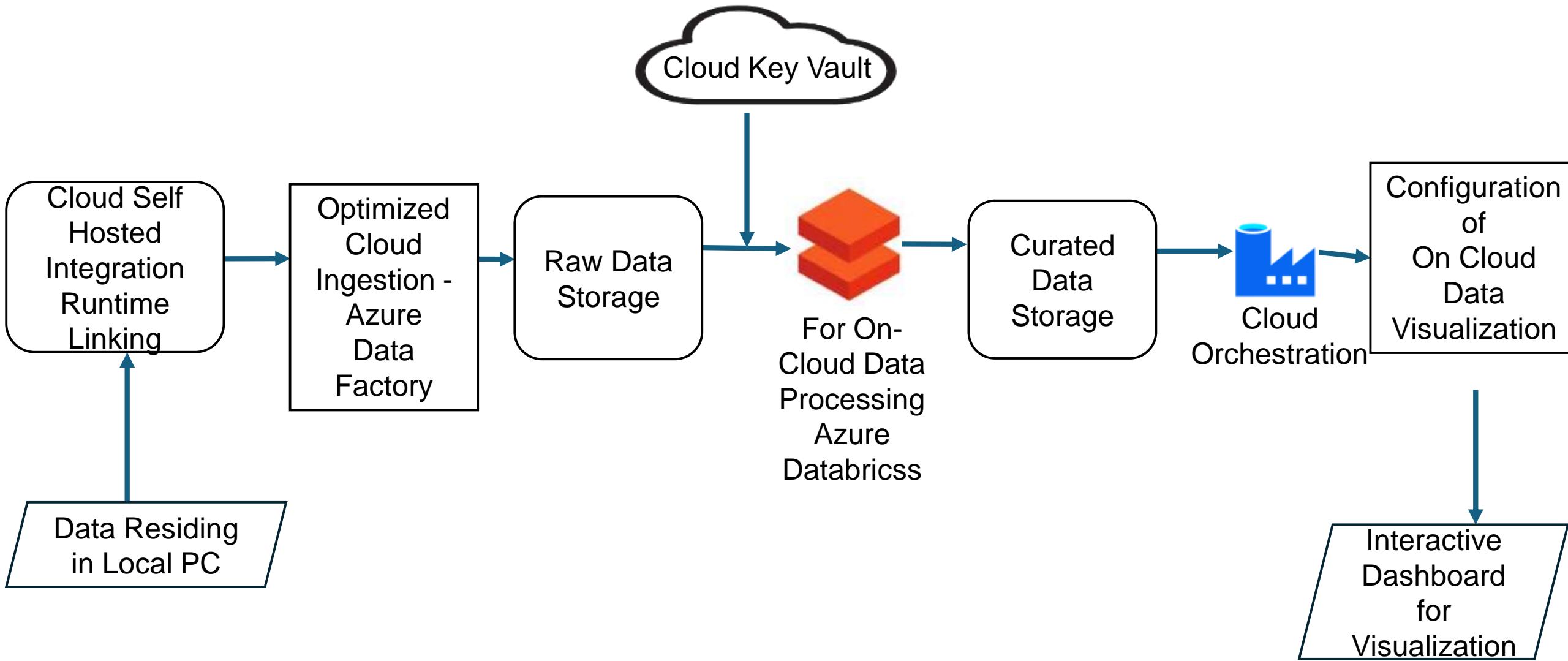


# Azure Data Engineering Pipeline with Medallion Architecture

# Architecture



# Deployment of Resource Group

Home >

## vehicletheftproject\_1758439475069 | Overview

Deployment

Search Delete Cancel Redeploy Download Refresh

Overview Inputs Outputs Template

Your deployment is complete

Deployment name: vehicletheftproject\_1758439475069 Start time: 21/09/2025, 12:54:45  
Subscription: Azure for Students Correlation ID: 3656615e-74ec-4650  
Resource group: vehiclethefttrg

Deployment details

Next steps

Go to resource

# Deployment of Azure Data Factory

Home >



## Microsoft.DataFactory-20250921130324 | Overview

Deployment

Search

X <<



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   ... 

What are the best practices for managing ·

Resource group

»  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 Add container Upload Refresh Delete Change access level

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Partner solutions

Resource visualizer

...

Showing all 4 items

	Name	Last modified	Anonymous ac
<input type="checkbox"/>	\$logs	21/09/2025, 12:55:16	Private
<input type="checkbox"/>	bronze	21/09/2025, 12:59:35	Private
<input type="checkbox"/>	gold	21/09/2025, 12:59:57	Private
<input type="checkbox"/>	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 
 <a href="#">VehicleTheftFileServer1</a>	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

trash can icon /> icon square icon right arrow icon

A screenshot of the Azure Data Factory pipeline editor interface. At the top, it shows the pipeline name "vehicletheftpipeline1". Below that is a toolbar with "Activities", a search bar containing "copy", and sections for "Move and transform" and "Copy data". The main area shows a "Copy data" activity named "Copy data1". The activity has a blue header with the title "Copy data" and a white body containing the name "Copy data1" next to a blue cylinder icon. At the bottom of the activity card are four icons: a trash can, a slash and greater-than symbol, a square, and a right-pointing arrow. Above the activity card, there are validation status indicators: a green checkmark for "Validate" and "Validate copy runtime", a red circle with a minus sign for "Debug", and a lightning bolt icon for "Add trigger".

