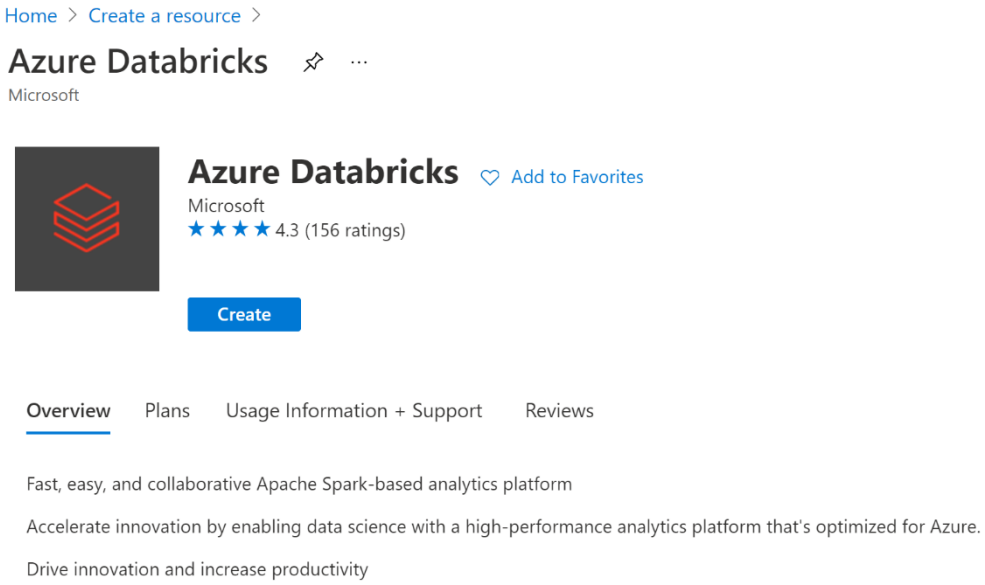
Chapter 1: Creating an Azure Databricks Service



Home > Create a resource > Azure Databricks >

Create an Azure Databricks workspace ...

Basics | Networking | Advanced | Tags | Review + create

Project Details

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

Subscription * ⓘ

Resource group * ⓘ

Create new

A resource group is a container that holds related resources for an Azure solution.

Name *

CookbookRG ✓

OK | Cancel

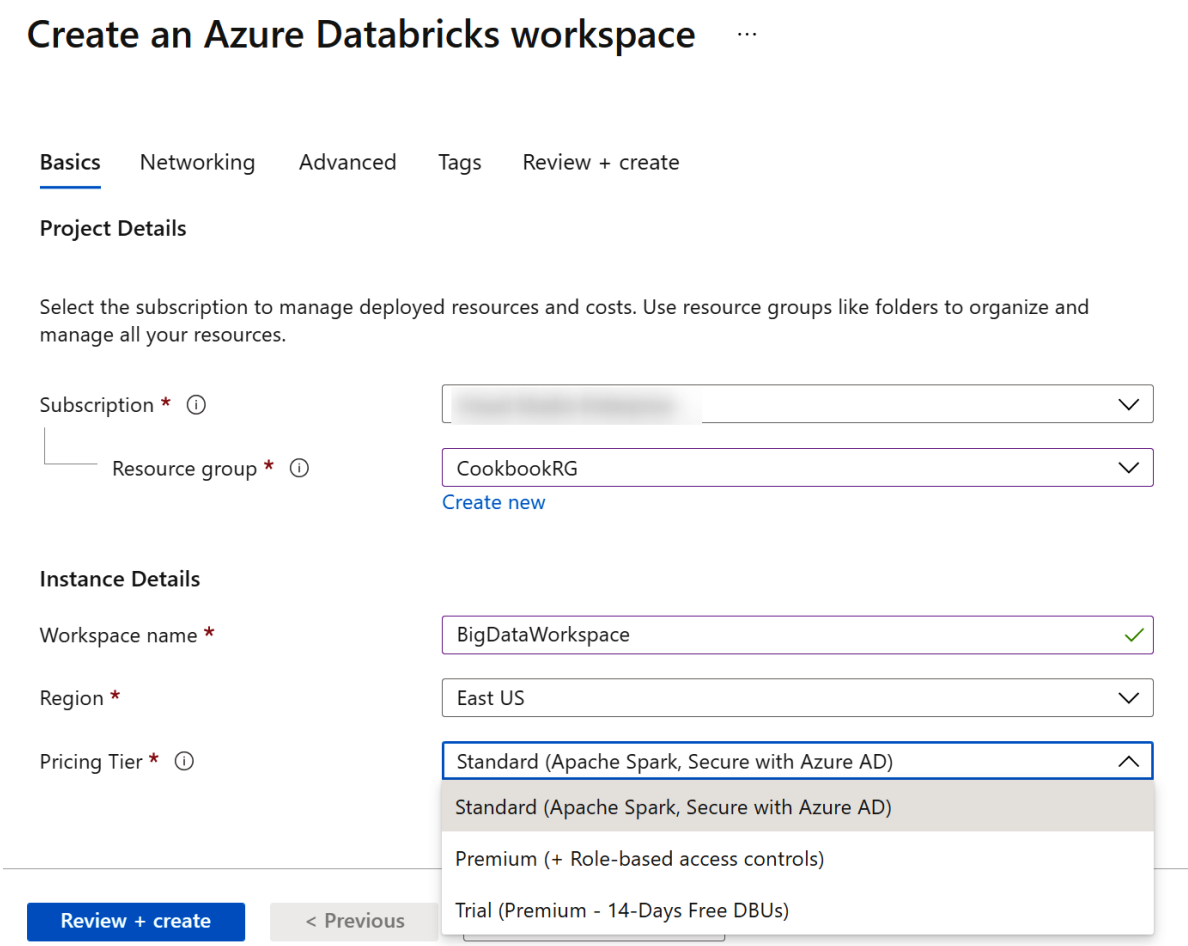
Instance Details

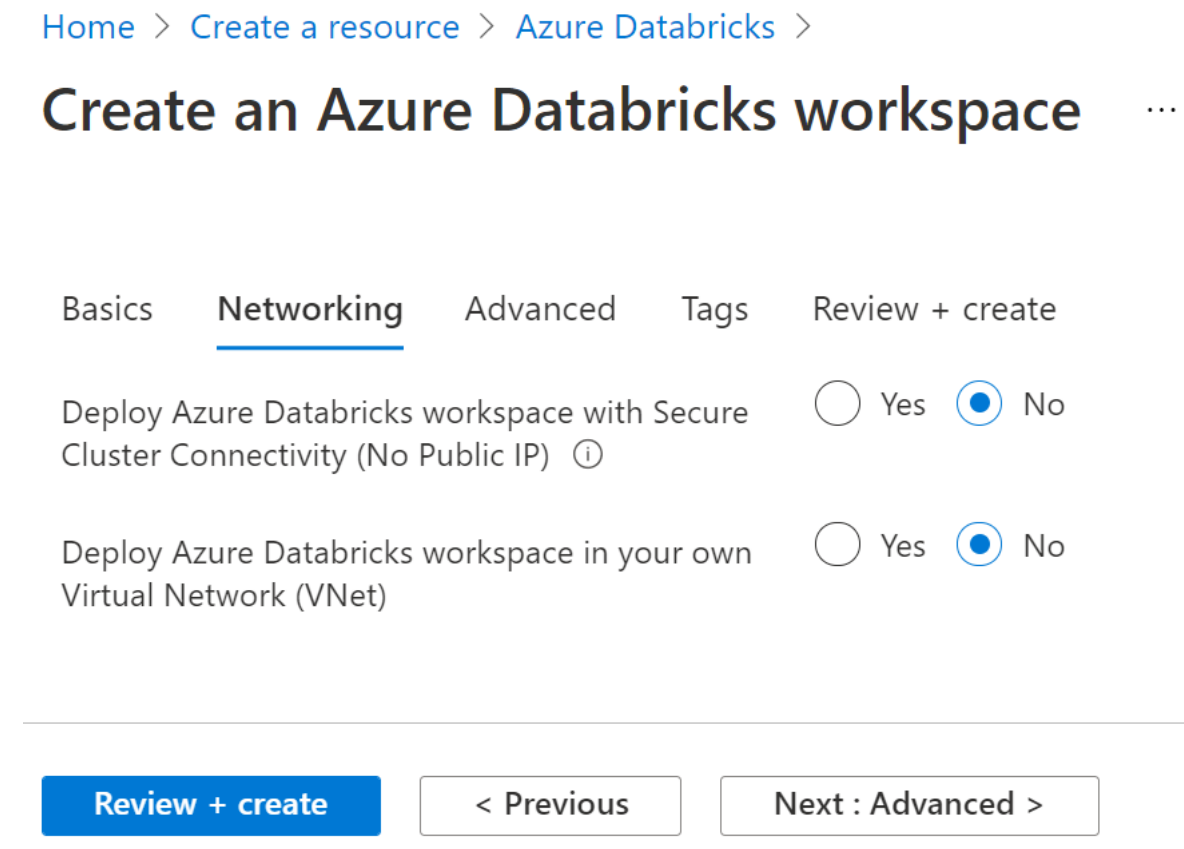
Workspace name *

Region *

Pricing Tier * ⓘ

Review + create | < Previous | Next : Networking >


Create an Azure Databricks workspace
Basics
Networking
Advanced
Tags
Review + create
Project Details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription *
Resource group *
CookbookRG
Create new
Instance Details
Workspace name *
BigDataWorkspace
Region *
East US
Pricing Tier *
Standard (Apache Spark, Secure with Azure AD)
Standard (Apache Spark, Secure with Azure AD)
Premium (+ Role-based access controls)
Trial (Premium - 14-Days Free DBUs)
Review + create
< Previous


Home > Create a resource > Azure Databricks >
Create an Azure Databricks workspace
Basics
Networking
Advanced
Tags
Review + create
Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)
Yes
No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)
Yes
No
Review + create
< Previous
Next : Advanced >

Create an Azure Databricks workspace ...

Validation Succeeded

Basics Networking Advanced Tags **Review + create**

Summary

Basics

Workspace name	BigDataWorkspace
Subscription	[redacted]
Resource group	CookbookRG
Region	East US
Pricing Tier	trial

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)	No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)	No

---

Create **< Previous** Download a template for automation

BigDataWorkspace

Azure Databricks Service

Search (Ctrl+/)

- Overview
- Activity log
- Access control (IAM)
- Tags

Settings

- Virtual Network Peerings
- Encryption
- Properties
- Locks

Automation

- Tasks (preview)
- Export template

Support + troubleshooting

- New support request

Delete

Essentials

Status
Active

Resource group
CookbookRG

Location
East US

Subscription
[redacted]

Subscription ID
[redacted]

