




## 1) Storage Setup

+ Add container
↶ Upload
↻ Refresh
🗑 Delete
🔒 Change access level
↶ Restore containers
🔧 Edit columns

Only show active containers

Showing all 3 items

<input type="checkbox"/>	Name	Last modified	Anonymous access level	Lease state
<input type="checkbox"/>	 \$logs	6/25/2025, 11:06:45 AM	Private	Available <span>⋮</span>
<input type="checkbox"/>	 fraud-data	7/11/2025, 9:41:01 AM	Private	Available <span>⋮</span>
<input type="checkbox"/>	 sparkhexa	6/25/2025, 11:40:40 AM	Private	Available <span>⋮</span>

+ Add Directory
↑ Upload
🔒 Change access level
🔄 Refresh

🗑️ Delete
📄 Copy
📄 Paste
🔄 Rename
🔑 Acquire lease
🔑 Break lease
🔧 Edit columns

📁 fraud-data

**Authentication method:** Access key ([Switch to Microsoft Entra user account](#))

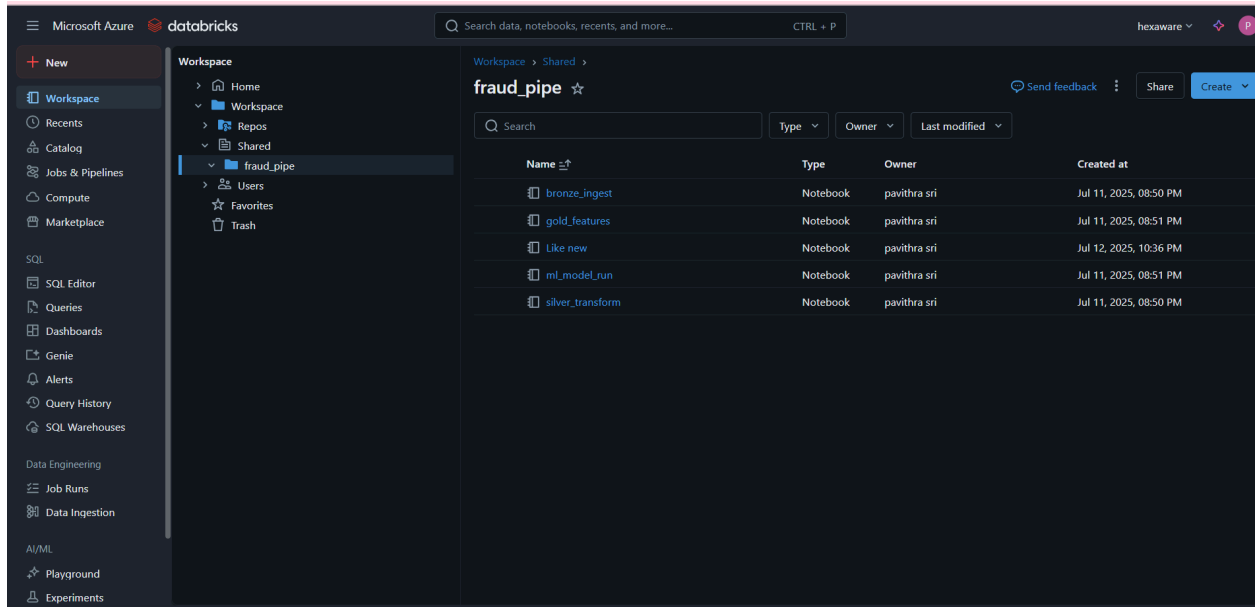
🔍 Add filter

Only show active blobs

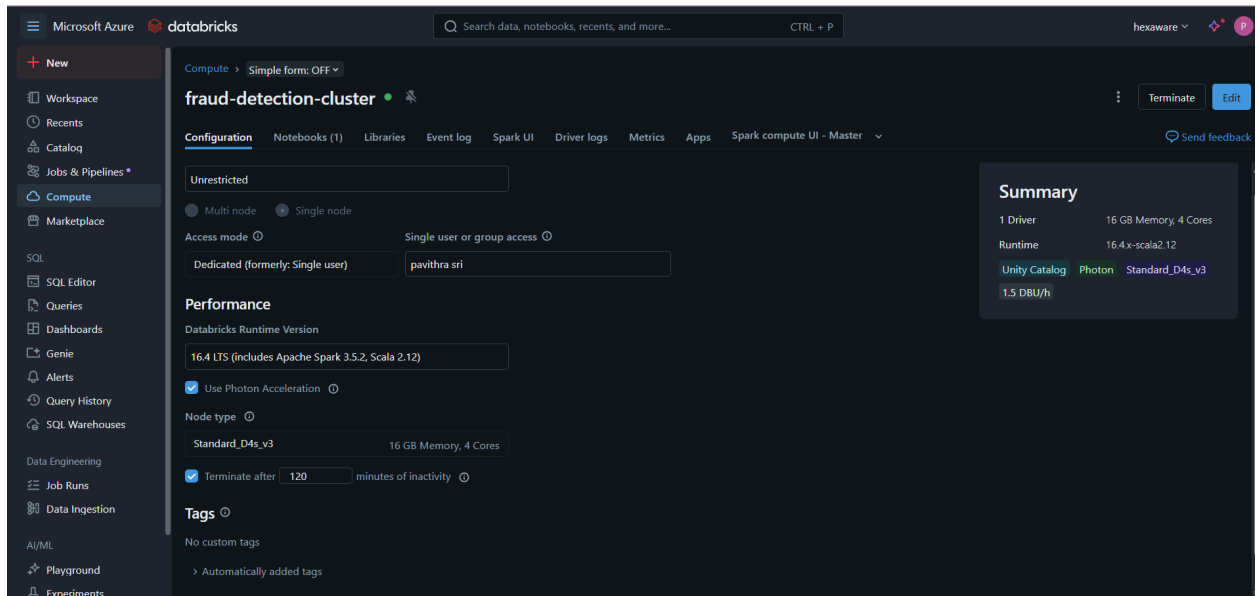
Showing all 1 items

<input type="checkbox"/>	Name	Last modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>	📄 PS_20174392719_1491204439457_log.csv	7/11/2025, 9:48:08 AM	Hot (Inferred)	Block blob	470.67 MiB	Available