# Successful Data Ingestion

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

Pr



vehicletheftpipeline1 ●

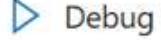


Activities



Validate

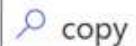
Validate copy runtime



Debug



Add trigger



copy

Validate the current resource

Move and transform



Copy data

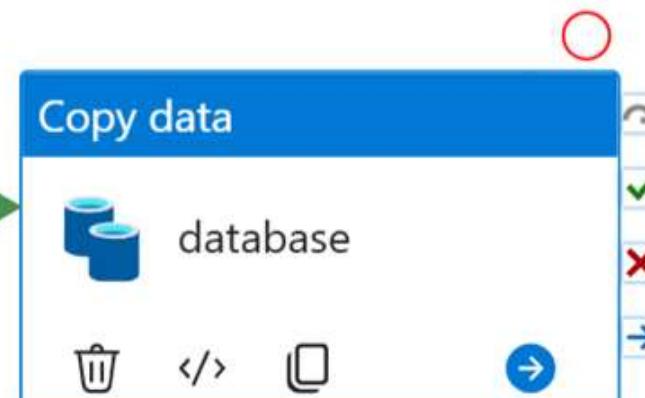


location

Copy data



make details





+ Add Directory

Upload

Refresh

Delete

Copy

Paste

Rename

🔗 Acquire lease

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"/>	<a href="#">BankTransactionDataset_1K (1).csv</a>	26/09/2025, 09:03:23	Hot (Inferred)
<input type="checkbox"/>	<a href="#">locations.csv</a>	26/09/2025, 09:03:23	Hot (Inferred)
<input type="checkbox"/>	<a href="#">make_details.csv</a>	21/09/2025, 20:54:54	Hot (Inferred)
<input type="checkbox"/>	<a href="#">stolen_vehicles.csv</a>	21/09/2025, 20:55:19	Hot (Inferred)
<input type="checkbox"/>	<a href="#">stolen_vehicles_db_data_dictionary.csv</a>	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 shows the Databricks workspace interface. At the top left, there's a Microsoft Azure logo and a Databricks logo. On the far right, there's a dropdown menu for 'vehciletheftadb' and a settings icon. The main header says 'Welcome to Databricks'. Below it is a search bar with the placeholder 'Search data, notebooks, recents, and more...' and a 'CTRL + P' keyboard shortcut. A large callout box in the center says 'Set up your workspace' with a 'Get started' button. Below this, a section titled 'Start your journey' encourages users to try the 'New' menu. The left sidebar has several sections: 'New' (highlighted), 'Workspace', 'Recents', 'Catalog', 'Jobs & Pipelines' (with a blue asterisk), 'Compute', 'Data Engineering', 'Job Runs', 'AI/ML', 'Playground', 'Experiments', 'Features', 'Models', and 'Serving'. The 'Recents' tab in the bottom navigation bar is highlighted with a blue border.

Microsoft Azure databricks

vehciletheftadb

+ New

Workspace

Recents

Catalog

Jobs & Pipelines \*

Compute

Data Engineering

Job Runs

AI/ML

Playground

Experiments

Features

Models

Serving

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

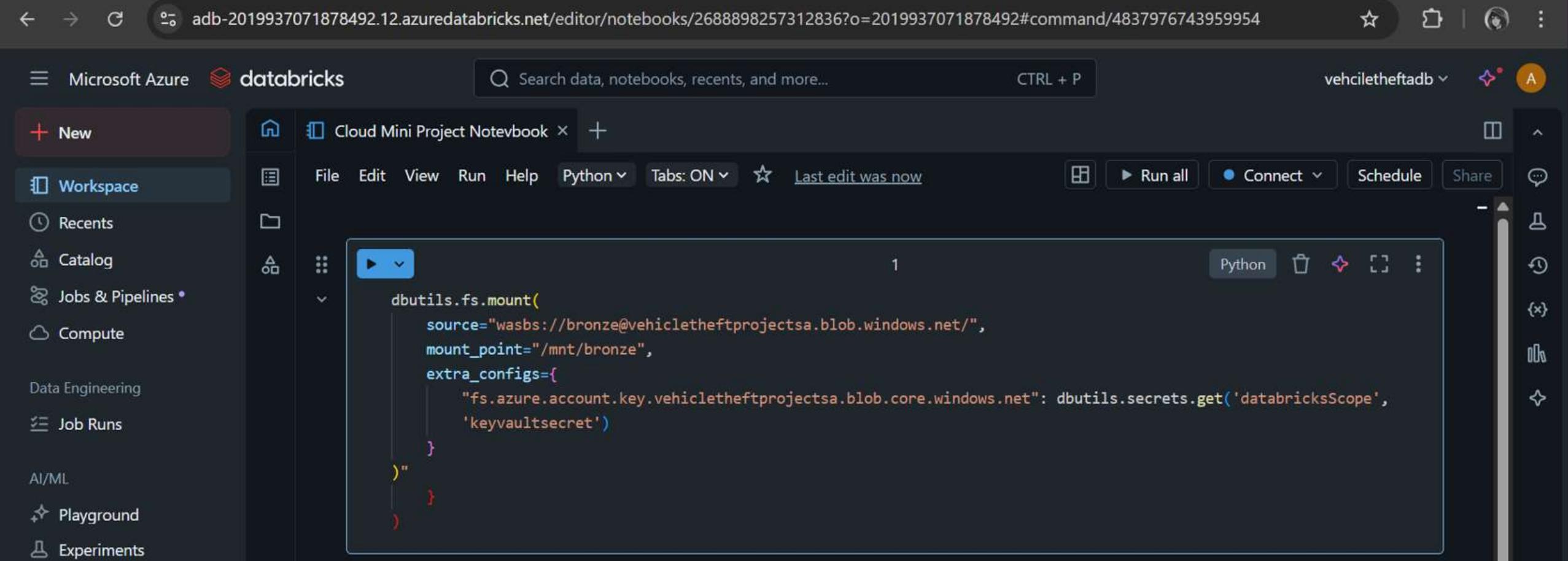
Recent Favorites Popular Mosaic AI What's new

