Azure Data Engineering Pipeline with Medallion Architecture

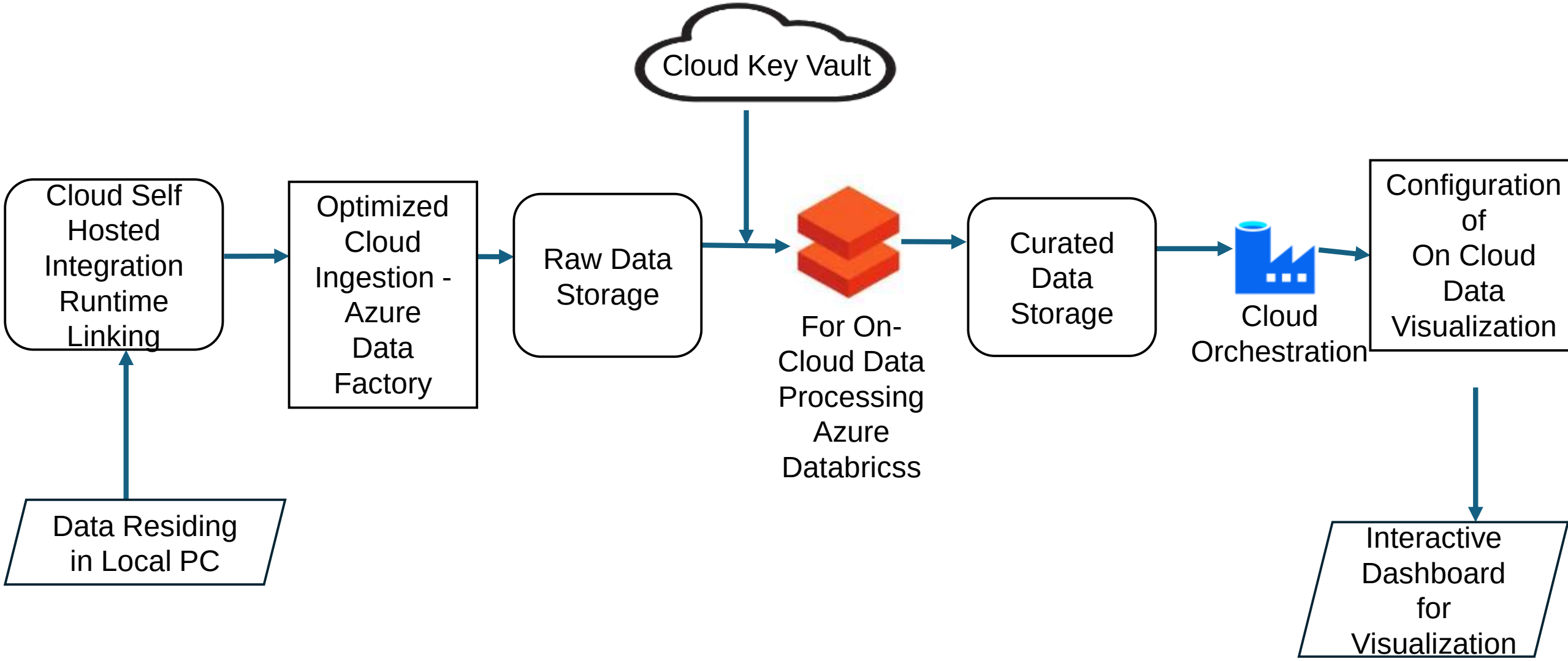
Problem statement

- Enterprises struggle to ingest and integrate large volumes of on-premises and external data into the cloud in a secure and efficient manner.
- Designing scalable cloud pipelines that can dynamically handle both full and incremental data loads remains a challenge.
- Lack of a structured data architecture in the cloud leads to inconsistent, unreliable, and hard-to-govern data.
- Businesses increasingly demand end-to-end cloud solutions that provide secure ingestion, transformation, storage, and analytics for decision-making.

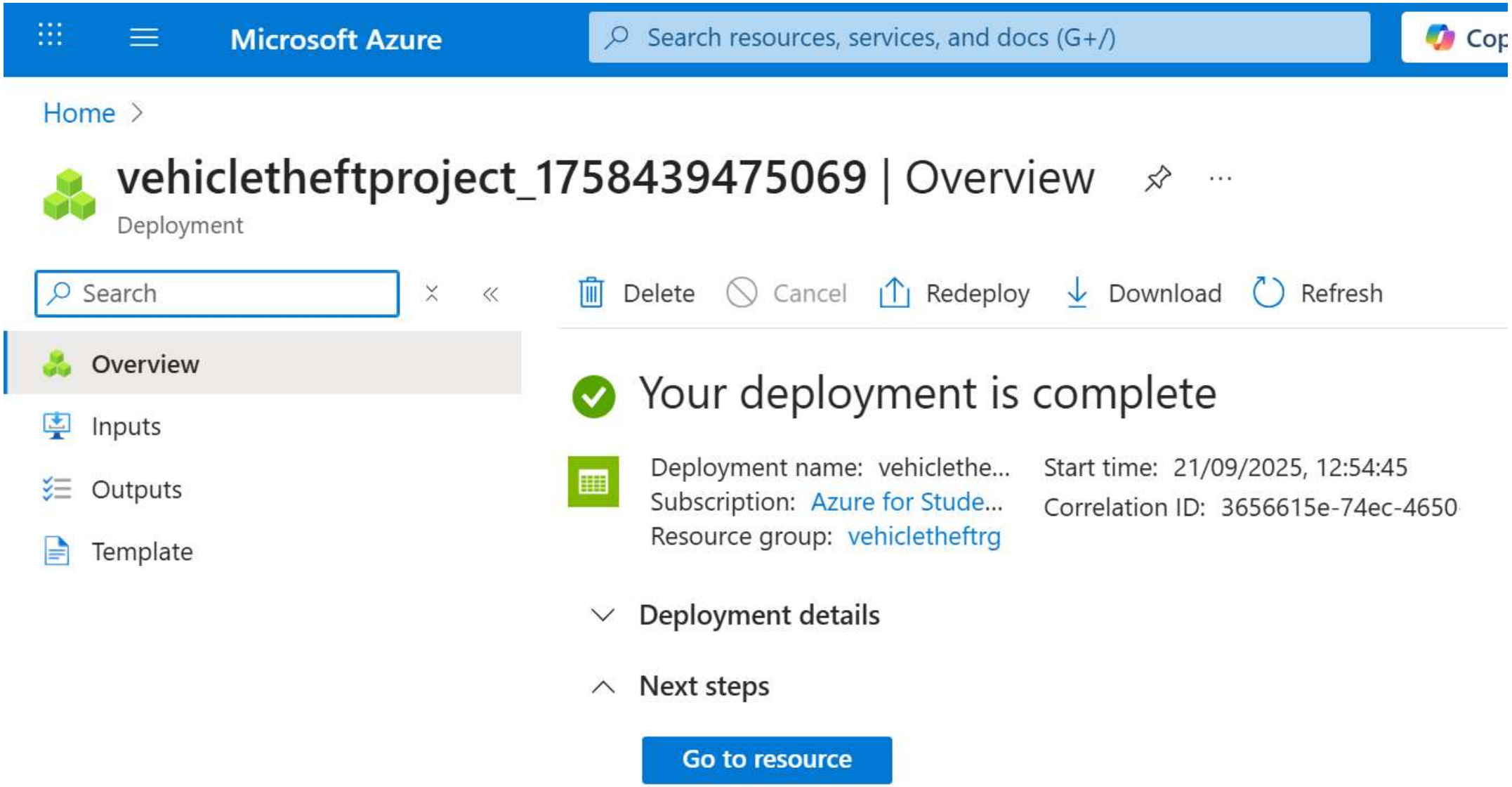
Objectives

- Implement secure and scalable cloud-based data ingestion using Azure Data Factory and Data Lake Gen2.
- Apply metadata-driven ingestion patterns to optimize full and incremental data loads in the cloud.
- Establish a structured Medallion Architecture on Azure for data reliability, quality, and governance.
- Deliver cloud-hosted business intelligence by integrating curated data with Power BI for real-time insights.