## 2) Notebooks Overview



## 3) Cluster in databricks



#### 4) Reading and printing initial bronze table

```
df_bronze: pyspark.sql.dataframe.DataFrame = [step: integer, type: string ... 9 more fields]
|-- type: string (nullable = true)
|-- amount: double (nullable = true)
|-- nameOrig: string (nullable = true)
|-- oldbalanceOrig: double (nullable = true)
|-- newbalanceOrig: double (nullable = true)
|-- nameDest: string (nullable = true)
|-- oldbalanceDest: double (nullable = true)
|-- newbalanceDest: double (nullable = true)
|-- isFraud: integer (nullable = true)
|-- isFlaggedFraud: integer (nullable = true)

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|step|  type| amount| nameOrig|oldbalanceOrig|newbalanceOrig| nameDest|oldbalanceDest|newbalanceDest|isFraud|isFlaggedFraud|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|  1| PAYMENT| 9839.64|C1231006815| 170136.0| 160296.36|M1979787155|      0.0|      0.0|      0|      0|
|  1| PAYMENT| 1864.28|C1666544295|  21249.0|  19384.72|M2044282225|      0.0|      0.0|      0|      0|
|  1| TRANSFER|  181.0|C1305486145|    181.0|      0.0| C553264065|      0.0|      0.0|      1|      0|
|  1| CASH_OUT|  181.0| C840083671|    181.0|      0.0| C38997010|  21182.0|      0.0|      1|      0|
|  1| PAYMENT|11668.14|C2048537720|  41554.0|  29885.86|M1230701703|      0.0|      0.0|      0|      0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows
```

#### 5) Initial dataframe

```
df_raw: pyspark.sql.dataframe.DataFrame
  step: string
  type: string
  amount: string
  nameOrig: string
  oldbalanceOrig: string
  newbalanceOrig: string
  nameDest: string
  oldbalanceDest: string
  newbalanceDest: string
  isFraud: string
  isFlaggedFraud: string
  _rescued_data: string

<pyspark.sql.streaming.query.StreamingQuery at 0x7f0de0db7890>
```

## 6) Bronze table:-

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest
1	249	PAYMENT	3119.51	C1515157403	15175.2	12055.69	M880864628	0.0
2	249	PAYMENT	11494.08	C1197653668	12055.69	561.61	M1158991064	0.0
3	249	PAYMENT	14220.43	C1682344014	561.61	0.0	M1461592831	0.0
4	249	PAYMENT	334.8	C89862438	0.0	0.0	M1716146009	0.0
5	249	PAYMENT	13467.2	C907984678	0.0	0.0	M271674048	0.0
6	249	PAYMENT	34657.33	C1279245281	0.0	0.0	M2017138104	0.0
7	249	PAYMENT	4064.65	C991912470	0.0	0.0	M551774431	0.0
8	249	PAYMENT	10756.58	C288501649	0.0	0.0	M830766810	0.0
9	249	PAYMENT	2027.2	C1900786990	0.0	0.0	M1247120968	0.0
10	249	PAYMENT	6208.09	C1243315764	0.0	0.0	M1492892419	0.0
11	249	PAYMENT	497.74	C1949669367	0.0	0.0	M309620983	0.0
12	249	PAYMENT	403.47	C59721299	0.0	0.0	M103401090	0.0
13	249	PAYMENT	1338.56	C1585148831	0.0	0.0	M1866023791	0.0
14	249	PAYMENT	287.29	C385694981	0.0	0.0	M1579738254	0.0

## 7) Clean and transform the data and saving to the delta table:-

SchemaDetailsHistory

```
step: string
type: string
amount: string
nameOrig: string
oldbalanceOrg: string
newbalanceOrig: string
nameDest: string
oldbalanceDest: string
newbalanceDest: string
isFraud: string
isFlaggedFraud: string
_rescued_data: string
```

df\_silver: pyspark.sql.dataframe.DataFrame

```
step: string
type: string
amount: string
originator: string
oldbalanceOrg: string
newbalanceOrig: string
receiver: string
oldbalanceDest: string
newbalanceDest: string
_rescued_data: string
is_fraud: integer
is_flagged_fraud: integer
```

<pyspark.sql.streaming.query.StreamingQuery at 0x7f0de0db5910>

## 8) Displaying silver transformed table:-

	A <sup>B</sup> <sub>C</sub> amount	A <sup>B</sup> <sub>C</sub> originator	A <sup>B</sup> <sub>C</sub> oldbalanceOrg	A <sup>B</sup> <sub>C</sub> newbalanceOrig	A <sup>B</sup> <sub>C</sub> receiver	A <sup>B</sup> <sub>C</sub> oldbalanceDest	A <sup>B</sup> <sub>C</sub> newbalanceDest
1	3119.51	C1515157403	15175.2	12055.69	M880864628	0.0	0.0
2	11494.08	C1197653668	12055.69	561.61	M1158991064	0.0	0.0
3	14220.43	C1682344014	561.61	0.0	M1461592831	0.0	0.0
4	334.8	C89862438	0.0	0.0	M1716146009	0.0	0.0
5	13467.2	C907984678	0.0	0.0	M271674048	0.0	0.0
6	34657.33	C1279245281	0.0	0.0	M2017138104	0.0	0.0
7	4064.65	C991912470	0.0	0.0	M551774431	0.0	0.0
8	10756.58	C288501649	0.0	0.0	M830766810	0.0	0.0
9	2027.2	C1900786990	0.0	0.0	M1247120968	0.0	0.0
10	6208.09	C1243315764	0.0	0.0	M1492892419	0.0	0.0
11	497.74	C1949669367	0.0	0.0	M309620983	0.0	0.0
12	403.47	C59721299	0.0	0.0	M103401090	0.0	0.0
13	1338.56	C1585148831	0.0	0.0	M1866023791	0.0	0.0
14	287.29	C385694981	0.0	0.0	M1579738254	0.0	0.0

## 9) Feature engineering - example aggregations and saving it to the gold delta table:-

df_gold: pyspark.sql.dataframe.DataFrame
originator: string
total_transactions: long
total_amount_sent: double
average_amount_sent: double
fraud_count: long
df_silver: pyspark.sql.dataframe.DataFrame
Schema
Details
History
step: string
type: string
amount: string
originator: string
oldbalanceOrg: string
newbalanceOrig: string
receiver: string
oldbalanceDest: string
newbalanceDest: string
_rescued_data: string
is_fraud: integer
is_flagged_fraud: integer

## 10) Displaying gold table:-

	<sup>B</sup> <sub>C</sub> originator	<sup>1</sup> <sub>3</sub> total_transactions	<sup>1</sup> <sub>2</sub> total_amount_sent	<sup>1</sup> <sub>2</sub> average_amount_sent	<sup>1</sup> <sub>3</sub> fraud_count
1	C876714021	1	443010.87	443010.87	0
2	C224938823	1	2959.54	2959.54	0
3	C1676650606	1	34739.91	34739.91	0
4	C289153199	1	371.23	371.23	0
5	C1181048539	1	17411.44	17411.44	0
6	C1503438438	1	196626.77	196626.77	0
7	C765877097	1	63191.19	63191.19	0
8	C47278966	1	287868.3	287868.3	0
9	C326407304	1	250341.94	250341.94	0
10	C1285801989	1	90349.08	90349.08	0
11	C1424608077	1	13656.17	13656.17	0
12	C714358180	1	401340.14	401340.14	0
13	C909652139	1	29606.84	29606.84	0
14	C2124981192	1	313025.88	313025.88	0
15	C17536176	1	203336.28	203336.28	0