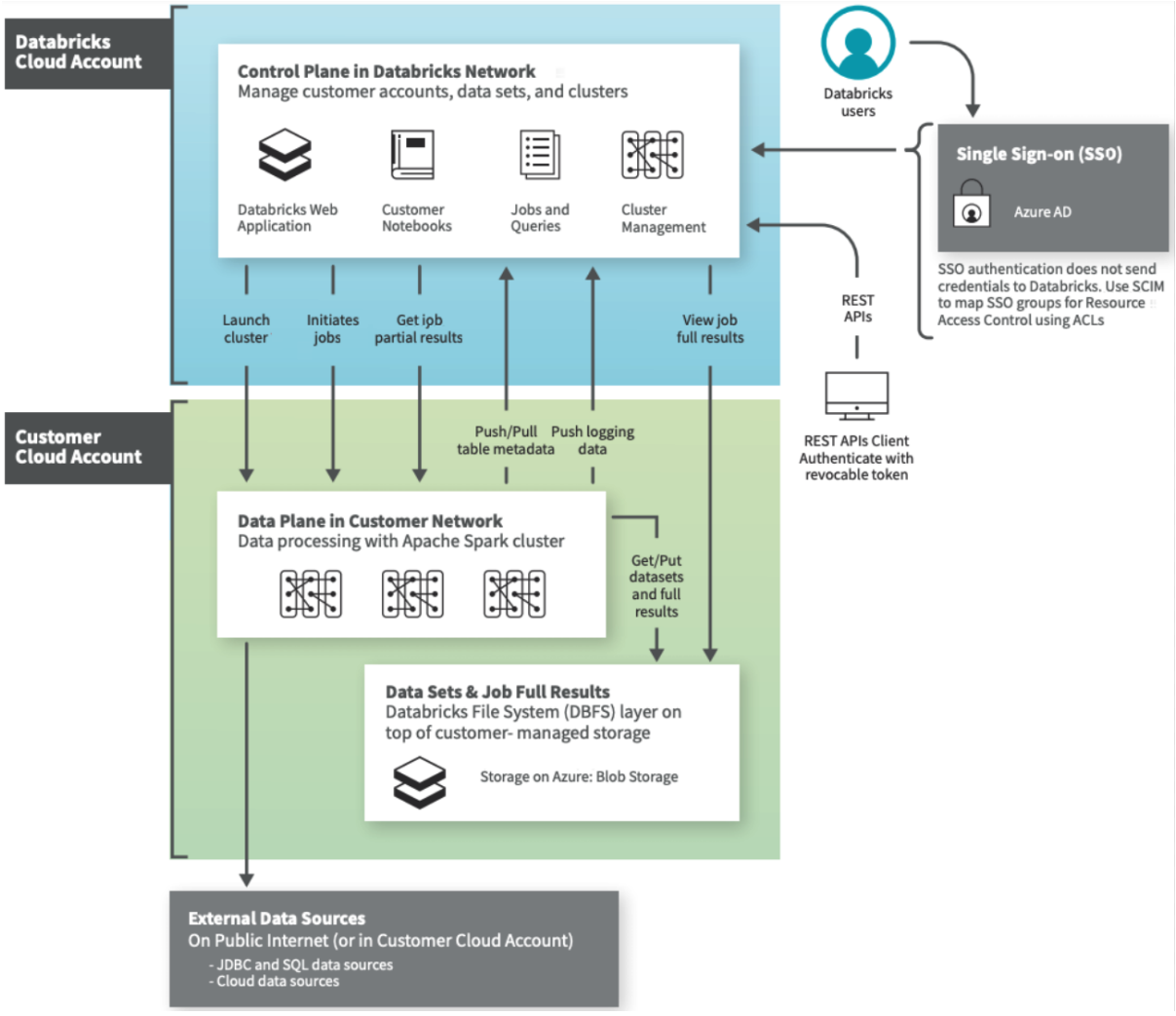
Tags (change)
Click here to add tags

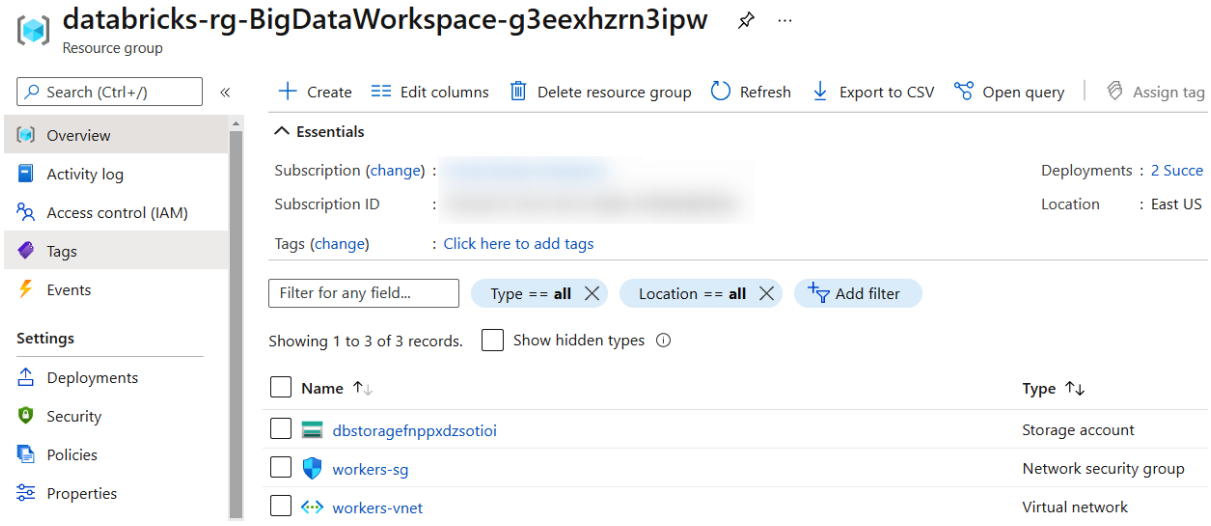
Managed Resource Group
databricks-rg-BigDataWorkspace-g3eexhzrn3ipw

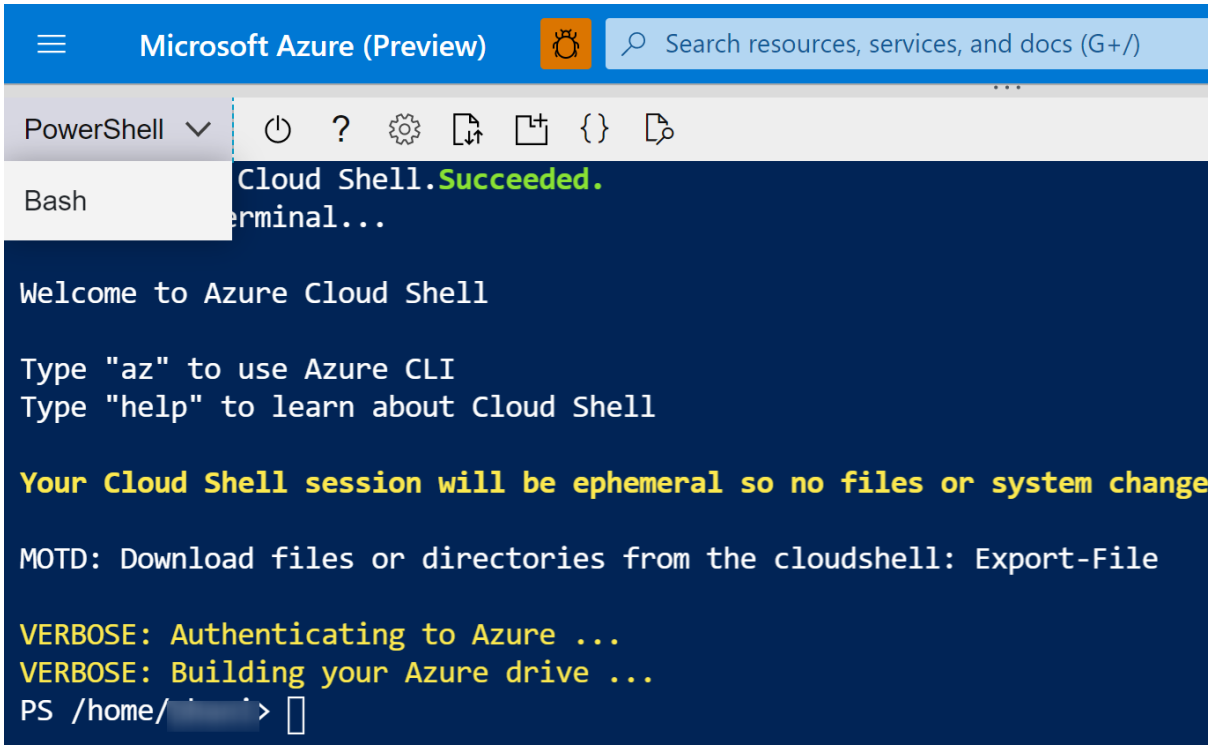
URL
https://adb-[redacted]7.azuredatabricks.net

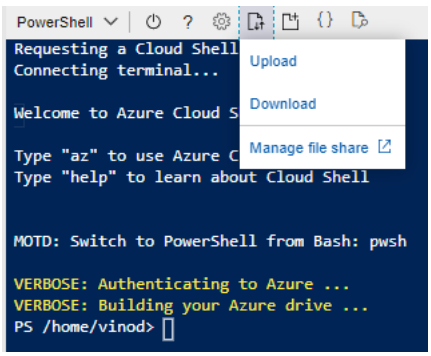
Pricing Tier
Trial (Premium - 14-Days Free DBUs)

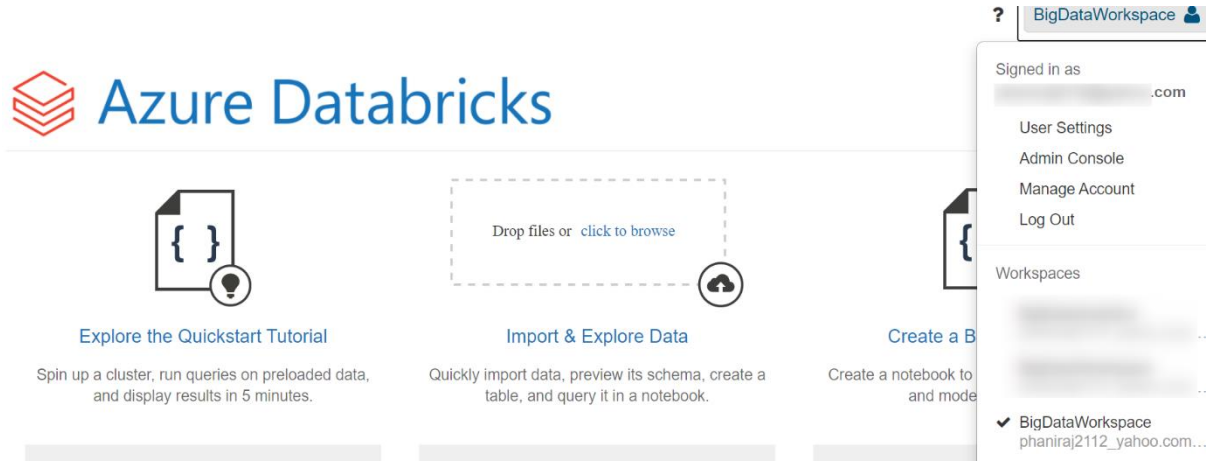
Launch Workspace

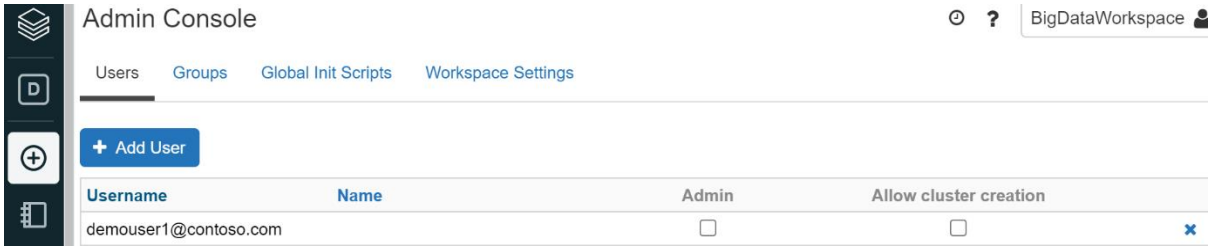


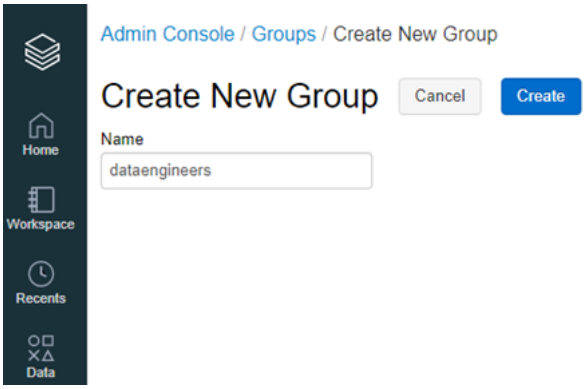

databricks-rg-BigDataWorkspace-g3eexhzrn3ipw
Resource group
Search (Ctrl+/)
Create
Edit columns
Delete resource group
Refresh
Export to CSV
Open query
Assign tag
Overview
Activity log
Access control (IAM)
Tags
Events
Settings
Deployments
Security
Policies
Properties
Essentials
Subscription (change) :
Subscription ID :
Tags (change) : Click here to add tags
Deployments : 2 Succe
Location : East US
Filter for any field...
Type == all
Location == all
Add filter
Showing 1 to 3 of 3 records.
Show hidden types
Name
Type
dbstoragefnppxdzsotioi
Storage account
workers-sg
Network security group
workers-vnet
Virtual network

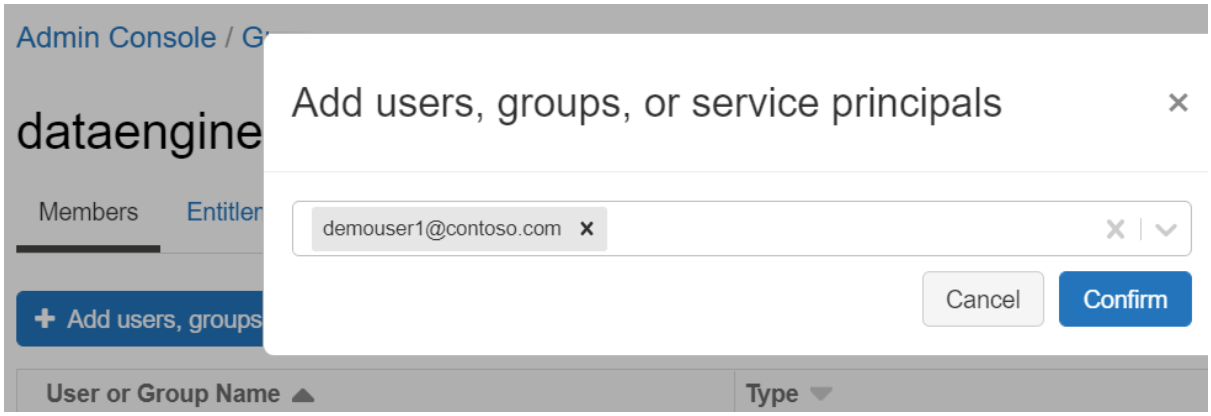

Microsoft Azure (Preview)
Search resources, services, and docs (G+/)
PowerShell
Bash
Cloud Shell.Succeeded.
rminal...
Welcome to Azure Cloud Shell
Type "az" to use Azure CLI
Type "help" to learn about Cloud Shell
Your Cloud Shell session will be ephemeral so no files or system change
MOTD: Download files or directories from the cloudshell: Export-File
VERBOSE: Authenticating to Azure ...
VERBOSE: Building your Azure drive ...
PS /home/