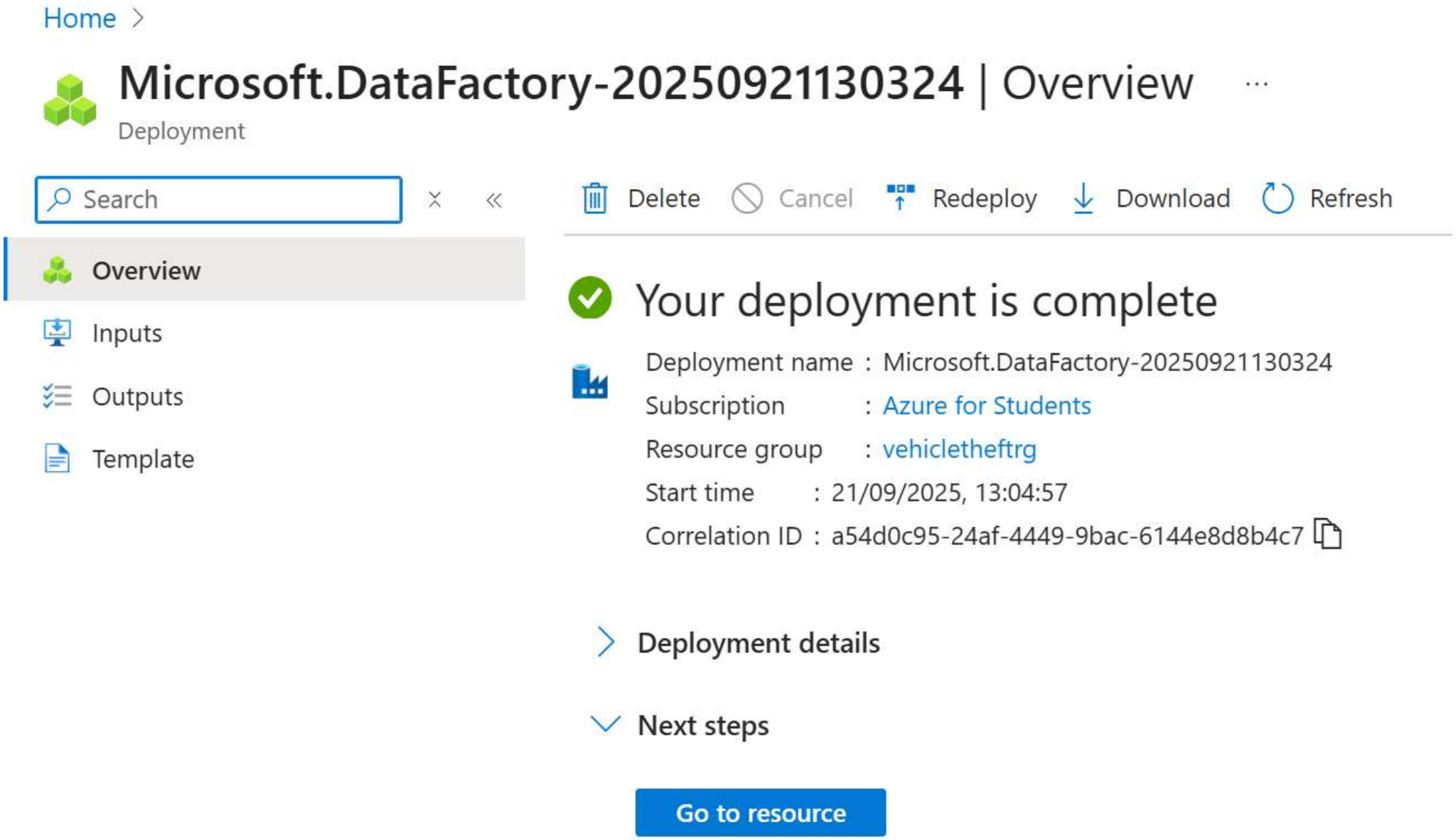
Architecture



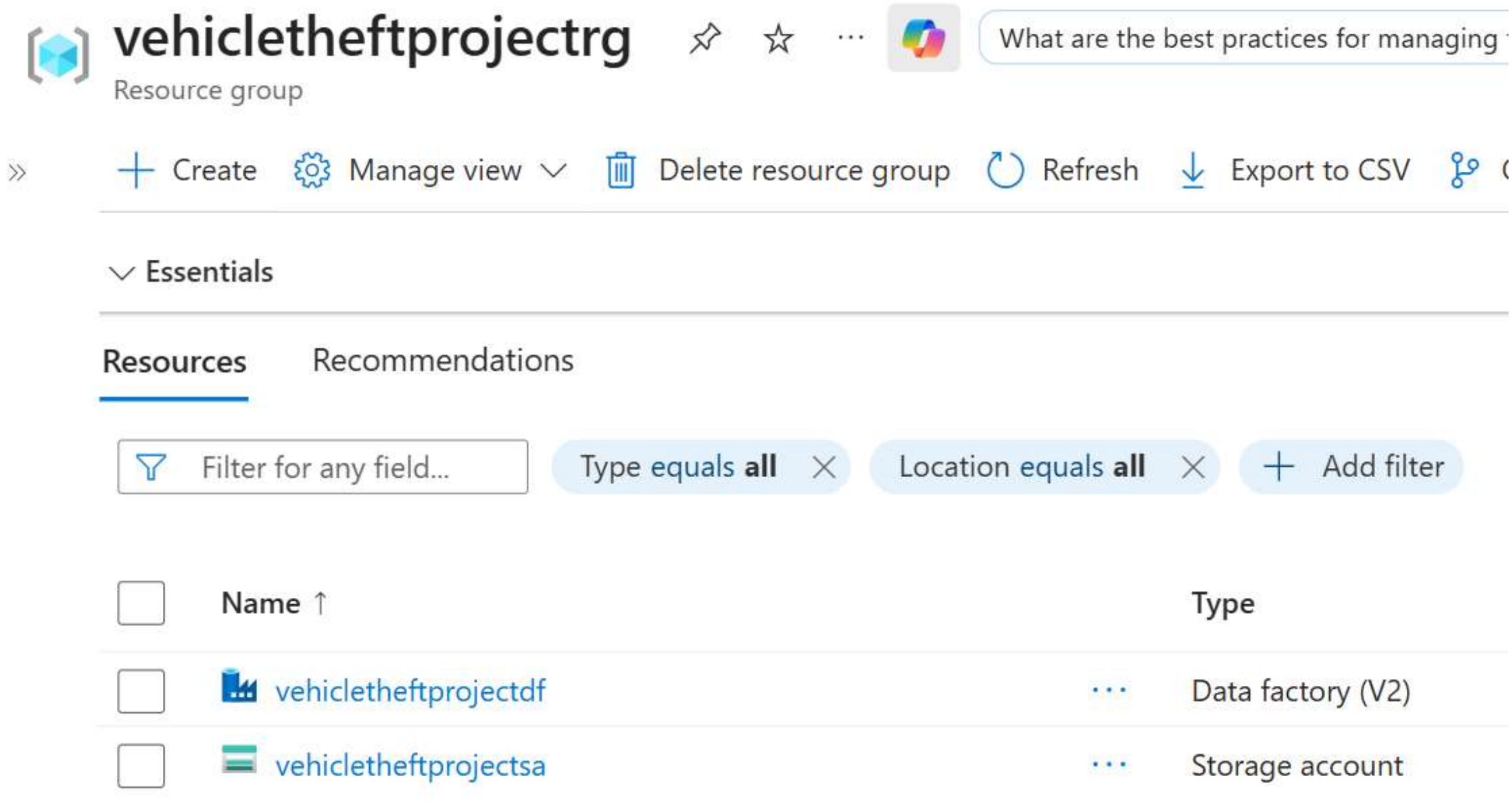
Deployment of Resource Group



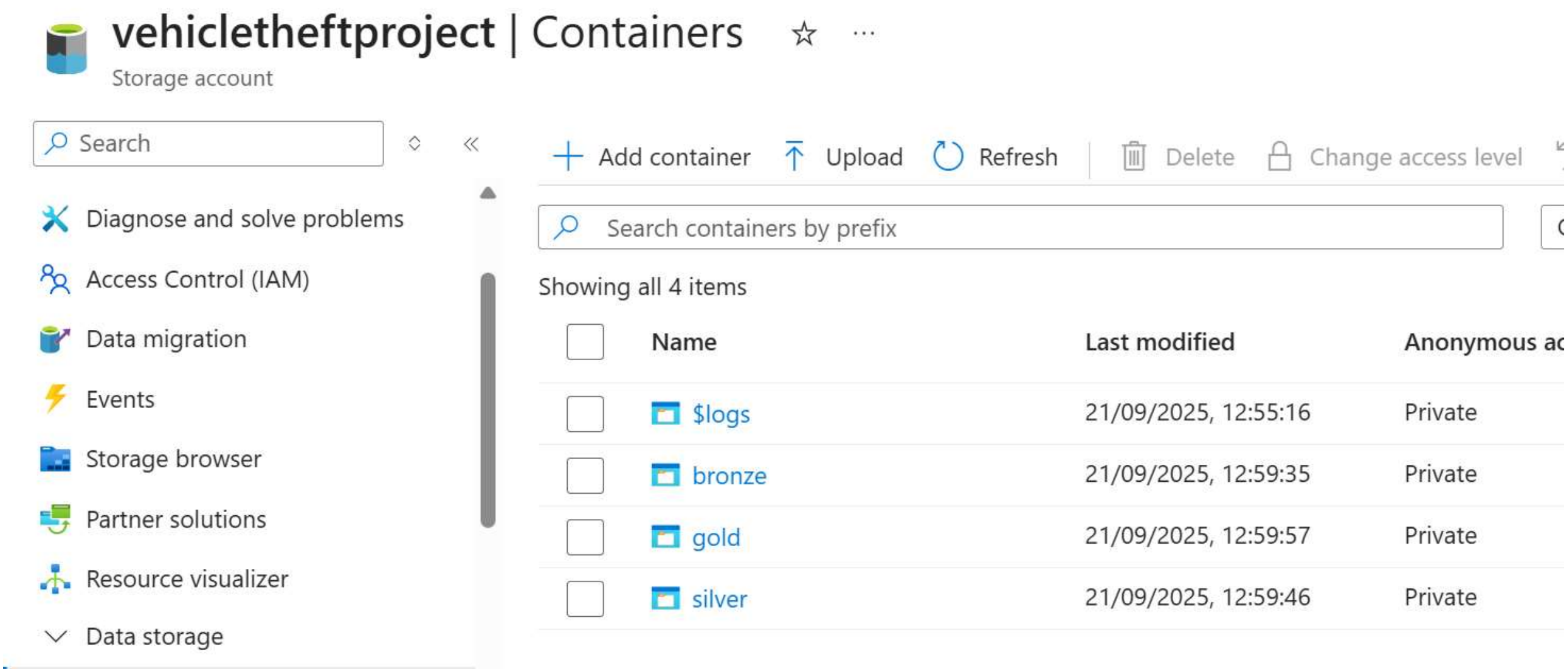
Deployment of Azure Data Factory



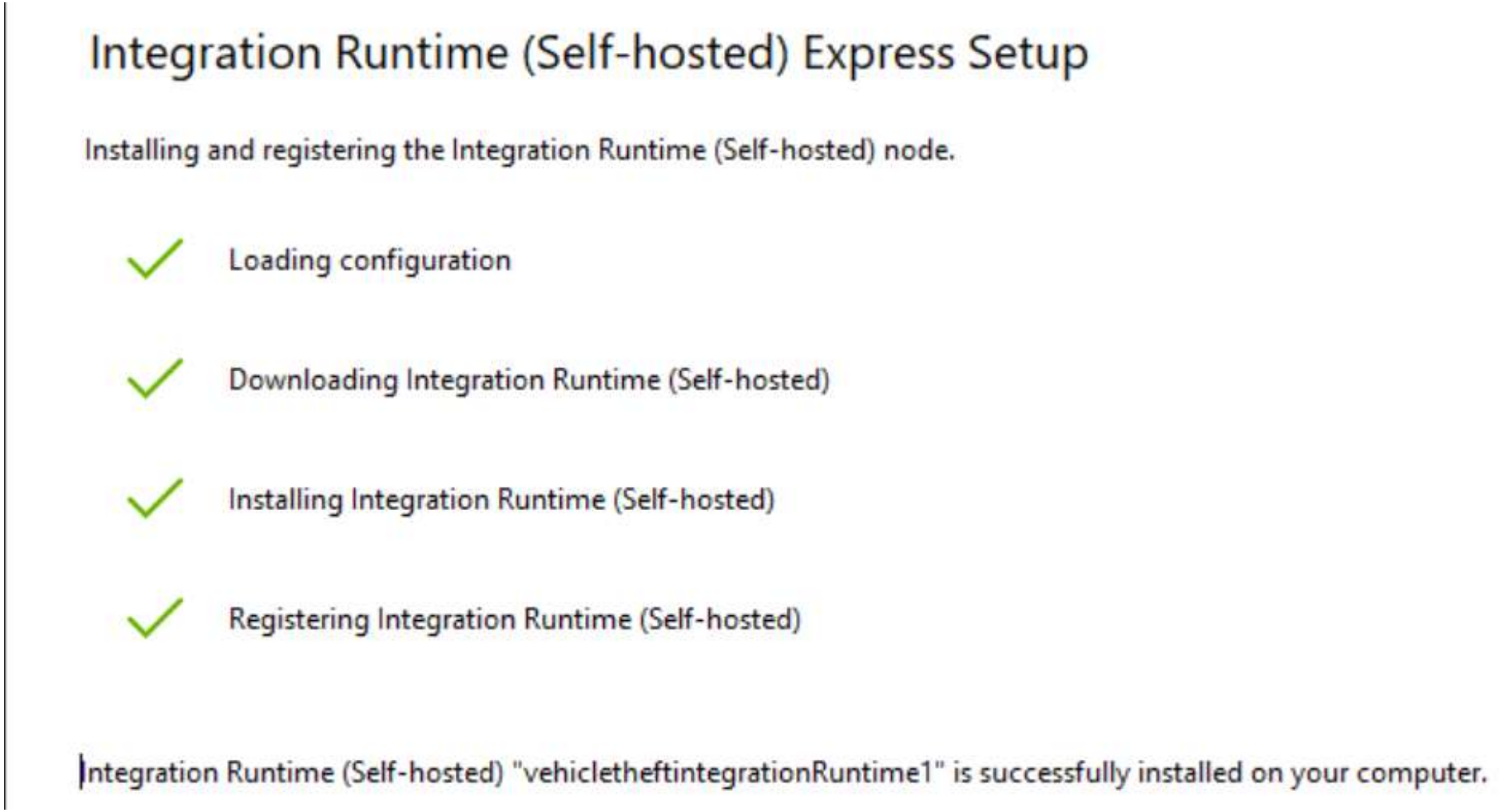
Linking of Storage Account and ADF



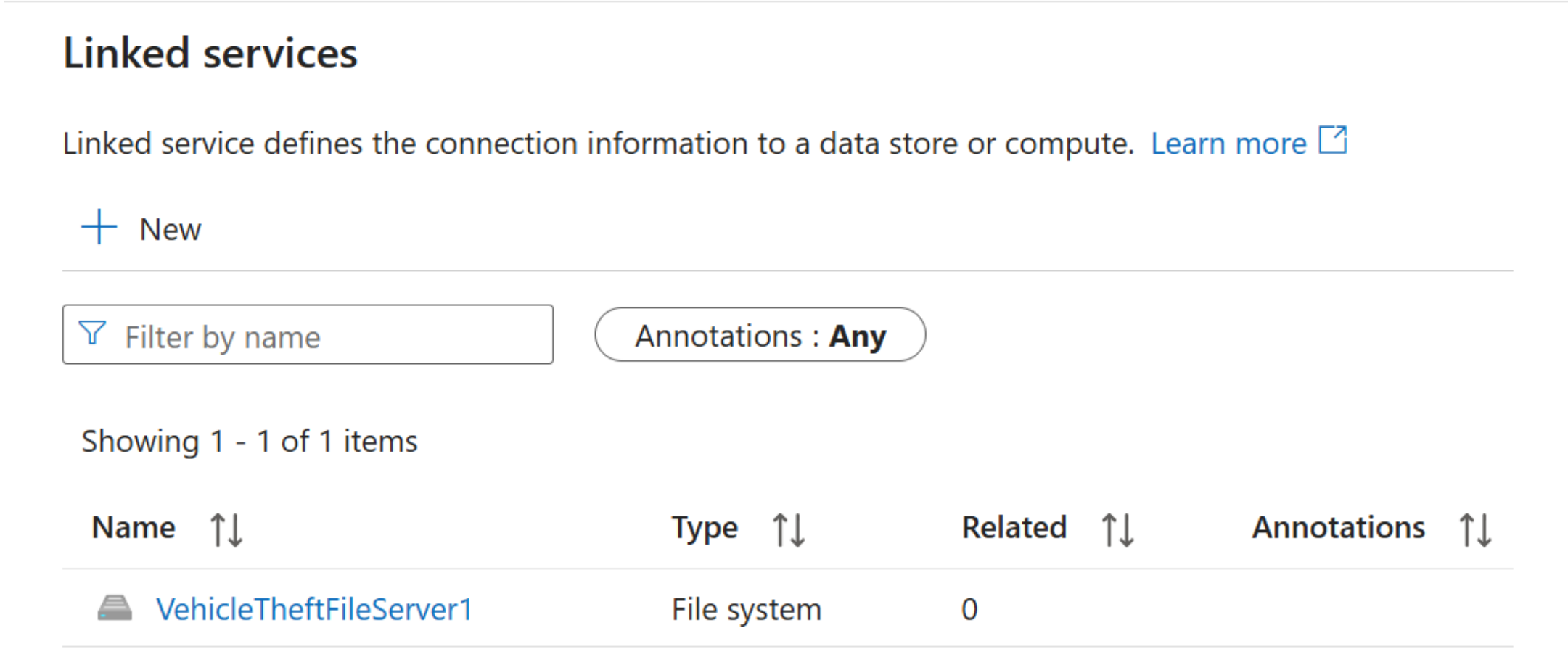
Initialization of Containers for Medallion



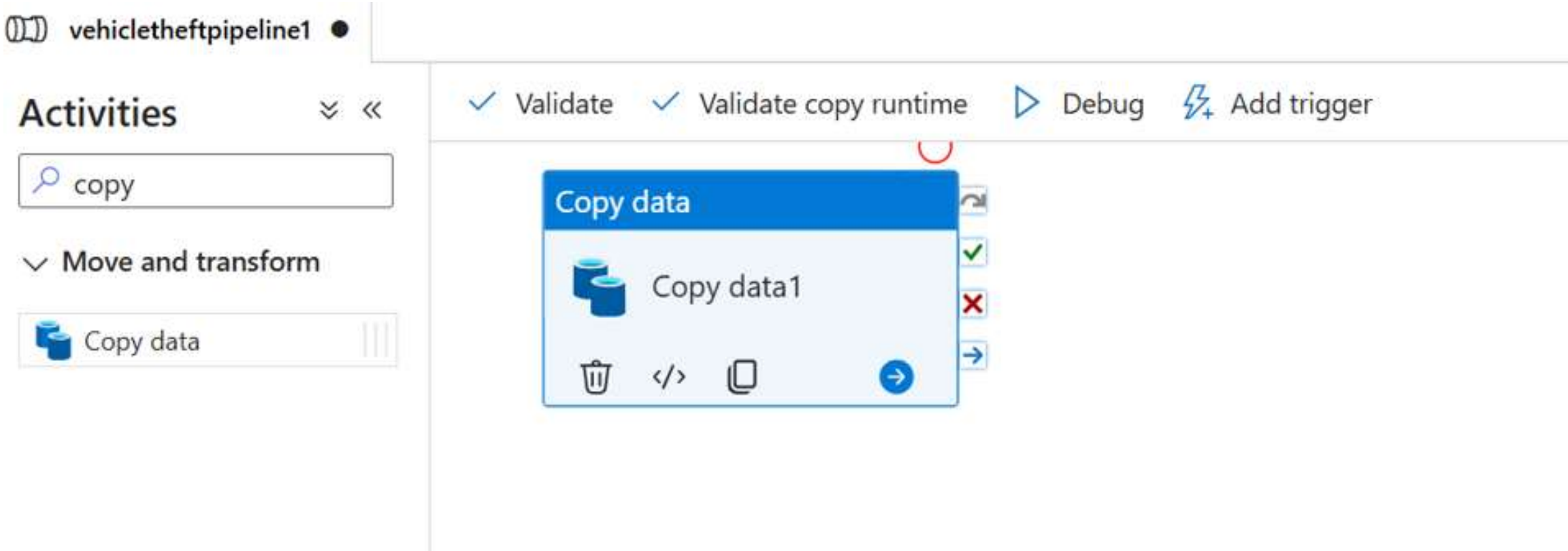
Configuration of Self Hosted IR



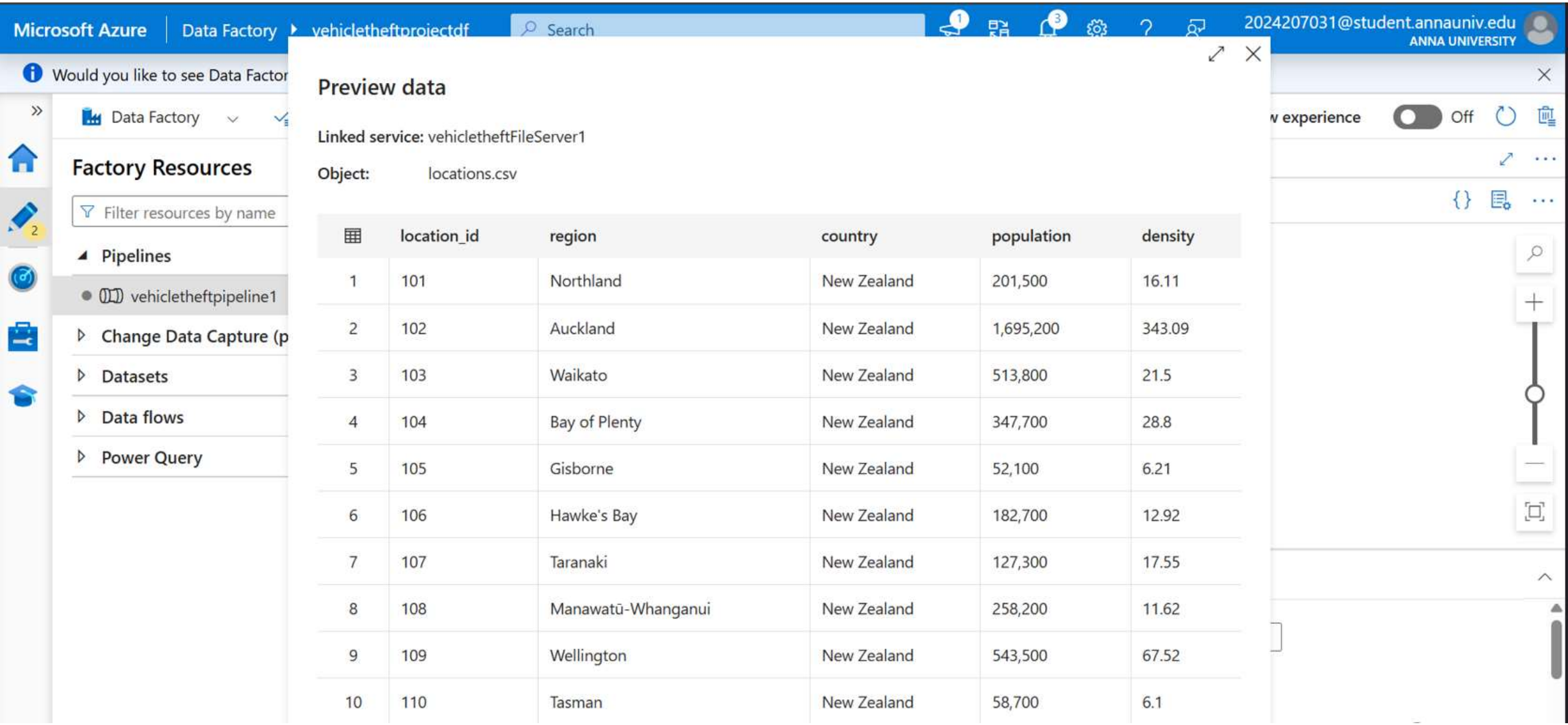
Created Linked Service for File System

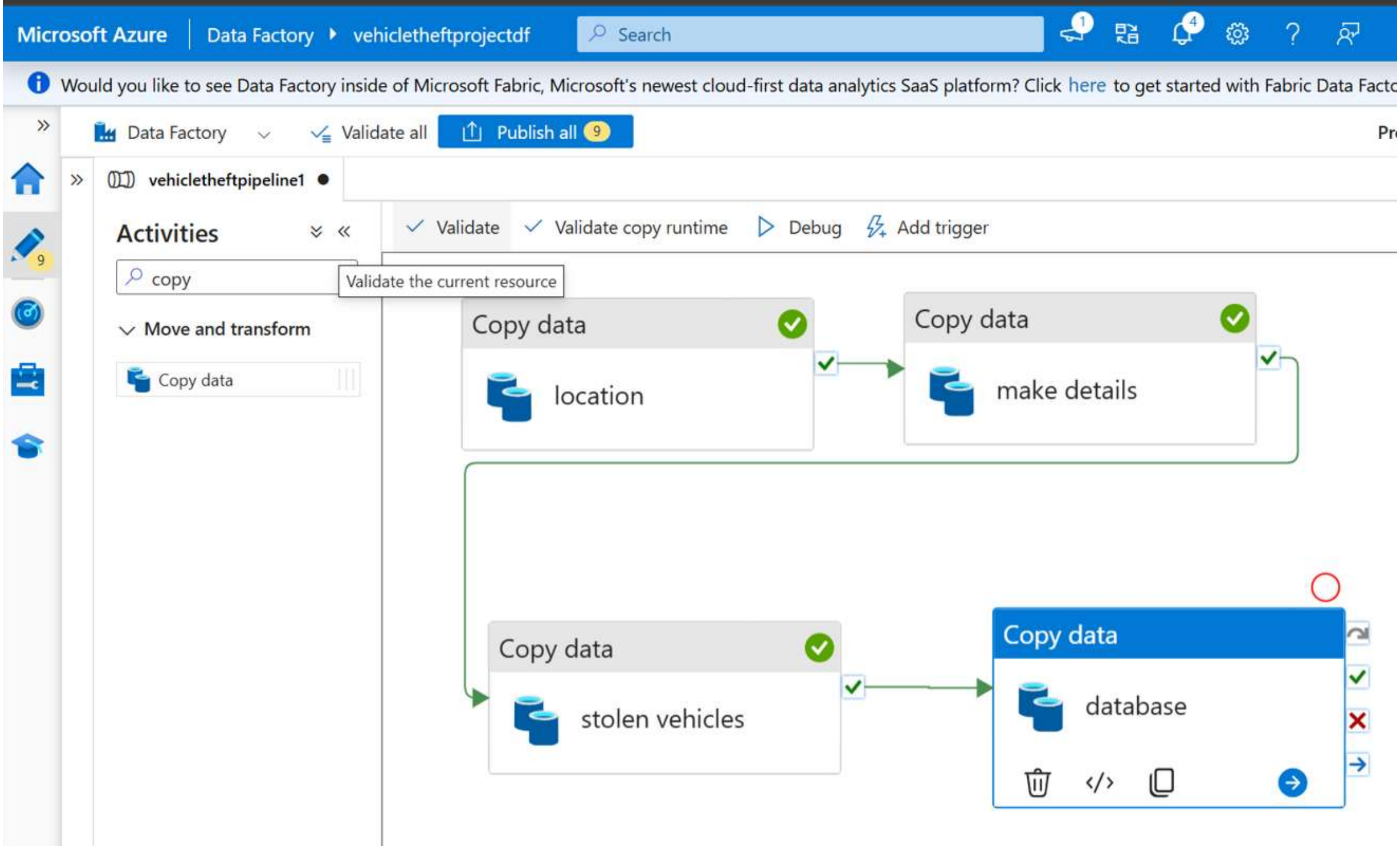


Initialization of Data Copy in Pipeline



Successful Data Ingestion




Microsoft Azure
Data Factory
vehicletheftprojectdf
Search
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started with Fabric Data Facto
Data Factory
Validate all
Publish all 9
vehicletheftpipeline1
Activities
copy
Validate the current resource
Move and transform
Copy data
Validate
Validate copy runtime
Debug
Add trigger
Copy data
location
Copy data
make details
Copy data
stolen vehicles
Copy data
database