## 11)Gold table sample, schema, and count of records:-

	<sup>B</sup> <sub>C</sub> originator	<sup>1</sup> <sub>3</sub> total_transactions	<sup>1</sup> <sub>2</sub> total_amount_sent	<sup>1</sup> <sub>2</sub> average_amount_sent	<sup>1</sup> <sub>3</sub> fraud_count
1	C876714021	1	443010.87	443010.87	0
2	C224938823	1	2959.54	2959.54	0
3	C1676650606	1	34739.91	34739.91	0
4	C289153199	1	371.23	371.23	0
5	C1181048539	1	17411.44	17411.44	0

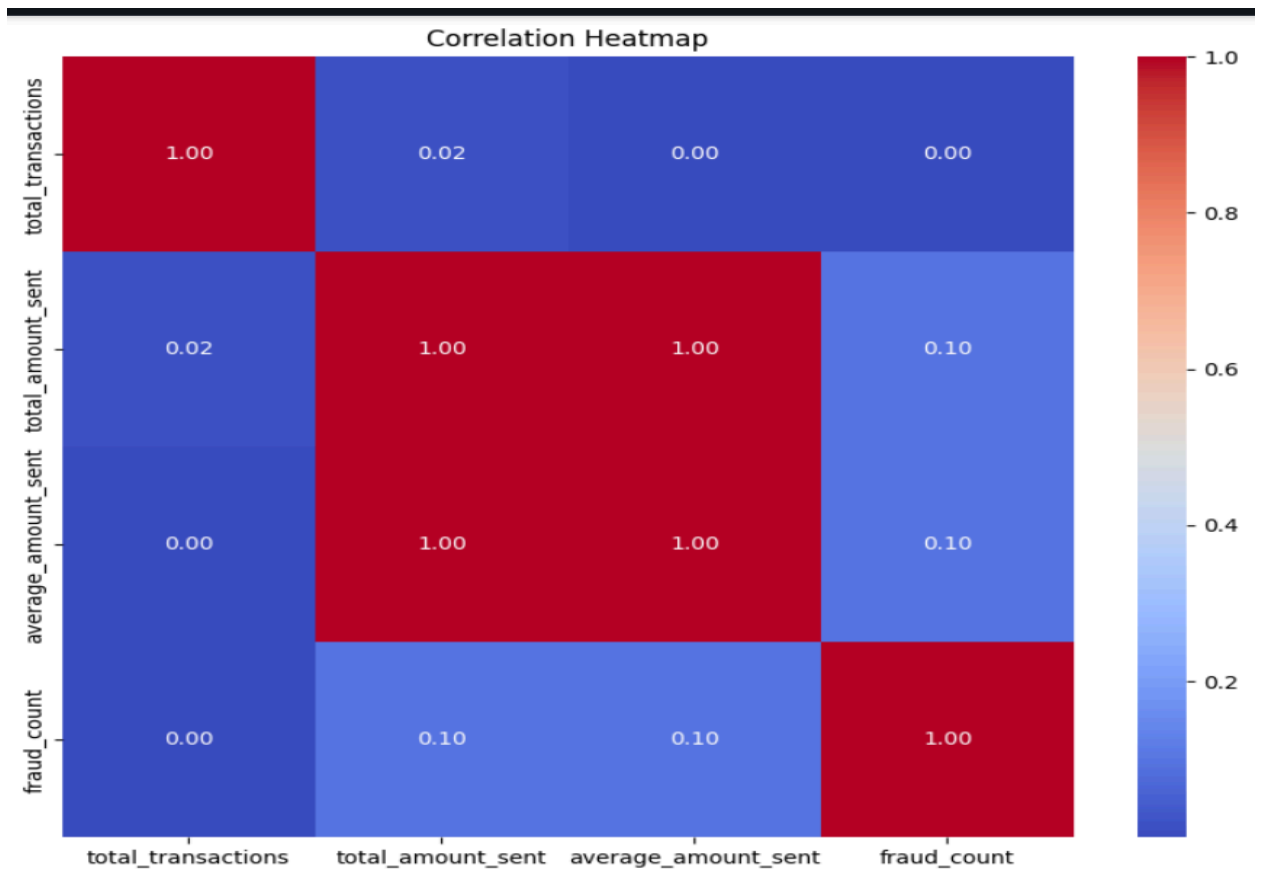
↓ 5 rows

```
root
|-- originator: string (nullable = true)
|-- total_transactions: long (nullable = true)
|-- total_amount_sent: double (nullable = true)
|-- average_amount_sent: double (nullable = true)
|-- fraud_count: long (nullable = true)
```

Gold Table Record Count: 6347542



## 15) Correlation Heatmap



## 16) Train model

	total_transactions	total_amount_sent	average_amount_sent	fraud_count	anomaly_flag
0	1	5014.17	5014.17	0	0
1	1	15188.56	15188.56	0	0
2	1	9349.98	9349.98	0	0
3	1	35423.27	35423.27	0	0
4	1	71906.86	71906.86	0	0
5	1	80557.71	80557.71	0	0
6	1	7762.46	7762.46	0	0
7	1	8759.81	8759.81	0	0
8	1	405.30	405.30	0	0
9	1	8600.77	8600.77	0	0