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 shows a Microsoft Azure Databricks workspace. The left sidebar contains navigation links for Microsoft Azure, Databricks, New, Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Data Engineering, Job Runs, AI/ML, Playground, and Experiments. The main area displays a notebook titled "Cloud Mini Project Notebook". The notebook interface includes a toolbar with File, Edit, View, Run, Help, Python dropdown, Tabs: ON, Last edit was now, and buttons for Run all, Connect, Schedule, and Share. A search bar at the top says "Search data, notebooks, recents, and more...". The code cell contains the following Python code:

```
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 a Databricks workspace interface. On the left, there's a sidebar with various navigation options like Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Data Engineering, Job Runs, AI/ML, Playground, and Experiments. The main area is a Python notebook cell containing the following code:

```
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')  
    }  
)  
)
```

The notebook cell has a play button icon and a number '1' indicating it's the first cell. The Python language icon is visible at the top right of the cell. The browser tabs at the top show the URL of the current notebook.

# Creation of DatabricksScope

The screenshot shows the Databricks web interface with a dark theme. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, a search bar with placeholder text 'Search data, notebooks, recents, and more...', and a dropdown menu for the workspace name 'vehciletheftadb'. On the far right, there are icons for star, copy, refresh, and settings.

The left sidebar contains a 'New' button and several navigation items: Workspace, Recents, Catalog, Jobs & Pipelines (with a red asterisk), Compute, Data Engineering, Job Runs, AI/ML, Playground, Experiments, Features, Models, and Serving.

The main content area is titled 'Create Secret Scope' with 'Cancel' and 'Create' buttons. It provides a brief description: 'A store for secrets that is identified by a name and backed by a specific store type'. A 'Learn more' link is available. The form fields include:

- Scope Name**: An input field currently empty.
- Manage Principal**: A dropdown menu set to 'Creator'.
- Azure Key Vault**: Fields for 'DNS Name' containing 'https://xxx.vault.azure.net/' and 'Resource ID' containing '/subscriptions/xxxxxx/...'.

# Creation of Azure Key Vault

## Create a key vault ...

X

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](#)

Name ↑	Email ↑	Key Permissions	Secret Permissions	Certificate Permissions
USER	AAFREEN SANA H	Get, List, Update, Create, Import, Delete...	Get, List, Set, Delete, Recover, Backup,...	Get, List, Update, Create, Import, Delete...

Previous

Next

Review + create

 Give feedback

# Azure Key Vault

Microsoft Azure Search resources, services, and docs (G/) Copilot 1 ? User icon 2024207031@student.annauniv.in ANNA UNIVERSITY (ANNAUNIV.EDU)

Home > vehicletheftproject-kv

## vehicletheftproject-kv | Properties

Key vault

Search Save Discard changes Refresh

Diagnose and solve problems	Sku (Pricing tier)	Standard
Access policies	Location	uaenorth
Resource visualizer	Vault URI	<a href="https://vehicletheftproject-kv.vault.azure.net/">https://vehicletheftproject-kv.vault.azure.net/</a>
Events	Resource ID	/subscriptions/5d998e73-b6f4-454e-b869-4a7c060e09c0/resourceGroups/vehicletheftprojectrg/providers/Microsoft...
Objects	Subscription ID	5d998e73-b6f4-454e-b869-4a7c060e09c0
Settings	Subscription Name	Azure for Students
Access configuration	Directory ID	6e804f24-0209-4dcd-ac89-97525eddbd30
Networking	Directory Name	Anna University

# Secret Scope using Azure Key Vault

The screenshot shows the Databricks interface with a dark theme. At the top, there is a navigation bar with the Microsoft Azure logo, the Databricks logo, a search bar containing "Search data, notebooks, recents, and more...", and a "CTRL + P" keyboard shortcut. To the right of the search bar are icons for a profile, a star, and a gear.

The main content area is titled "Create Secret Scope" and shows a progress bar indicating "Verifying...". Below the title, a description states: "A store for secrets that is identified by a name and backed by a specific store type." A "Learn more" link is provided. The "Scope Name" field contains "dbScope". The "Manage Principal" dropdown is set to "All workspace users".

Below the scope configuration, there is an "Azure Key Vault" section. It includes fields for "DNS Name" (containing "https://vehicletheftproject-kv.vault.azure.net/") and "Resource ID" (containing "/subscriptions/5d998e73-b6f4-454e-b869-4a7c060e09c0/resourceGroups/vehicletheft").

The left sidebar lists various Databricks features: Workspace, Recents, Catalog, Jobs & Pipelines\*, Compute, Data Engineering, Job Runs, AI/ML, Playground, Experiments, Features, Models, and Serving. The "Jobs & Pipelines" item has a red asterisk next to it, indicating it is a required step.

# Access Key from Storage Account

Microsoft Azure Search resources, services, and docs (G/) Copilot Home > Storage center | Storage accounts (Blobs) > vehicletheftprojectsa 2024207031@student.a... ANNA UNIVERSITY (ANNAUNIV....)

## vehicletheftprojectsa | Access keys

Storage account

Search Set rotation reminder Refresh Give feedback

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.

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

key1 Rotate key  
Last rotated: 21/09/2025 (20 days ago)  
Key: 1LYnXvPqhyMLhgCa8sF/+oTx6C+2oC9rjIIIL6G+EZCoqV4Qbnh6UiB/wf/uuJB3F3q... Copy to clipboard Hide

Connection string:

Networking Access keys

# 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

The screenshot shows a Jupyter Notebook cell with the following details:

- Cell number: 2
- Language: Python
- Last run: Just now (<1s)
- Code: `dbutils.fs.ls("/mnt/bronze")`
- Output:

```
[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

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

A screenshot of a Jupyter Notebook cell. The cell has a green checkmark icon and the text "Just now (1s)" indicating it was run recently. The cell number is 7. The language is set to Python. The code executed is `location_df.show()`. The output shows a Spark DataFrame with 11 rows and 5 columns: location\_id, region, country, population, and density. The data represents locations in New Zealand.

```
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



A screenshot of a Jupyter Notebook cell. The cell contains the following Python code:

```
location_df.printSchema()
```

The output of the code is:

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

The cell has a status bar at the top indicating "Just now (<1s)" and a number "8". The language tab shows "Python". There are standard notebook controls like play/pause, stop, and refresh.

# 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

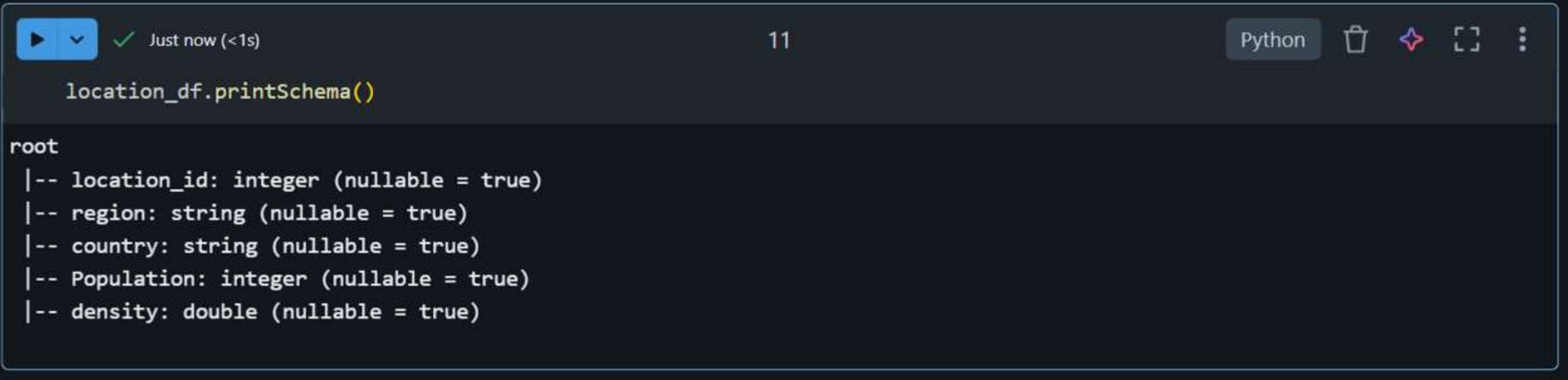


```
location_df.show()
```

▶ (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

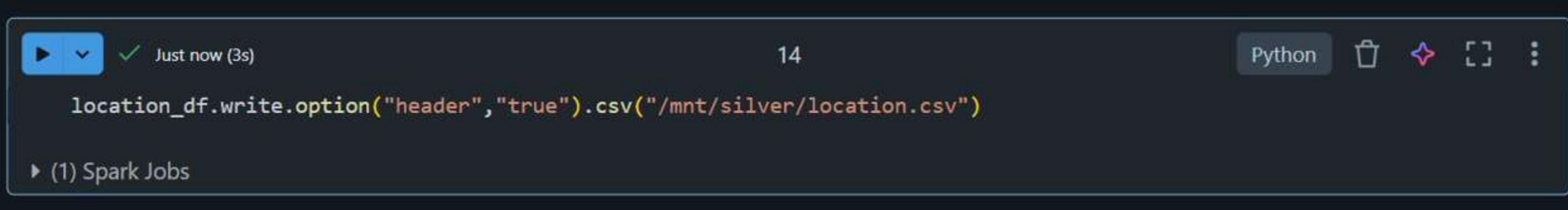
# After Cloud Data Transformation in ADF