bronze

Container

Add Directory | Upload | Refresh | Delete | Copy | Paste | Rename | Acquire lease

bronze

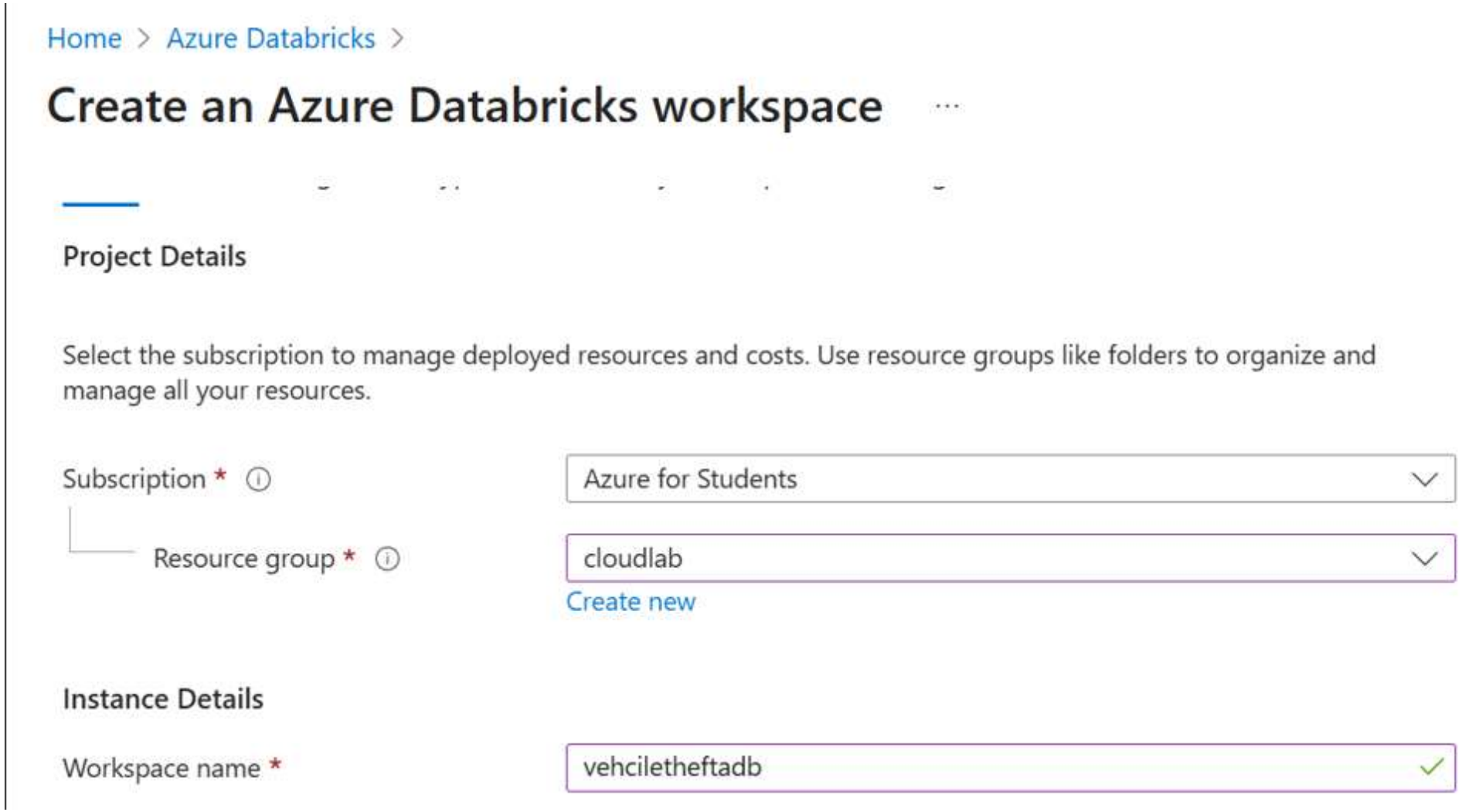
Authentication method: Access key (Switch to Microsoft Entra user account)

Search blobs by prefix (case-sensitive)

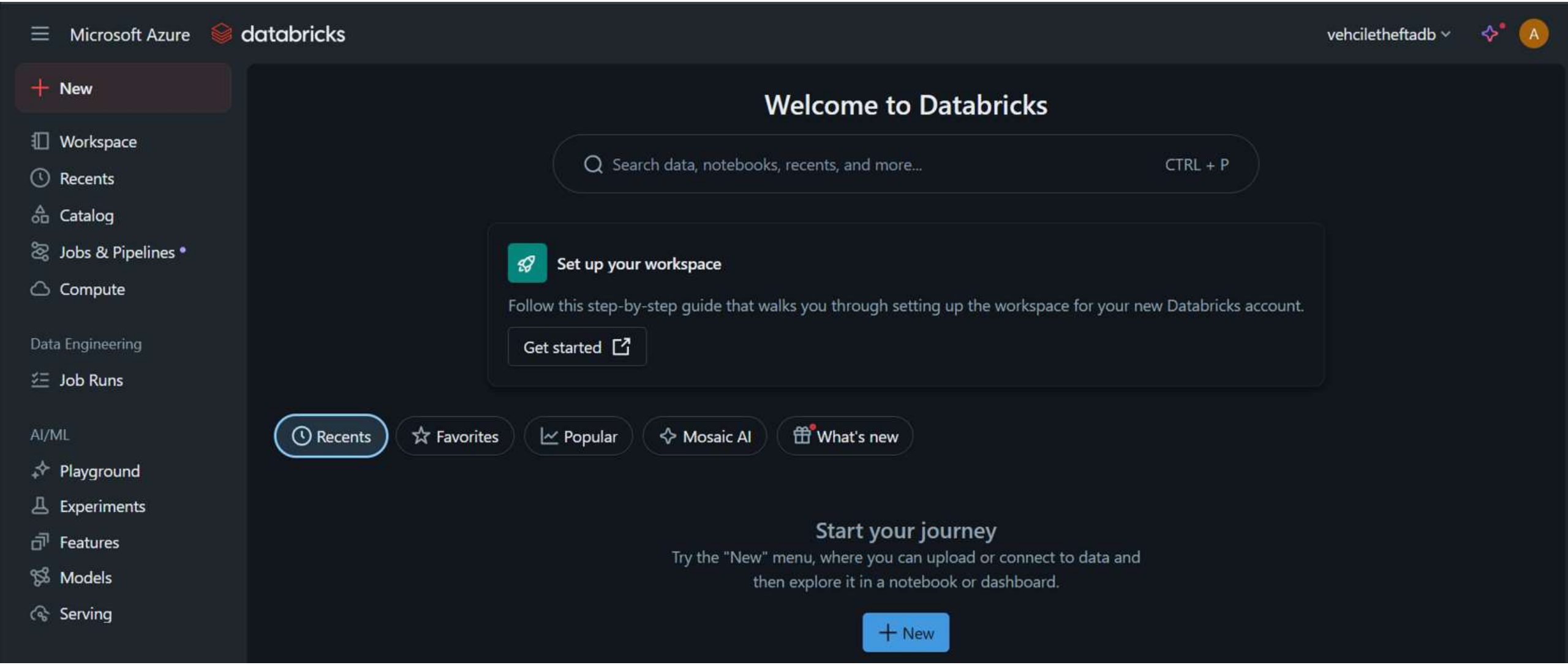
Showing all 5 items

Name	Last modified	Access tier
BankTransactionDataset_1K (1).csv	26/09/2025, 09:03:23	Hot (Inferred)
locations.csv	26/09/2025, 09:03:23	Hot (Inferred)
make_details.csv	21/09/2025, 20:54:54	Hot (Inferred)
stolen_vehicles.csv	21/09/2025, 20:55:19	Hot (Inferred)
stolen_vehicles_db_data_dictionary.csv	21/09/2025, 20:55:40	Hot (Inferred)

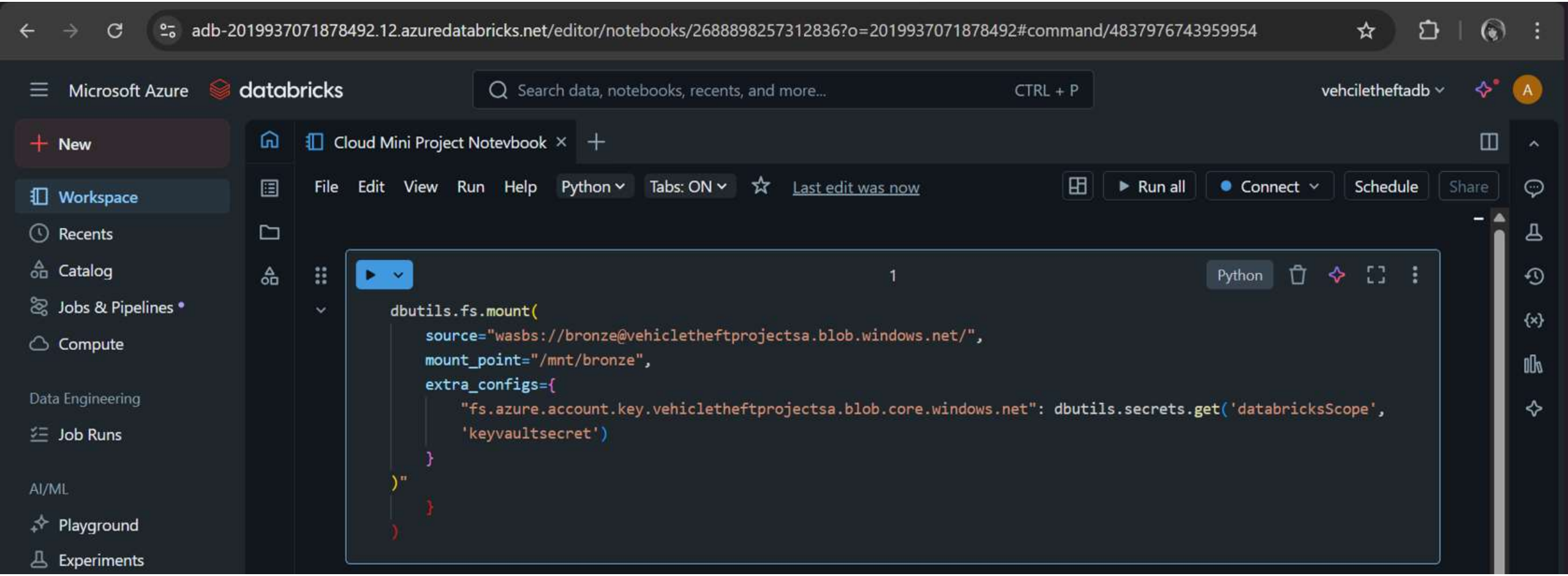
Creation of Azure Databricks WS



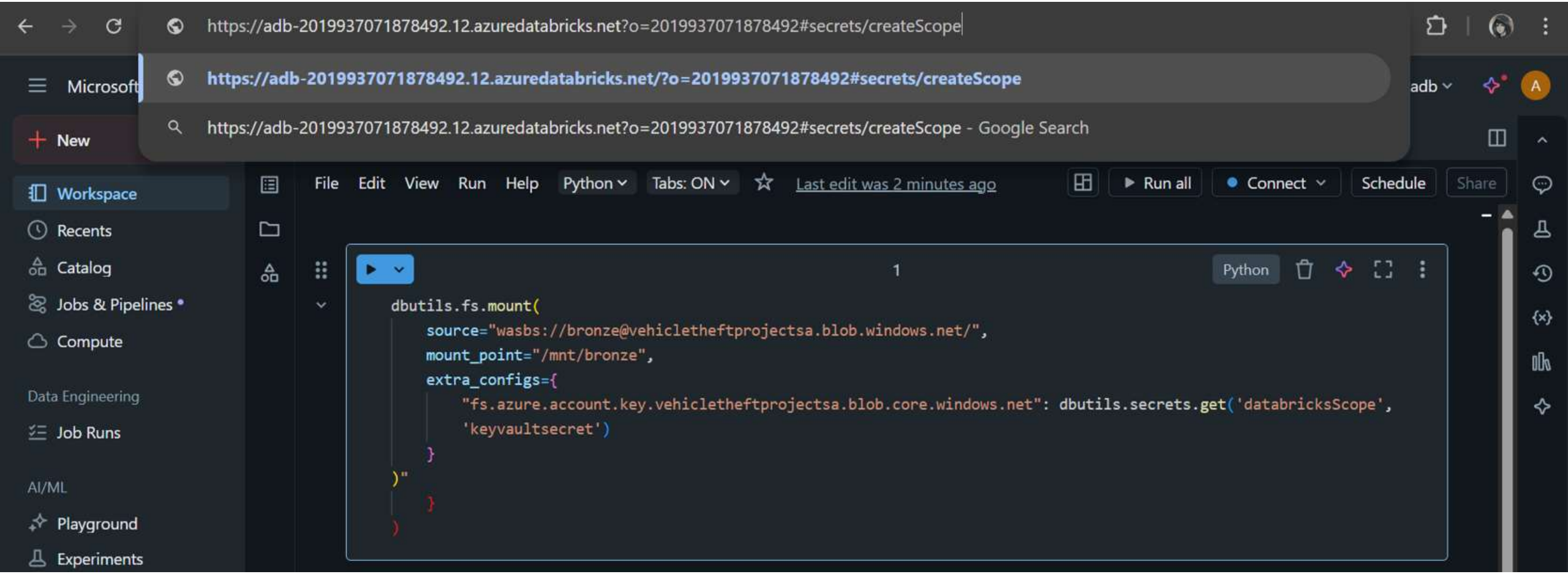
Deployed Azure Data Bricks



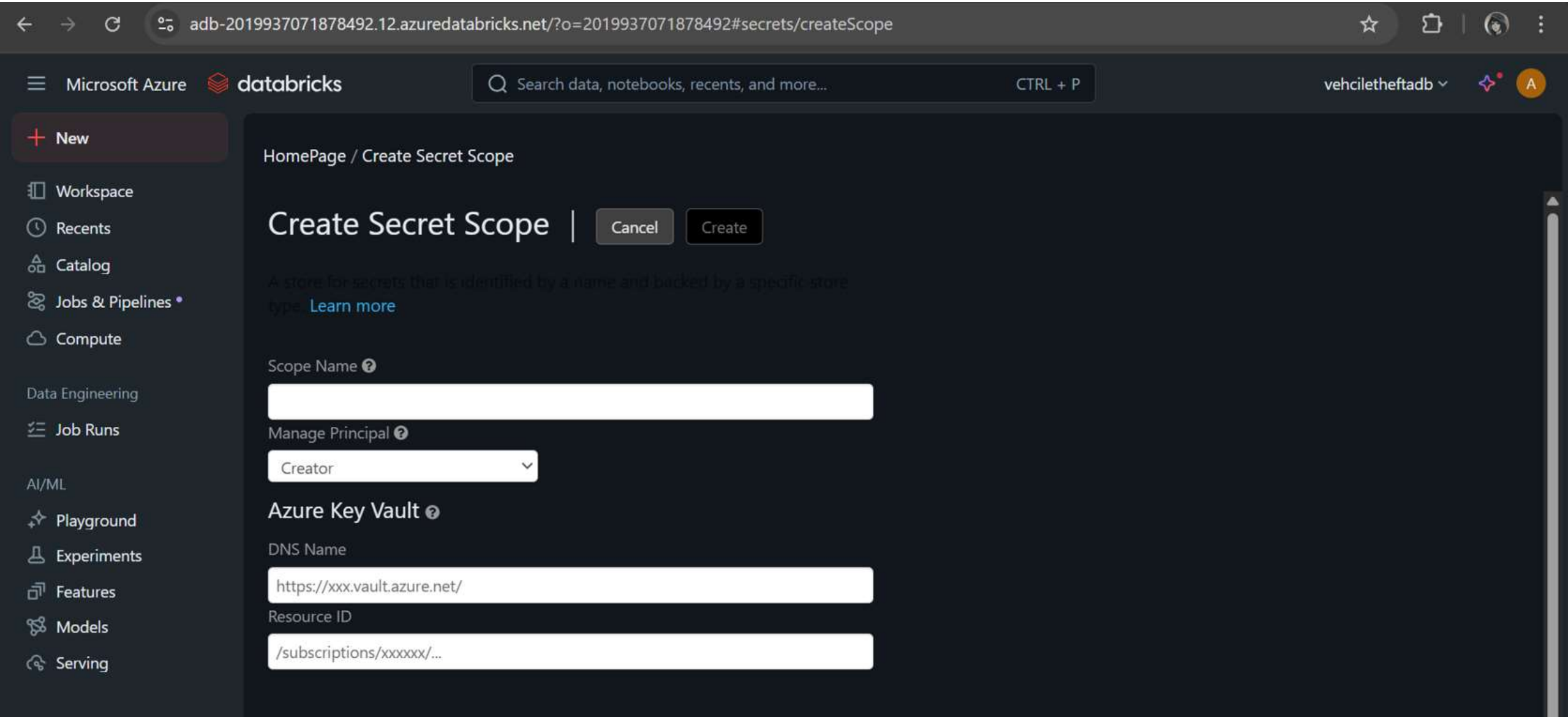
Mounting of Bronze Container into Azure DataBricks



