

Azure Data Engineering Pipeline with Medallion Architecture

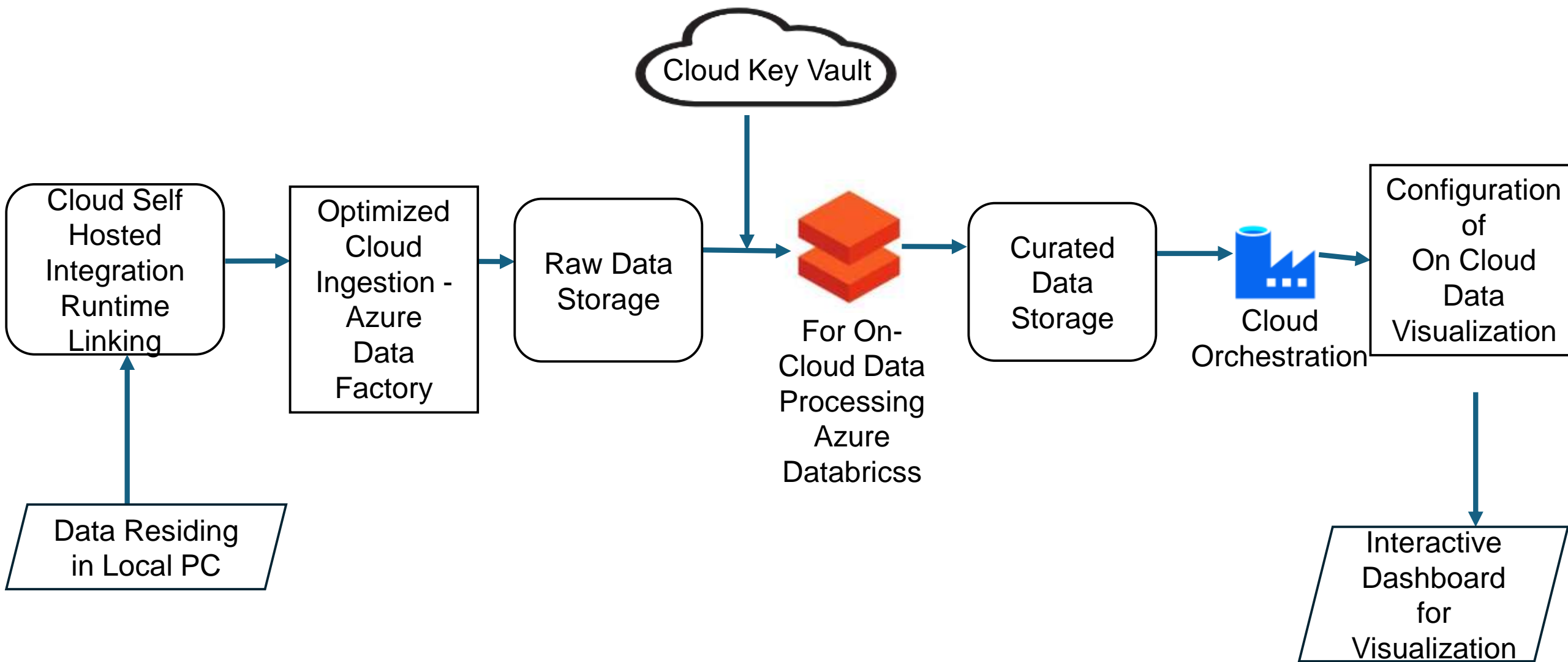
Problem statement

- 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

 Microsoft Azure



Search resources, services, and docs (G+ /)






 Copilot


Home >


 **vehicletheftproject_1758439475069** | Overview  ...


Deployment


 


 Delete  Cancel  Redeploy  Download  Refresh


 Overview

 Inputs

 Outputs

 Template

 Your deployment is complete




Deployment name: vehiclethe...


Subscription: [Azure for Stude...](#)

Resource group: [vehicletheftg](#)

Start time: 21/09/2025, 12:54:45

Correlation ID: 3656615e-74ec-4650

 Deployment details

 Next steps

Go to resource

Deployment of Azure Data Factory

[Home](#) >



Microsoft.DataFactory-20250921130324 | Overview ...

Deployment



Delete



Cancel



Redeploy



Download



Refresh



Overview



Inputs



Outputs



Template



Your deployment is complete



Deployment name : Microsoft.DataFactory-20250921130324

Subscription : [Azure for Students](#)

Resource group : [vehicletheftrg](#)

Start time : 21/09/2025, 13:04:57

Correlation ID : a54d0c95-24af-4449-9bac-6144e8d8b4c7




Deployment details







Next steps

[Go to resource](#)

Linking of Storage Account and ADF


 **vehicletheftprojectrg**
Resource group



  





What are the best practices for managing


>>


 Create


 Manage view 

 Delete resource group

 Refresh


 Export to CSV





 Essentials


Resources






Recommendations

 Filter for any field...

Type equals **all** 

Location equals **all** 

 Add filter

| <input type="checkbox"/> | Name  | Type |
|--------------------------|---|---|
| <input type="checkbox"/> |  vehicletheftprojectdf |  Data factory (V2) |
| <input type="checkbox"/> |  vehicletheftprojectsa |  Storage account |

Initialization of Containers for Medallion



vehicletheftproject | Containers



Storage account



Search



Diagnose and solve problems



Access Control (IAM)



Data migration



Events



Storage browser



Partner solutions



Resource visualizer



Data storage



Add container



Upload



Refresh



Delete



Change access level



Search containers by prefix

Showing all 4 items



Name

Last modified

Anonymous ac



 \$logs

21/09/2025, 12:55:16

Private



 bronze

21/09/2025, 12:59:35

Private



 gold

21/09/2025, 12:59:57

Private



 silver

21/09/2025, 12:59:46

Private

Configuration of Self Hosted IR

Integration Runtime (Self-hosted) Express Setup


Installing and registering the Integration Runtime (Self-hosted) node.


- ✓ Loading configuration
- ✓ Downloading Integration Runtime (Self-hosted)
- ✓ Installing Integration Runtime (Self-hosted)
- ✓ Registering Integration Runtime (Self-hosted)


Integration Runtime (Self-hosted) "vehicletheftintegrationRuntime1" is successfully installed on your computer.

Created Linked Service for File System

Linked services






Linked service defines the connection information to a data store or compute. [Learn more](#) 

 New

 Filter by name

Annotations : **Any**

Showing 1 - 1 of 1 items

| Name  | Type  | Related  | Annotations  |
|---|--|---|---|
|  VehicleTheftFileServer1 | File system | 0 | |

Initialization of Data Copy in Pipeline

vehicletheftpipeline1 ●

Activities ⌵ ⏪

copy

✓ Move and transform

Copy data

✓ Validate ✓ Validate copy runtime ▶ Debug ⚡ Add trigger

Copy data

Copy data1

🗑️ </> 📄 ➡

✓ ✗ ➡

The screenshot displays the Azure Data Factory (ADF) user interface. At the top, a pipeline named 'vehicletheftpipeline1' is selected. Below the pipeline name, there's a toolbar with buttons for 'Validate', 'Validate copy runtime', 'Debug', and 'Add trigger'. The 'Activities' pane on the left shows a search bar with 'copy' entered, and a list of activities under the 'Move and transform' category, with 'Copy data' selected. The main canvas shows the 'Copy data' activity, labeled 'Copy data1', which is represented by a blue icon of two cylinders. To the right of the activity name are three status icons: a green checkmark, a red 'X', and a blue arrow. Below the activity name is a toolbar with icons for deleting the activity, viewing its code, copying it, and a blue arrow icon. The activity is currently in a state where it can be added to the pipeline, as indicated by the blue arrow icon in the bottom right corner of the activity box.