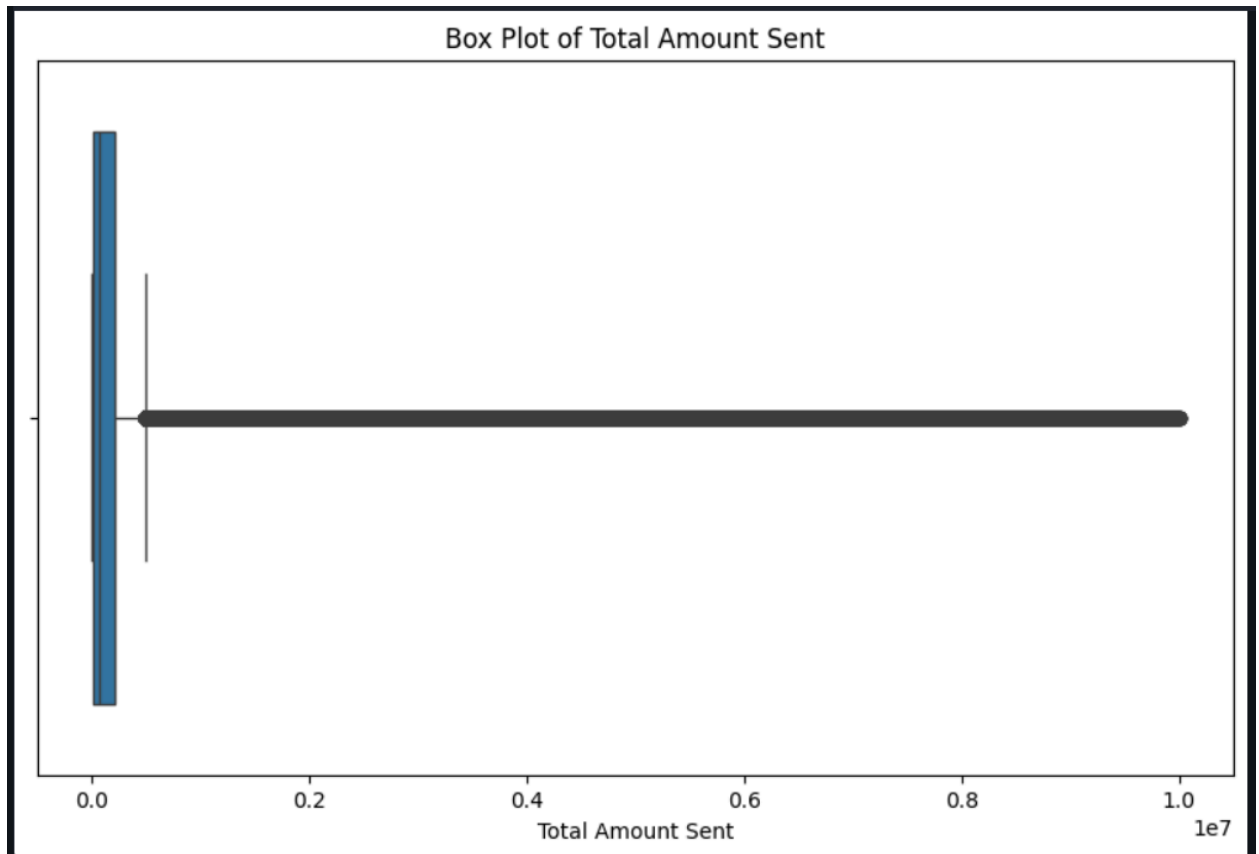


## 17) Save final output with anomaly flag to Delta Lake

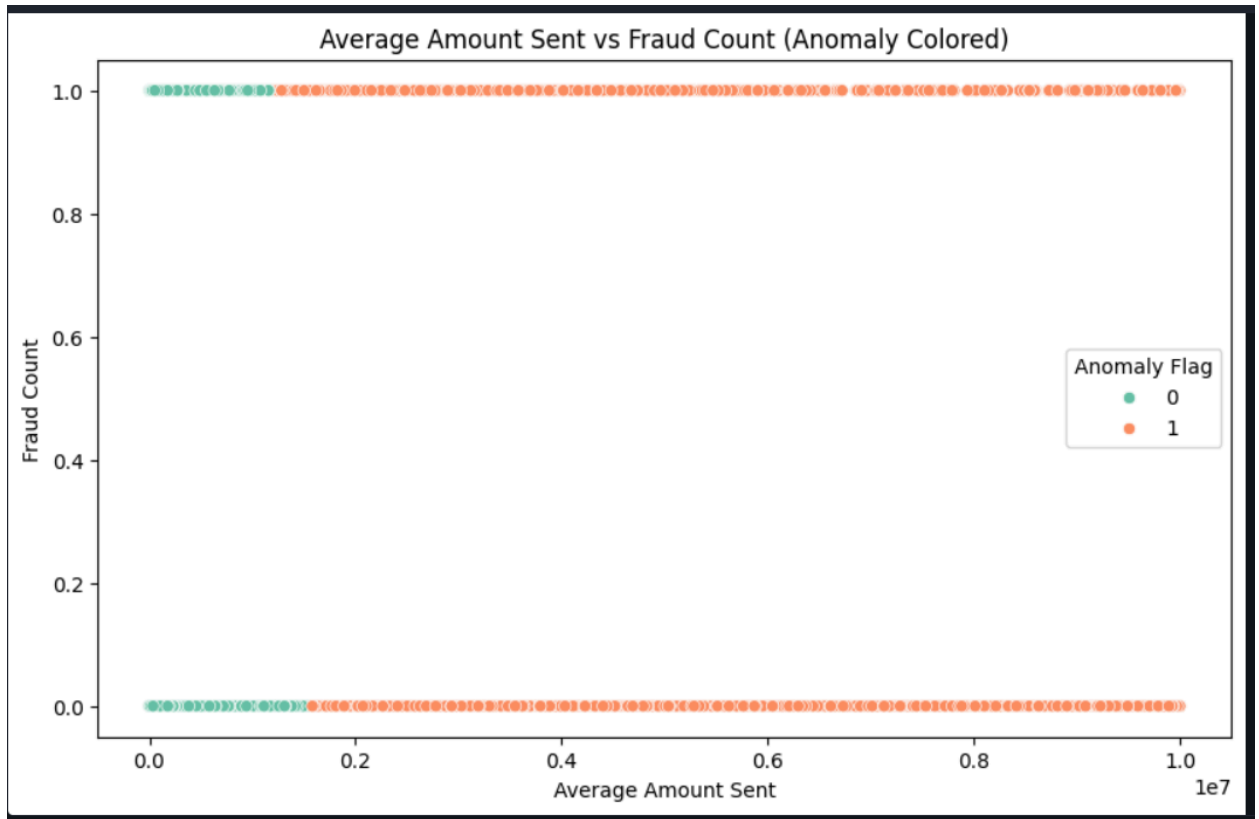
```
anomaly_sdf: pyspark.sql.dataframe.DataFrame = [total_transactions: long, total_amount_sent: double ... 3 more fields]
```

Anomaly detection results saved to Delta successfully.

