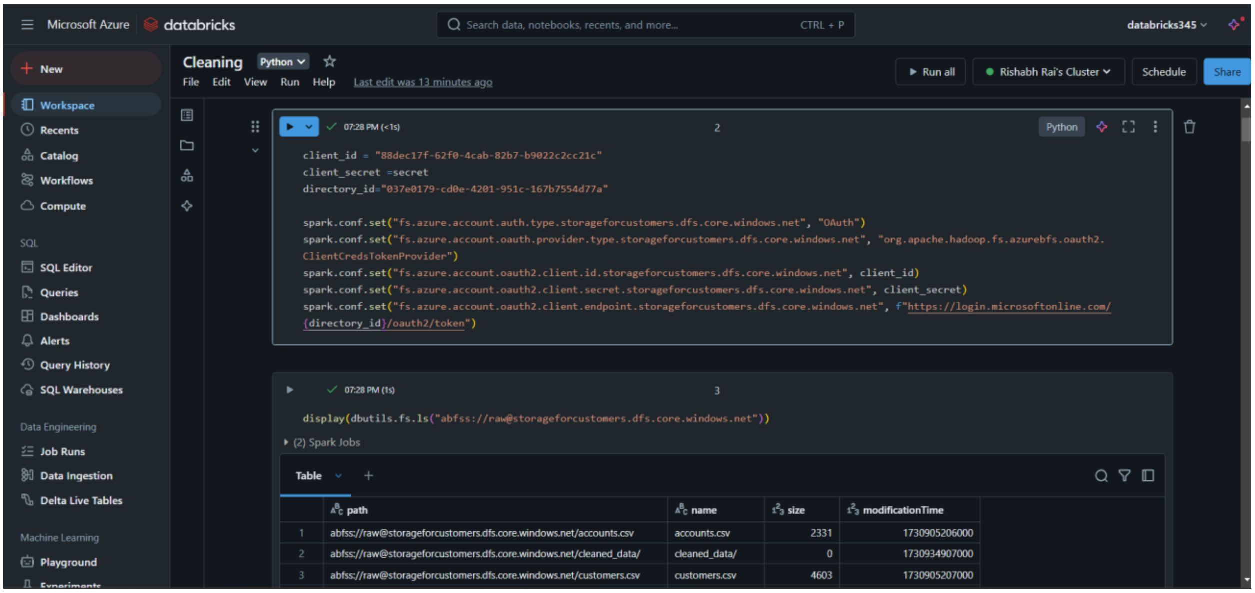


Authentication method: Access key (Switch to Microsoft Entra user account)
Location: raw

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
cleaned_data						-
accounts.csv	11/6/2024, 10:00:06 ...	Hot (Inferred)		Block blob	2.28 KiB	Available
cleaned_data	11/6/2024, 6:15:07 PM	Hot (Inferred)		Block blob	0 B	Available
customers.csv	11/6/2024, 10:00:07 ...	Hot (Inferred)		Block blob	4.5 KiB	Available
loan_payments.csv	11/6/2024, 10:00:05 ...	Hot (Inferred)		Block blob	2.55 KiB	Available
loans.csv	11/6/2024, 10:00:09 ...	Hot (Inferred)		Block blob	2.29 KiB	Available
transactions.csv	11/6/2024, 10:00:07 ...	Hot (Inferred)		Block blob	3.43 KiB	Available

Step 2: Databricks Activity (Incremental/Delta Processing)

Here we read data and clean all the data and move them to cleaned_data



```
client_id = "88dec17f-62f0-4cab-82b7-b9022c2cc21c"
client_secret =secret
directory_id="037e0179-cd0e-4201-951c-167b7554d77a"

spark.conf.set("fs.azure.account.auth.type.storageforcustomers.dfs.core.windows.net", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type.storageforcustomers.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.
ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.storageforcustomers.dfs.core.windows.net", client_id)
spark.conf.set("fs.azure.account.oauth2.client.secret.storageforcustomers.dfs.core.windows.net", client_secret)
spark.conf.set("fs.azure.account.oauth2.client.endpoint.storageforcustomers.dfs.core.windows.net", f"https://login.microsoftonline.com/
{directory_id}/oauth2/token")
```