Successful Data Ingestion

The screenshot displays the Microsoft Azure Data Factory portal. The top navigation bar includes the Microsoft Azure logo, the Data Factory name 'vehicletheftprojectdf', a search bar, and a user profile for '2024207031@student.annauniv.edu' from 'ANNA UNIVERSITY'. A notification banner asks 'Would you like to see Data Factory...'. The left sidebar shows 'Factory Resources' with a filter and a list of items: Pipelines (vehicletheftpipeline1), Change Data Capture (p), Datasets, Data flows, and Power Query. The main area shows a 'Preview data' window for the linked service 'vehicletheftFileServer1' and object 'locations.csv'. The preview displays a table with 10 rows of location data from New Zealand.

Preview data

Linked service: vehicletheftFileServer1

Object: locations.csv

| | location_id | region | country | population | density |
|----|-------------|--------------------|-------------|------------|---------|
| 1 | 101 | Northland | New Zealand | 201,500 | 16.11 |
| 2 | 102 | Auckland | New Zealand | 1,695,200 | 343.09 |
| 3 | 103 | Waikato | New Zealand | 513,800 | 21.5 |
| 4 | 104 | Bay of Plenty | New Zealand | 347,700 | 28.8 |
| 5 | 105 | Gisborne | New Zealand | 52,100 | 6.21 |
| 6 | 106 | Hawke's Bay | New Zealand | 182,700 | 12.92 |
| 7 | 107 | Taranaki | New Zealand | 127,300 | 17.55 |
| 8 | 108 | Manawatū-Whanganui | New Zealand | 258,200 | 11.62 |
| 9 | 109 | Wellington | New Zealand | 543,500 | 67.52 |
| 10 | 110 | Tasman | New Zealand | 58,700 | 6.1 |



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 Factory.



Data Factory



Validate all



Publish all

9

Pr



vehicletheftpipeline1



Activities



Validate

Validate copy runtime

Debug

Add trigger

copy

Validate the current resource

Move and transform

Copy data

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

| <input type="checkbox"/> | Name | Last modified | Access tier |
|--------------------------|--|----------------------|----------------|
| <input type="checkbox"/> | BankTransactionDataset_1K (1).csv | 26/09/2025, 09:03:23 | Hot (Inferred) |
| <input type="checkbox"/> | locations.csv | 26/09/2025, 09:03:23 | Hot (Inferred) |
| <input type="checkbox"/> | make_details.csv | 21/09/2025, 20:54:54 | Hot (Inferred) |
| <input type="checkbox"/> | stolen_vehicles.csv | 21/09/2025, 20:55:19 | Hot (Inferred) |
| <input type="checkbox"/> | stolen_vehicles_db_data_dictionary.csv | 21/09/2025, 20:55:40 | Hot (Inferred) |

Creation of Azure Databricks WS

[Home](#) > [Azure Databricks](#) >

Create an Azure Databricks workspace ...

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

Azure for Students



Resource group * ⓘ

cloudlab



[Create new](#)

Instance Details

Workspace name *

vehciletheftadb



Deployed Azure Data Bricks

The screenshot displays the Databricks interface within the Microsoft Azure ecosystem. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, the user name 'vehciletheftadb', and a profile icon. A left-hand sidebar lists various navigation options: '+ New', 'Workspace', 'Recents', 'Catalog', 'Jobs & Pipelines', 'Compute', 'Data Engineering', 'Job Runs', 'AI/ML', 'Playground', 'Experiments', 'Features', 'Models', and 'Serving'. The main content area is titled 'Welcome to Databricks' and features a search bar with the placeholder text 'Search data, notebooks, recents, and more...' and a keyboard shortcut 'CTRL + P'. Below the search bar is a 'Set up your workspace' section with a rocket icon, a description: 'Follow this step-by-step guide that walks you through setting up the workspace for your new Databricks account.', and a 'Get started' button with an external link icon. At the bottom of the main area, there is a row of five filter buttons: 'Recents' (selected), 'Favorites', 'Popular', 'Mosaic AI', and 'What's new'. The bottom section is titled 'Start your journey' and contains the text: 'Try the "New" menu, where you can upload or connect to data and then explore it in a notebook or dashboard.' A blue '+ New' button is positioned at the bottom right of this section.

Microsoft Azure databricks vehciletheftadb

Welcome to Databricks

Search data, notebooks, recents, and more... CTRL + P

Set up your workspace

Follow this step-by-step guide that walks you through setting up the workspace for your new Databricks account.

Get started

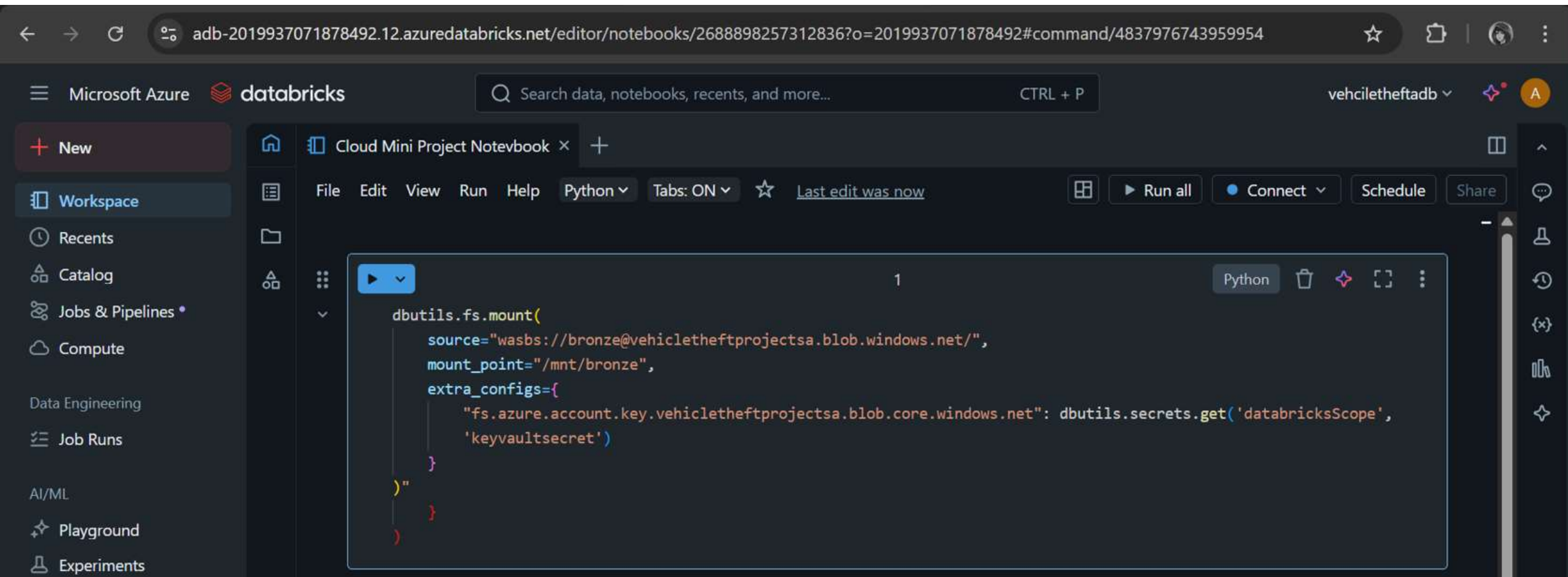
Recents Favorites Popular Mosaic AI What's new