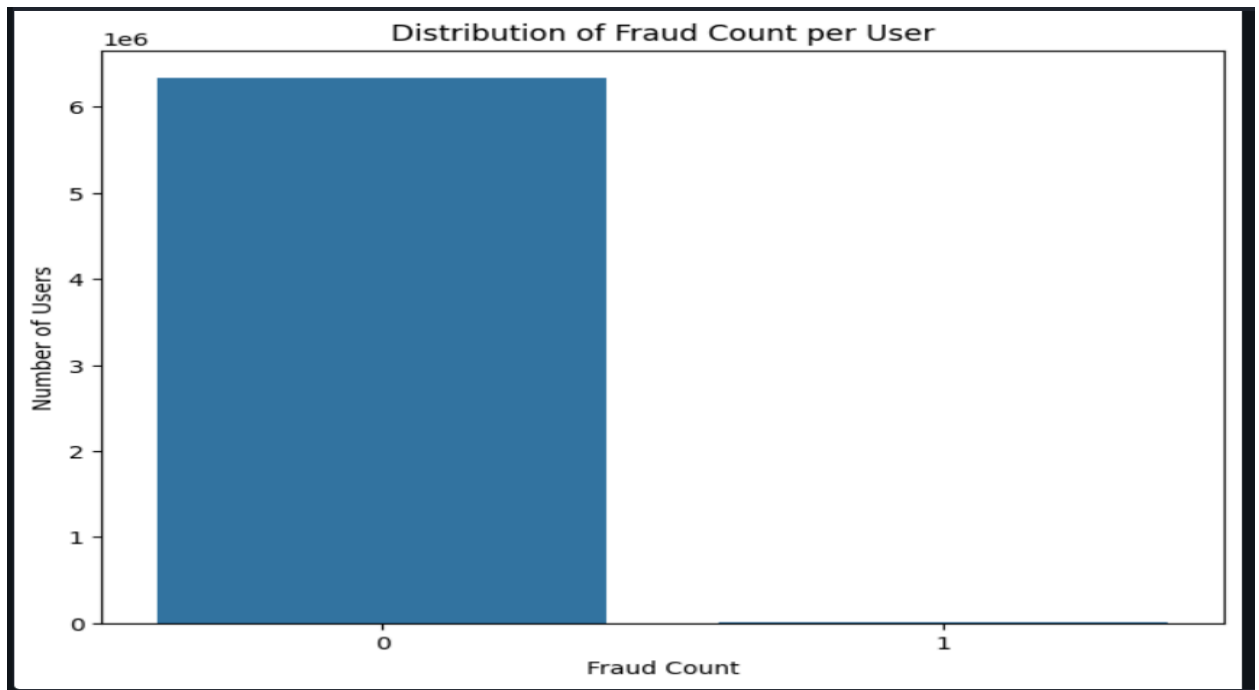
## 18) Box plot



## 19) Anomaly colored



## 20) Bar graph



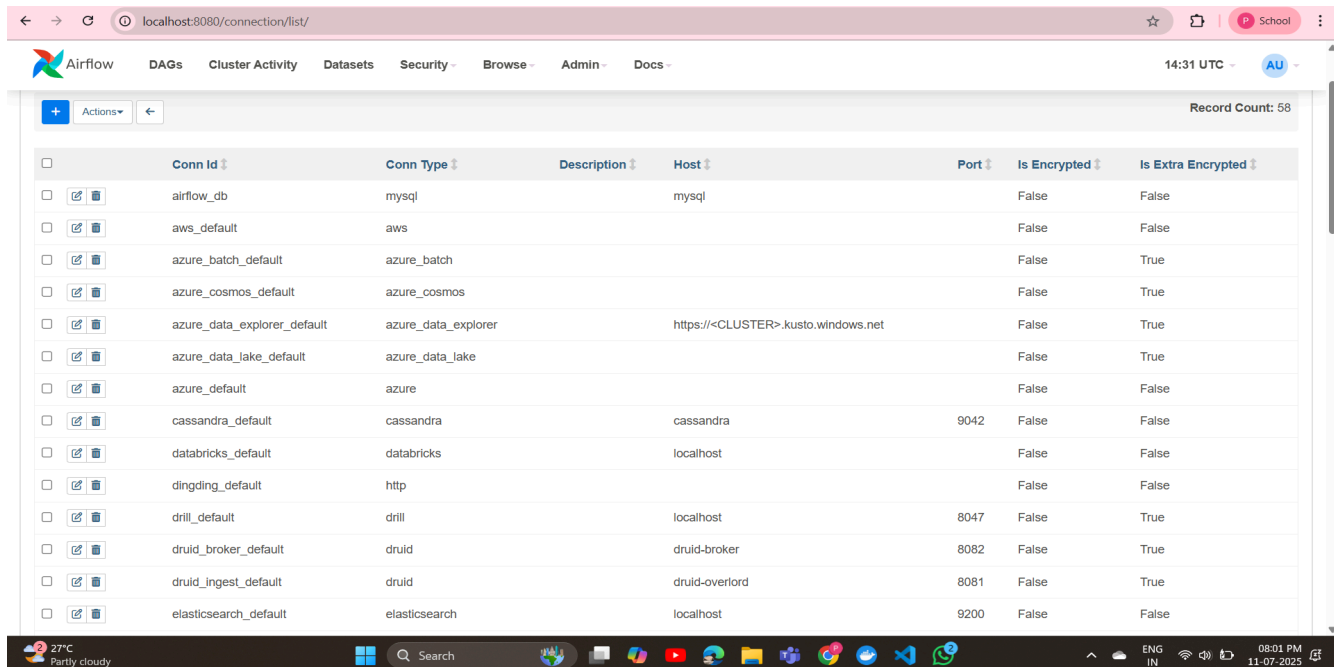
## 21) Airflow DAG's creation

The screenshot shows the Airflow web interface. At the top, there's a navigation bar with links: Airflow, DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs. The user is logged in as 'AU' and the time is 08:54 UTC. A yellow warning banner states: "The scheduler does not appear to be running. The DAGs list may not update, and new tasks will not be scheduled." Below this, the 'DAGs' section is visible. It includes filters for 'All' (1), 'Active' (0), and 'Paused' (0). There are also buttons for 'Running' (0) and 'Failed' (0), a 'Filter DAGs by tag' input, and a 'Search DAGs' input. An 'Auto-refresh' toggle is set to 'On'. The table below has columns: DAG, Owner, Runs, Schedule, Last Run, Next Run, Recent Tasks, Actions, and Links. The table is empty with the text 'No results'. At the bottom, it says 'Showing 0-0 of 0 DAGs'. The footer shows 'Version: v2.7.2'.