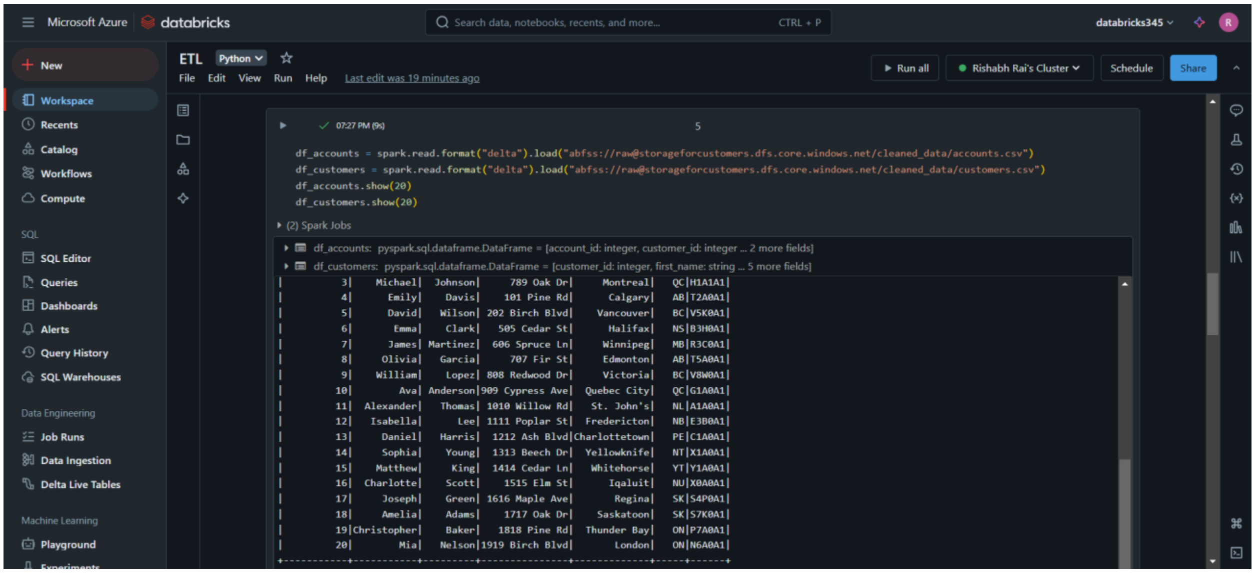
```
display(dbutils.fs.ls("abfss://raw@storageforcustomers.dfs.core.windows.net"))
```

	path	name	size	modificationTime
1	abfss://raw@storageforcustomers.dfs.core.windows.net/accounts.csv	accounts.csv	2331	1730905206000
2	abfss://raw@storageforcustomers.dfs.core.windows.net/cleaned_data/	cleaned_data/	0	1730934907000
3	abfss://raw@storageforcustomers.dfs.core.windows.net/customers.csv	customers.csv	4603	1730905207000