Start your journey

Try the "New" menu, where you can upload or connect to data and then explore it in a notebook or dashboard.

+ New

Mounting of Bronze Container into Azure DataBricks



The screenshot displays the Azure Databricks web interface. The browser address bar shows the URL: `adb-2019937071878492.12.azuredatabricks.net/editor/notebooks/2688898257312836?o=2019937071878492#command/4837976743959954`. The interface includes a sidebar with navigation options: New, Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Data Engineering, Job Runs, AI/ML, Playground, and Experiments. The main workspace area shows a notebook titled "Cloud Mini Project Notevbook". The notebook's menu bar includes File, Edit, View, Run, Help, Python, Tabs: ON, and a star icon. The code editor displays a Python snippet for mounting a storage container:

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

Creation of DatabricksScope

The screenshot displays the Databricks web interface. The browser's address bar shows the URL: `https://adb-2019937071878492.12.azure.databricks.net/?o=2019937071878492#secrets/createScope`. The left sidebar contains navigation options: Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Data Engineering, Job Runs, AI/ML, Playground, and Experiments. The top toolbar includes buttons for File, Edit, View, Run, Help, Python, Tabs: ON, and a star icon. The main editor area shows a Python script for mounting a storage location and creating a DatabricksScope. The script is as follows:

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

Creation of DatabricksScope

The screenshot shows the Databricks web interface for creating a secret scope. The browser address bar shows the URL: `adb-2019937071878492.12.azuredatabricks.net/?o=2019937071878492#secrets/createScope`. The top navigation bar includes the Microsoft Azure and Databricks logos, a search bar with the text "Search data, notebooks, recents, and more...", and a user profile icon labeled "vehciletheftadb". The left sidebar contains a "New" button and a list of navigation items: Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Data Engineering, Job Runs, AI/ML, Playground, Experiments, Features, Models, and Serving. The main content area is titled "HomePage / Create Secret Scope" and features a "Create Secret Scope" heading with "Cancel" and "Create" buttons. Below the heading is a descriptive text: "A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)". The form includes three input fields: "Scope Name" (empty), "Manage Principal" (a dropdown menu showing "Creator"), and "Azure Key Vault" (a section with two sub-fields: "DNS Name" containing "https://xxx.vault.azure.net/" and "Resource ID" containing "/subscriptions/xxxxxx/...").

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

vehciletheftadb

+ New

Workspace

Recents

Catalog

Jobs & Pipelines

Compute

Data Engineering

Job Runs

AI/ML

Playground

Experiments

Features

Models

Serving

HomePage / Create Secret Scope

Create Secret Scope

Cancel Create

A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)

Scope Name

Manage Principal

Creator

Azure Key Vault

DNS Name

`https://xxx.vault.azure.net/`

Resource ID

`/subscriptions/xxxxxx/...`

Creation of Azure Key Vault

Create a key vault ...

Grant data plane access by using a [Azure RBAC](#) or [Key Vault access policy](#)

- ☐ Azure role-based access control (recommended) ⓘ
- ☒ Vault access policy ⓘ

Resource access

- ☐ Azure Virtual Machines for deployment ⓘ
- ☐ Azure Resource Manager for template deployment ⓘ
- ☐ Azure Disk Encryption for volume encryption ⓘ

Access policies

Access policies enable you to have fine grained control over access to vault items. [Learn more](#)

+ Create ✎ Edit 🗑 Delete

| <input type="checkbox"/> Name ↑↓ | Email ↑↓ | Key Permissions | Secret Permissions | Certificate Permissions |
|---|---------------------------------|--|---|--|
| ▼ USER | | | | |
| <input type="checkbox"/> AAFREEN SANA H | 2024207031@student.annauniv.edu | Get, List, Update, Create, Import, Dele... | Get, List, Set, Delete, Recover, Backup,... | Get, List, Update, Create, Import, Dele... |

Previous

Next

Review + create

[Give feedback](#)

Azure Key Vault

The screenshot displays the Microsoft Azure portal interface. At the top, the navigation bar includes the Microsoft Azure logo, a search bar, and a Copilot button. The user's profile, '2024207031@student.a... ANNA UNIVERSITY (ANNAUNIV...', is visible in the top right corner.

The main content area shows the 'vehicletheftproject-kv' Key vault properties. The left sidebar contains a navigation menu with options like 'Diagnose and solve problems', 'Access policies', 'Resource visualizer', 'Events', 'Objects', 'Settings', 'Access configuration', 'Networking', and 'Microsoft Defender for Cloud'.

The properties table lists the following details:

| | |
|--------------------|---|
| Sku (Pricing tier) | Standard |
| Location | uaenorth |
| Vault URI | https://vehicletheftproject-kv.vault.azure.net/ |
| Resource ID | /subscriptions/5d998e73-b6f4-454e-b869-4a7c060e09c0/resourceGroups/vehicletheftprojectrg/providers/Microsoft... |
| Subscription ID | 5d998e73-b6f4-454e-b869-4a7c060e09c0 |
| Subscription Name | Azure for Students |
| Directory ID | 6e804f24-0209-4dcd-ac89-97525eddbd30 |
| Directory Name | Anna University |

Secret Scope using Azure Key Vault

The screenshot shows the Databricks web interface with the 'Create Secret Scope' dialog open. The left sidebar contains navigation links: New, Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Data Engineering, Job Runs, AI/ML, Playground, Experiments, Features, Models, and Serving. The top header includes the Microsoft Azure and Databricks logos, a search bar, and the user profile 'vehciletheftadb'. The dialog title is 'Create Secret Scope' with a 'Cancel' button and a 'Verifying...' status indicator. Below the title, a description states: 'A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)'. The form fields are: 'Scope Name' with the value 'dbScope', 'Manage Principal' with a dropdown set to 'All workspace users', 'Azure Key Vault' with a help icon, 'DNS Name' with the value 'https://vehciletheftproject-kv.vault.azure.net/', and 'Resource ID' with the value '/subscriptions/5d998e73-b6f4-454e-b869-4a7c060e09c0/resourceGroups/vehcileth'.