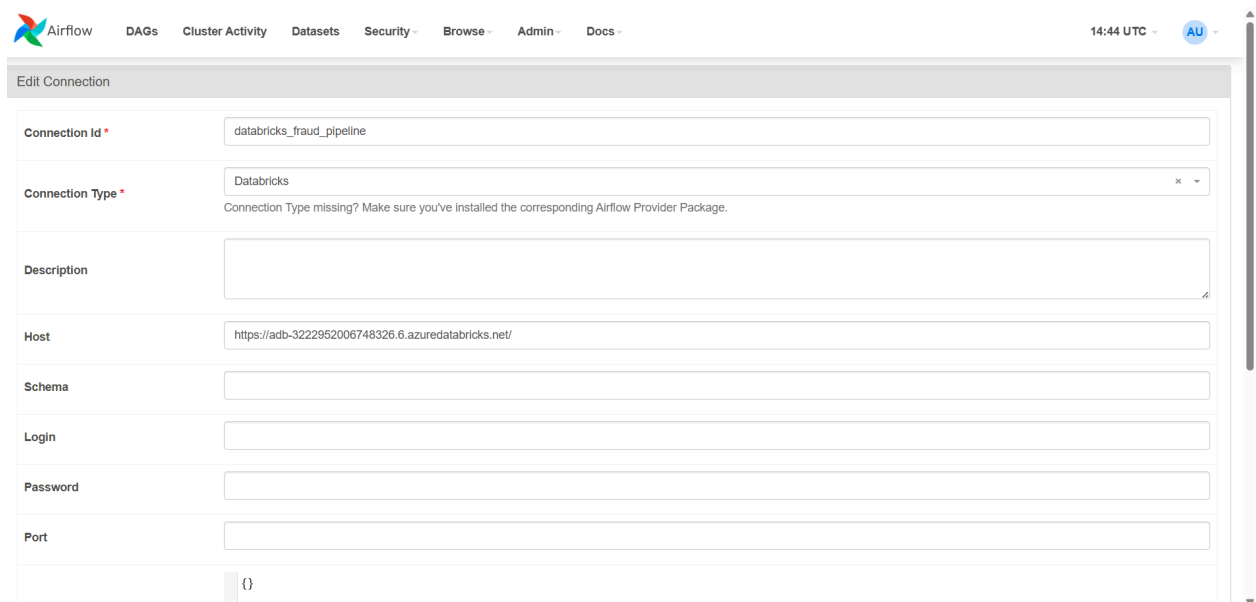
## 22) Dag created successfully

The screenshot shows the Airflow web interface with a successfully created DAG. The navigation bar is the same. The time is now 14:27 UTC. The 'DAGs' section shows filters for 'All' (1), 'Active' (0), and 'Paused' (1). The 'Running' (0) and 'Failed' (0) buttons are still present. The 'Auto-refresh' toggle is 'On'. The table now contains one entry: 'fraud\_etl\_pipeline' with owner 'airflow'. The 'Runs' column shows four empty circles, and the 'Schedule' column shows '@daily'. The 'Last Run' column shows '2025-07-10, 00:00:00'. The 'Next Run' column shows a play button icon. The 'Recent Tasks' column shows a series of empty circles. The 'Actions' column has a play button, a stop button, and a menu icon. At the bottom, it says 'Showing 1-1 of 1 DAGs'. The footer shows 'Version: v2.7.2' and 'Git Version: .release:c8b25cb3eea2bcd951ed7c1d7d0a1f9f04db206'.

## 23) Creating DAG schedule



	Conn Id	Conn Type	Description	Host	Port	Is Encrypted	Is Extra Encrypted
<input type="checkbox"/>	airflow_db	mysql		mysql		False	False
<input type="checkbox"/>	aws_default	aws				False	False
<input type="checkbox"/>	azure_batch_default	azure_batch				False	True
<input type="checkbox"/>	azure_cosmos_default	azure_cosmos				False	True
<input type="checkbox"/>	azure_data_explorer_default	azure_data_explorer		https://<CLUSTER>.kusto.windows.net		False	True
<input type="checkbox"/>	azure_data_lake_default	azure_data_lake				False	True
<input type="checkbox"/>	azure_default	azure				False	False
<input type="checkbox"/>	cassandra_default	cassandra		cassandra	9042	False	False
<input type="checkbox"/>	databricks_default	databricks		localhost		False	False
<input type="checkbox"/>	dingding_default	http				False	False
<input type="checkbox"/>	drill_default	drill		localhost	8047	False	True
<input type="checkbox"/>	druid_broker_default	druid		druid-broker	8082	False	True
<input type="checkbox"/>	druid_ingest_default	druid		druid-overlord	8081	False	True
<input type="checkbox"/>	elasticsearch_default	elasticsearch		localhost	9200	False	False



Connection Id \*

Connection Type \*

Description

Host

Schema

Login

Password

Port

{}

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

14:44 UTC

AU

Added Row

List Connection



















Search

+

Actions

←

Record Count: 59

<input type="checkbox"/>	Conn Id	Conn Type	Description	Host	Port	Is Encrypted	Is Extra Encrypted
<input type="checkbox"/>  	airflow_db	mysql		mysql		False	False
<input type="checkbox"/>  	aws_default	aws				False	False
<input type="checkbox"/>  	azure_batch_default	azure_batch				False	True
<input type="checkbox"/>  	azure_cosmos_default	azure_cosmos				False	True
<input type="checkbox"/>  	azure_data_explorer_default	azure_data_explorer		https://<CLUSTER>.kusto.windows.net		False	True
<input type="checkbox"/>  	azure_data_lake_default	azure_data_lake				False	True
<input type="checkbox"/>  	azure_default	azure				False	False
<input type="checkbox"/>  	cassandra_default	cassandra		cassandra	9042	False	False
<input type="checkbox"/>  	databricks_default	databricks		localhost		False	False

← → ↺


localhost:8080/home

🏠

🔖

📄

👤 School



Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

14:28 UTC

👤

AU

DAGs

All 1

Active 0

Paused 1

Running 0

Failed 0

Filter DAGs by tag

Search DAGs

Auto-refresh

↺

DAG ↕	Owner ↕	Runs ↕	Schedule	Last Run ↕	Next Run ↕	Recent Tasks ↕	Actions	Links
<div>🔵</div> <div>fraud_ott_pipeline</div> <div>example</div>	airflow	<div>🟢 1</div> <div>⬜</div> <div>⬜</div>	@daily	2025-07-10, 00:00:00	2025-07-10, 00:00:00	<div>⬜</div> <div>⬜</div> <div>⬜</div> <div>🟢 2</div> <div>⬜</div> <div>⬜</div> <div>⬜</div> <div>⬜</div> <div>⬜</div> <div>⬜</div> <div>⬜</div> <div>⬜</div>	<div>▶</div> <div>🗑️</div> <div>⋮</div>	