also used key vault to access id secret and you can see it in cleaning. Ipynb

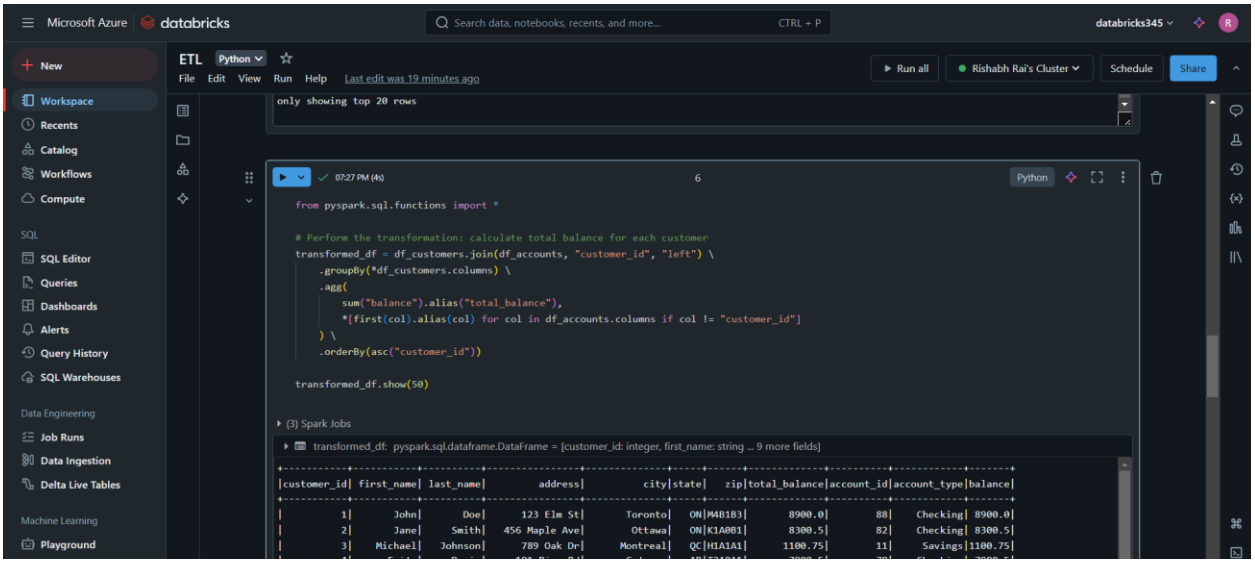
Step 3: Databricks Activity (ETL Processing)

We created another databricks notebook for ETL Processing here we read all the cleaned data we created in step 2 and wrote the query to calculate the total balance across all accounts for each customer, ensuring that all columns from the accounts and customers tables are selected and included and stored it in gold container.



```
df_accounts = spark.read.format("delta").load("abfss://raw@storageforcustomers.dfs.core.windows.net/cleaned_data/accounts.csv")
df_customers = spark.read.format("delta").load("abfss://raw@storageforcustomers.dfs.core.windows.net/cleaned_data/customers.csv")
df_accounts.show(20)
df_customers.show(20)
```

```
(2) Spark Jobs
df_accounts:  pyspark.sql.dataframe.DataFrame = [account_id: integer, customer_id: integer ... 2 more fields]
df_customers:  pyspark.sql.dataframe.DataFrame = [customer_id: integer, first_name: string ... 5 more fields]
|          3|    Michael|  Johnson|      789 Oak Dr|      Montreal|   QC|H1A1A1|
|          4|      Emily|    Davis|     101 Pine Rd|       Calgary|   AB|T2A0A1|
|          5|      David|   Wilson|  202 Birch Blvd|     Vancouver|   BC|V5K0A1|
|          6|       Emma|    Clark|   505 Cedar St|       Halifax|   NS|B3H0A1|
|          7|      James| Martinez|  606 Spruce Ln|      Winnipeg|   MB|R3C0A1|
|          8|     Olivia|   Garcia|      707 Fir St|      Edmonton|   AB|T5A0A1|
|          9|    William|    Lopez| 808 Redwood Dr|      Victoria|   BC|V8W0A1|
|         10|        Ava| Anderson|909 Cypress Ave|   Quebec City|   QC|G1A0A1|
|         11|  Alexander|   Thomas| 1010 Willow Rd|    St. John's|   NL|A1A0A1|
|         12|   Isabella|      Lee| 1111 Poplar St|   Fredericton|   NB|E3B0A1|
|         13|     Daniel|   Harris|  1212 Ash Blvd|Charlottetown|   PE|C1A0A1|
|         14|     Sophia|    Young|  1313 Beech Dr|   Yellowknife|   NT|X1A0A1|
|         15|    Matthew|     King|  1414 Cedar Ln|    Whitehorse|   YT|Y1A0A1|
|         16|  Charlotte|    Scott|    1515 Elm St|       Iqaluit|   NU|X0A0A1|
|         17|     Joseph|    Green| 1616 Maple Ave|        Regina|   SK|S4P0A1|
|         18|     Amelia|    Adams|    1717 Oak Dr|     Saskatoon|   SK|S7K0A1|
|         19|Christopher|    Baker|   1818 Pine Rd|   Thunder Bay|   ON|P7A0A1|
|         20|        Mia|   Nelson|1919 Birch Blvd|        London|   ON|N6A0A1|
+-----------+-----------+---------+---------------+-------------+-----+------+
```