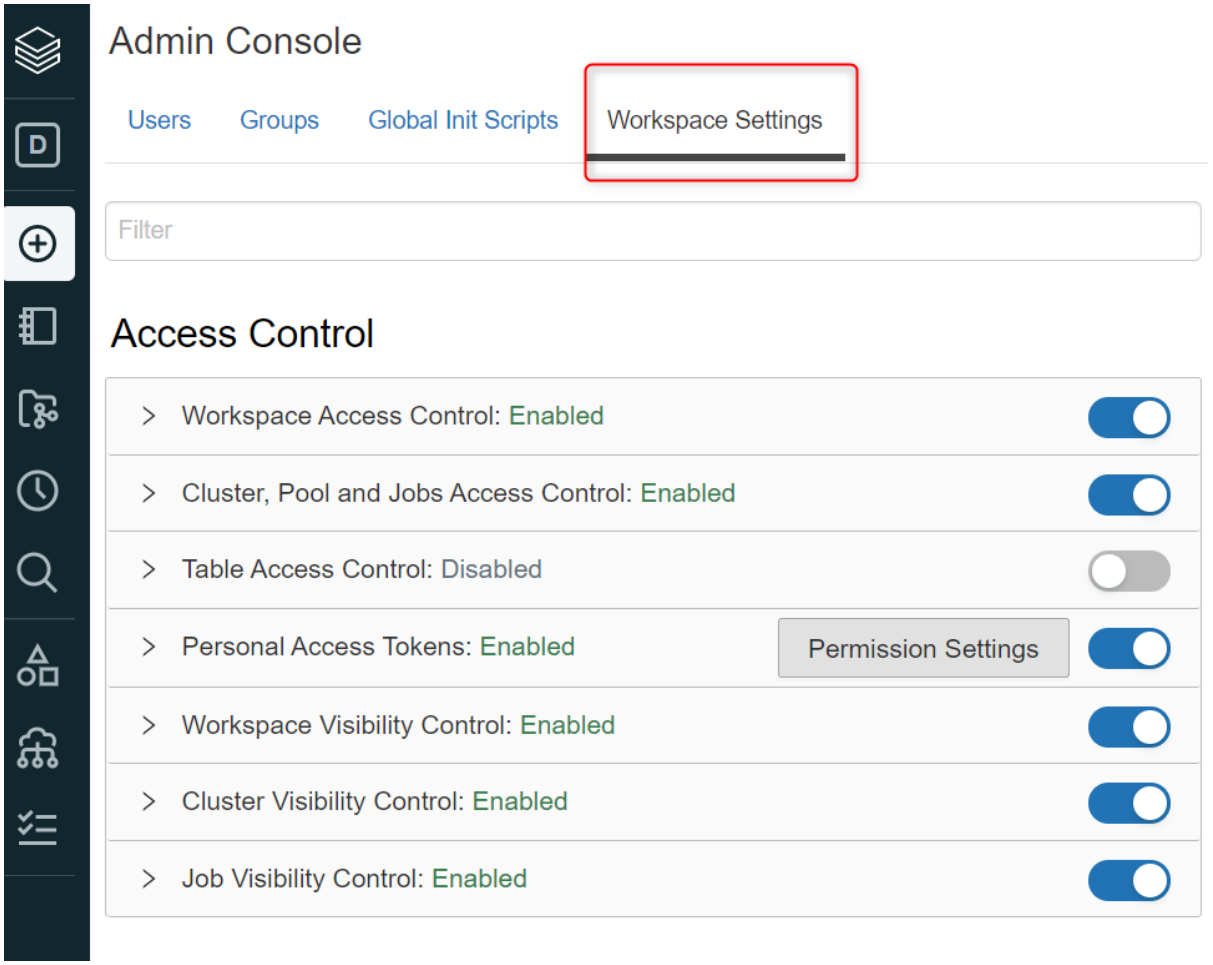

PowerShell
Requesting a Cloud Shell
Connecting terminal...
Welcome to Azure Cloud S
Type "az" to use Azure C
Type "help" to learn about Cloud Shell
MOTD: Switch to PowerShell from Bash: pwsh
VERBOSE: Authenticating to Azure ...
VERBOSE: Building your Azure drive ...
PS /home/vinod>
Upload
Download
Manage file share

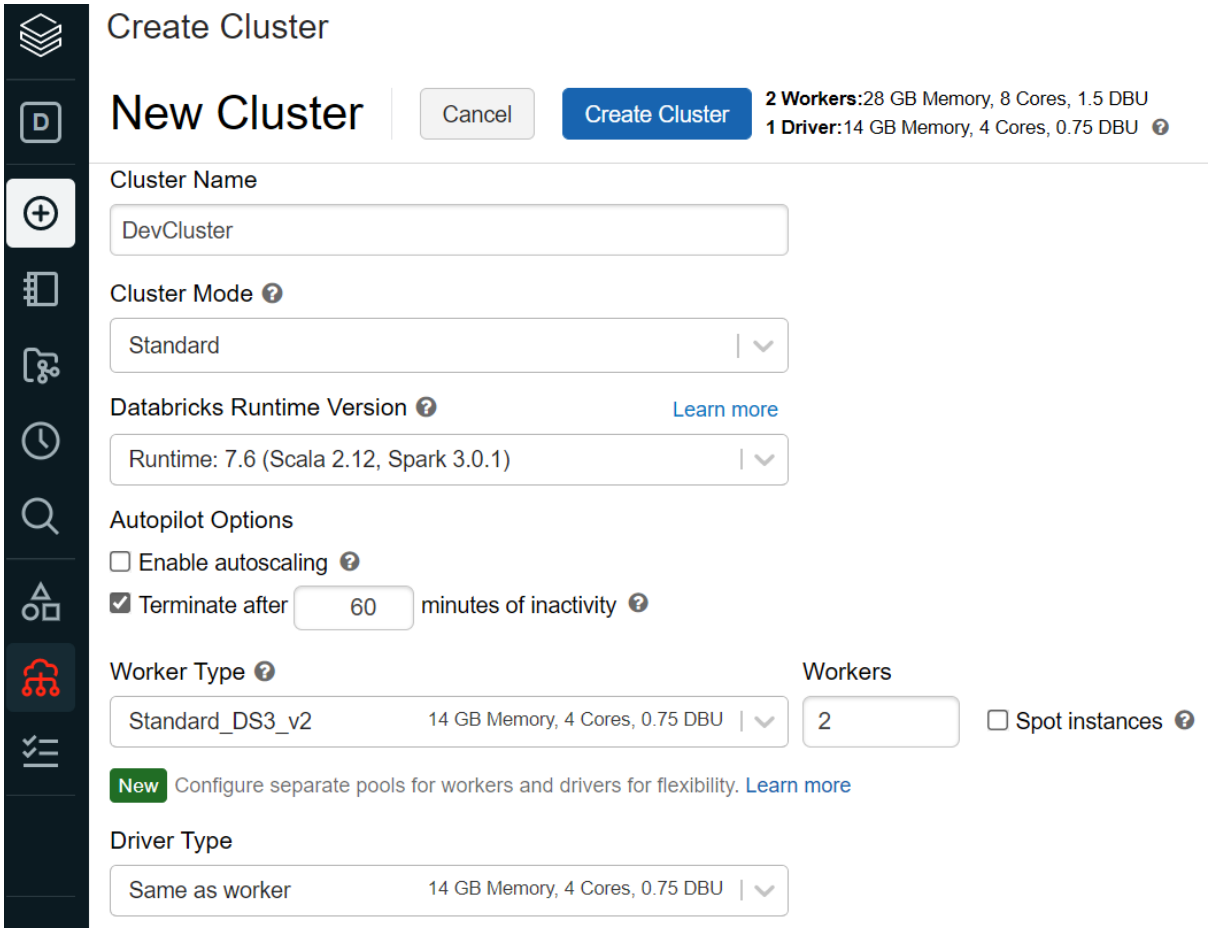

Azure Databricks
BigDataWorkspace
Signed in as
.com
User Settings
Admin Console
Manage Account
Log Out
Workspaces
BigDataWorkspace
phaniraj2112_yahoo.com...
Drop files or click to browse
Explore the Quickstart Tutorial
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.
Import & Explore Data
Quickly import data, preview its schema, create a table, and query it in a notebook.
Create a B
Create a notebook to and mode

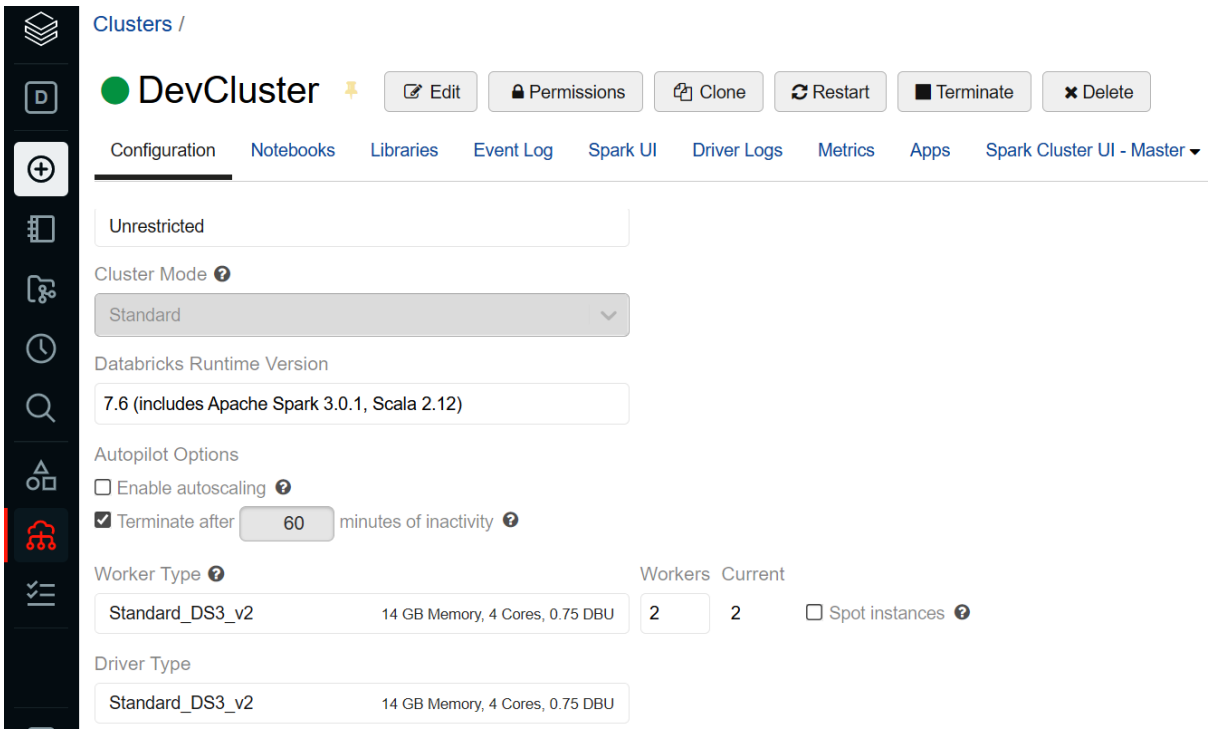

Admin Console
BigDataWorkspace
Users
Groups
Global Init Scripts
Workspace Settings
Add User
Username
Name
Admin
Allow cluster creation
demouser1@contoso.com