«

<

1

>

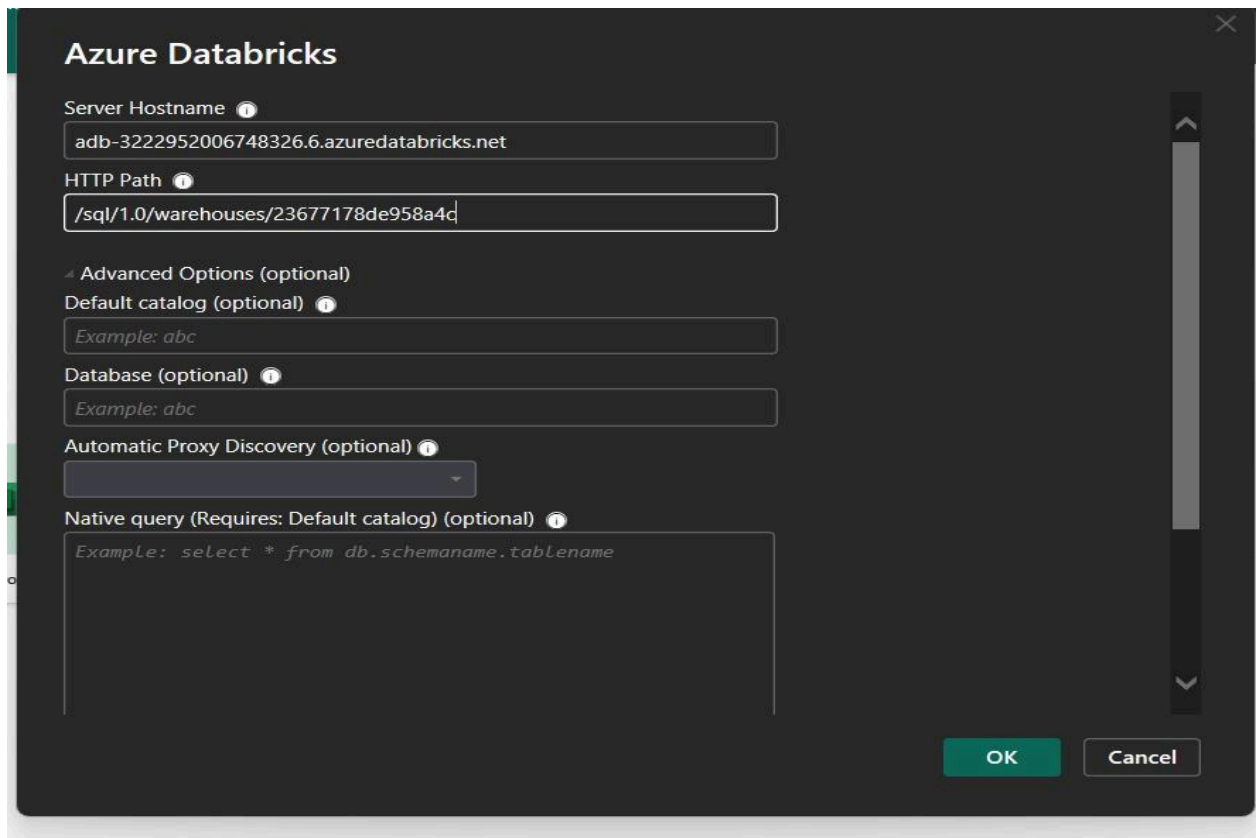
»

Showing 1-1 of 1 DAGs

Version: v2.7.2

Git Version: .release:c8b25cb3eea2bcd951ed7c1d7d0a1f9f04db206

## 25) Power BI and Azure databricks connection



**Azure Databricks**

Server Hostname ⓘ  
adb-3222952006748326.6.azuredatabricks.net

HTTP Path ⓘ  
/sql/1.0/warehouses/23677178de958a4c

Advanced Options (optional)

Default catalog (optional) ⓘ  
Example: abc

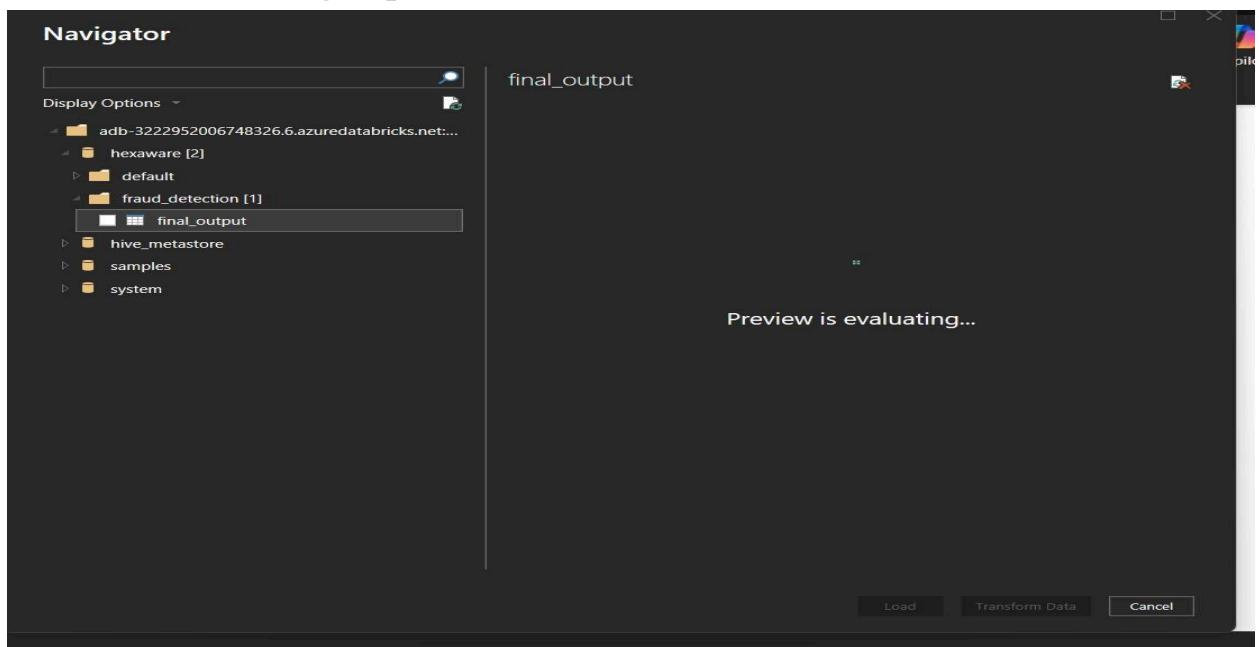
Database (optional) ⓘ  
Example: abc

Automatic Proxy Discovery (optional) ⓘ  
▼

Native query (Requires: Default catalog) (optional) ⓘ  
Example: `select * from db.schemaname.tablename`