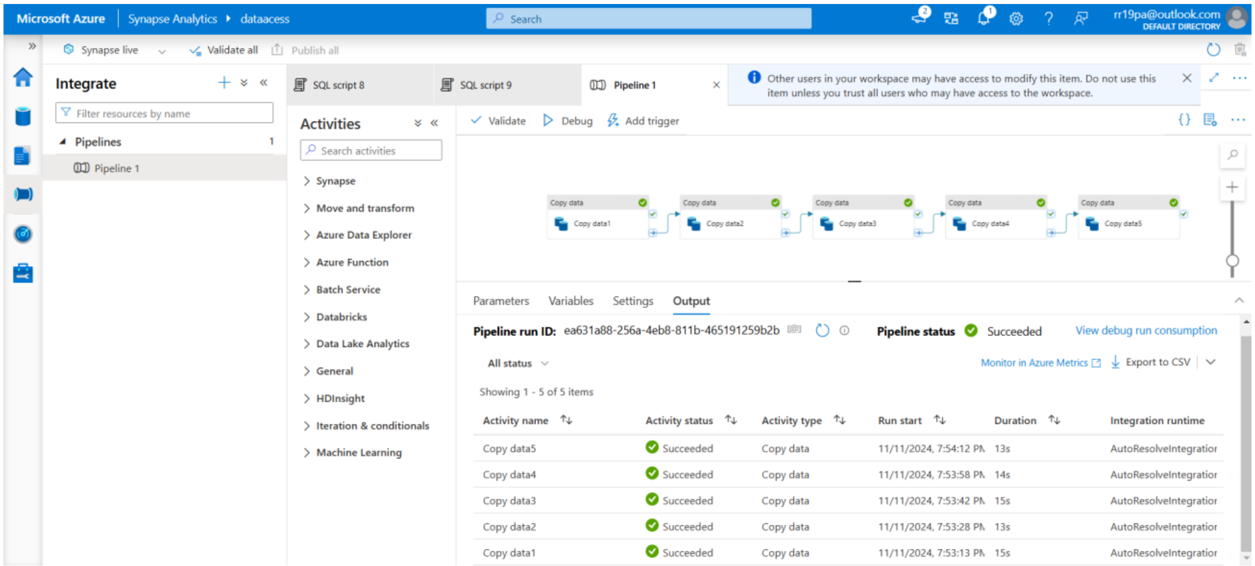
```
only showing top 20 rows
```

```
from pyspark.sql.functions import *

# Perform the transformation: calculate total balance for each customer
transformed_df = df_customers.join(df_accounts, "customer_id", "left") \
    .groupBy(*df_customers.columns) \
    .agg(
        sum("balance").alias("total_balance"),
        *[first(col).alias(col) for col in df_accounts.columns if col != "customer_id"]
    ) \
    .orderBy(asc("customer_id"))

transformed_df.show(50)
```

```
(3) Spark Jobs
transformed_df:  pyspark.sql.dataframe.DataFrame = [customer_id: integer, first_name: string ... 9 more fields]
+-----------+----------+---------+---------------+-------------+-----+------+-------------+----------+------------+-------+
|customer_id| first_name| last_name|        address|         city|state|   zip|total_balance|account_id|account_type|balance|
+-----------+----------+---------+---------------+-------------+-----+------+-------------+----------+------------+-------+
|          1|      John|      Doe|     123 Elm St|      Toronto|   ON|M4B1B3|       8900.0|        88|    Checking| 8900.0|
|          2|      Jane|    Smith|  456 Maple Ave|       Ottawa|   ON|K1A0B1|       8300.5|        82|    Checking| 8300.5|
|          3|   Michael|  Johnson|     789 Oak Dr|     Montreal|   QC|H1A1A1|      1100.75|        11|     Savings|1100.75|
```

Step 4: Azure Synapse Analytics

Create external tables in Azure Synapse Analytics to map to the data stored in the Curated(silver) and Refined(gold) containers of your data lake. This allows data analysts and business intelligence teams to access and query the data directly using tools like Synapse Studio or notebooks.



