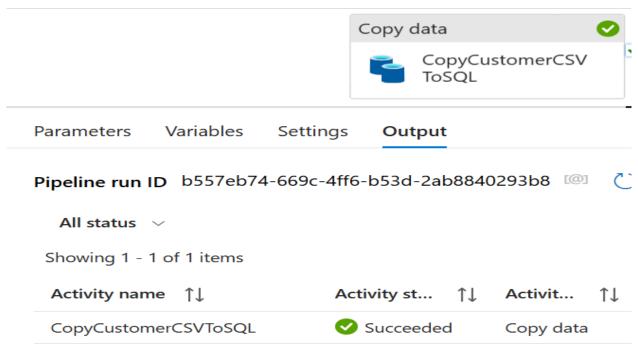
## **Azure Data Quality Validation Project**

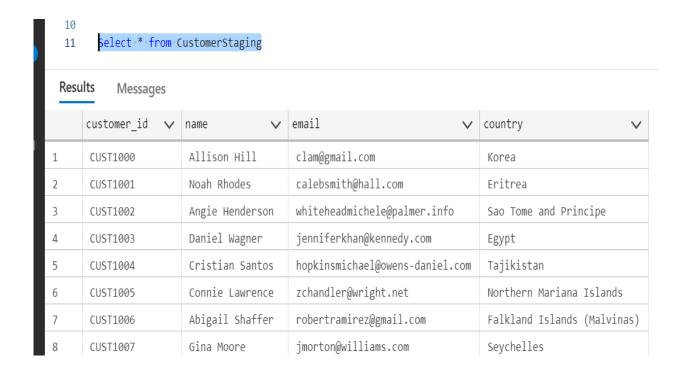
**Step 1: Raw Data Source in Azure Data Lake (ADLS)** 

🗐 raw-banking-data							
Authentication method: Access key (Switch to Microsoft Entra user account)							
Search blobs by prefix (case-sensitive)							
Showing all 5 items							
	Name	Last modified	Access tier				
	Customers	7/27/2025, 2:06:00 PM					
	Transactions	7/27/2025, 2:06:07 PM					
	customers.csv	8/7/2025, 2:59:40 PM	Hot (Inferred)				
	products.csv	8/7/2025, 2:59:40 PM	Hot (Inferred)				
	transactions.csv	8/7/2025, 2:59:40 PM	Hot (Inferred)				

**Step 2: Ingestion using Azure Data Factory (ADF)** 



Step 3: Staging Table (CustomerStaging) in Azure SQL



**Step 4: Data Validation in Azure Databricks** 

```
databricks
                               Q Search data, notebooks, recents, and more.
                                                                                                                                   adbro
  dq_batch_validation.py Python ➤ Tabs: OFF ➤ ☆
                                                                             田
                                                                                    ▶ Run all
                                                                                                 Data_Transform_Cluster ~
                                                                                                                              Schedule
  File Edit View Run Help
                               Last edit was 4 minutes ago
  ⊞
                       4 minutes ago (18s)
 from pyspark.sql import SparkSession
                   from pyspark.sql.functions import col, lit, when
  品
                   # Define DQValidator class
  class DQValidator:
                       def __init__(self, df):
                       def check_nulls(self, columns):
                           return self.df.filter(
                               " OR ".join([f"{col} IS NULL" for col in columns])
                       def check_duplicates(self, subset):
                           return self.df.groupBy(subset).count().filter("count > 1")
```

Step 5: Cleaned & Validated Data Stored to Azure SQL (CustomerValidated)

## 19 Select \* from CustomerValidated 20 Select \* from CustomerRejected

Results Messages

	customer_id 🗸	name 🗸	email 🗸	country		
1	CUST1000	Allison Hill	clam@gmail.com	Korea		
2	CUST1001	Noah Rhodes	calebsmith@hall.com	Eritrea		
3	CUST1002	Angie Henderson	whiteheadmichele@palmer.info	Sao Tome and Principe		
4	CUST1003	Daniel Wagner	jenniferkhan@kennedy.com	Egypt		
5	CUST1004	Cristian Santos	hopkinsmichael@owens-daniel.com	Tajikistan		
6	CUST1005	Connie Lawrence	zchandler@wright.net	Northern Mariana Islands		
7	CUST1006	Abigail Shaffer	robertramirez@gmail.com	Falkland Islands (Malvinas)		
8	CUST1007	Gina Moore	jmorton@williams.com	Seychelles		
9	CUST1008	Gabrielle Davis	gabrieltucker@hancock.com	Slovakia (Slovak Republic)		
10	CUST1009	Ryan Munoz	gallowayjoseph@yahoo.com	Maldives		
11	CUST1010	Monica Herrera	uhorton@hotmail.com	Libyan Arab Jamahiriya		
12	CUST1011	Jamie Arnold	jamesrobinson@gmail.com	Nigeria		
13	CUST1012	Lisa Hensley	brian97@calhoun.net	Dominican Republic		