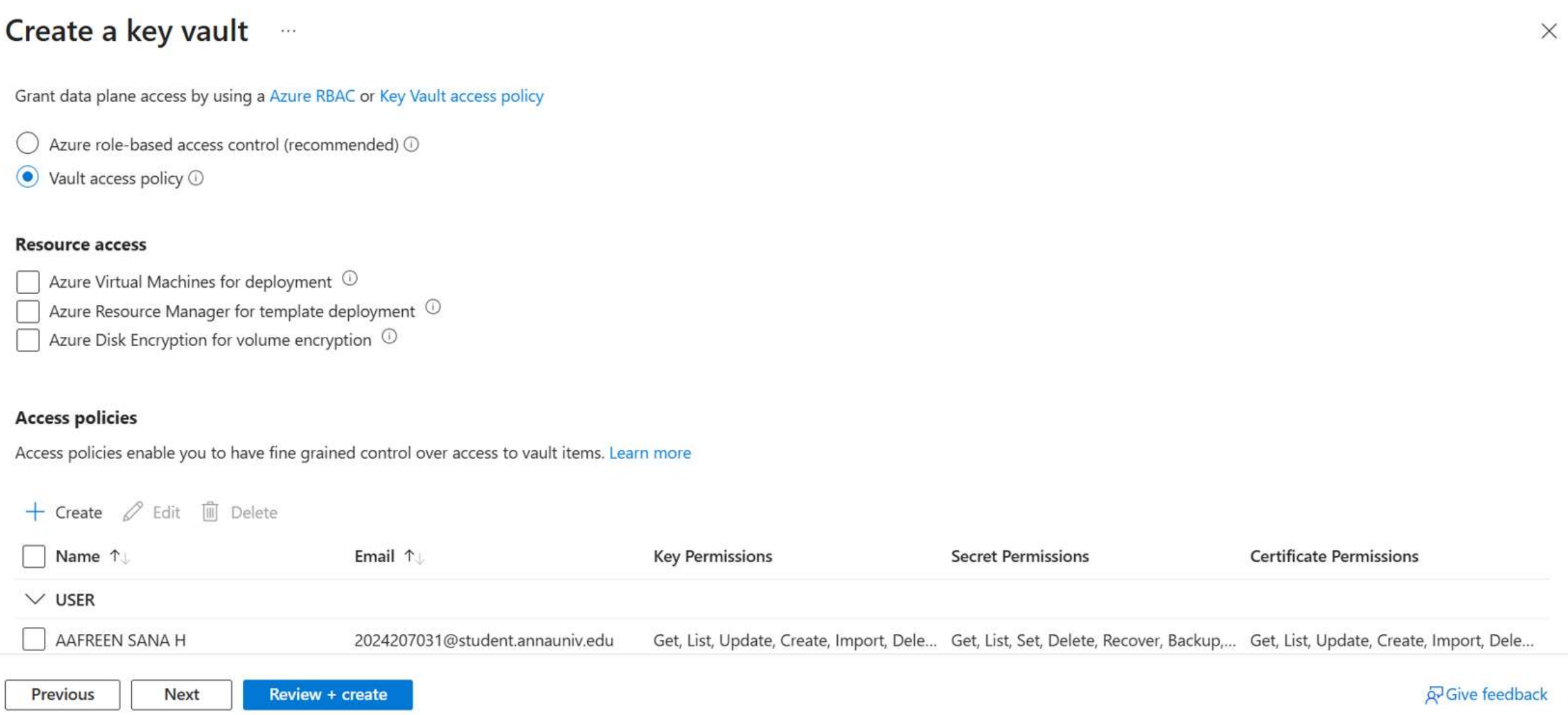
Creation of DatabricksScope



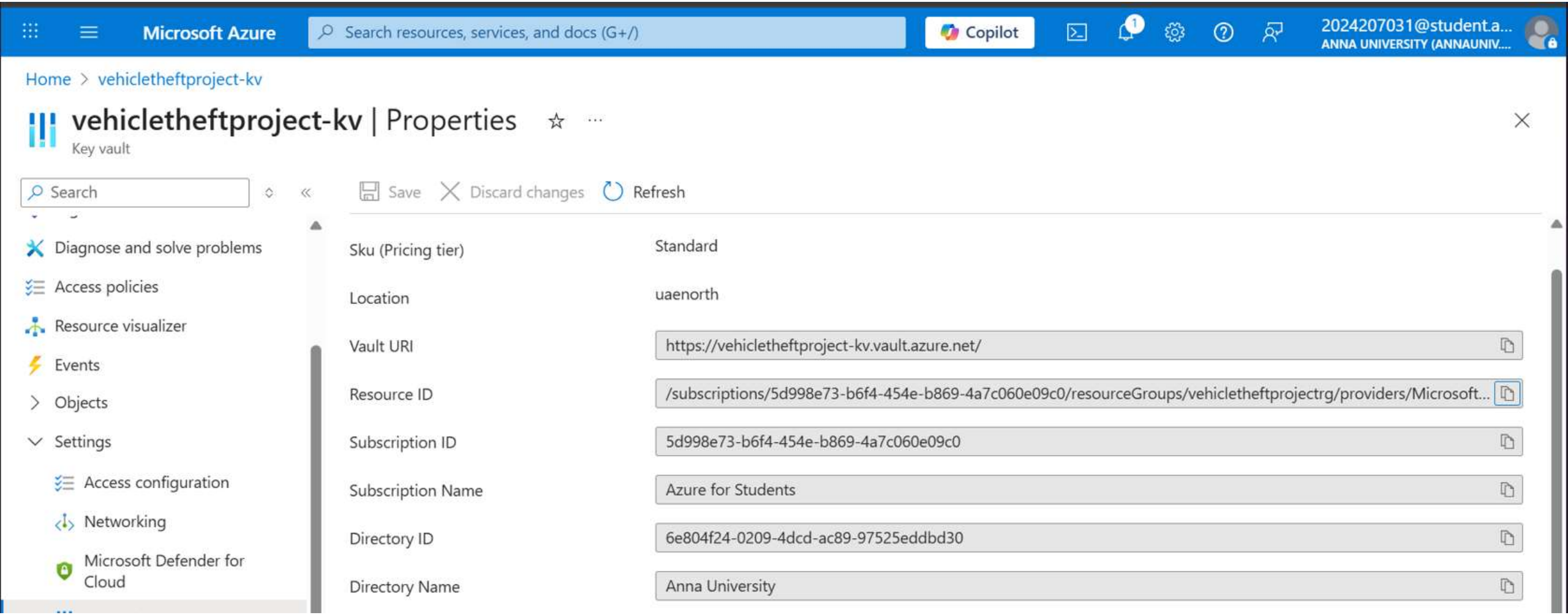
Creation of DatabricksScope



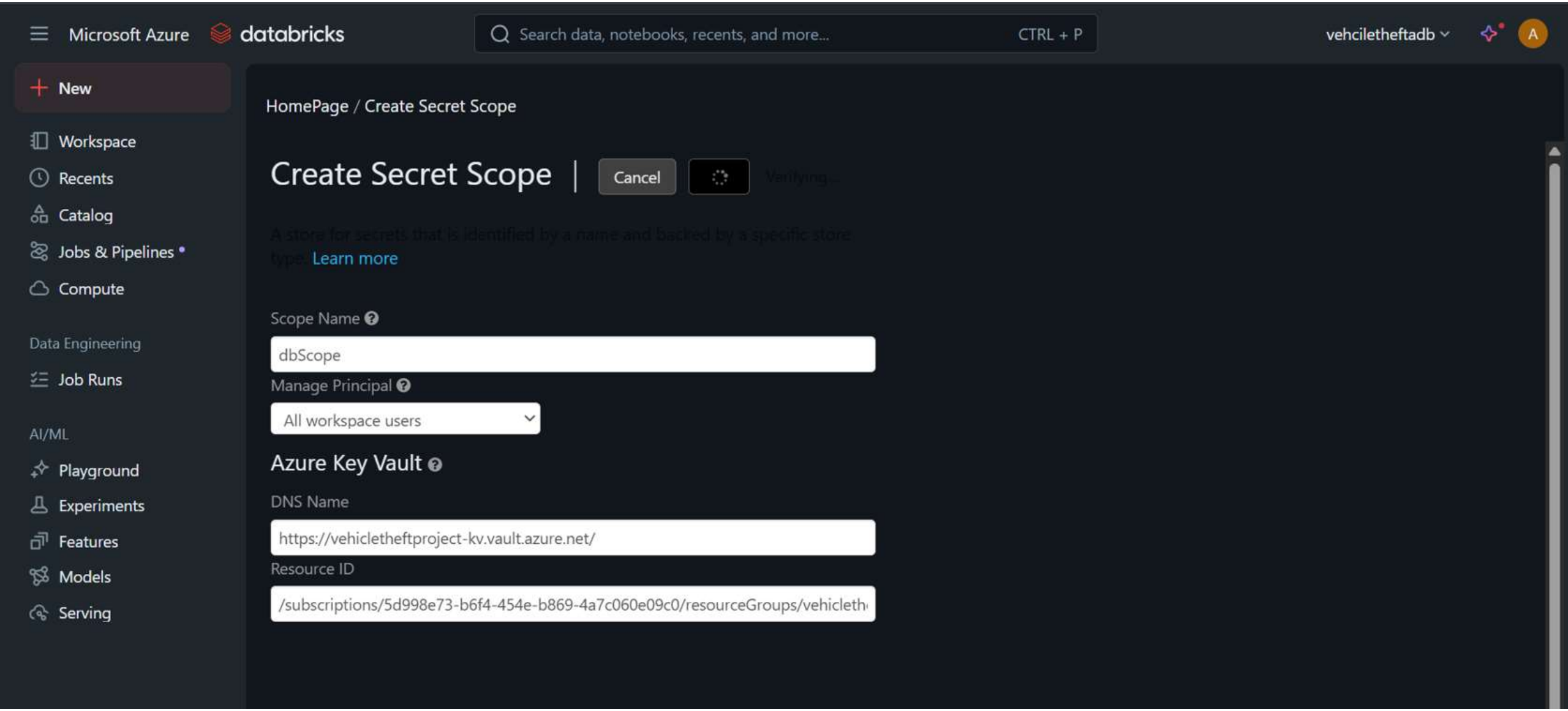
Creation of Azure Key Vault



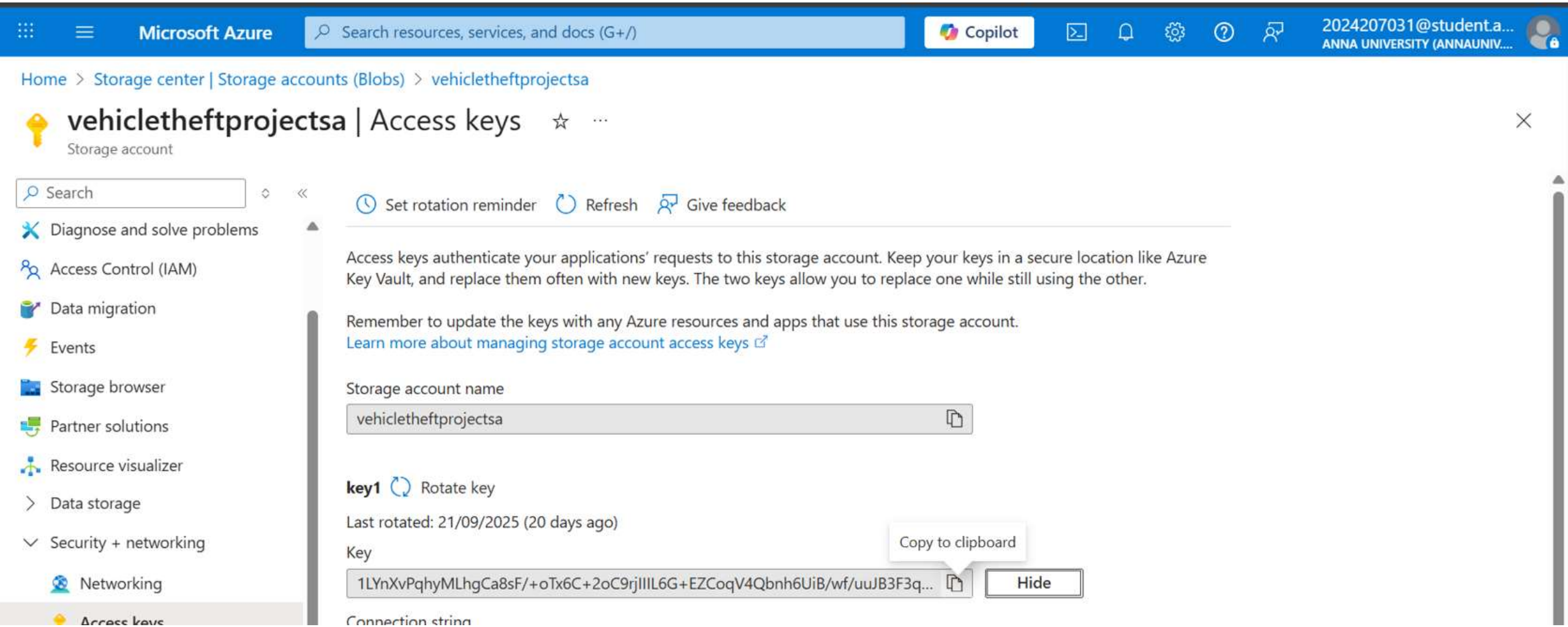
Azure Key Vault



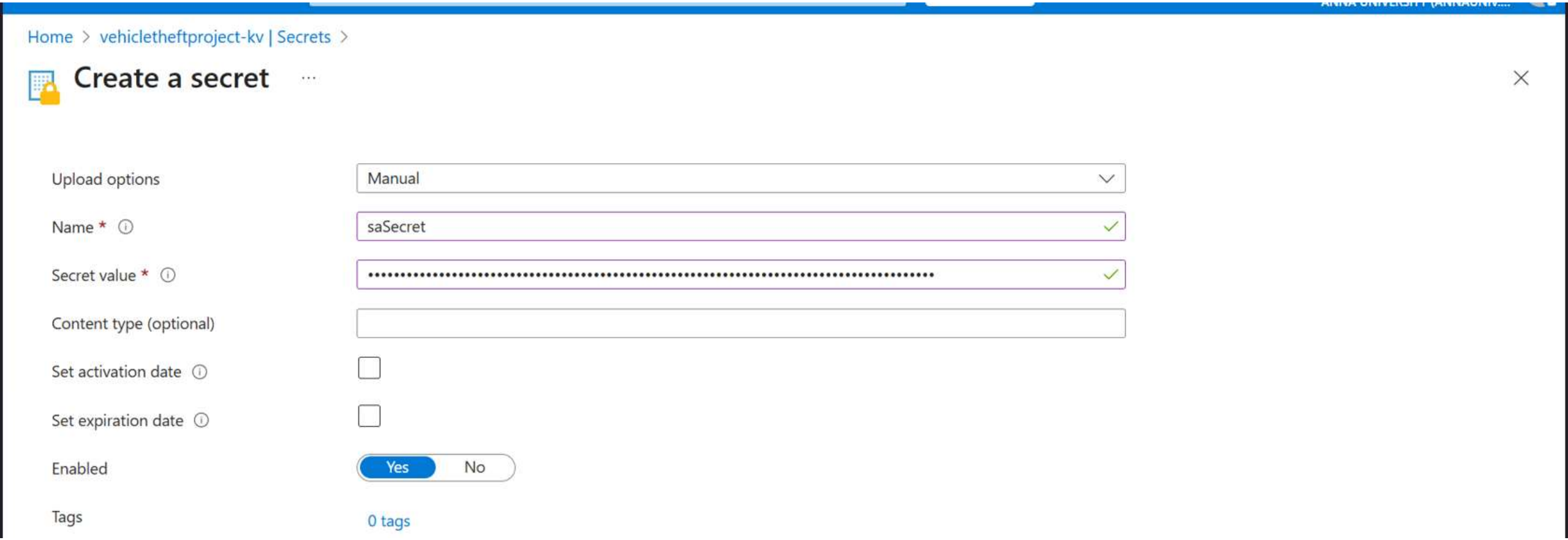
Secret Scope using Azure Key Vault



Access Key from Storage Account



Secret Creation



Mounting of the Medallion Containers

Just now (25s) 2 Python

```
dbutils.fs.mount(
    source = "wasbs://silver@vehicletheftprojectsa.blob.core.windows.net/",
    mount_point="/mnt/silver",
    extra_configs={
        "fs.azure.account.key.vehicletheftprojectsa.blob.core.windows.net": dbutils.secrets.get('dbScope','saSecret')
    }
)

dbutils.fs.mount(
    source = "wasbs://gold@vehicletheftprojectsa.blob.core.windows.net/",
    mount_point="/mnt/gold",
    extra_configs={
        "fs.azure.account.key.vehicletheftprojectsa.blob.core.windows.net": dbutils.secrets.get('dbScope','saSecret')
    }
)
```

```
True
```