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

vehciletheftadb

+ New

Workspace

Recents

Catalog

Jobs & Pipelines

Compute

Data Engineering

Job Runs

AI/ML

Playground

Experiments

Features

Models

Serving

HomePage / Create Secret Scope

Create Secret Scope | Cancel Verifying...

A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)

Scope Name ?

dbScope

Manage Principal ?

All workspace users

Azure Key Vault ?

DNS Name

https://vehciletheftproject-kv.vault.azure.net/

Resource ID

/subscriptions/5d998e73-b6f4-454e-b869-4a7c060e09c0/resourceGroups/vehcileth

Access Key from Storage Account

The screenshot displays the Microsoft Azure portal interface. At the top, the header includes the Microsoft Azure logo, a search bar, and the user profile '2024207031@student.a... ANNA UNIVERSITY (ANNAUNIV...)'.

The breadcrumb navigation path is: Home > Storage center | Storage accounts (Blobs) > vehicletheftprojectsa.

The main heading is 'vehicletheftprojectsa | Access keys', with a star icon and a menu icon to its right.

On the left sidebar, the 'Access keys' option is highlighted. Other options include 'Diagnose and solve problems', 'Access Control (IAM)', 'Data migration', 'Events', 'Storage browser', 'Partner solutions', 'Resource visualizer', 'Data storage', 'Security + networking', 'Networking', and 'Access keys'.

The main content area contains the following elements:

- A search bar.
- Actions: 'Set rotation reminder', 'Refresh', and 'Give feedback'.
- Text: 'Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.'
- Text: 'Remember to update the keys with any Azure resources and apps that use this storage account. [Learn more about managing storage account access keys](#)'.
- Storage account name: 'vehicletheftprojectsa'.
- Key management: 'key1' with a 'Rotate key' button.
- Last rotated: '21/09/2025 (20 days ago)'.
- Key: '1LYnXvPqhyMLhgCa8sF/+oTx6C+2oC9rjIIL6G+EZCoqV4Qbnh6UiB/wf/uuJB3F3q...'. A 'Copy to clipboard' button is visible above the key text, and a 'Hide' button is to its right.
- Connection string: 'Connection string'.

Secret Creation

Home > vehicletheftproject-kv | Secrets >



Create a secret

...



Upload options

Manual



Name * ⓘ

saSecret



Secret value * ⓘ

.....



Content type (optional)

Set activation date ⓘ

☐

Set expiration date ⓘ

☐

Enabled

Yes

No

Tags

0 tags

Mounting of the Medallion Containers

▶ ▾ ✓ Just now (25s)

2

Python

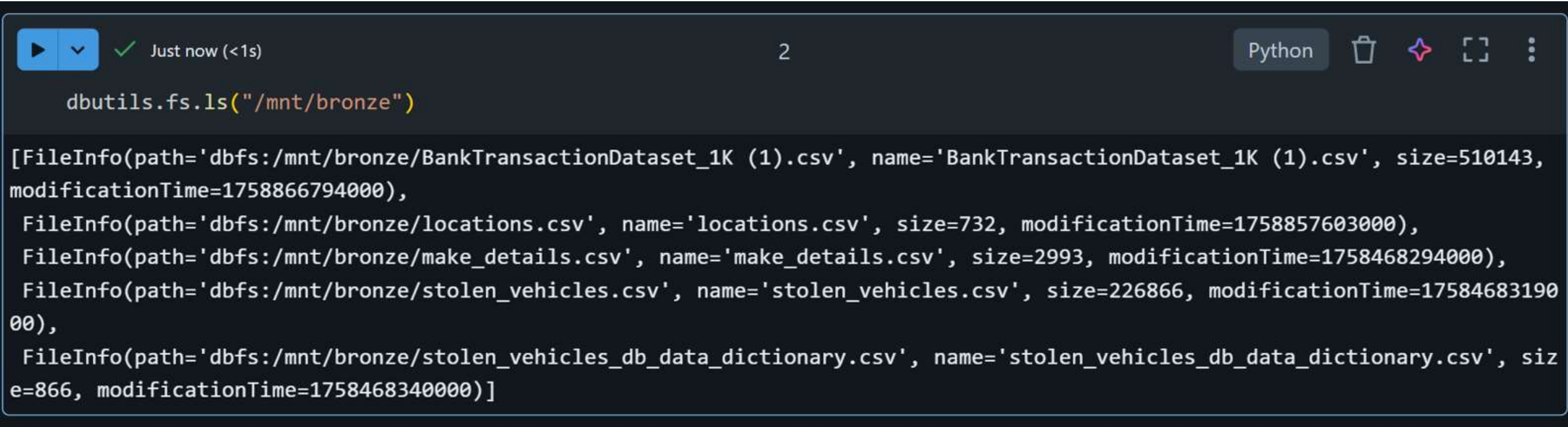


```
dbutils.fs.mount(  
    source = "wasbs://silver@vehicletheftprojectsa.blob.core.windows.net/",  
    mount_point="/mnt/silver",  
    extra_configs={  
        "fs.azure.account.key.vehicletheftprojectsa.blob.core.windows.net": dbutils.secrets.get('dbScope', 'saSecret')  
    }  
)
```

```
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



A terminal window with a dark background. The top bar shows a play button, a dropdown arrow, a green checkmark, the text "Just now (<1s)", the number "2", a "Python" label, and icons for trash, star, expand, and menu. The command `dbutils.fs.ls("/mnt/bronze")` is entered. The output is a list of file information objects.

```
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=1758468319000),  
FileInfo(path='dbfs:/mnt/bronze/stolen_vehicles_db_data_dictionary.csv', name='stolen_vehicles_db_data_dictionary.csv', size=866, modificationTime=1758468340000)]
```

Loading Files from Cloud into SPARK

```
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



✓ Just now (1s)

7

Python



```
location_df.show()
```

► (1) Spark Jobs

| +-----+-----+-----+-----+-----+ | | | | |
|---------------------------------|--------------------|-------------|------------|---------|
| location_id | region | country | population | density |
| +-----+-----+-----+-----+-----+ | | | | |
| 101 | Northland | New Zealand | 201,500 | 16.11 |
| 102 | Auckland | New Zealand | 1,695,200 | 343.09 |
| 103 | Waikato | New Zealand | 513,800 | 21.5 |
| 104 | Bay of Plenty | New Zealand | 347,700 | 28.8 |
| 105 | Gisborne | New Zealand | 52,100 | 6.21 |
| 106 | Hawke's Bay | New Zealand | 182,700 | 12.92 |
| 107 | Taranaki | New Zealand | 127,300 | 17.55 |
| 108 | Manawatū-Whanganui | New Zealand | 258,200 | 11.62 |
| 109 | Wellington | New Zealand | 543,500 | 67.52 |
| 110 | Tasman | New Zealand | 58,700 | 6.1 |
| 111 | Nelson | New Zealand | 54,500 | 129.15 |

Cloud Data Transformation in ADF



✓ Just now (<1s)

8

Python



```
location_df.printSchema()
```

```
root
```

```
|-- location_id: integer (nullable = true)
|-- region: string (nullable = true)
|-- country: string (nullable = true)
|-- population: string (nullable = true)
|-- density: double (nullable = true)
```