```
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)
```

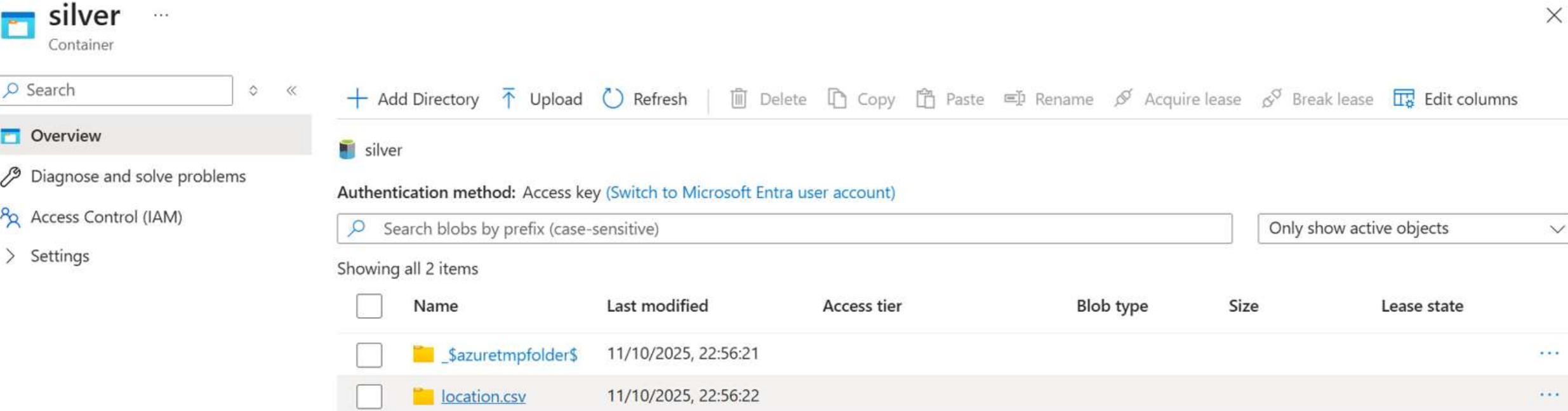
# Migration into Silver Container with Metadata



A screenshot of a Jupyter Notebook cell. The cell has just run a Python command to write a DataFrame to a CSV file in the Silver container. The output shows a green checkmark and the message "Just now (3s)". The cell number is 14. The code is:

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

The status bar at the bottom indicates "(1) Spark Jobs".



A screenshot of the Azure Storage Explorer interface. The left sidebar shows a tree view with a "silver" container selected. The main area displays the contents of the "silver" blob container.

Container details:

- Overview (selected)
- Diagnose and solve problems
- Access Control (IAM)
- Settings

Toolbar:

- Add Directory
- Upload
- Refresh
- Delete
- Copy
- Paste
- Rename
- Acquire lease
- Break lease
- Edit columns

Search and filters:

- Search: silver
- Search blobs by prefix (case-sensitive):
- Only show active objects

Items listed:

	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 shows the Microsoft Azure Storage Explorer interface for a container named "silver". The left sidebar displays navigation options like Home, vehicletheftprojects, Containers, Overview, Diagnose and solve problems, Access Control (IAM), and Settings. The main area shows a list of blobs: ".Sazuretmpfolder\$" and "location.csv". A delete confirmation dialog box is overlaid on the screen, containing 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.

**Delete**   **Cancel**

At the bottom of the page, there is a footer note: "Add or remove favorites by pressing Ctrl+Shift+F".

# Data Cleanup and Re-Ingestion Preparation

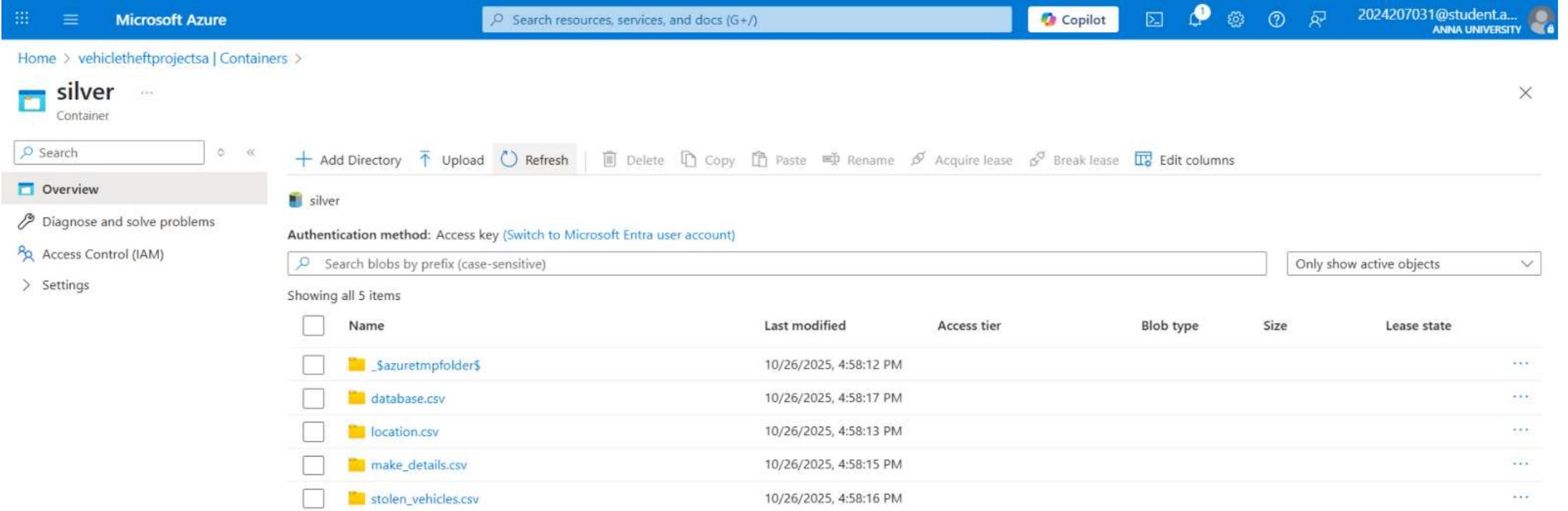
The image shows two screenshots related to data management in Azure.

The top screenshot is a Microsoft Azure Storage Explorer interface. It displays a container named "silver" under the "Containers" section of a project named "vehicletheftprojects". The "Overview" tab is selected. The main pane shows a search bar, a toolbar with actions like Add Directory, Upload, Refresh, Delete, Copy, Paste, Rename, Acquire lease, Break lease, and Edit columns, and a blob named "silver". Below the toolbar, there's a search bar for blobs by prefix and a dropdown for "Only show active objects". A message indicates the authentication method is "Access key" and provides a link to "Switch to Microsoft Entra user account". The table below shows 0 items found. The bottom screenshot is a Jupyter Notebook cell in Python. It shows a code block with four lines of code using the pandas library to write DataFrames to CSV files in the "/mnt/silver" directory. The cell has a status bar indicating it was run "Just now (6s)" and took 14 seconds. The code is:

```
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")
```

Below the code, a collapsible section titled "(4) Spark Jobs" is shown.

# Data Cleanup and Re-Ingestion Preparation



The screenshot shows the Microsoft Azure Storage Explorer interface. The top navigation bar includes the Microsoft Azure logo, a search bar, Copilot, and user information (2024207031@student.annauniv.edu, ANNA UNIVERSITY). Below the navigation bar, the breadcrumb path indicates the current location: Home > vehicletheftprojects | Containers > silver.

The left sidebar for the 'silver' container shows:

- Overview (selected)
- Diagnose and solve problems
- Access Control (IAM)
- Settings

The main content area displays the contents of the 'silver' blob container:

- Authentication method: Access key (Switch to Microsoft Entra user account)
- Search bar: Search blobs by prefix (case-sensitive)
- Filter: Only show active objects
- Table showing 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				...

Add or remove favorites by pressing **Ctrl+Shift+F**.

# Data Validation and Quality Check in Databricks



Just now (2s) 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_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

```
Just now (<1s) 17 Python   
```

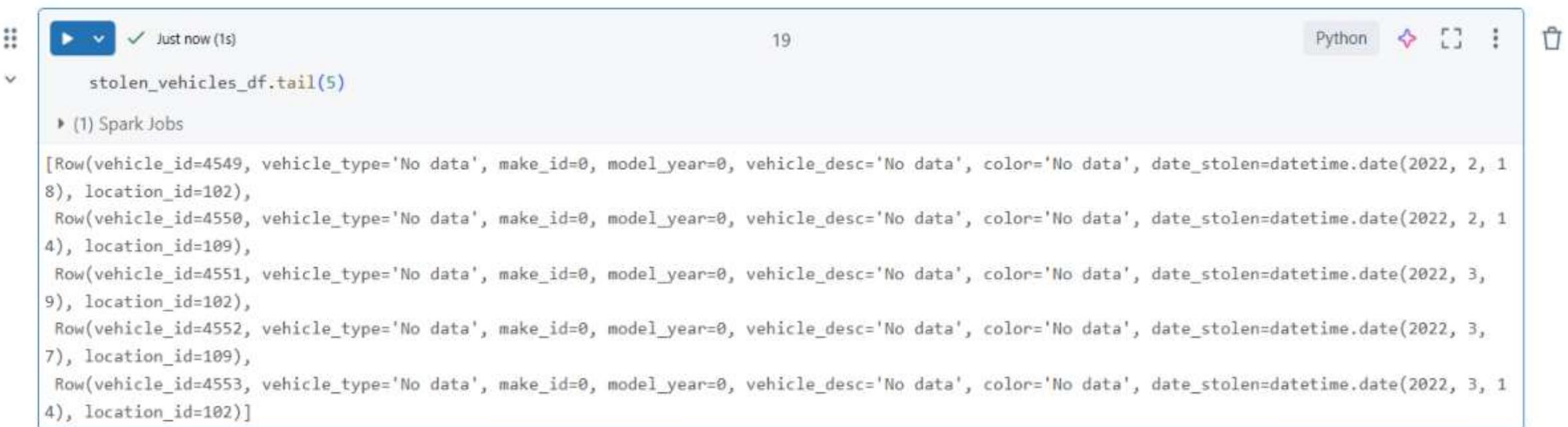
stolen\_vehicles\_df.printSchema()

```
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)
```

```
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



A screenshot of a Jupyter Notebook cell. The cell has a status bar at the top indicating "Just now (1s)" and a code editor below it. The code editor contains the following Python code:

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

The output of the cell shows five rows of data from the `stolen_vehicles_df` DataFrame, each represented as a Row object with various fields set to 'No data' or 0.

```
[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")
```

The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar displays the 'gold' container under 'Containers'. The main area shows five CSV files listed in a table:

	Name	Last modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>	\$_Sazuretmpfolder\$	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				...

The 'Overview' tab is selected in the sidebar. The top navigation bar includes 'Copilot', '1 notifications', and the user '2024207031@student.a... ANNA UNIVERSITY'.

# Data Querying in Databricks (Gold Layer Analysis)

The screenshot shows a Databricks workspace interface with two code cells and a table view.