Successful Mount

Just now (<1s) 2 Python

```
dbutils.fs.ls("/mnt/bronze")
```

```
[FileInfo(path='dbfs:/mnt/bronze/BankTransactionDataset_1K (1).csv', name='BankTransactionDataset_1K (1).csv', size=510143,
modificationTime=1758866794000),
 FileInfo(path='dbfs:/mnt/bronze/locations.csv', name='locations.csv', size=732, modificationTime=1758857603000),
 FileInfo(path='dbfs:/mnt/bronze/make_details.csv', name='make_details.csv', size=2993, modificationTime=1758468294000),
 FileInfo(path='dbfs:/mnt/bronze/stolen_vehicles.csv', name='stolen_vehicles.csv', size=226866, modificationTime=17584683190
00),
 FileInfo(path='dbfs:/mnt/bronze/stolen_vehicles_db_data_dictionary.csv', name='stolen_vehicles_db_data_dictionary.csv', siz
e=866, modificationTime=1758468340000)]
```

Loading Files from Cloud into SPARK

Just now (2s) 6 Python

```
location_df=spark.read.format("csv").option("header","true").option("inferSchema","true").load("/mnt/bronze/locations.csv")
make_details_df=spark.read.format("csv").option("header","true").option("inferSchema","true").load("/mnt/bronze/make_details.csv")
stolen_vehicles_df=spark.read.format("csv").option("header","true").option("inferSchema","true").load("/mnt/bronze/stolen_vehicles.csv")
database_df=spark.read.format("csv").option("header","true").option("inferSchema","true").load("/mnt/bronze/locations.csv")
```

▸ (8) Spark Jobs

▸ database_df: pyspark.sql.dataframe.DataFrame = [location_id: integer, region: string ... 3 more fields]
▸ location_df: pyspark.sql.dataframe.DataFrame = [location_id: integer, region: string ... 3 more fields]
▸ make_details_df: pyspark.sql.dataframe.DataFrame = [make_id: integer, make_name: string ... 1 more field]
▸ stolen_vehicles_df: pyspark.sql.dataframe.DataFrame = [vehicle_id: integer, vehicle_type: string ... 6 more fields]

Cloud Data Transformation in ADF

Cloud Data Transformation in ADF

After Cloud Data Transformation in ADF

Just now (<1s) 9 Python

```
location_df=location_df.withColumn("Population",regexp_replace(col("Population"),",","").cast("integer"))
```

location_df: pyspark.sql.dataframe.DataFrame = [location_id: integer, region: string ... 3 more fields]

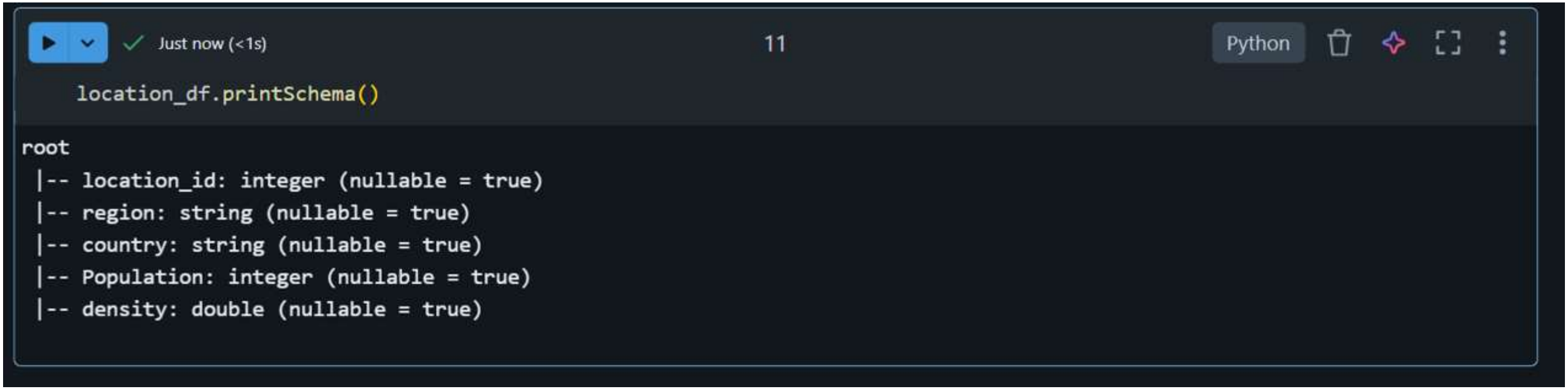
Just now (<1s) 10 Python

```
location_df.show()
```

(1) Spark Jobs

```
+-----------+------------------+-----------+----------+-------+
|location_id|            region|    country|Population|density|
+-----------+------------------+-----------+----------+-------+
|        101|         Northland|New Zealand|    201500|  16.11|
|        102|          Auckland|New Zealand|   1695200| 343.09|
|        103|           Waikato|New Zealand|    513800|   21.5|
|        104|     Bay of Plenty|New Zealand|    347700|   28.8|
|        105|          Gisborne|New Zealand|     52100|   6.21|
```

After Cloud Data Transformation in ADF

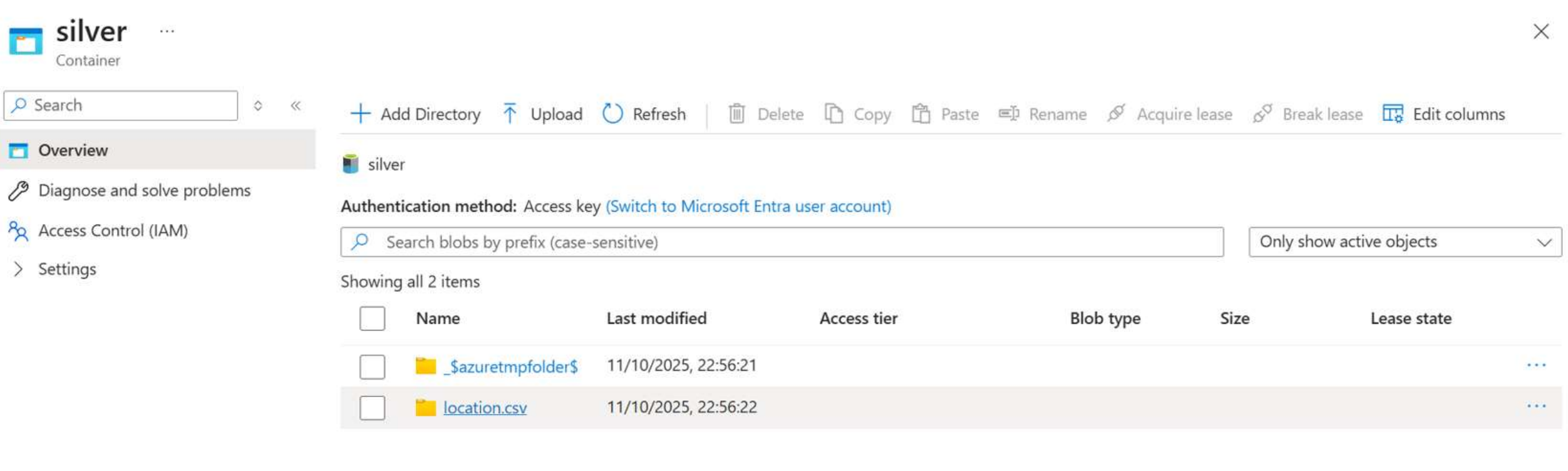


Migration into Silver Container with Metadata

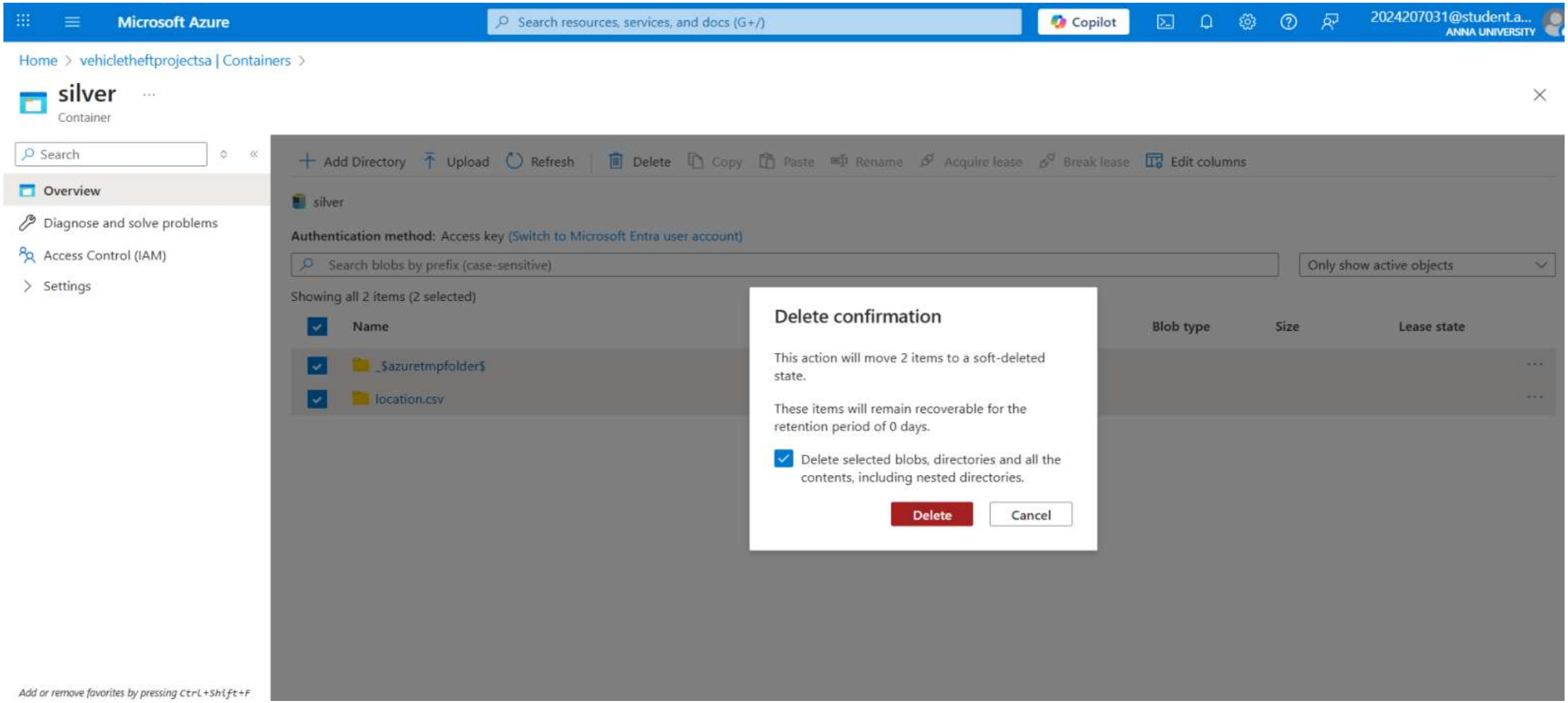
Just now (3s) 14 Python

```
location_df.write.option("header","true").csv("/mnt/silver/location.csv")
```

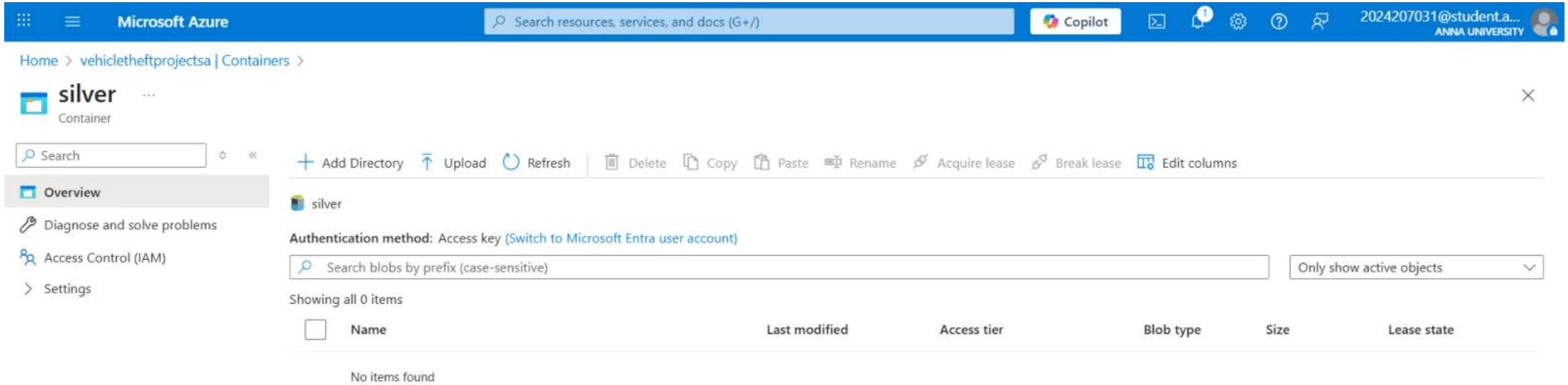
(1) Spark Jobs



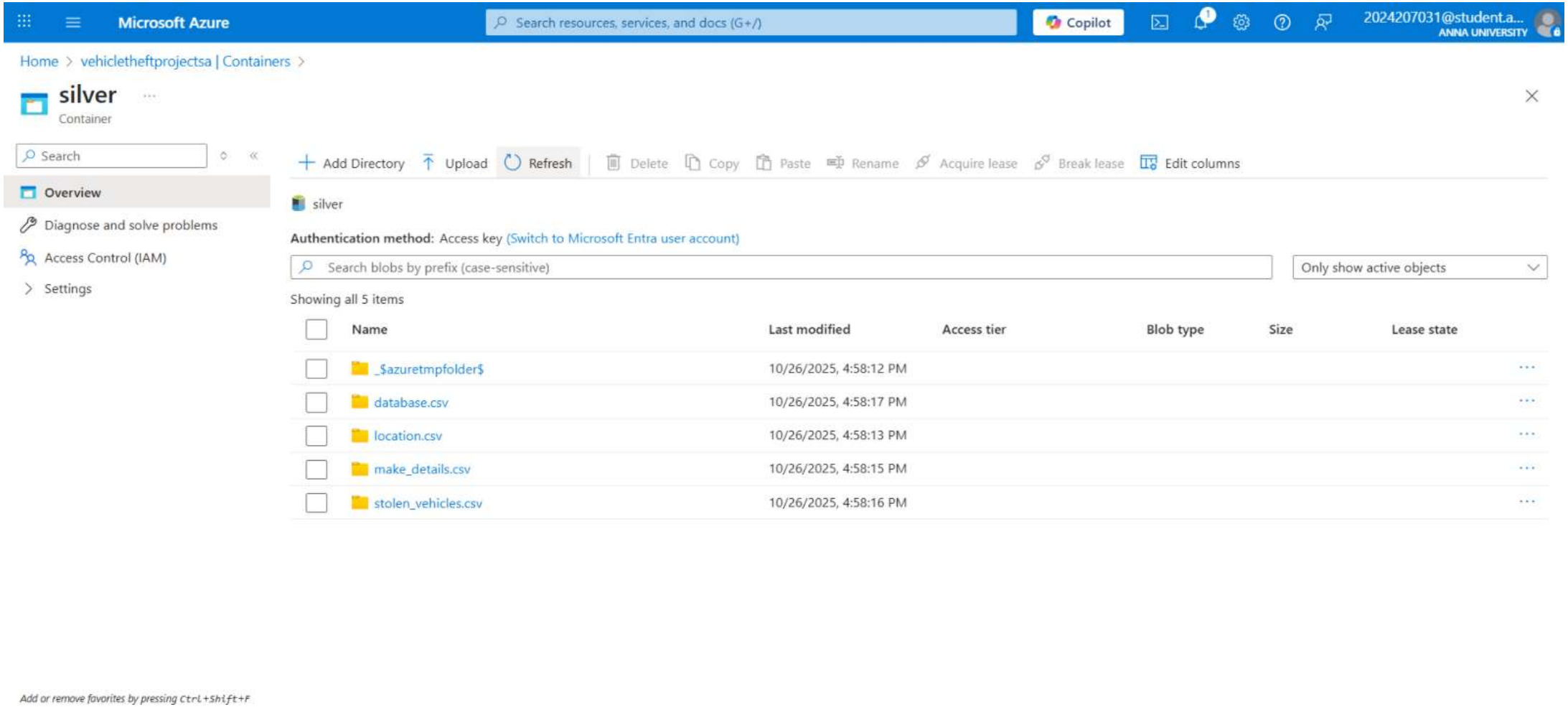
Data Cleanup and Re-Ingestion Preparation