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



✓ Just now (<1s)

11

Python



```
location_df.printSchema()
```

root

```
|-- location_id: integer (nullable = true)
|-- region: string (nullable = true)
|-- country: string (nullable = true)
|-- Population: integer (nullable = true)
|-- density: double (nullable = true)
```

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

silver

Container

Search

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

+ 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 2 items

| <input type="checkbox"/> | Name | Last modified | Access tier | Blob type | Size | Lease state |
|--------------------------|-------------------------------------|----------------------|-------------|-----------|------|-------------|
| <input type="checkbox"/> | 📁 _azuretmpfolder\$ | 11/10/2025, 22:56:21 | | | | ... |
| <input type="checkbox"/> | 📁 location.csv | 11/10/2025, 22:56:22 | | | | ... |

Data Cleanup and Re-Ingestion Preparation

The screenshot displays the Microsoft Azure portal interface. At the top, the header bar includes the Microsoft Azure logo, a search bar, and the user profile '2024207031@student.a... ANNA UNIVERSITY'. The breadcrumb navigation shows 'Home > vehicletheftprojectsa | Containers >'. The left sidebar contains a search bar and a list of options: 'Overview' (selected), 'Diagnose and solve problems', 'Access Control (IAM)', and 'Settings'.

The main content area shows a container named 'silver'. The authentication method is 'Access key (Switch to Microsoft Entra user account)'. Below this is a search bar for blobs and a dropdown menu set to 'Only show active objects'. The container shows 'Showing all 2 items (2 selected)'. The items are:

| ✓ | Name |
|---|-------------------|
| ✓ | _azuretmpfolder\$ |
| ✓ | location.csv |

A 'Delete confirmation' dialog box is open in the center. It contains the following text:

Delete confirmation

This action will move 2 items to a soft-deleted state.

These items will remain recoverable for the retention period of 0 days.

☒ Delete selected blobs, directories and all the contents, including nested directories.

At the bottom of the dialog are two buttons: 'Delete' and 'Cancel'.

At the bottom left of the screen, there is a small note: 'Add or remove favorites by pressing Ctrl+Shift+F'.

Data Cleanup and Re-Ingestion Preparation

The screenshot displays the Microsoft Azure portal interface. At the top, the header shows the Microsoft Azure logo, a search bar, and the user profile for '2024207031@student.a... ANNA UNIVERSITY'. The breadcrumb navigation indicates the current location: 'Home > vehicletheftprojectsa | Containers > silver'. The 'silver' container is selected, showing its 'Overview' tab. The container's authentication method is 'Access key', and it is currently empty, displaying 'Showing all 0 items'. Below the container details, a recent script execution is shown, titled 'Just now (6s)' with a status of '14'. The script is written in Python and contains four lines of code that write CSV files to the container's root directory. The script is labeled as '(4) Spark Jobs'.

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

2024207031@student.a... ANNA UNIVERSITY

Home > vehicletheftprojectsa | Containers >

silver Container

Search

+ Add Directory ↑ Upload ↻ Refresh 🗑 Delete 📄 Copy 📄 Paste 🔄 Rename 🔄 Acquire lease 🔄 Break lease 📄 Edit columns

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

silver

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

Search blobs by prefix (case-sensitive)

Only show active objects

Showing all 0 items

| <input type="checkbox"/> | Name | Last modified | Access tier | Blob type | Size | Lease state |
|--------------------------|------|---------------|-------------|-----------|------|-------------|
|--------------------------|------|---------------|-------------|-----------|------|-------------|

No items found

Just now (6s) 14 Python

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

(4) Spark Jobs

Data Cleanup and Re-Ingestion Preparation

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

2024207031@student.a... ANNA UNIVERSITY

Home > vehicletheftprojectsa | Containers >

silver
Container

Search

+ 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 5 items

| <input type="checkbox"/> | Name | Last modified | Access tier | Blob type | Size | Lease state |
|--------------------------|---------------------|------------------------|-------------|-----------|------|-------------|
| <input type="checkbox"/> | _\$azuretmpfolder\$ | 10/26/2025, 4:58:12 PM | | | | ... |
| <input type="checkbox"/> | database.csv | 10/26/2025, 4:58:17 PM | | | | ... |
| <input type="checkbox"/> | location.csv | 10/26/2025, 4:58:13 PM | | | | ... |
| <input type="checkbox"/> | make_details.csv | 10/26/2025, 4:58:15 PM | | | | ... |
| <input type="checkbox"/> | stolen_vehicles.csv | 10/26/2025, 4:58:16 PM | | | | ... |

Data Validation and Quality Check in Databricks



The image shows a Databricks notebook interface. At the top, there's a status bar with a play button, a checkmark, the text "Just now (2s)", the number "16", and a "Python" language selector. Below this, the code cell contains two lines of Python code: `null_count_location = location_df.select([sum(when(col(column).isNull(),1).otherwise (0)).alias(column) for column in location_df.columns])` and `null_count_location.show()`. The output section shows "(2) Spark Jobs" and a table header for `null_count_location`. The table has five columns: `location_id`, `region`, `country`, `population`, and `density`. The first row of data shows all five columns with the value `0`.

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

null_count_location.show()
```

▶ (2) Spark Jobs

▶ null_count_location: pyspark.sql.dataframe.DataFrame = [location_id: long, region: long ... 3 more fields]

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

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

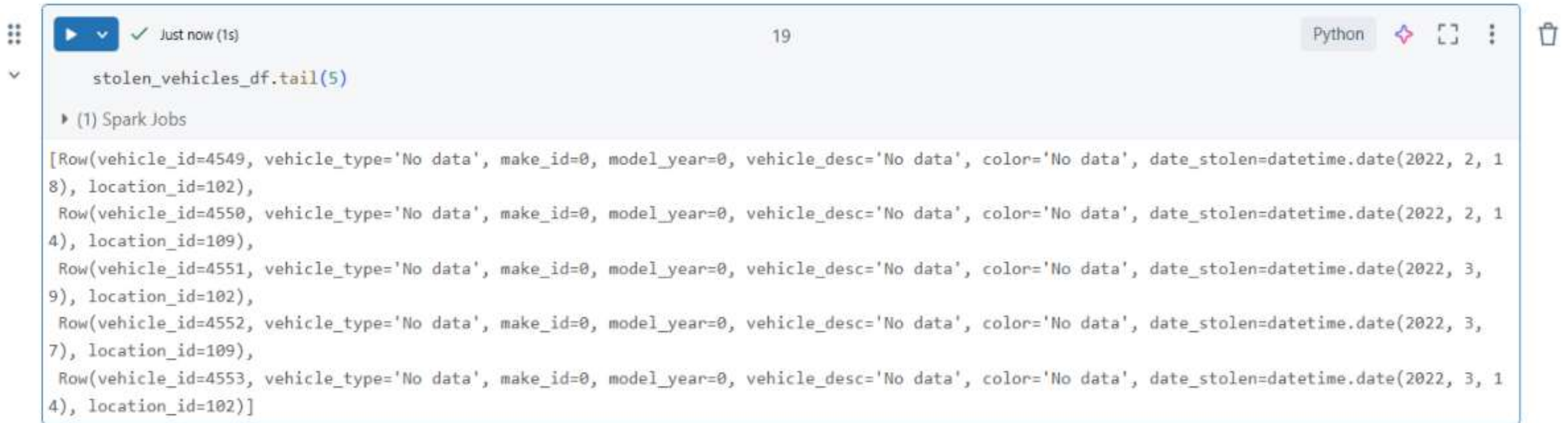
```
stolen_vehicles_df.printSchema()
```