OK Cancel

## 26) Connection string in power BI



**Navigator**

Display Options ▼

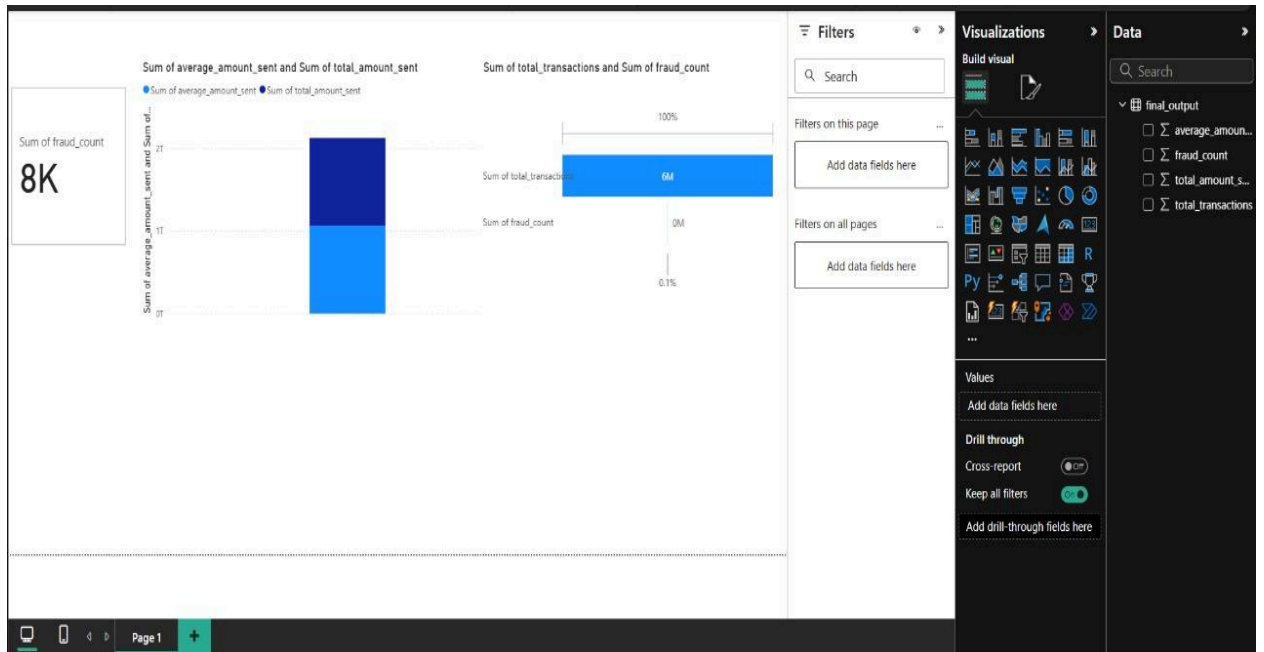
- adb-3222952006748326.6.azuredatabricks.net...
- hexaware [2]
  - default
  - fraud\_detection [1]
    - final\_output**
- hive\_metastore
- samples
- system

final\_output

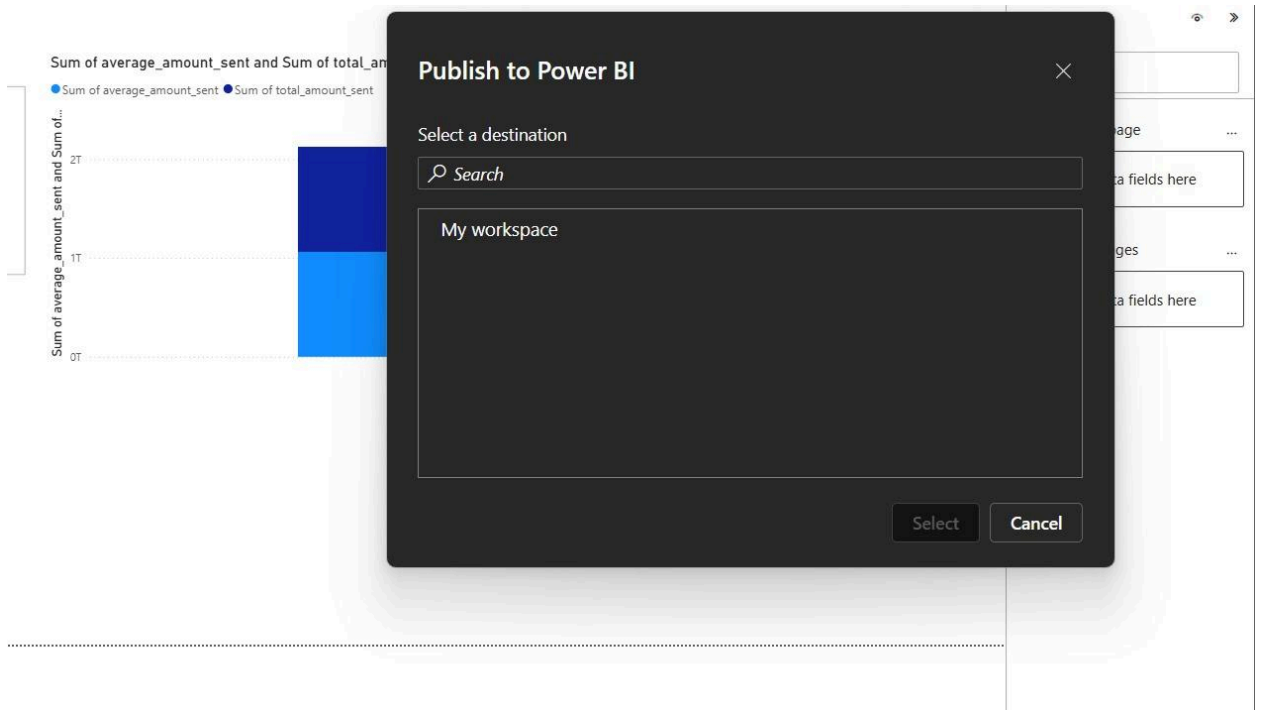
Preview is evaluating...

Load Transform Data Cancel

## 27) Power Bi visualization creation



## 28) Publishing to our workspace



## 29) Creating alerts

The screenshot shows the 'Alerts' configuration window in Power BI. At the top, there's a bell icon and the title 'Alerts'. Below this, it says '1 alert on this report' and a green '+ Add alert' button. The main configuration area has two sections: 'Changes' and 'Becomes'. The 'Becomes' section is selected with a green radio button. Under 'Becomes', there's a 'Condition' dropdown set to 'Greater than' and a 'Value' input field set to '0'. Below this is a 'Send notification' section with a 'Via' dropdown set to 'Teams' and a 'Send to' field containing the email 'azuser3607\_mml.local'. An 'Apply' button is at the bottom right of the configuration area. At the very bottom, there's a status bar with a red dot icon and the text 'My Power BI Activator Alerts'.

Alerts

1 alert on this report

+ Add alert

☐ Changes

☒ Becomes

Condition: Greater than

Value: 0

Send notification

Via: Teams

Send to: azuser3607\_mml.local

Apply

My Power BI Activator Alerts



30) Power BI workspace

My workspace

+ New item

New folder

→ Import

Migrate

Filter by keyword

Filter

Workspace settings

Choose from predefined task flows or add a task to build one

Select from one of Microsoft's predefined task flows or add a task to start building one yourself.

Select a predefined task flow

Add a task

→ Import a task flow

	Name	Type	Task	Owner	Refreshed	Next refresh	Endorsement	Sensitivity
	My Power BI Activator Alerts	Activator	—	azuser3607_m...	—	—	—	—
	PowerBI	Report	—	azuser3607_m...	7/16/2025, 9:18:03 ...	—	—	—
	PowerBI	Semantic model	—	azuser3607_m...	7/16/2025, 9:18:0...	7/16/2025, 10:18:...	—	—