Data Cleanup and Re-Ingestion Preparation



Data Cleanup and Re-Ingestion Preparation



Data Validation and Quality Check in Databricks

Data Validation and Quality Check in Databricks

Just now (1s) 16 Python

```
null_count_location = location_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in location_df.columns])
null_count_make_details = make_details_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in make_details_df.
columns])
null_count_stolen_vehicles = stolen_vehicles_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in
stolen_vehicles_df.columns])
null_count_database = database_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in database_df.columns])

null_count_make_details.show()
```

▸ (2) Spark Jobs

▸ null_count_database: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]
▸ null_count_location: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]
▸ null_count_make_details: pyspark.sql.dataframe.DataFrame = [make_id: long, make_name: long ... 1 more field]
▸ null_count_stolen_vehicles: pyspark.sql.dataframe.DataFrame = [vehicle_id: long, vehicle_type: long ... 6 more fields]

```
+-------+---------+---------+
|make_id|make_name|make_type|
+-------+---------+---------+
|      0|        0|        0|
+-------+---------+---------+
```

Null Value Analysis and Detection

Just now (1s) 16 Python

```
null_count_location = location_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in location_df.columns])
null_count_make_details = make_details_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in make_details_df.
columns])
null_count_stolen_vehicles = stolen_vehicles_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in
stolen_vehicles_df.columns])
null_count_database = database_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in database_df.columns])

null_count_database.show()
```

(2) Spark Jobs

- null_count_database: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]
- null_count_location: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]
- null_count_make_details: pyspark.sql.dataframe.DataFrame = [make_id: long, make_name: long ... 1 more field]
- null_count_stolen_vehicles: pyspark.sql.dataframe.DataFrame = [vehicle_id: long, vehicle_type: long ... 6 more fields]

```
+-----------+------+-------+----------+-------+
|location_id|region|country|population|density|
+-----------+------+-------+----------+-------+
|          0|     0|      0|         0|      0|
+-----------+------+-------+----------+-------+
```

Null Value Analysis and Detection

Just now (1s) 16 Python

```
null_count_location = location_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in location_df.columns])
null_count_make_details = make_details_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in make_details_df.
columns])
null_count_stolen_vehicles = stolen_vehicles_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in
stolen_vehicles_df.columns])
null_count_database = database_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in database_df.columns])

null_count_stolen_vehicles.show()
```

▸ (2) Spark Jobs

▸ null_count_database: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]
▸ null_count_location: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]
▸ null_count_make_details: pyspark.sql.dataframe.DataFrame = [make_id: long, make_name: long ... 1 more field]
▸ null_count_stolen_vehicles: pyspark.sql.dataframe.DataFrame = [vehicle_id: long, vehicle_type: long ... 6 more fields]

vehicle_id	vehicle_type	make_id	model_year	vehicle_desc	color	date_stolen	location_id
0	26	15	15	33	15	0	0

Schema Validation and Missing Value Handling

Just now (<1s) 17 Python

```
stolen_vehicles_df.printSchema()
```

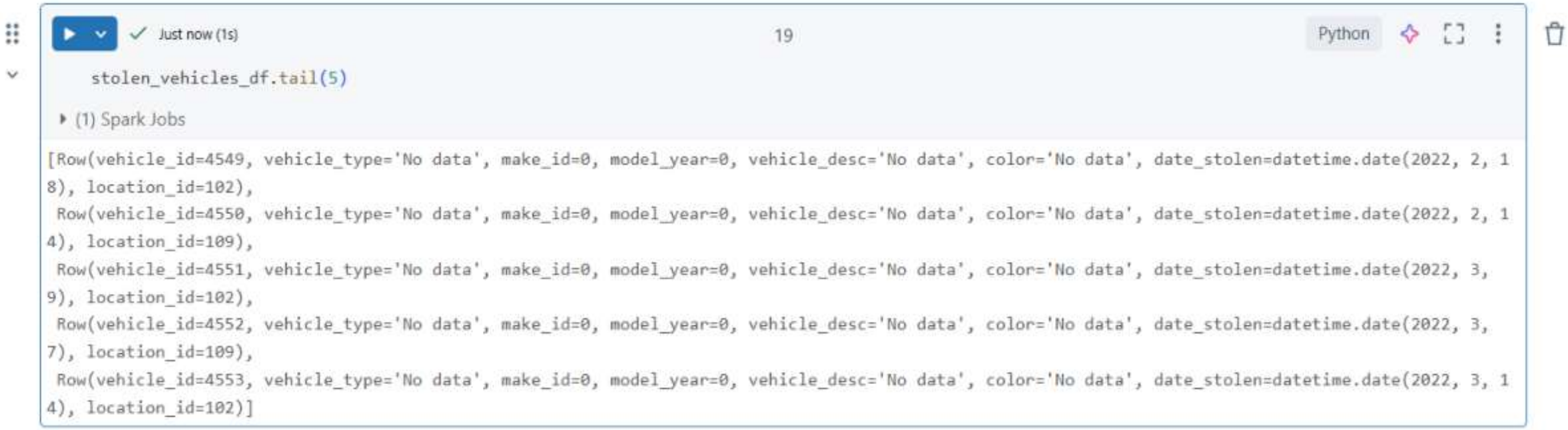
```
root
 |-- vehicle_id: integer (nullable = true)
 |-- vehicle_type: string (nullable = true)
 |-- make_id: integer (nullable = true)
 |-- model_year: integer (nullable = true)
 |-- vehicle_desc: string (nullable = true)
 |-- color: string (nullable = true)
 |-- date_stolen: date (nullable = true)
 |-- location_id: integer (nullable = true)
```

Just now (<1s) 18 Python

```
stolen_vehicles_df = stolen_vehicles_df.fillna({
    "vehicle_type" : "No data",
    "make_id" : 0,
    "model_year" : 0,
    "vehicle_desc": "No data",
    "color": "No data"
})
```

stolen_vehicles_df: pyspark.sql.dataframe.DataFrame = [vehicle_id: integer, vehicle_type: string ... 6 more fields]

Schema Validation and Missing Value Handling



Data Cleaning Validation and Final Output

Just now (1s) 20 Python

```
null_count_stolen_vehicles = stolen_vehicles_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in
stolen_vehicles_df.columns])
null_count_stolen_vehicles.show()
```

(2) Spark Jobs

null_count_stolen_vehicles: pyspark.sql.dataframe.DataFrame = [vehicle_id: long, vehicle_type: long ... 6 more fields]

vehicle_id	vehicle_type	make_id	model_year	vehicle_desc	color	date_stolen	location_id
0	0	0	0	0	0	0	0

Just now (<1s) 21 Python

```
stolen_vehicles_df.show(5)
```

(1) Spark Jobs

vehicle_id	vehicle_type	make_id	model_year	vehicle_desc	color	date_stolen	location_id
1	Trailer	623	2021	BST2021D	Silver	2021-11-05	102
2	Boat Trailer	623	2021	OUTBACK BOATS FT470	Silver	2021-12-13	105
3	Boat Trailer	623	2021	ASD JETSKI	Silver	2022-02-13	102
4	Trailer	623	2021	MSC 7X4	Silver	2021-11-13	106
5	Trailer	623	2018	D-MAX 8X5	Silver	2022-01-10	102

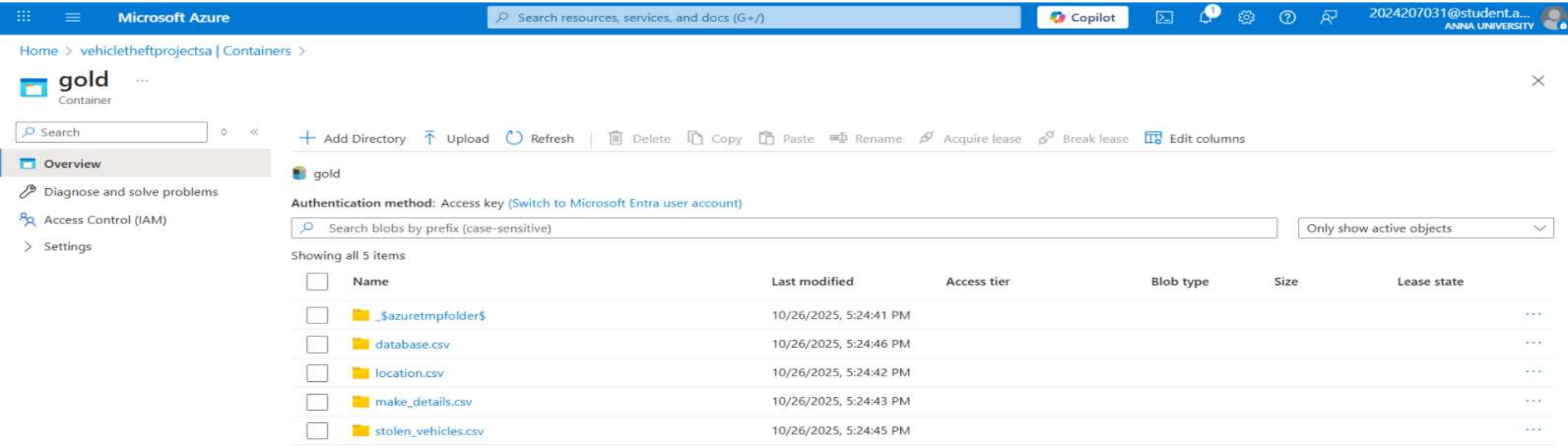
only showing top 5 rows

Exporting Curated Data to Gold Laver

```
Just now (6s)                                                    22                    Python

location_df.write.option("header", "true").csv("/mnt/gold/location.csv")
make_details_df.write.option("header", "true").csv("/mnt/gold/make_details.csv")
stolen_vehicles_df.write.option("header", "true").csv("/mnt/gold/stolen_vehicles.csv")
database_df.write.option("header", "true").csv("/mnt/gold/database.csv")

(4) Spark Jobs
```



Data Querying in Databricks (Gold Layer Analysis)

Just now (1s) 23 Python

```
location_df.createOrReplaceTempView("location")
make_details_df.createOrReplaceTempView("make_details")
stolen_vehicles_df.createOrReplaceTempView("stolen_vehicles")
database_df.createOrReplaceTempView("database")
```

Just now (2s) 24 SQL

```
%sql

SELECT model_year, count(*) AS number_of_vehicles_stolen
FROM stolen_vehicles
GROUP BY model_year
ORDER BY number_of_vehicles_stolen DESC
```

▸ (2) Spark Jobs

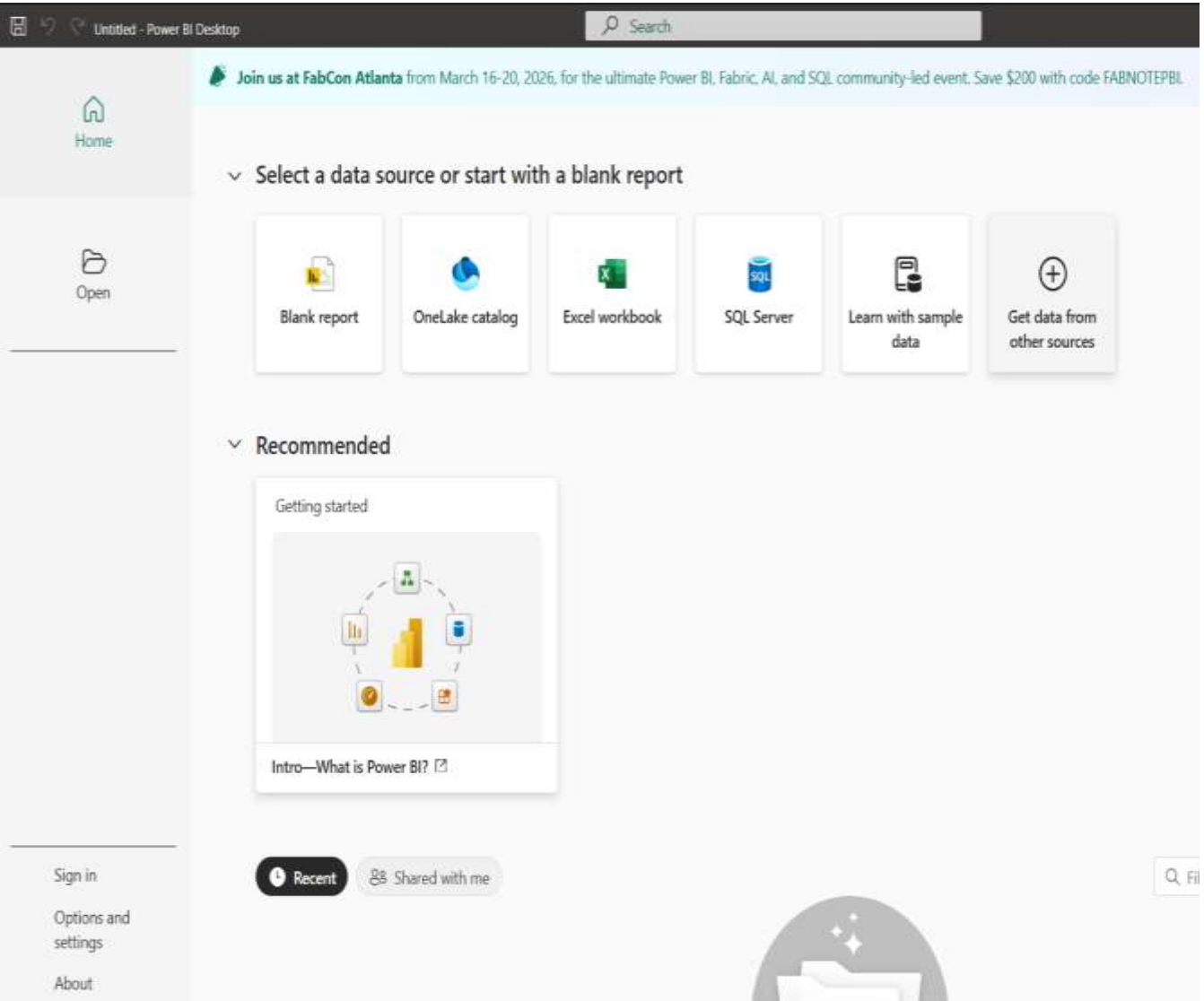
▸ _sqldf: pyspark.sql.dataframe.DataFrame = [model_year: integer, number_of_vehicles_stolen: long]

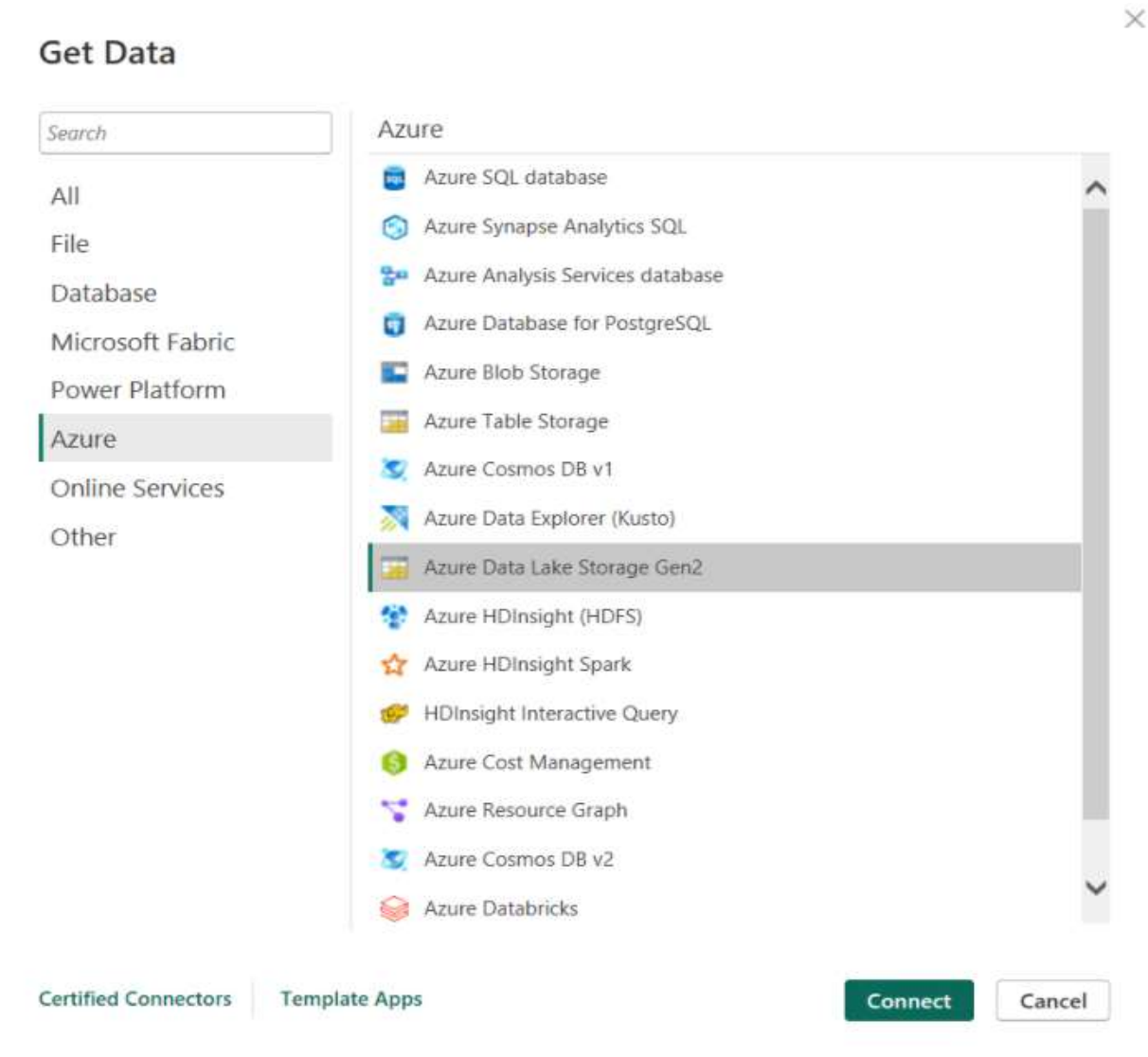
Table +

	model_year	number_of_vehicles_stolen
1	2005	347
2	2006	333
3	2007	251
4	2004	238
5	2008	190
6	2002	181
7	2003	173
8	1998	159
9	1996	156
10	2001	152
11	2021	148
12	1997	146
13	2000	145
14	1999	137
15	2009	125