**Code Cell 23 (Python):**

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

**Code Cell 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
```

**Table View:**

#	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

The screenshot shows the Power BI Desktop interface with the 'Get Data' dialog open. The main window displays a navigation bar with 'Untitled - Power BI Desktop', a search bar, and a promotional message about FabCon Atlanta. Below the navigation bar are sections for 'Select a data source or start with a blank report' and 'Recommended'. The 'Get Data' dialog is centered, showing a search bar and a list of data sources categorized under 'Azure'. The 'Azure Data Lake Storage Gen2' connector is selected and highlighted.

Untitled - Power BI Desktop

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 FABNOTEPEBL.

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

Q File

Get Data

Search

Azure

- 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

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

Copilot    1    ?    🔍    2024207031@student.a... ANNA UNIVERSITY

Home > Storage center | Storage accounts (Blobs) > vehicletheftprojectsa

### vehicletheftprojectsa | Containers

Storage account

Search    Add container    Upload    Refresh    Delete    Change access level    Edit columns

Search containers by prefix

Showing all 4 items

	Name	Last modified	Anon
<input type="checkbox"/>	\$logs	9/21/2025, 7:12:19 PM	Private
<input type="checkbox"/>	bronze	9/21/2025, 7:12:41 PM	Private
<input checked="" type="checkbox"/>	gold	9/21/2025, 7:12:46 PM	Private
<input type="checkbox"/>	silver	9/21/2025, 7:12:51 PM	Private

**Container properties**

gold

Refresh    Give feedback

**NAME**  
gold

**URL**  
[Copy Copied](https://vehicletheftprojectsa.blob.c...)

**LAST MODIFIED**  
21/9/2025, 7:12:46 pm

**ETAG**  
0x8DDF914BF873E81

**LEASE STATUS**  
Unlocked

**LEASE STATE**  
Available

**LEASE DURATION**  
-

**ENCRYPTION SCOPE**

Containers    File shares    Queues    Tables

Security + networking

Add or remove favorites by pressing **Ctrl+Shift+F**

# Linking Power BI to Azure Storage Account

The screenshot shows the Power BI Desktop interface with the 'Home' tab selected in the ribbon. A dialog box titled 'Azure Data Lake Storage Gen2' is open in the center. It contains a URL input field with the value <https://vehicletheftprojectsa.dfs.core.windows.net/gold>, a 'Data View' section with 'File System View' selected, and 'OK' and 'Cancel' buttons. To the right of the dialog is the 'Visualizations' pane, which lists various chart and table options. Below the dialog, there is a link 'Get data from another source →'. At the bottom of the screen, the navigation bar shows 'Page 1'.