```
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)
```

```
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



The screenshot shows a Jupyter Notebook interface. At the top, there is a toolbar with a play button, a checkmark, and the text 'Just now (1s)'. To the right of the toolbar, the number '19' is displayed. Further right, there is a 'Python' language selector and several icons for notebook actions. Below the toolbar, the code cell contains the following Python code:

```
stolen_vehicles_df.tail(5)
```

Below the code cell, there is a section labeled '(1) Spark Jobs'. The output of the code is displayed as a list of five rows, each representing a vehicle record. The output is as follows:

```
[Row(vehicle_id=4549, vehicle_type='No data', make_id=0, model_year=0, vehicle_desc='No data', color='No data', date_stolen=datetime.date(2022, 2, 18), location_id=102),  
 Row(vehicle_id=4550, vehicle_type='No data', make_id=0, model_year=0, vehicle_desc='No data', color='No data', date_stolen=datetime.date(2022, 2, 14), location_id=109),  
 Row(vehicle_id=4551, vehicle_type='No data', make_id=0, model_year=0, vehicle_desc='No data', color='No data', date_stolen=datetime.date(2022, 3, 9), location_id=102),  
 Row(vehicle_id=4552, vehicle_type='No data', make_id=0, model_year=0, vehicle_desc='No data', color='No data', date_stolen=datetime.date(2022, 3, 7), location_id=109),  
 Row(vehicle_id=4553, vehicle_type='No data', make_id=0, model_year=0, vehicle_desc='No data', color='No data', date_stolen=datetime.date(2022, 3, 14), location_id=102)]
```

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 Layer

```
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

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

2024207031@student.a... ANNA UNIVERSITY

Home > vehicletheftprojectsa | Containers >

gold Container

Search

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

gold

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

Search blobs by prefix (case-sensitive)

Only show active objects

Showing all 5 items

| <input type="checkbox"/> | Name | Last modified | Access tier | Blob type | Size | Lease state |
|--------------------------|---------------------|------------------------|-------------|-----------|------|-------------|
| <input type="checkbox"/> | _\$azuretmpfolder\$ | 10/26/2025, 5:24:41 PM | | | | ... |
| <input type="checkbox"/> | database.csv | 10/26/2025, 5:24:46 PM | | | | ... |
| <input type="checkbox"/> | location.csv | 10/26/2025, 5:24:42 PM | | | | ... |
| <input type="checkbox"/> | make_details.csv | 10/26/2025, 5:24:43 PM | | | | ... |
| <input type="checkbox"/> | stolen_vehicles.csv | 10/26/2025, 5:24:45 PM | | | | ... |

Data Querying in Databricks (Gold Layer Analysis)

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

```
%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]`

| | 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

Untitled - Power BI Desktop

Search

Join us at FabCon Atlanta from March 16-20, 2026, for the ultimate Power BI, Fabric, AI, and SQL community-led event. Save \$200 with code FABNOTEPIBIL.

Home

Open

Select a data source or start with a blank report

Blank report

OneLake catalog

Excel workbook

SQL Server

Learn with sample data

Get data from other sources

Recommended

Getting started

Intro—What is Power BI? [?]

Recent

Shared with me

Sign in

Options and settings

About

Get Data

Search

All

File

Database

Microsoft Fabric

Power Platform

Azure

Online Services

Other

Azure

Azure SQL database

Azure Synapse Analytics SQL

Azure Analysis Services database

Azure Database for PostgreSQL

Azure Blob Storage

Azure Table Storage

Azure Cosmos DB v1

Azure Data Explorer (Kusto)

Azure Data Lake Storage Gen2

Azure HDInsight (HDFS)

Azure HDInsight Spark

HDInsight Interactive Query

Azure Cost Management

Azure Resource Graph

Azure Cosmos DB v2

Azure Databricks

Certified Connectors

Template Apps

Connect

Cancel

Linking Power BI to Azure Storage Account

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and a Copilot button. The breadcrumb trail indicates the path: Home > Storage center | Storage accounts (Blobs) > vehicletheftprojectsa. The main content area displays the 'Containers' section for the 'vehicletheftprojectsa' storage account. A search bar is present, and a toolbar offers actions like 'Add container', 'Upload', 'Refresh', 'Delete', 'Change access level', and 'Edit columns'. A table lists four containers: 'Name', '\$logs', 'bronze', 'gold' (selected), and 'silver'. The 'gold' container is highlighted, and its properties are shown in a sidebar on the right.