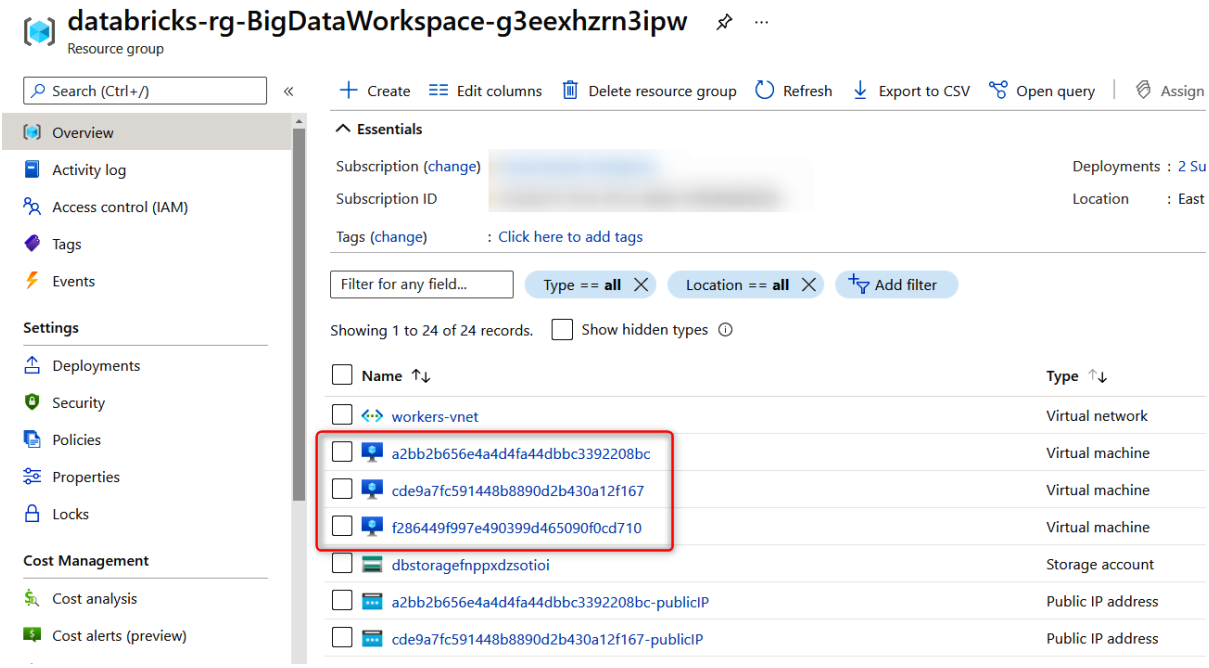

Admin Console / Groups / Create New Group
Create New Group
Cancel
Create
Name
dataengineers
Home
Workspace
Recents
Data

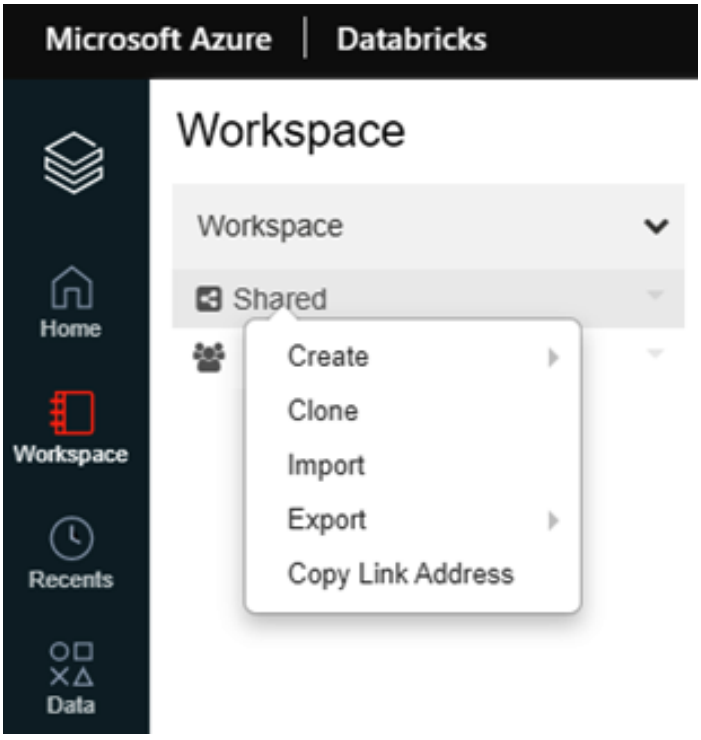

Admin Console / G
dataengine
Members
Entitler
Add users, groups
User or Group Name
Type
Add users, groups, or service principals
demouser1@contoso.com
Cancel
Confirm

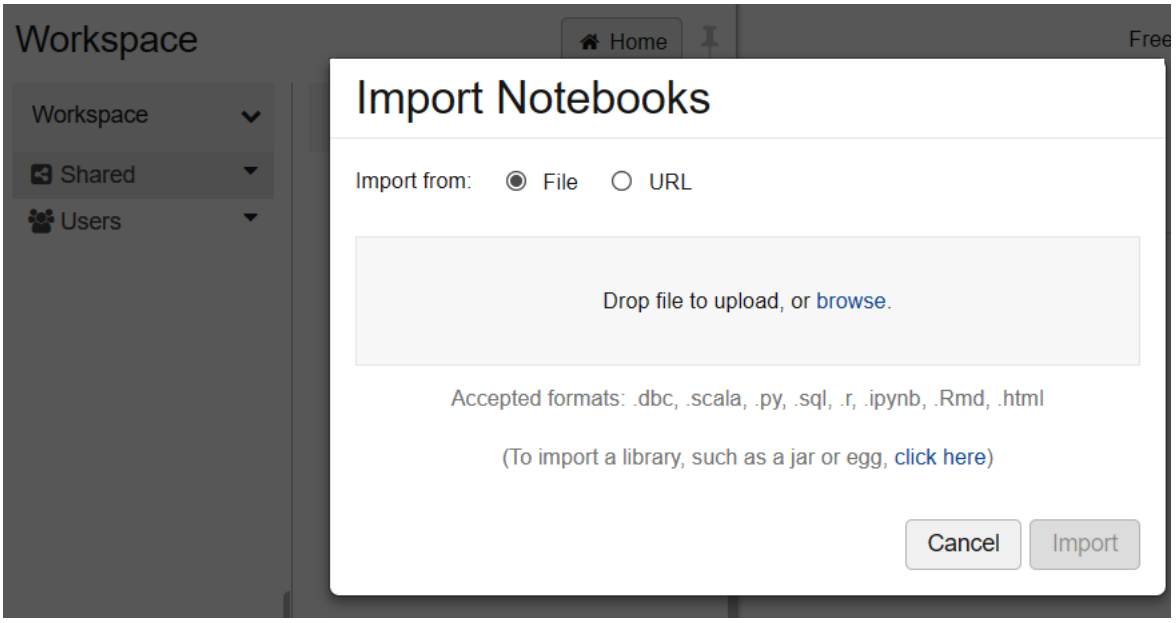

Admin Console
Users
Groups
Global Init Scripts
Workspace Settings
Filter
Access Control
Workspace Access Control: Enabled
Cluster, Pool and Jobs Access Control: Enabled
Table Access Control: Disabled
Personal Access Tokens: Enabled
Permission Settings
Workspace Visibility Control: Enabled
Cluster Visibility Control: Enabled
Job Visibility Control: Enabled

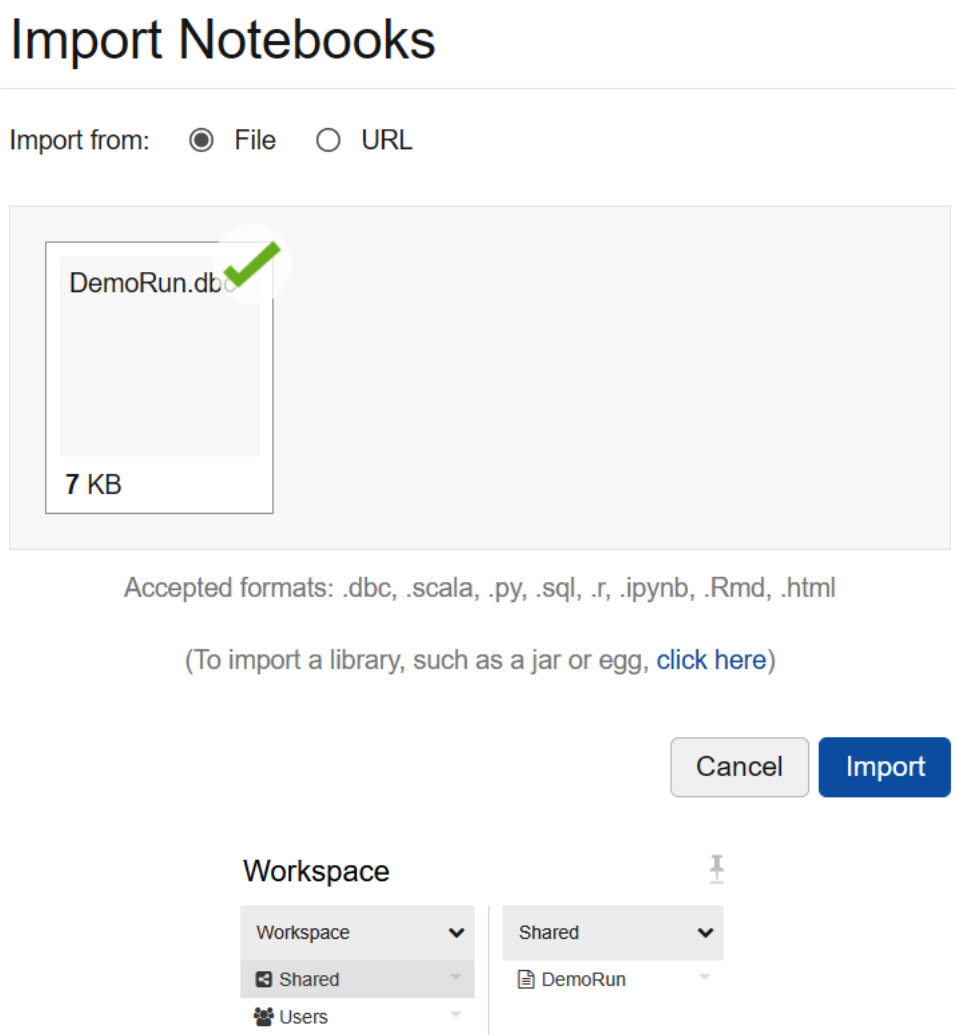

Create Cluster
New Cluster
Cancel
Create Cluster
2 Workers:28 GB Memory, 8 Cores, 1.5 DBU
1 Driver:14 GB Memory, 4 Cores, 0.75 DBU
Cluster Name
DevCluster
Cluster Mode
Standard
Databricks Runtime Version
Learn more
Runtime: 7.6 (Scala 2.12, Spark 3.0.1)
Autopilot Options
Enable autoscaling
Terminate after
60
minutes of inactivity
Worker Type
Workers
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU
2
Spot instances
New
Configure separate pools for workers and drivers for flexibility. Learn more
Driver Type
Same as worker
14 GB Memory, 4 Cores, 0.75 DBU


Clusters /
DevCluster
Edit
Permissions
Clone
Restart
Terminate
Delete
Configuration
Notebooks
Libraries
Event Log
Spark UI
Driver Logs
Metrics
Apps
Spark Cluster UI - Master
Unrestricted
Cluster Mode
Standard
Databricks Runtime Version
7.6 (includes Apache Spark 3.0.1, Scala 2.12)
Autopilot Options
Enable autoscaling
Terminate after
60
minutes of inactivity
Worker Type
Workers
Current
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU
2
2
Spot instances
Driver Type
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU

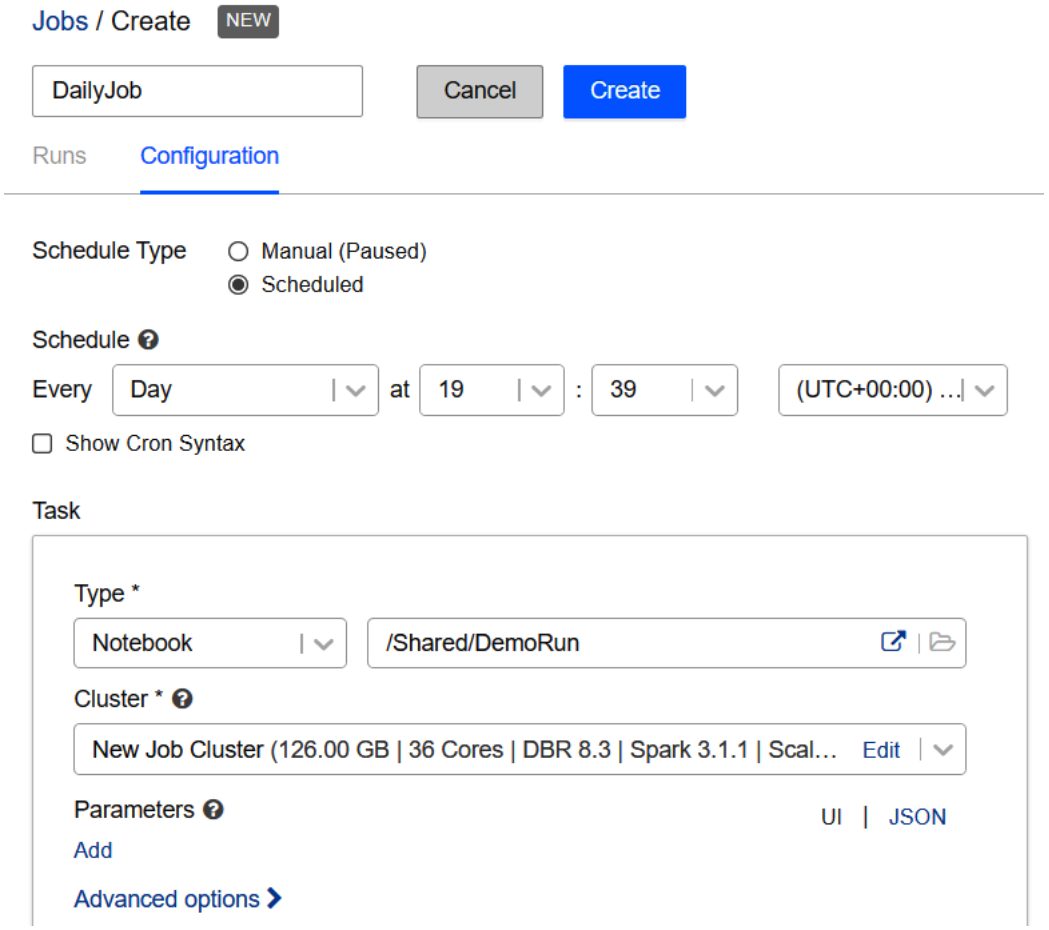

databricks-rg-BigDataWorkspace-g3eexhzrn3ipw
Resource group
Search (Ctrl+/)
Create
Edit columns
Delete resource group
Refresh
Export to CSV
Open query
Assign
Overview
Activity log
Access control (IAM)
Tags
Events
Settings
Deployments
Security
Policies
Properties
Locks
Cost Management
Cost analysis
Cost alerts (preview)
Essentials
Subscription (change)
Subscription ID
Deployments : 2 Su
Location : East
Tags (change) : Click here to add tags
Filter for any field...
Type == all
Location == all
Add filter
Showing 1 to 24 of 24 records.
Show hidden types
Name
Type
workers-vnet
Virtual network
a2bb2b656e4a4d4fa44dbbc3392208bc
Virtual machine
cde9a7fc591448b8890d2b430a12f167
Virtual machine
f286449f997e490399d465090f0cd710
Virtual machine
dbstoragefnppxdzsotioi
Storage account
a2bb2b656e4a4d4fa44dbbc3392208bc-publicIP
Public IP address
cde9a7fc591448b8890d2b430a12f167-publicIP
Public IP address

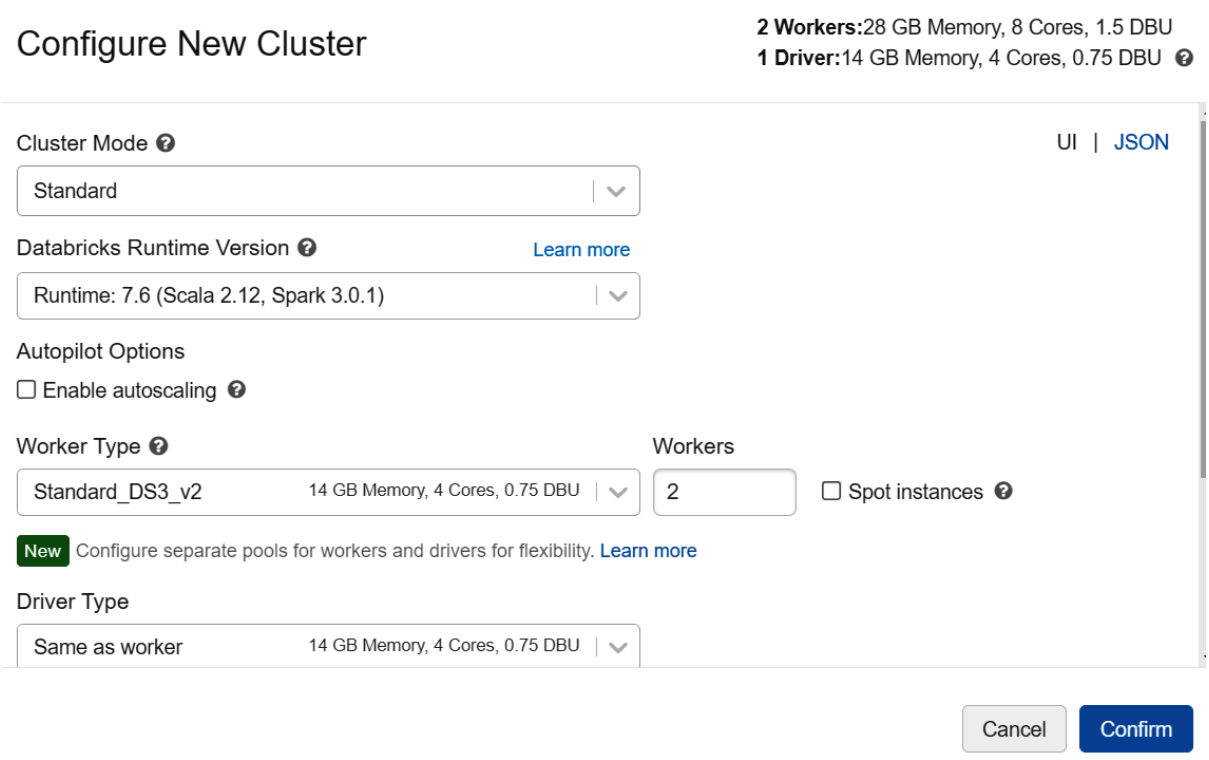

Microsoft Azure
Databricks
Workspace
Home
Workspace
Recents
Data
Workspace
Shared
Create
Clone
Import
Export
Copy Link Address

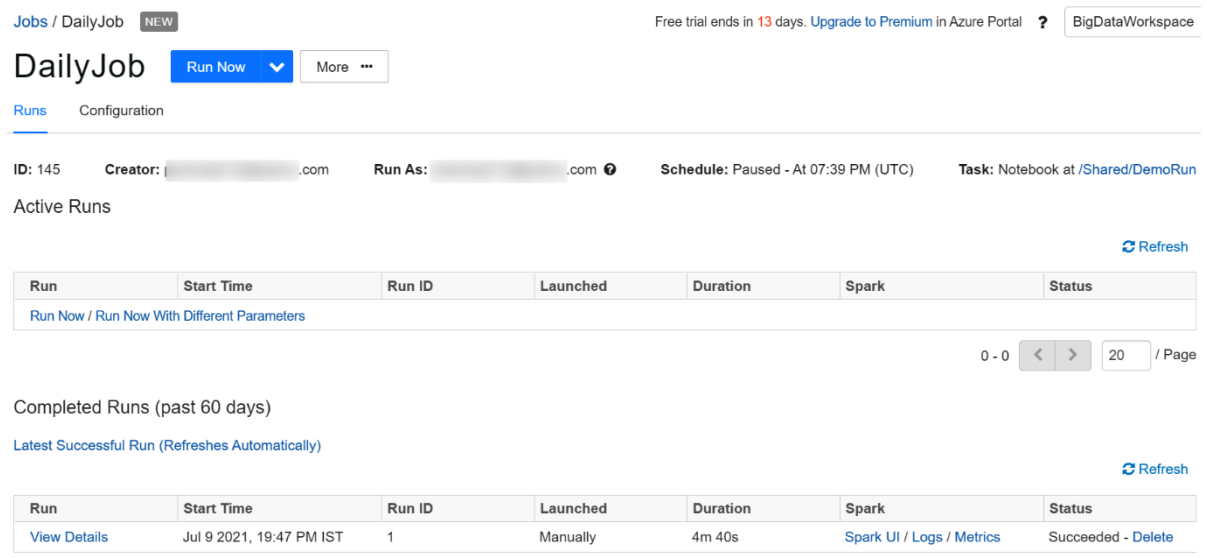

Workspace
Home
Workspace
Shared
Users
Import Notebooks
Import from: File URL
Drop file to upload, or browse.
Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html
(To import a library, such as a jar or egg, click here)
Cancel
Import

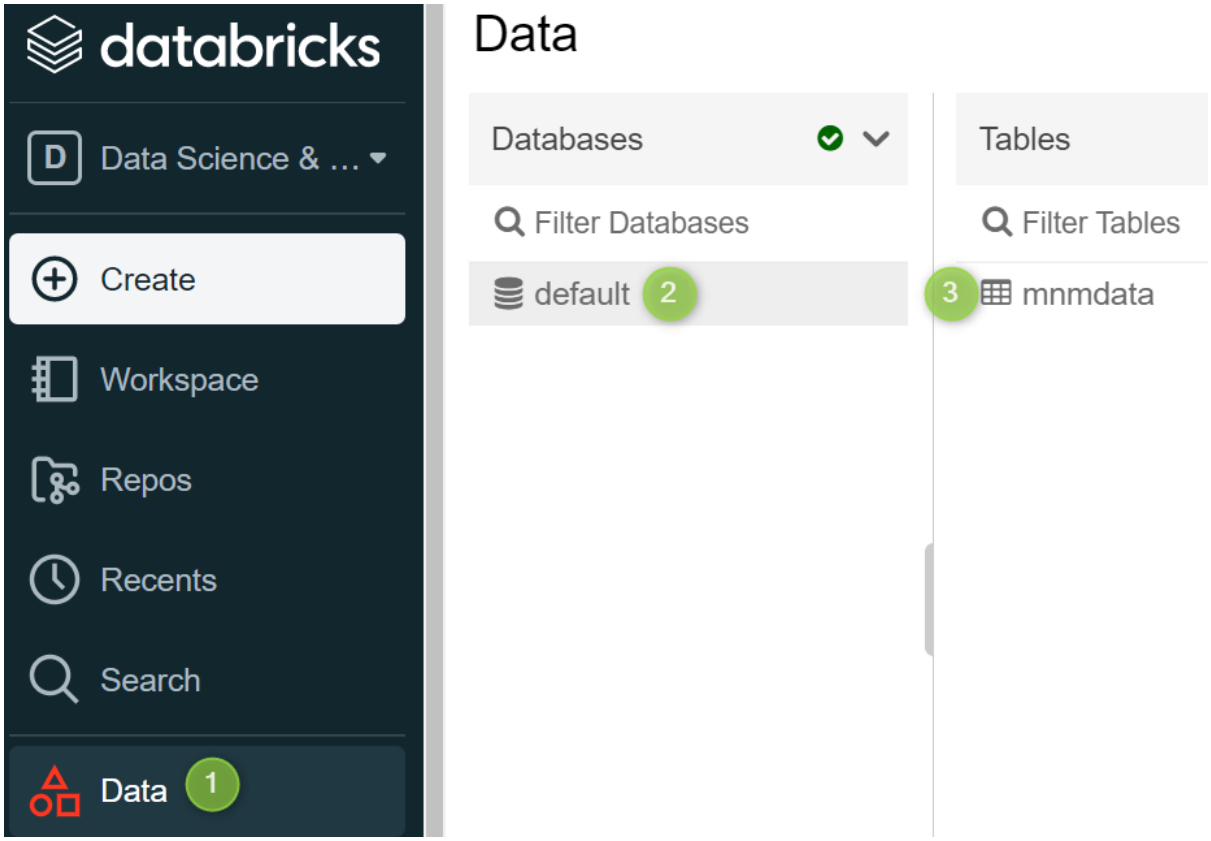

Import Notebooks
Import from: File URL
DemoRun.db
7 KB
Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html
(To import a library, such as a jar or egg, click here)
Cancel
Import