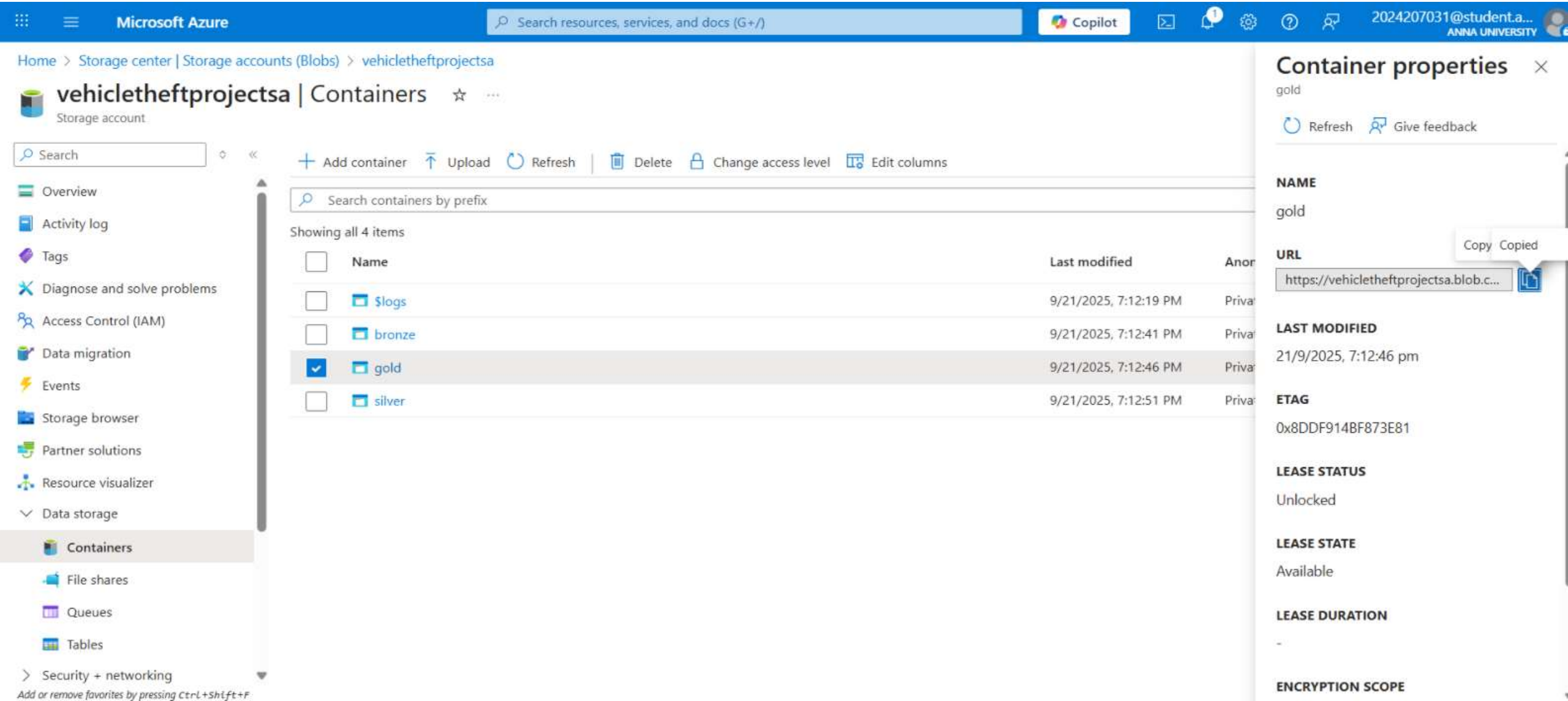
64 rows | 1.54s runtime

Power BI Integration – Initialization

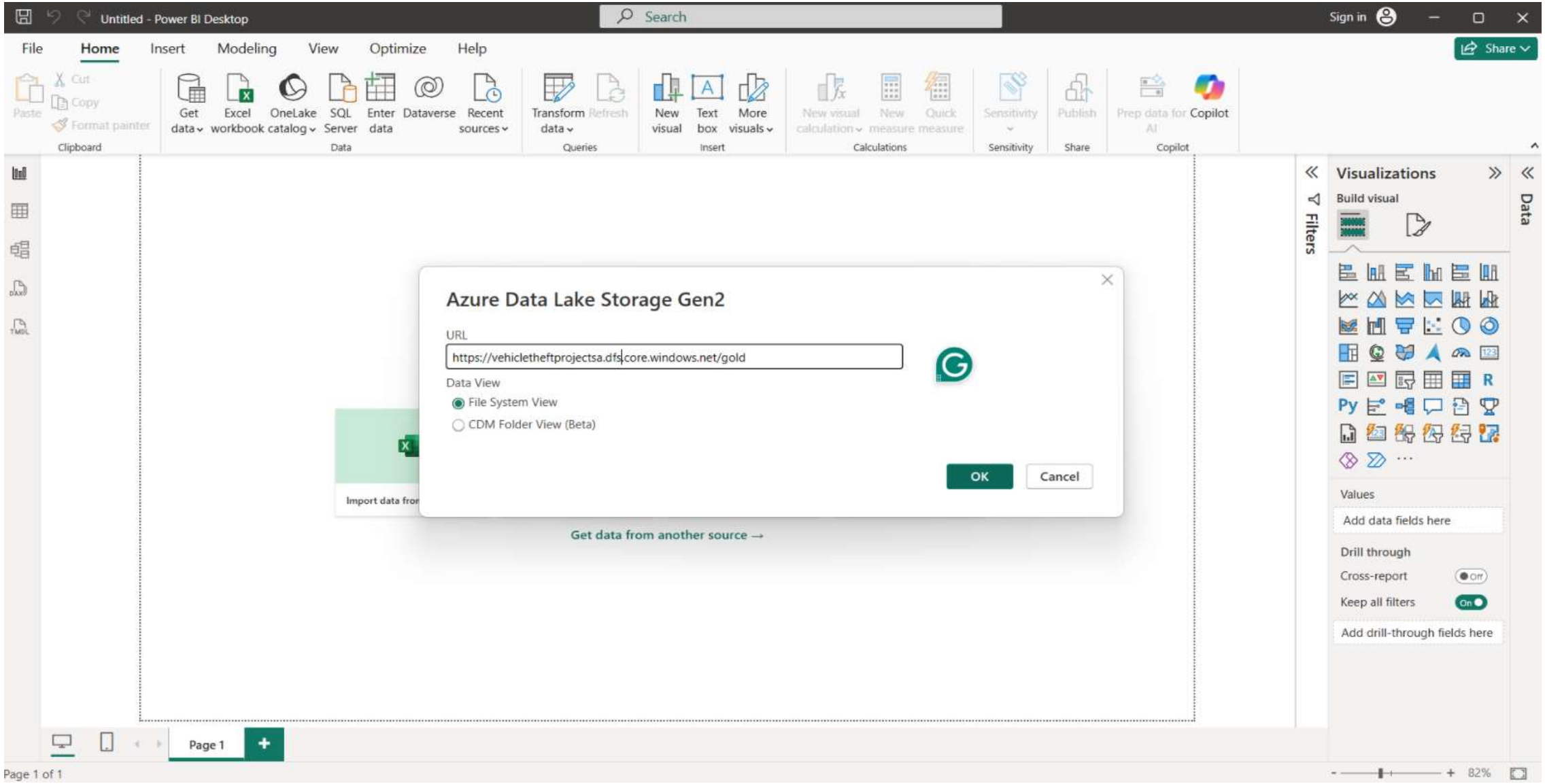




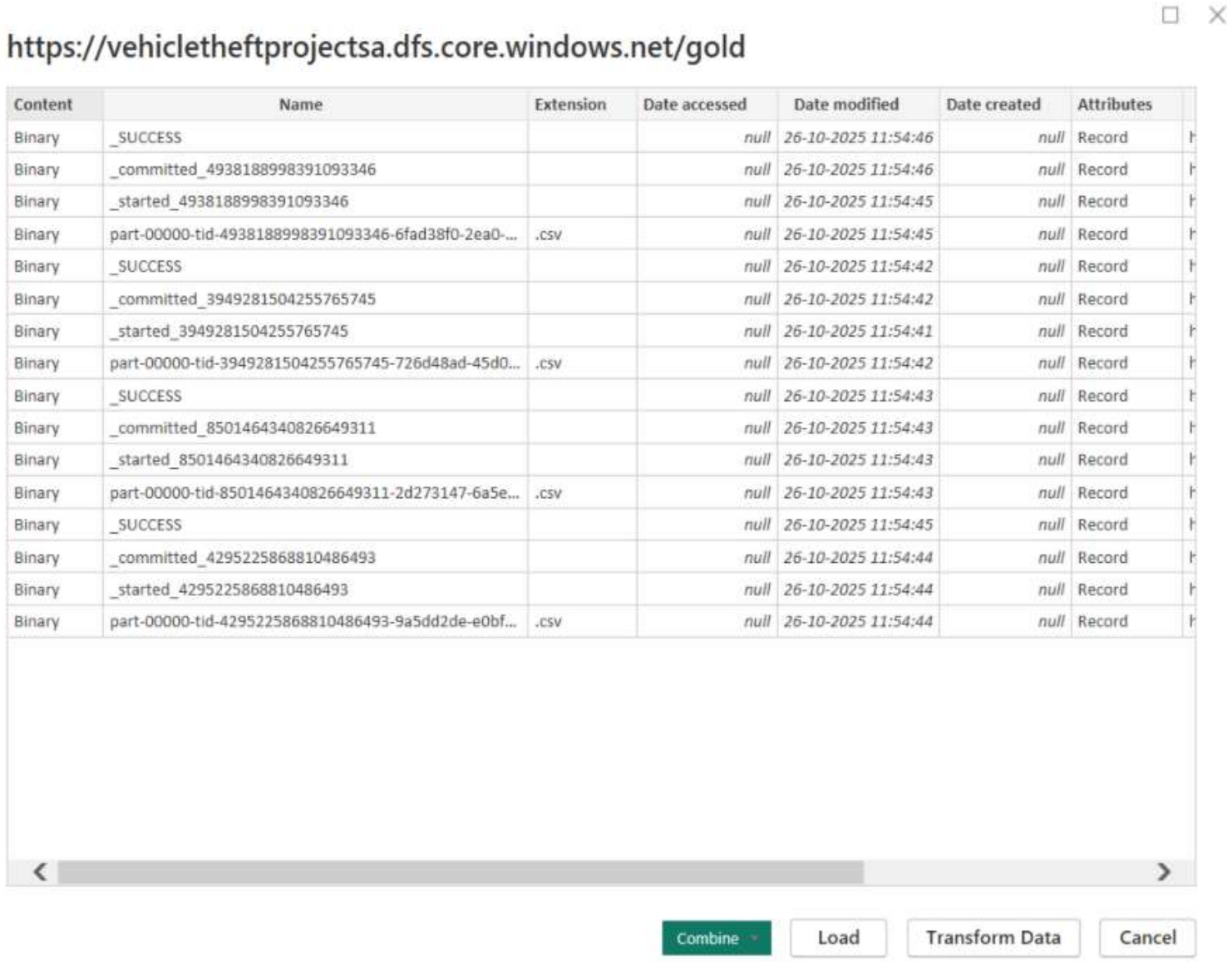
Linking Power BI to Azure Storage Account



Linking Power BI to Azure Storage Account

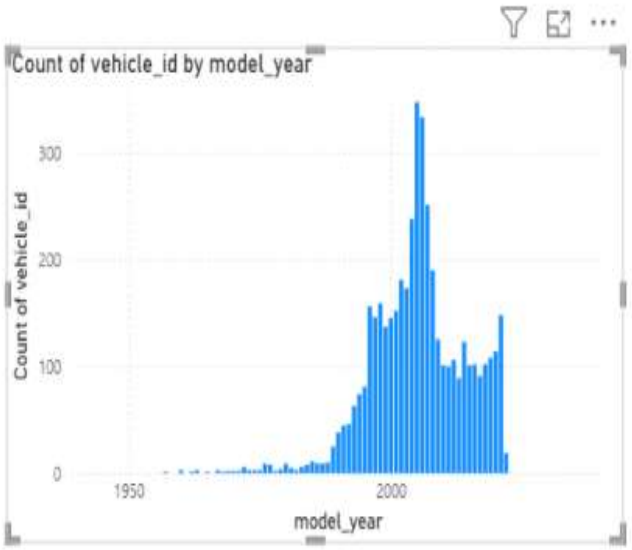


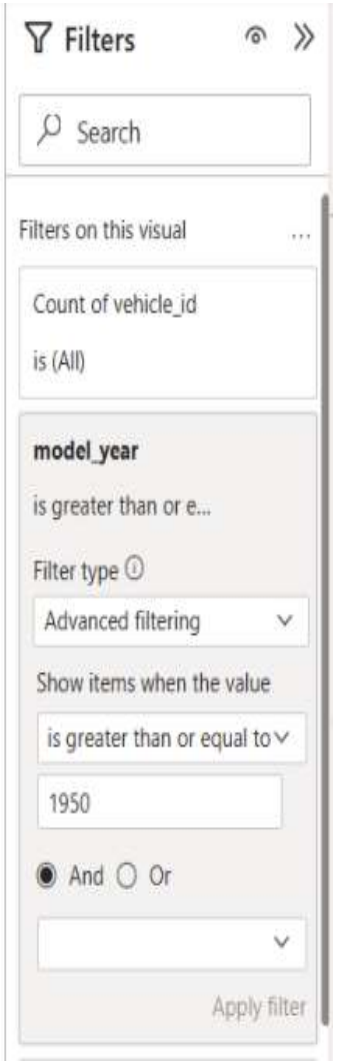
Data Import and Preview in Power BI

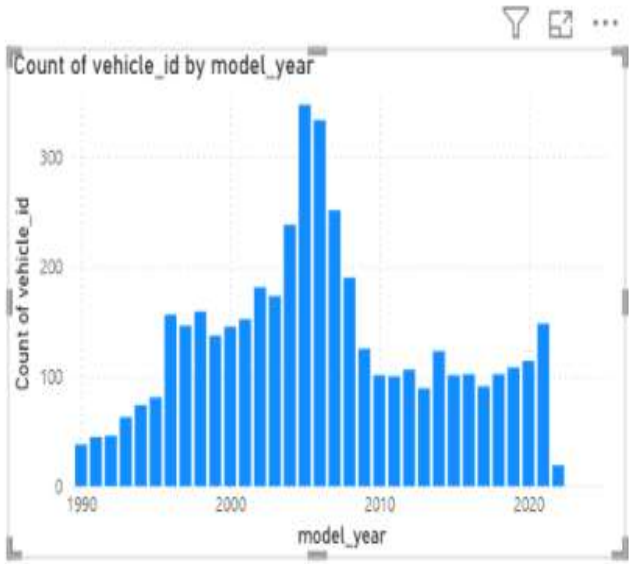


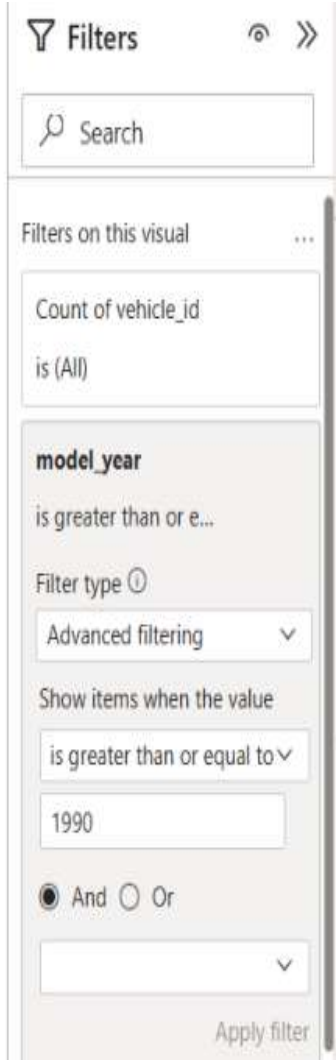
Content	Name	Extension	Date accessed	Date modified	Date created	Attributes
Binary	_SUCCESS		null	26-10-2025 11:54:46	null	Record
Binary	_committed_4938188998391093346		null	26-10-2025 11:54:46	null	Record
Binary	_started_4938188998391093346		null	26-10-2025 11:54:45	null	Record
Binary	part-00000-tid-4938188998391093346-6fad38f0-2ea0-...	.csv	null	26-10-2025 11:54:45	null	Record
Binary	_SUCCESS		null	26-10-2025 11:54:42	null	Record
Binary	_committed_3949281504255765745		null	26-10-2025 11:54:42	null	Record
Binary	_started_3949281504255765745		null	26-10-2025 11:54:41	null	Record
Binary	part-00000-tid-3949281504255765745-726d48ad-45d0...	.csv	null	26-10-2025 11:54:42	null	Record
Binary	_SUCCESS		null	26-10-2025 11:54:43	null	Record
Binary	_committed_8501464340826649311		null	26-10-2025 11:54:43	null	Record
Binary	_started_8501464340826649311		null	26-10-2025 11:54:43	null	Record
Binary	part-00000-tid-8501464340826649311-2d273147-6a5e...	.csv	null	26-10-2025 11:54:43	null	Record
Binary	_SUCCESS		null	26-10-2025 11:54:45	null	Record
Binary	_committed_4295225868810486493		null	26-10-2025 11:54:44	null	Record
Binary	_started_4295225868810486493		null	26-10-2025 11:54:44	null	Record
Binary	part-00000-tid-4295225868810486493-9a5dd2de-e0bf...	.csv	null	26-10-2025 11:54:44	null	Record