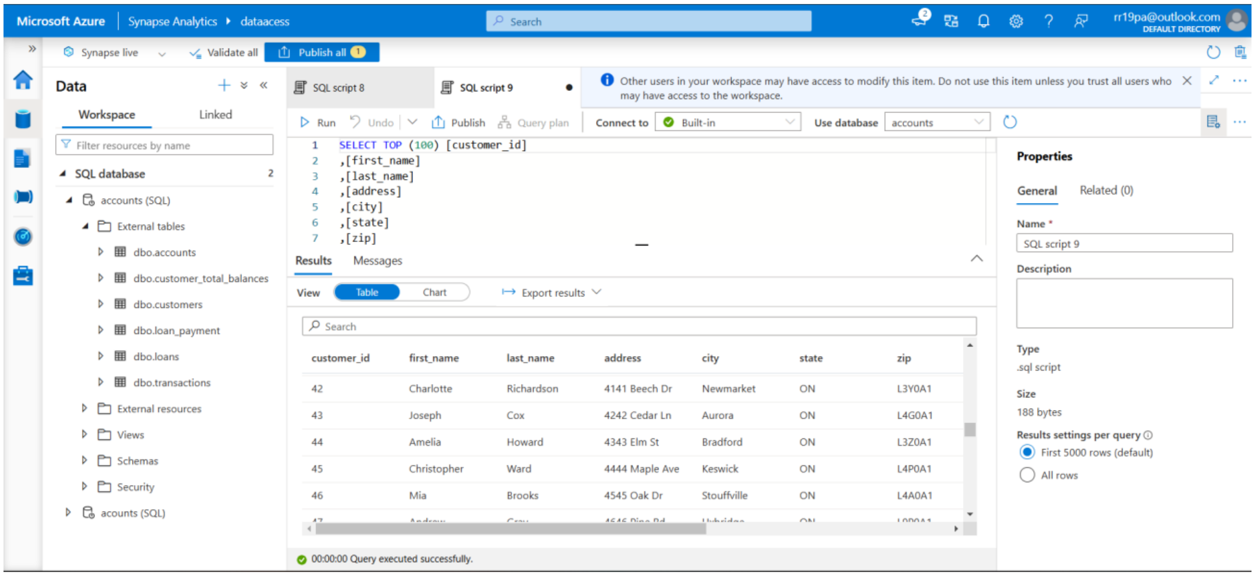
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Pipeline run ID: ea631a88-256a-4eb8-811b-465191259b2b

Pipeline status ✓ Succeeded

Showing 1 - 5 of 5 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
Copy data5	✓ Succeeded	Copy data	11/11/2024, 7:54:12 PM	13s	AutoResolveIntegratior
Copy data4	✓ Succeeded	Copy data	11/11/2024, 7:53:58 PM	14s	AutoResolveIntegratior
Copy data3	✓ Succeeded	Copy data	11/11/2024, 7:53:42 PM	15s	AutoResolveIntegratior
Copy data2	✓ Succeeded	Copy data	11/11/2024, 7:53:28 PM	13s	AutoResolveIntegratior
Copy data1	✓ Succeeded	Copy data	11/11/2024, 7:53:13 PM	15s	AutoResolveIntegratior



Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

```sql
SELECT TOP (100) [customer_id]
,[first_name]
,[last_name]
,[address]
,[city]
,[state]
,[zip]
```

customer_id	first_name	last_name	address	city	state	zip
42	Charlotte	Richardson	4141 Beech Dr	Newmarket	ON	L3Y0A1
43	Joseph	Cox	4242 Cedar Ln	Aurora	ON	L4G0A1
44	Amelia	Howard	4343 Elm St	Bradford	ON	L3Z0A1
45	Christopher	Ward	4444 Maple Ave	Keswick	ON	L4P0A1
46	Mia	Brooks	4545 Oak Dr	Stouffville	ON	L4A0A1
47	Andrew	Gray	4646 Pine Rd	Uxbridge	ON	L9P0A1

✓ 00:00:00 Query executed successfully.

Name *: SQL script 9

Description:

Type: .sql script

Size: 188 bytes

Results settings per query
● First 5000 rows (default)
○ All rows