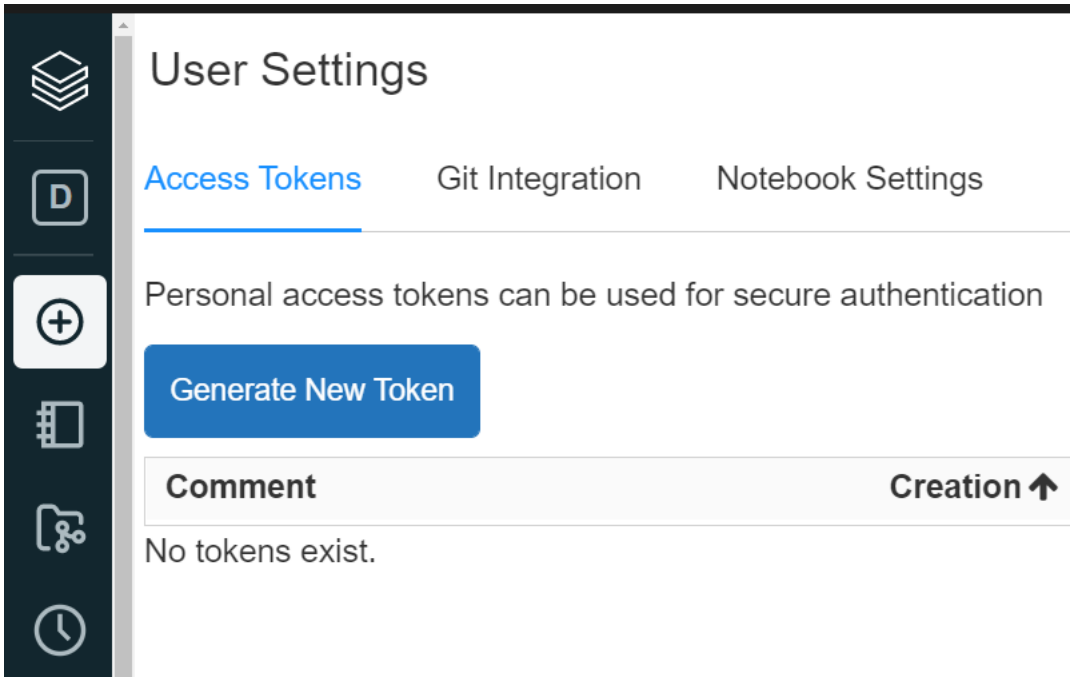
Workspace
Workspace
Shared
Users
Shared
DemoRun

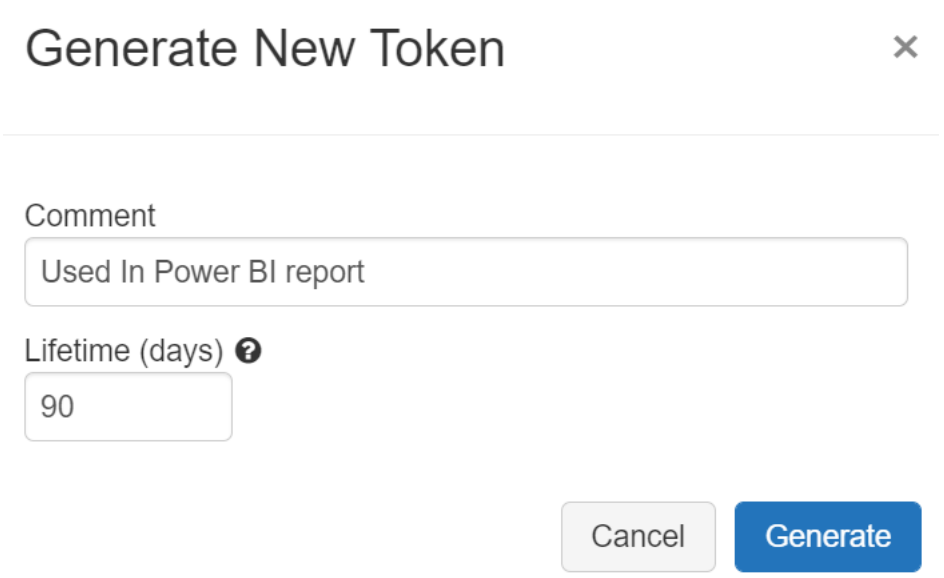

Jobs / Create
NEW
DailyJob
Cancel
Create
Runs
Configuration
Schedule Type
Manual (Paused)
Scheduled
Schedule
Every
Day
at
19
:
39
(UTC+00:00) ...
Show Cron Syntax
Task
Type *
Notebook
/Shared/DemoRun
Cluster *
New Job Cluster (126.00 GB | 36 Cores | DBR 8.3 | Spark 3.1.1 | Scal...
Edit
Parameters
UI | JSON
Add
Advanced options

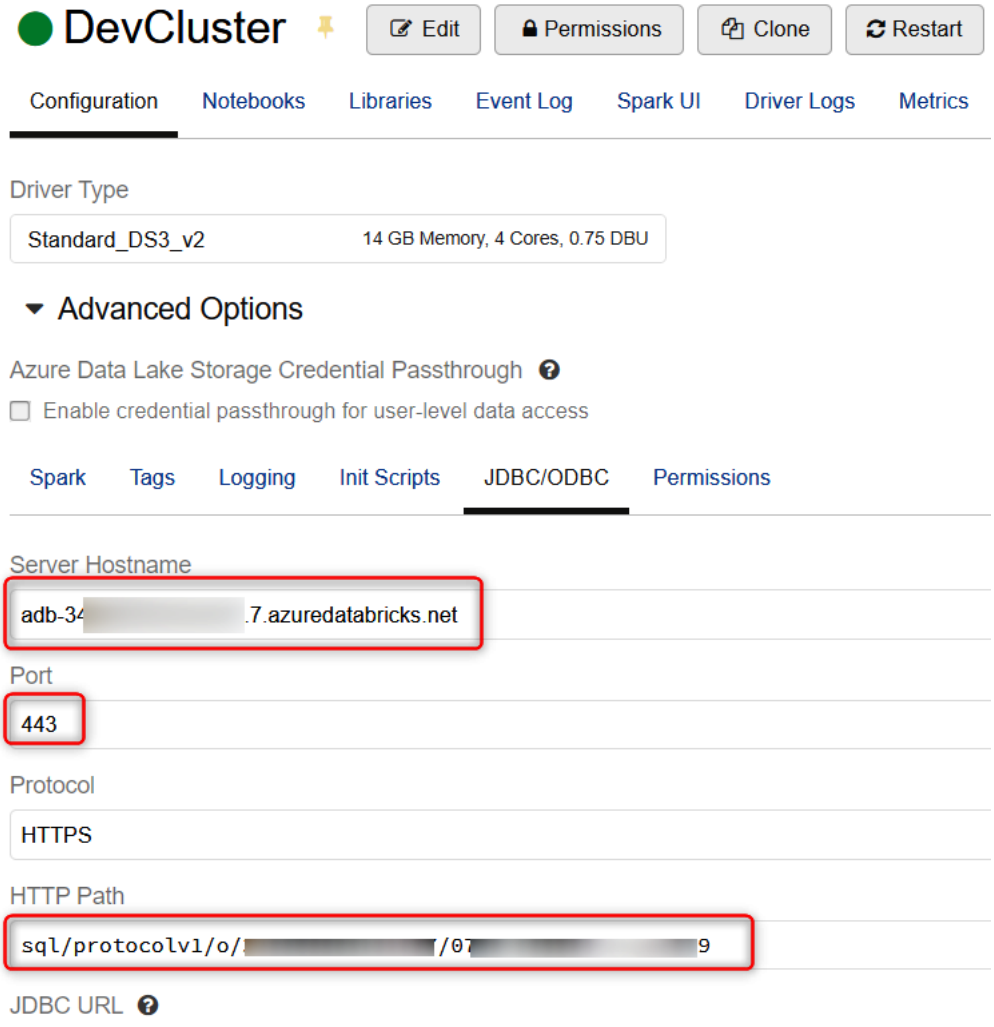

Configure New Cluster
2 Workers:28 GB Memory, 8 Cores, 1.5 DBU
1 Driver:14 GB Memory, 4 Cores, 0.75 DBU
Cluster Mode
UI | JSON
Standard
Databricks Runtime Version
Learn more
Runtime: 7.6 (Scala 2.12, Spark 3.0.1)
Autopilot Options
Enable autoscaling
Worker Type
Workers
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU
2
Spot instances
New
Configure separate pools for workers and drivers for flexibility. Learn more
Driver Type
Same as worker
14 GB Memory, 4 Cores, 0.75 DBU
Cancel
Confirm

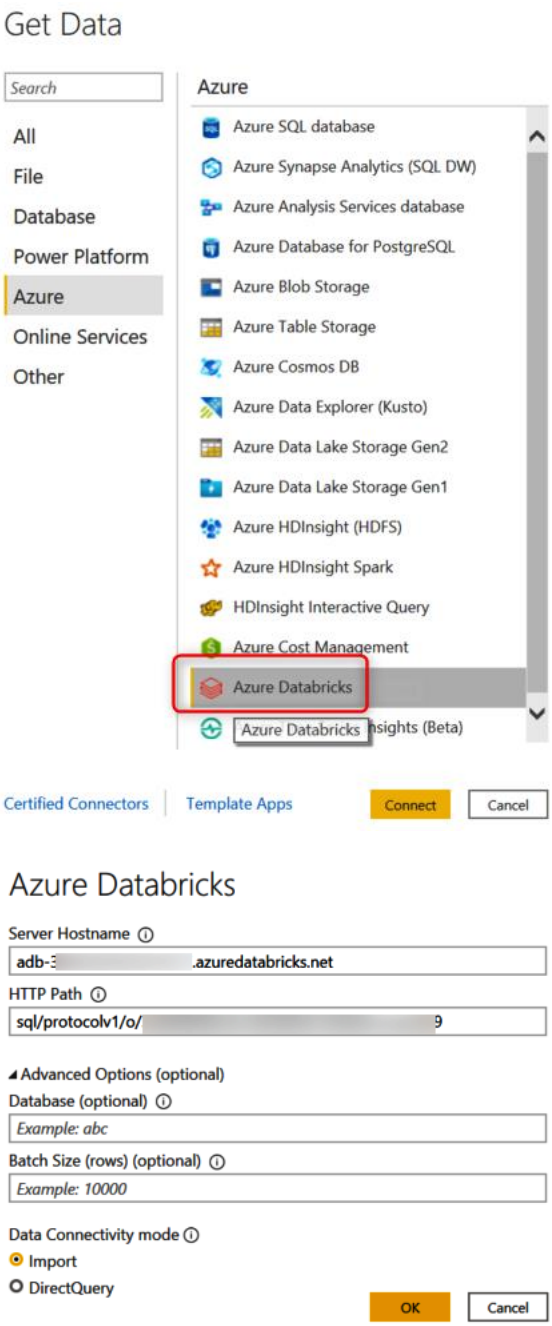

Jobs / DailyJob NEW
Free trial ends in 13 days. Upgrade to Premium in Azure Portal
BigDataWorkspace
DailyJob
Run Now
More
Runs
Configuration
ID: 145
Creator: .com
Run As: .com
Schedule: Paused - At 07:39 PM (UTC)
Task: Notebook at /Shared/DemoRun
Active Runs
Refresh
Run
Start Time
Run ID
Launched
Duration
Spark
Status
Run Now / Run Now With Different Parameters
0 - 0
20
/ Page
Completed Runs (past 60 days)
Latest Successful Run (Refreshes Automatically)
Refresh
Run
Start Time
Run ID
Launched
Duration
Spark
Status
View Details
Jul 9 2021, 19:47 PM IST
1
Manually
4m 40s
Spark UI / Logs / Metrics
Succeeded - Delete


databricks
Data Science & ...
Create
Workspace
Repos
Recents
Search
Data
1
Data
Databases
Filter Databases
default
2
Tables
Filter Tables
3
mnmdata

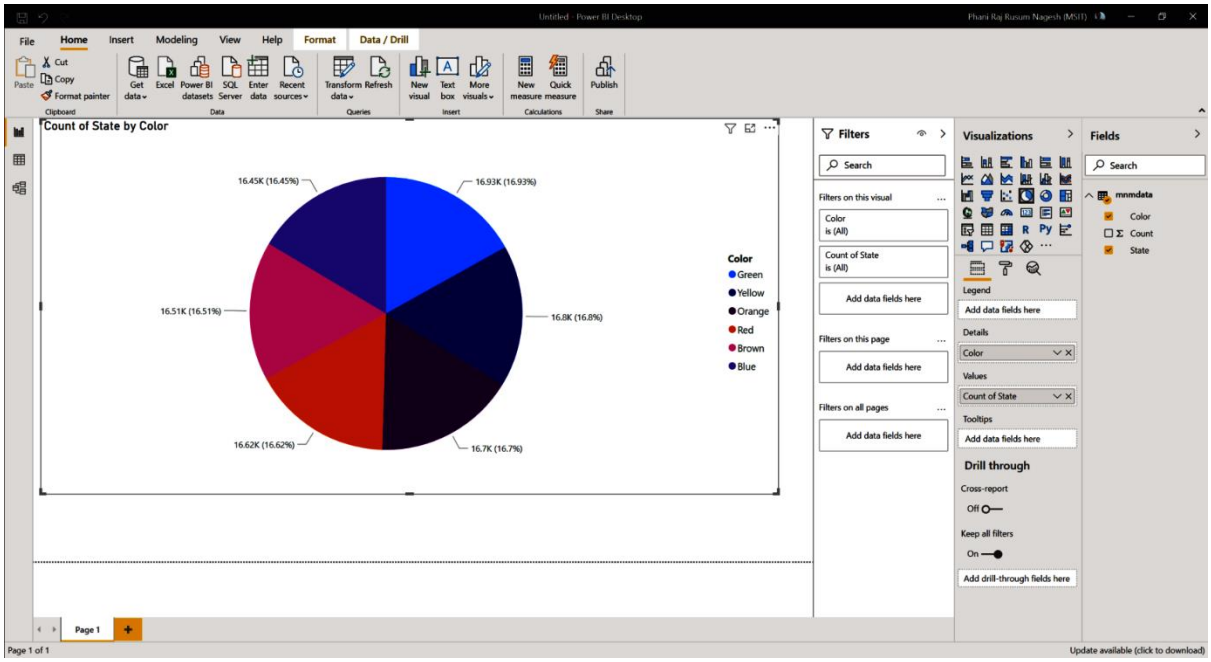

User Settings
Access Tokens
Git Integration
Notebook Settings
Personal access tokens can be used for secure authentication
Generate New Token
Comment
Creation
No tokens exist.