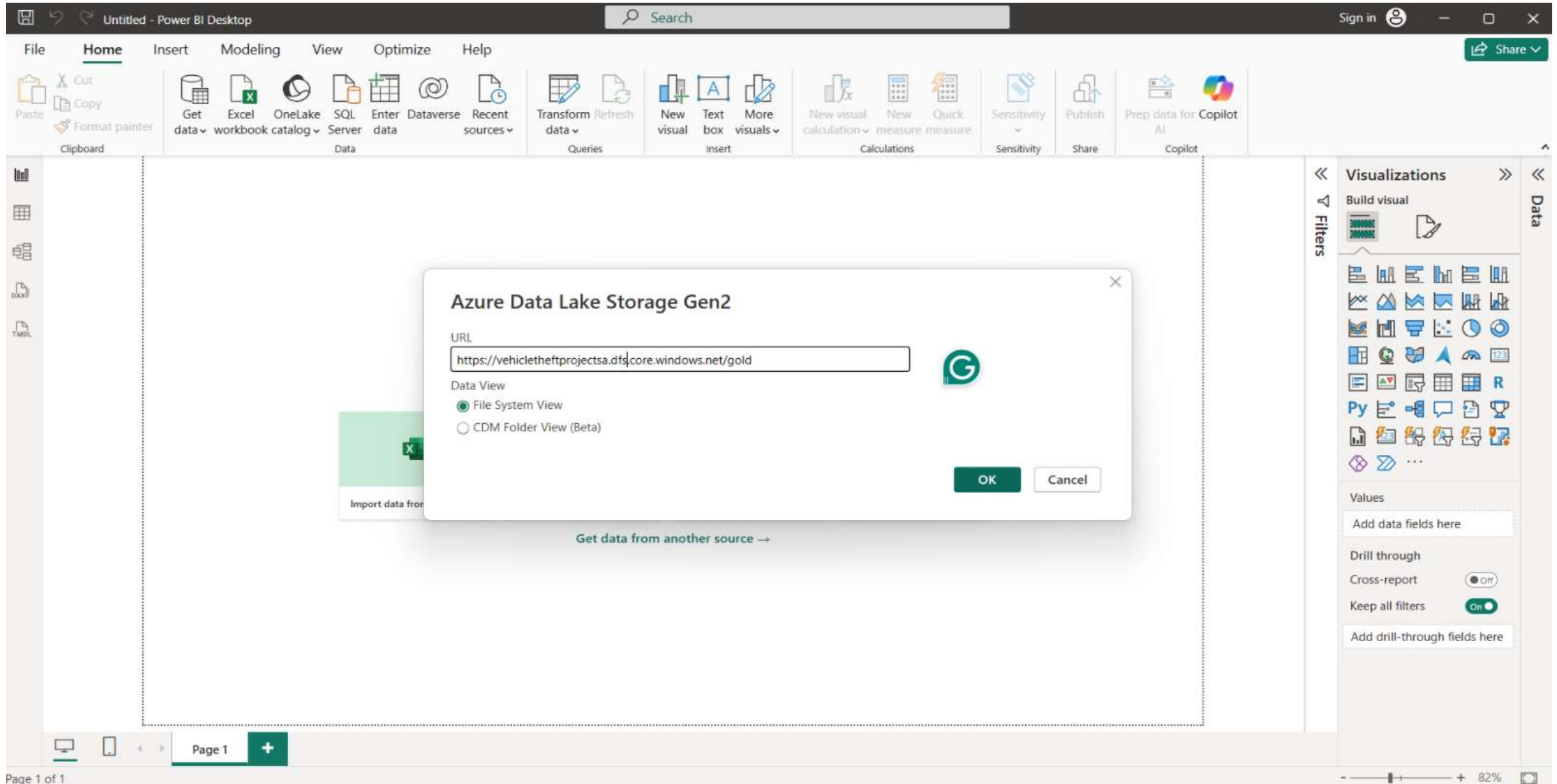
Container properties

gold



Refresh Give feedback

| NAME | URL | LAST MODIFIED | LEASE STATUS | LEASE STATE | LEASE DURATION | ENCRYPTION SCOPE |
|------|---|-----------------------|--------------|-------------|----------------|------------------|
| gold | https://vehicletheftprojectsa.blob.c... | 21/9/2025, 7:12:46 pm | Unlocked | Available | - | - |