# Data Import and Preview in Power BI

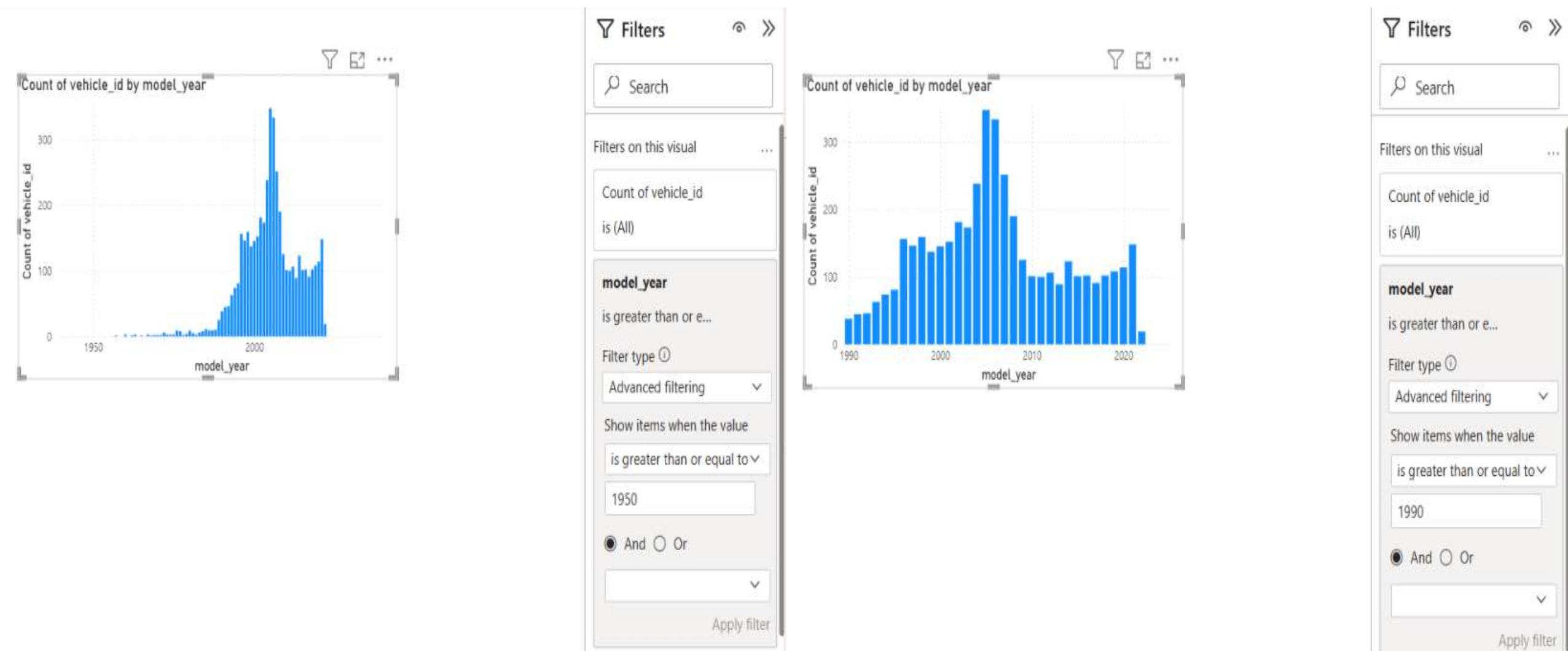
https://vehicletheftprojectsdfs.core.windows.net/gold

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	.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	.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	.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	.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

Untitled - Power BI Desktop

AAFREEN SANA H

File Home Insert Modeling View Optimize Help Format Data / Drill Table tools Column tools

Name: region Format: Text Summarization: Don't summarize Data category: Uncategorized Sort by column: Sort Groups: Data groups Relationships: Manage relationships New column: Calculations

Structure Formatting Properties

Count of vehicle\_id by region

Region	Count	Percentage
Auckland	1.64K	35.98%
Canterbury	0.66K	14.5%
Bay of Plenty	0.45K	9.8%
Wellington	0.42K	9.22%
Waikato	0.37K	8.1...
Northland	0.3K	6.7...
Gisborne	0.23K	5.14%
Manawatū-Whanganui	0.18K	3.9...
Otago	0.14K	3.0...
Taranaki	0.11K	2.46%
Hawke's Bay	0.03K	0.57%

Filters on this visual:

- Count of vehicle\_id: is (All)
- region: is (All)

Filter type: Top N, Show items: Top 5.

By value: Count of vehicle\_id.

Legend: region.

Values: Count of vehicle\_id.

Details: Add data fields here.

Tooltips: Add data fields here.

Drill through: Add data fields here.

Visualizations: Build visual

Data: Search Database1 Location: country, density, location\_id, population, region. Stolen Vehicles: color, date\_stolen, location\_id, make\_id, model\_year, vehicle\_desc, vehicle\_id, vehicle\_type.

Count of vehicle\_id by vehicle\_type

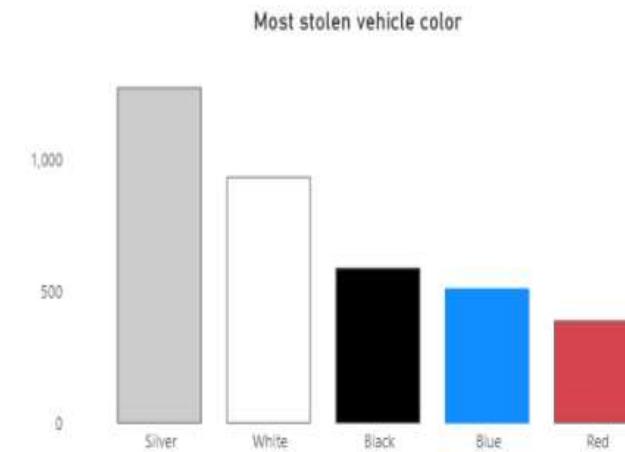
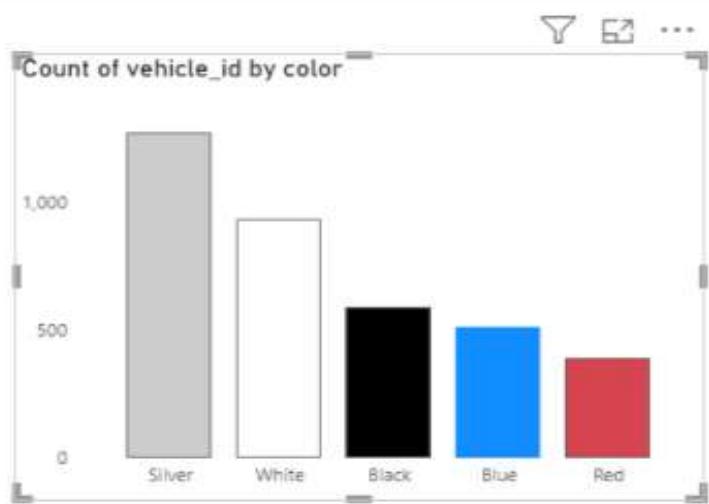
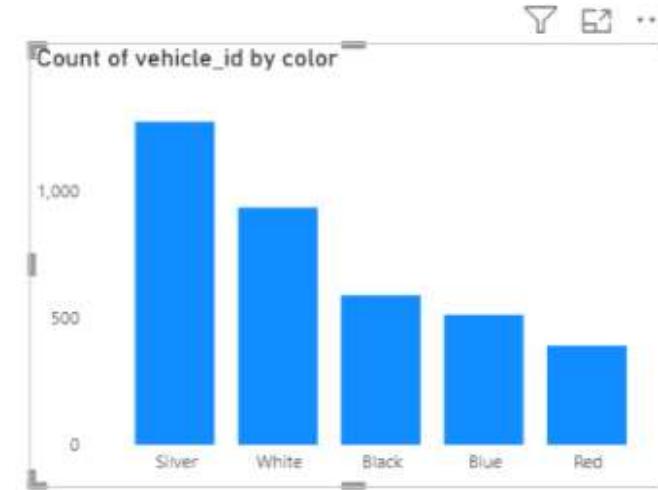
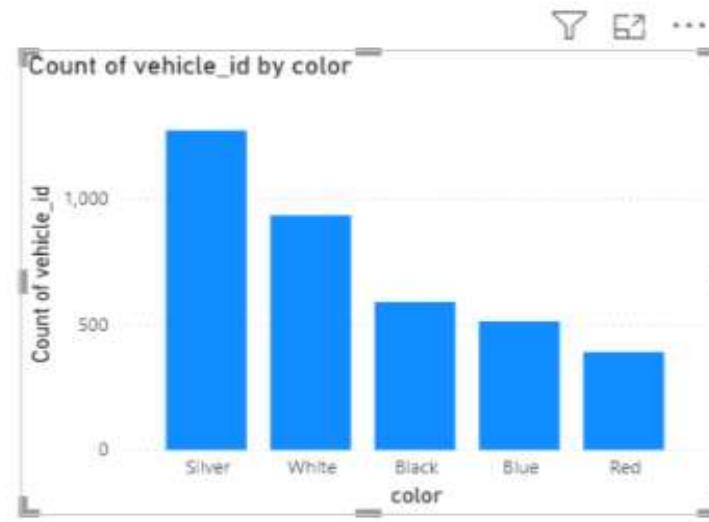
Vehicle Type	Count of vehicle_id
Stationwagon	~1000
Sedan	~800
Hatchback	~600
Trailer	~500
Utility	~400

Count of vehicle\_id by model\_year

Model Year	Count of vehicle_id
1990	~50
1995	~150
2000	~200
2005	~300
2010	~200
2015	~150
2020	~100

Page 1

# Color-Based Visualization – Bar Charts



# Dashboard