Generate New Token
Comment
Used In Power BI report
Lifetime (days)
90
Cancel
Generate

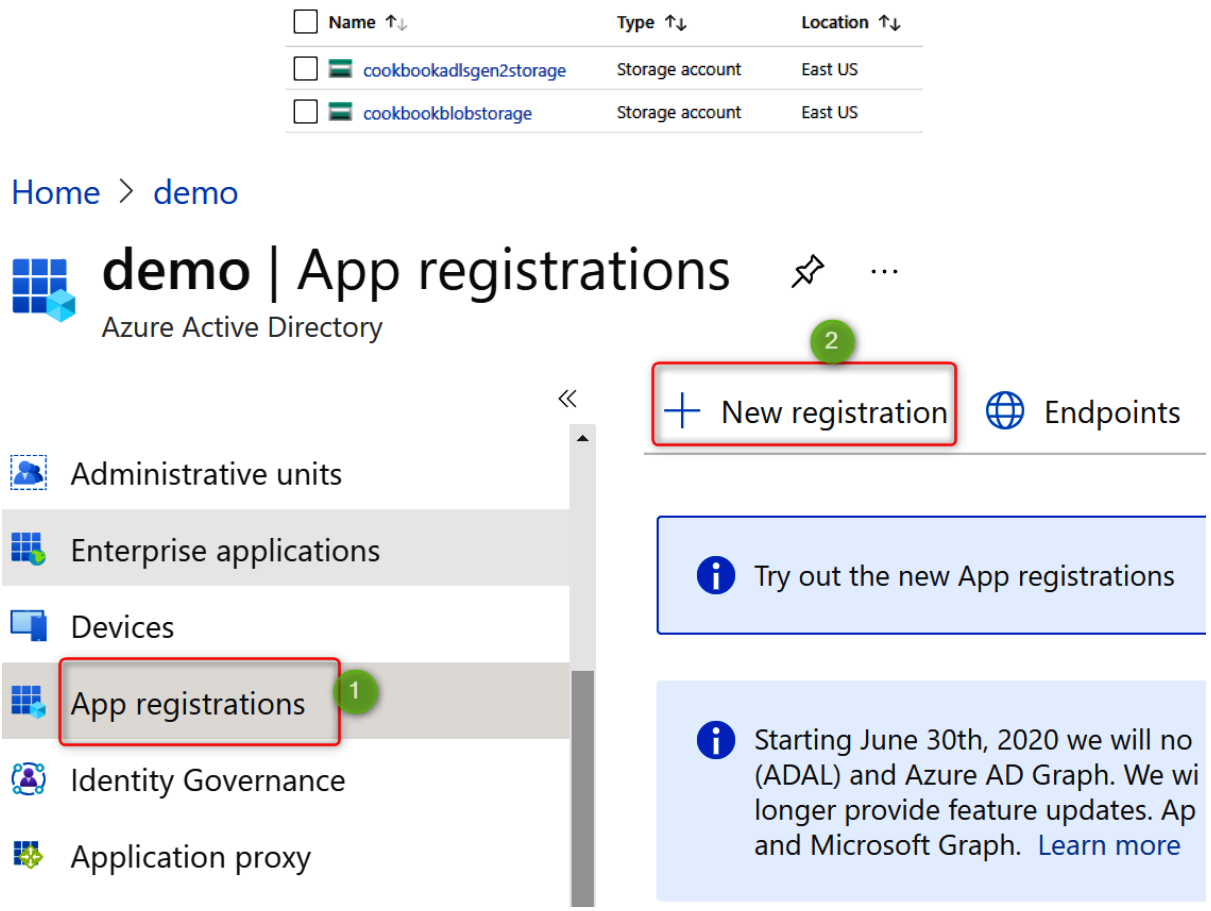

DevCluster
Edit
Permissions
Clone
Restart
Configuration
Notebooks
Libraries
Event Log
Spark UI
Driver Logs
Metrics
Driver Type
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU
Advanced Options
Azure Data Lake Storage Credential Passthrough
Enable credential passthrough for user-level data access
Spark
Tags
Logging
Init Scripts
JDBC/ODBC
Permissions
Server Hostname
adb-34 .7.azuredatabricks.net
Port
443
Protocol
HTTPS
HTTP Path
sql/protocolv1/o/ /07 9
JDBC URL


Get Data
Search
All
File
Database
Power Platform
Azure
Online Services
Other
Azure
Azure SQL database
Azure Synapse Analytics (SQL DW)
Azure Analysis Services database
Azure Database for PostgreSQL
Azure Blob Storage
Azure Table Storage
Azure Cosmos DB
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen2
Azure Data Lake Storage Gen1
Azure HDInsight (HDFS)
Azure HDInsight Spark
HDInsight Interactive Query
Azure Cost Management
Azure Databricks
Azure Databricks
nsights (Beta)
Certified Connectors
Template Apps
Connect
Cancel

Azure Databricks
Server Hostname
adb-3 .azuredatabricks.net
HTTP Path
sql/protocolv1/o/ 9
Advanced Options (optional)
Database (optional)
Example: abc
Batch Size (rows) (optional)
Example: 10000
Data Connectivity mode
Import
DirectQuery
OK
Cancel


Untitled - Power BI Desktop
Phani Raj Rusum Nagesh (MSIT)
File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Count of State by Color
16.45K (16.45%)
16.93K (16.93%)
16.51K (16.51%)
16.8K (16.8%)
16.62K (16.62%)
16.7K (16.7%)
Color
Green
Yellow
Orange
Red
Brown
Blue
Filters
Search
Filters on this visual
Color is (All)
Count of State is (All)
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Legend
Add data fields here
Details
Color
Values
Count of State
Tooltips
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Fields
Search
mnmdata
Color
Count
State
Page 1
Page 1 of 1
Update available (click to download)

Chapter 2: Reading and Writing Data from and to Various Azure Services and File Formats



Register an application ...

* Name

The user-facing display name for this application (this can be changed later).

ADLSGen2App

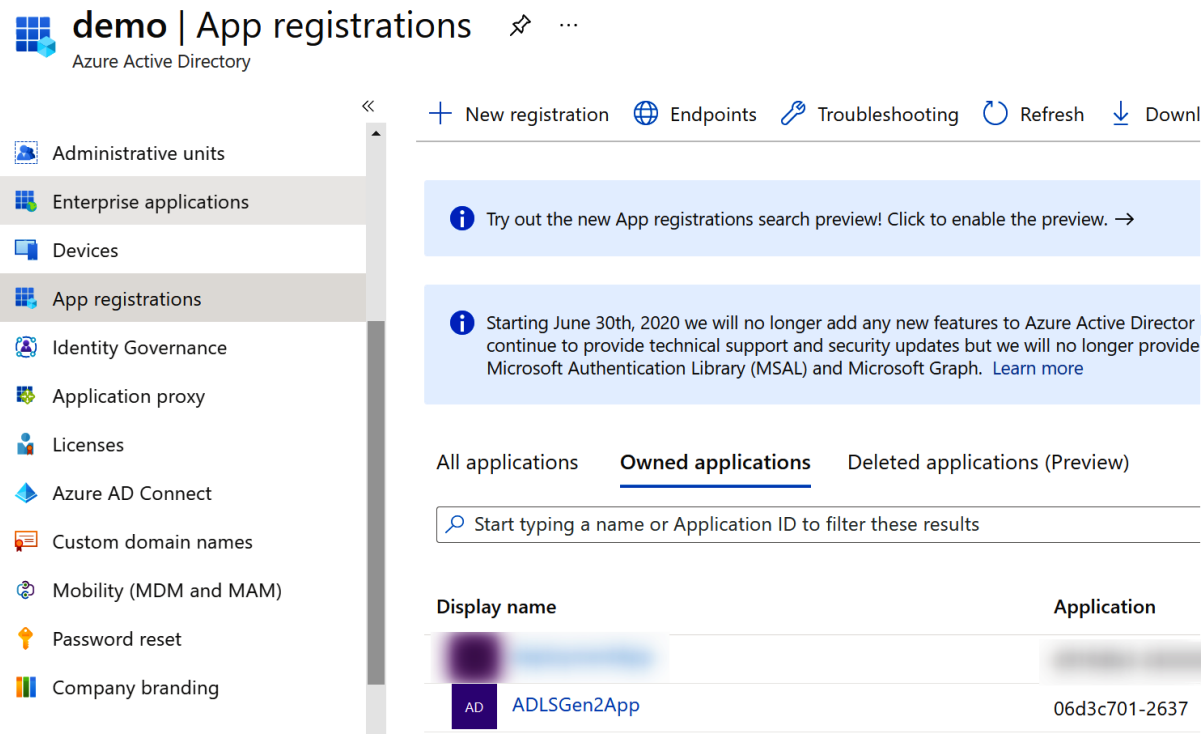
Supported account types

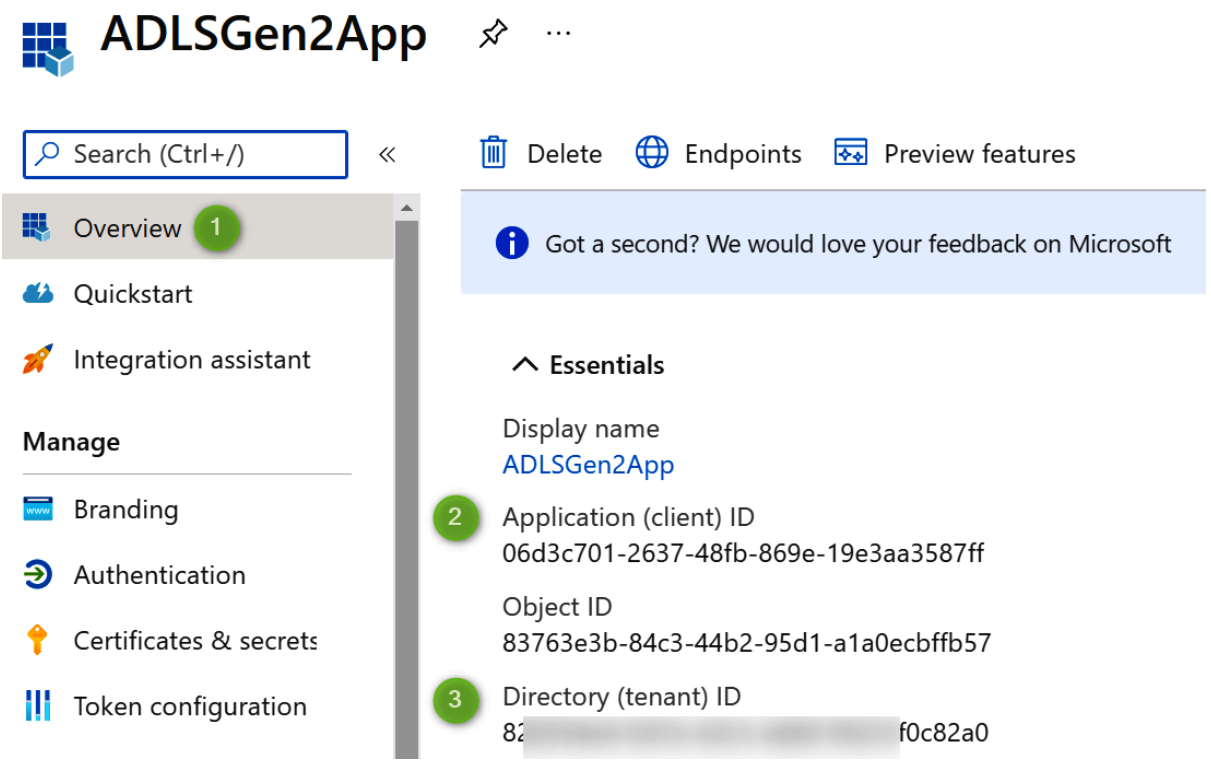
Who can use this application or access this API?