Linking Power BI to Azure Storage Account





Data Import and Preview in Power BI

<https://vehicletheftprojectsa.dfs.core.windows.net/gold>

| 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 |

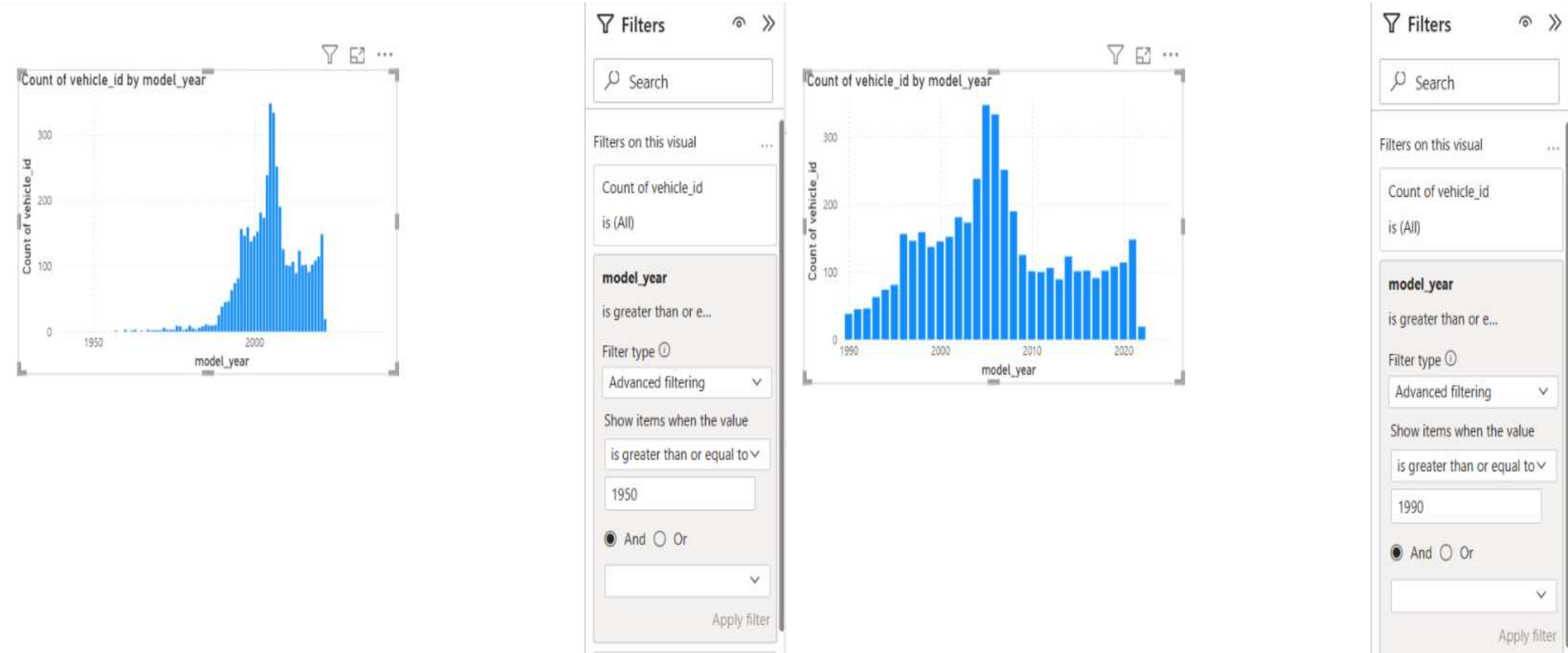
Combine

Load

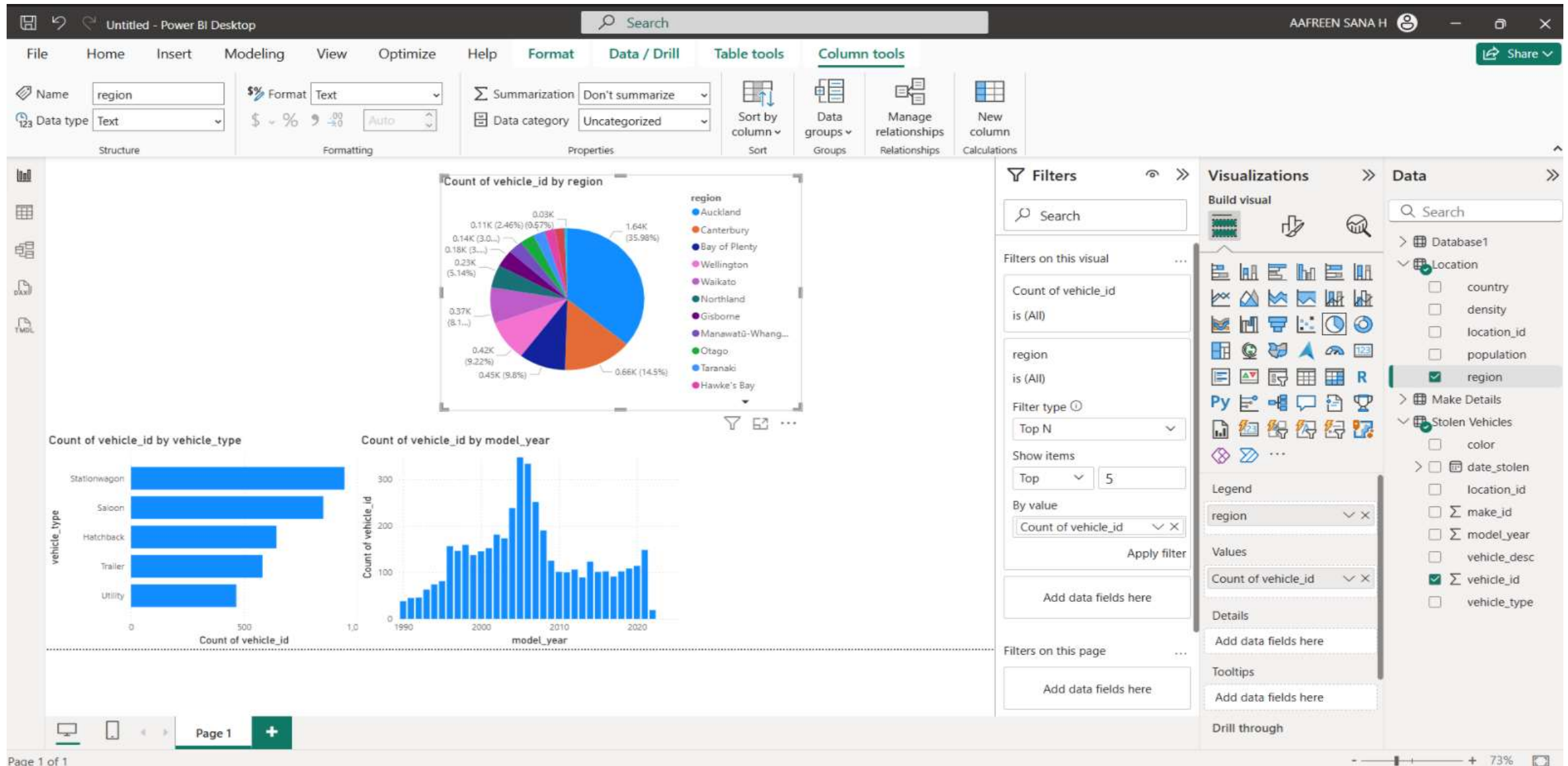
Transform Data

Cancel

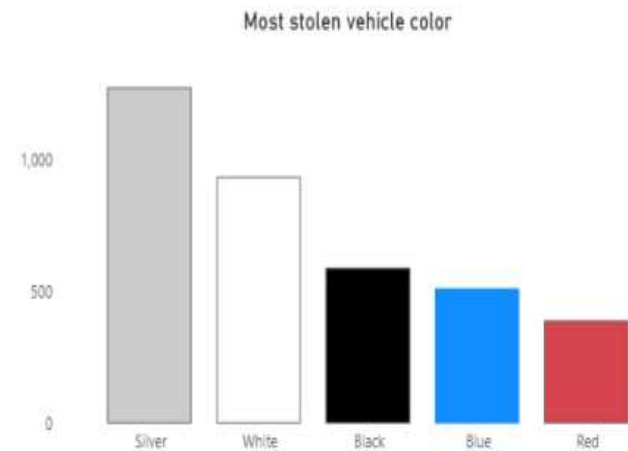
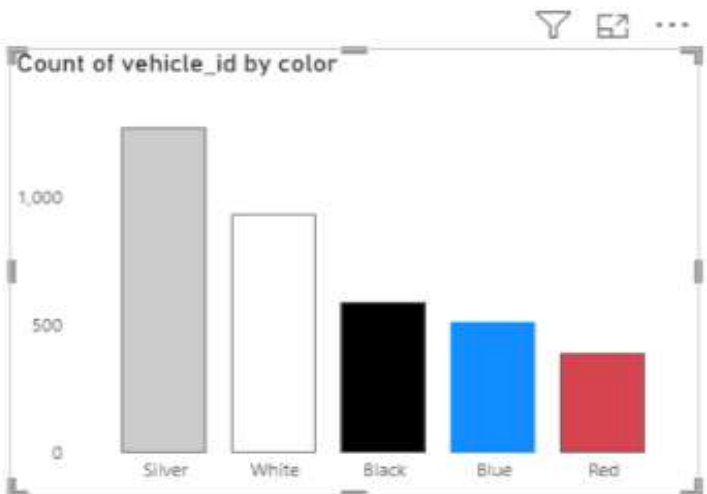
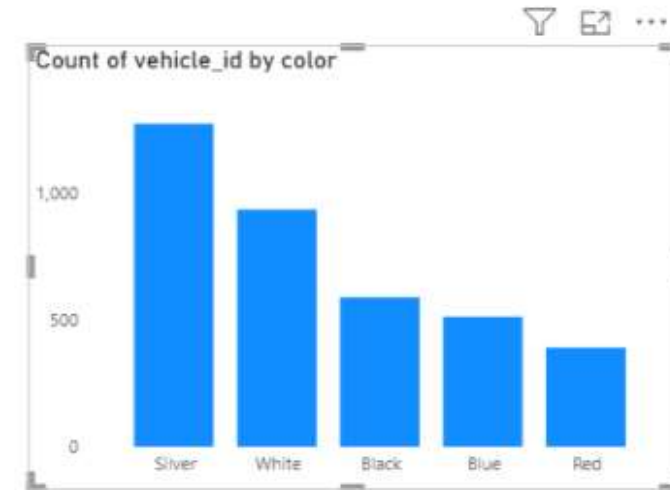
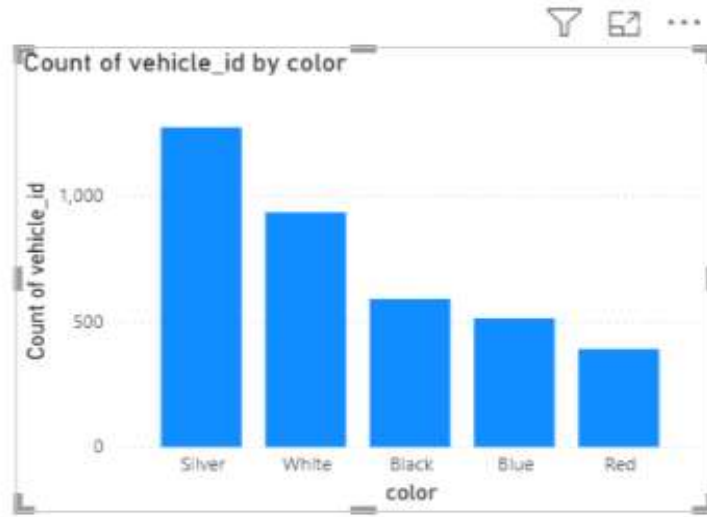
Model Year Visualization – Column Charts



Region-Wise Visualization – Pie Charts



Color-Based Visualization – Bar Charts



Dashboard