Model Year Visualization – Column Charts

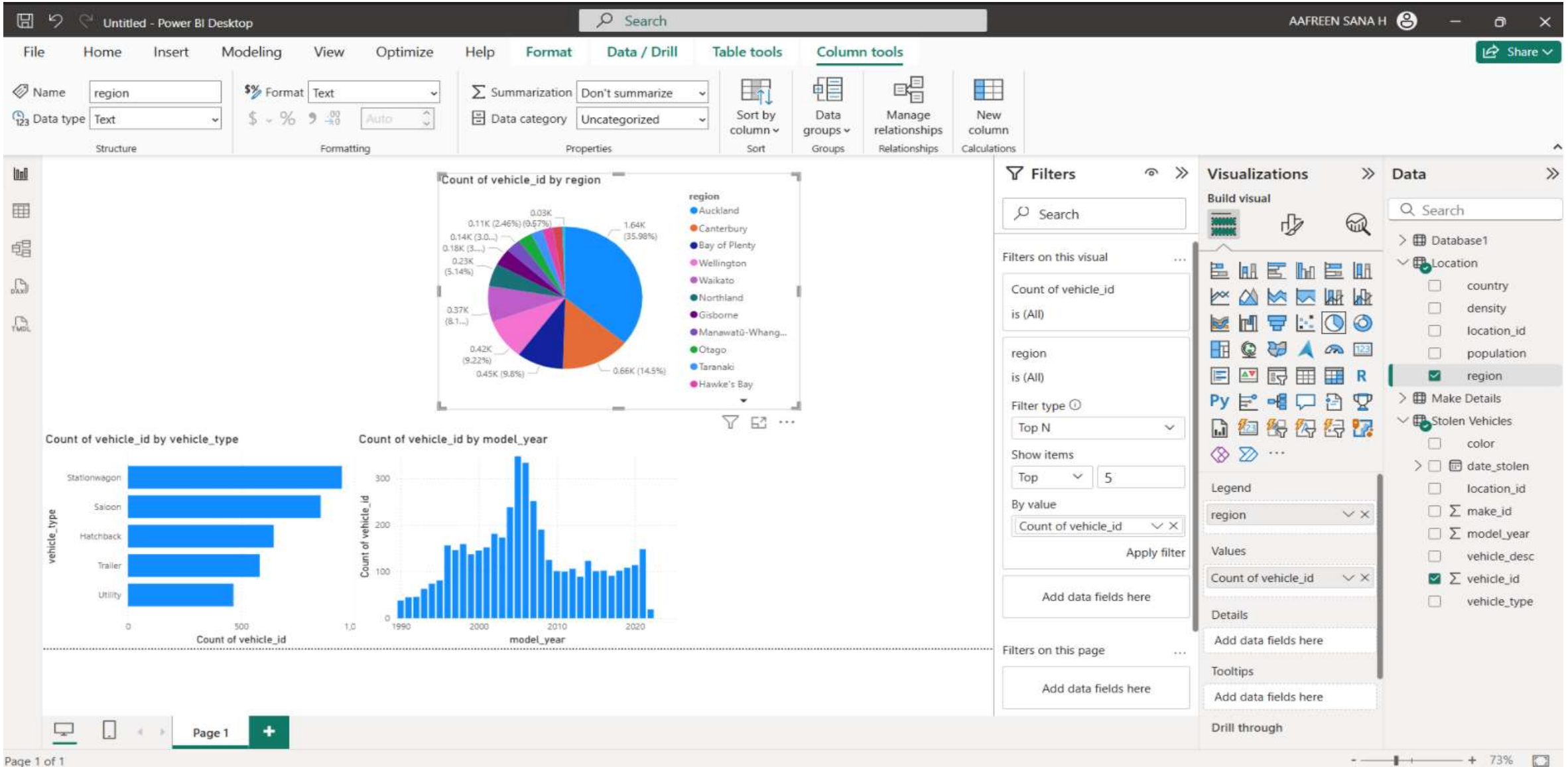




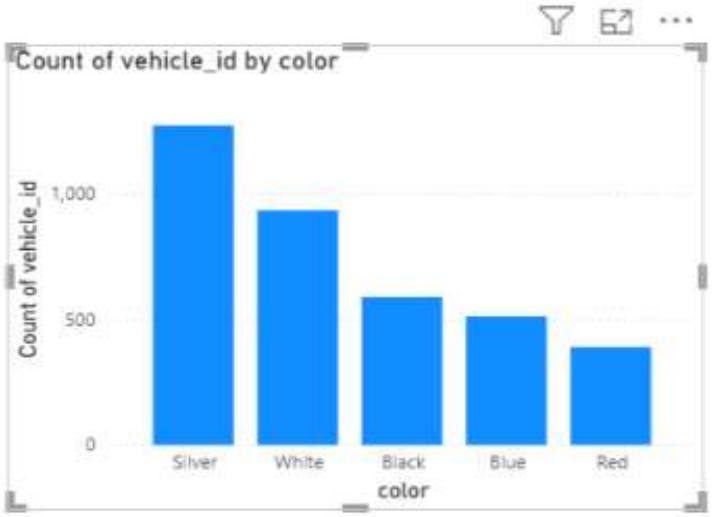


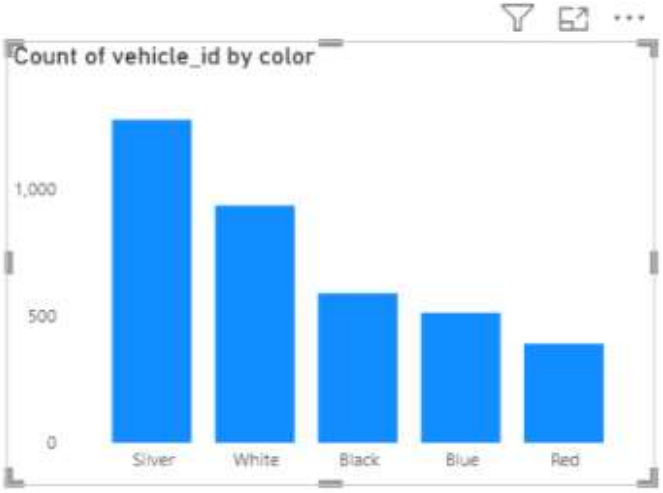


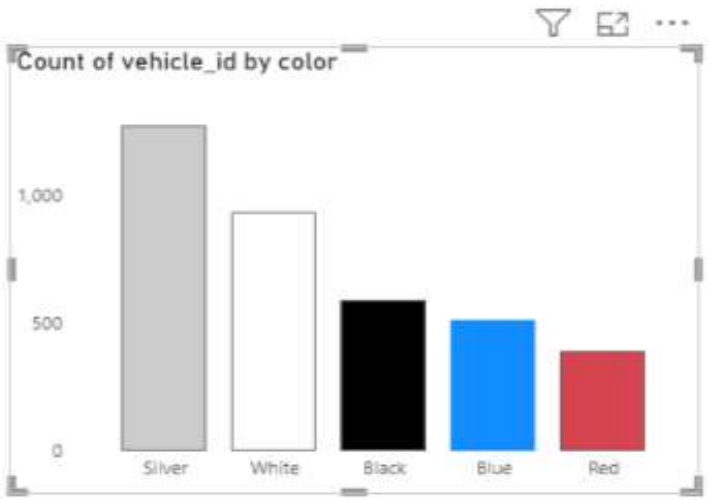
Region-Wise Visualization – Pie Charts

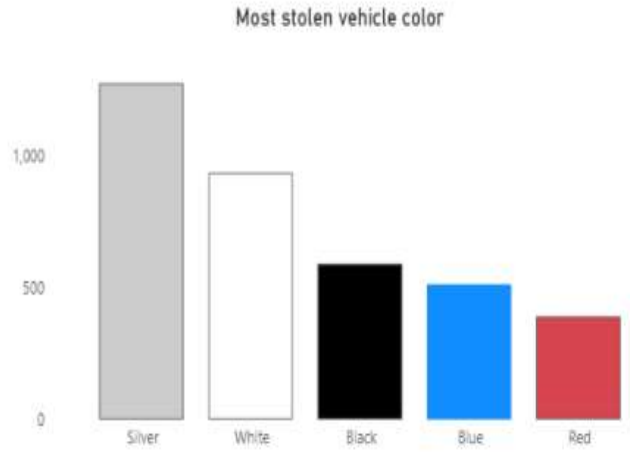


Color-Based Visualization – Bar Charts

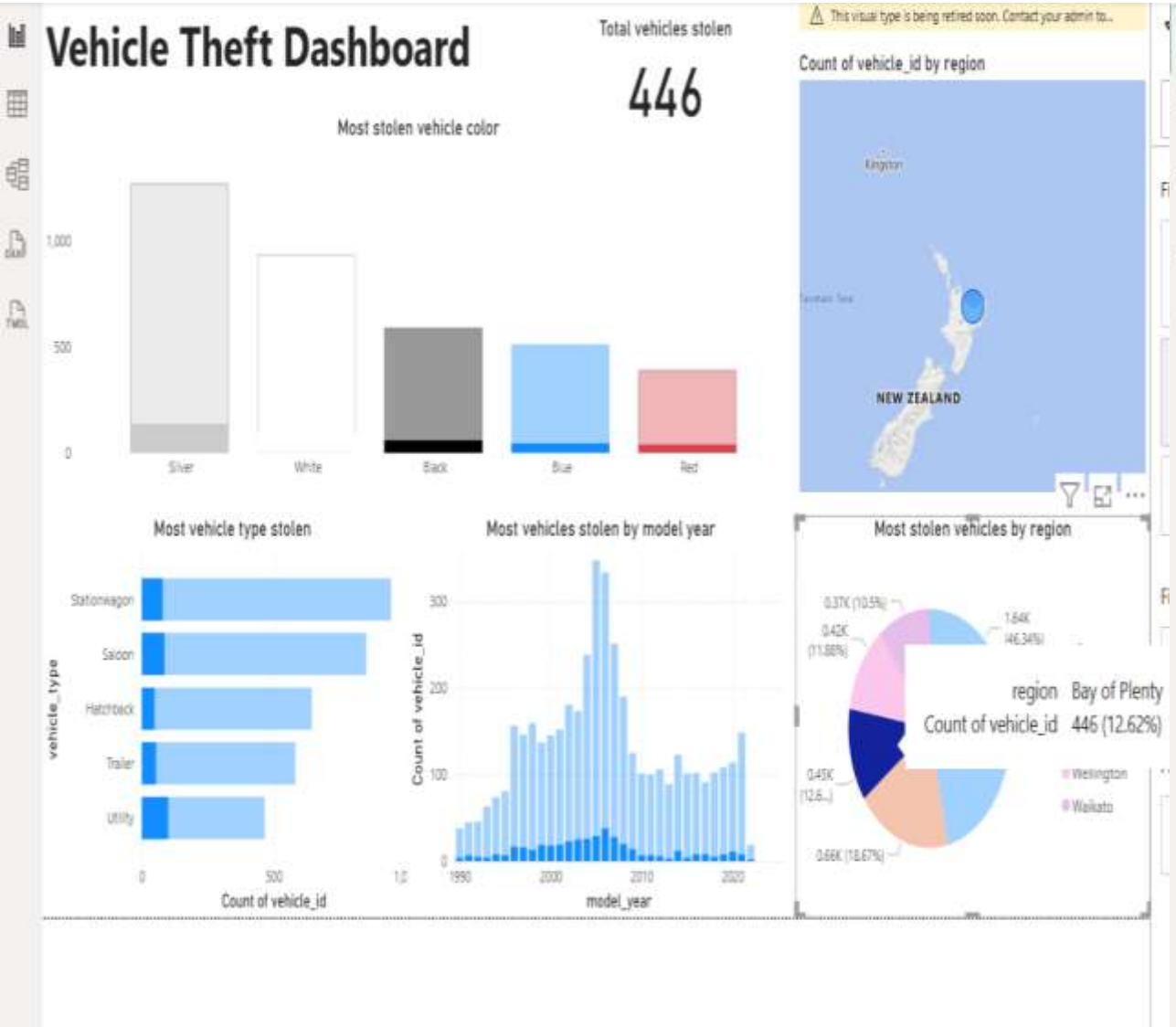


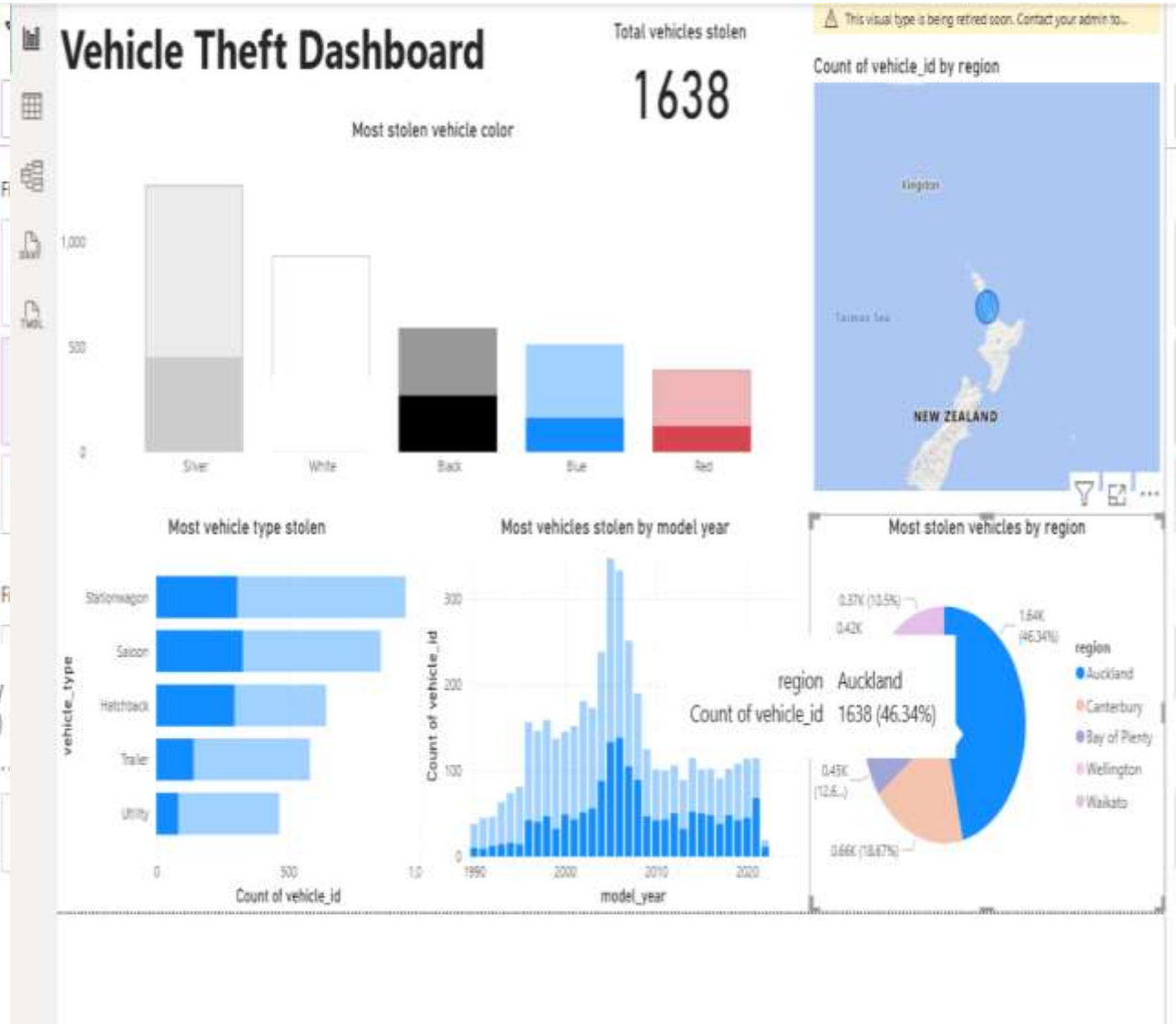






Dashboard


Vehicle Theft Dashboard
Total vehicles stolen
446
Most stolen vehicle color
Count of vehicle_id by region
NEW ZEALAND
Most vehicle type stolen
Most vehicles stolen by model year
Most stolen vehicles by region
region Bay of Plenty
Count of vehicle_id 446 (12.62%)


Vehicle Theft Dashboard
Total vehicles stolen
1638
Most stolen vehicle color
Count of vehicle_id by region
NEW ZEALAND
Most vehicle type stolen
Most vehicles stolen by model year
Most stolen vehicles by region
region Auckland
Count of vehicle_id 1638 (46.34%)