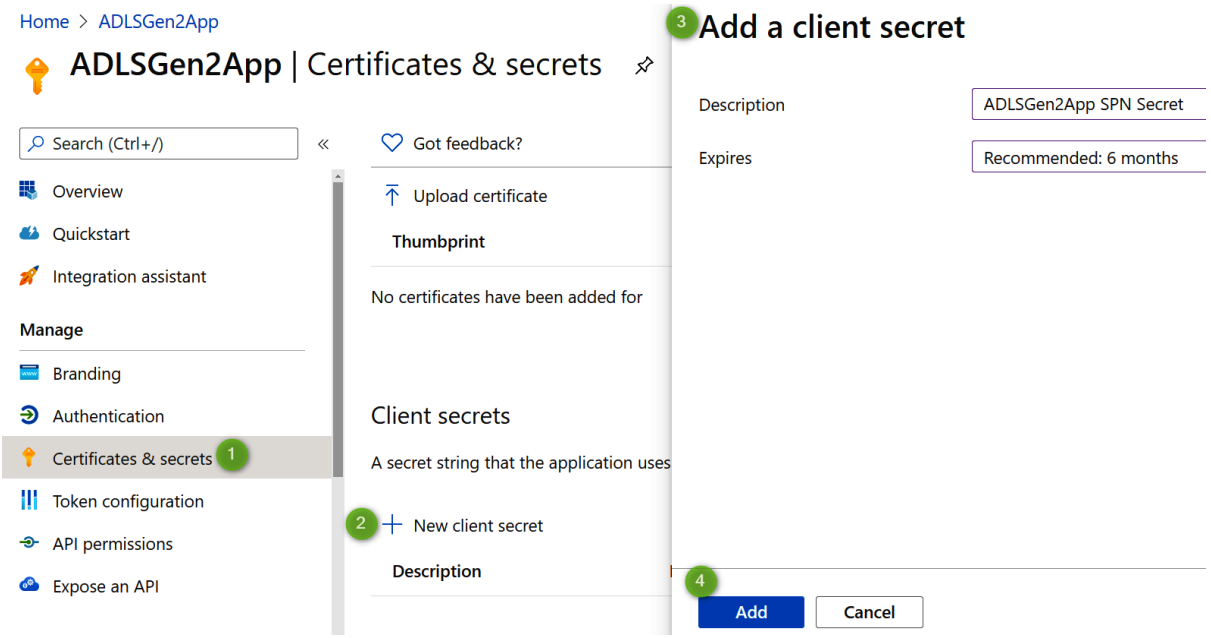
- (•) Accounts in this organizational directory only (demo only - Single tenant)
- ( ) Accounts in any organizational directory (Any Azure AD directory - Multitenant)
- ( ) Accounts in any organizational directory (Any Azure AD directory - Multitenant)
- ( ) Personal Microsoft accounts only

By proceeding, you agree to the Microsoft Platform Policies

Register





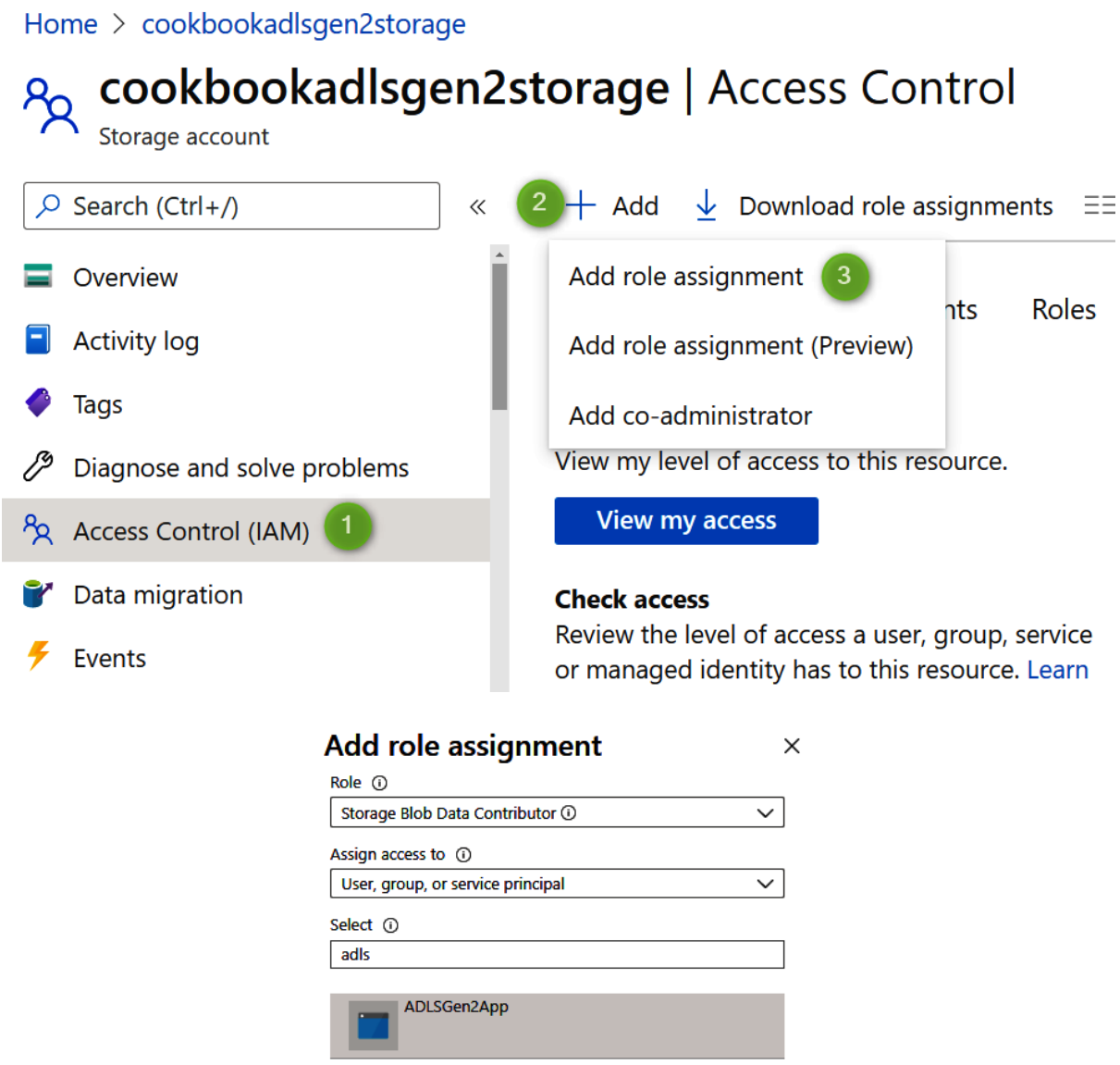


Client secrets

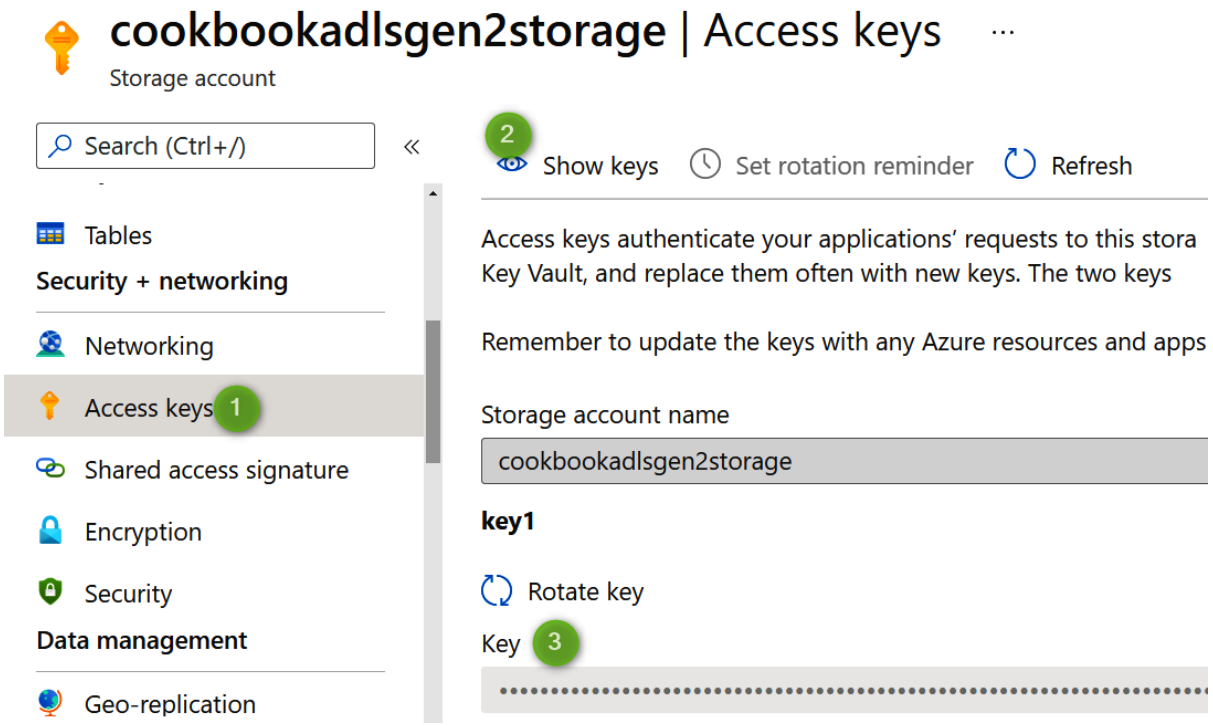
A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.

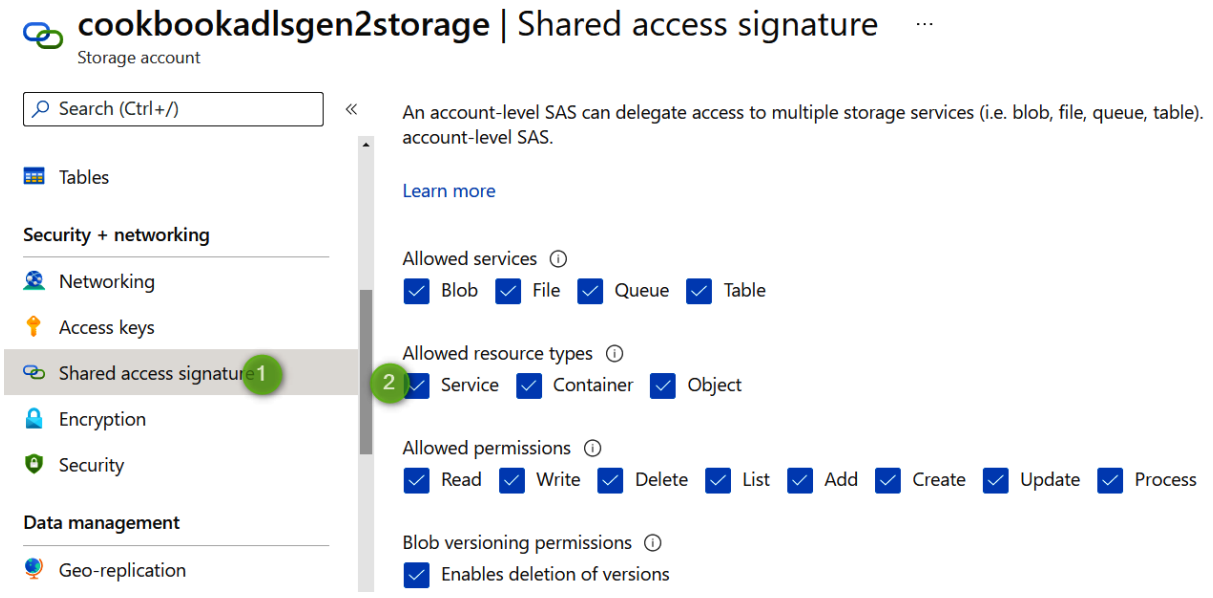
New client secret

Description	Expires	Value	Secret ID
ADLSGen2App SPN Secret	1/12/2022	H...	63e4d48d-cbb2-4a86-b52d-...


Home > cookbookadlsgen2storage
cookbookadlsgen2storage | Access Control
Storage account
Search (Ctrl+/)
2
Add
Download role assignments
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
1
Data migration
Events
Add role assignment
3
Add role assignment (Preview)
Add co-administrator
Roles
View my level of access to this resource.
View my access
Check access
Review the level of access a user, group, service or managed identity has to this resource. Learn

Add role assignment
Role
Storage Blob Data Contributor
Assign access to
User, group, or service principal
Select
adls
ADLSGen2App


cookbookadlsgen2storage | Access keys
Storage account
Search (Ctrl+/)
Tables
Security + networking
Networking
Access keys
1
Shared access signature
Encryption
Security
Data management
Geo-replication
2
Show keys
Set rotation reminder
Refresh
Access keys authenticate your applications' requests to this stora
Key Vault, and replace them often with new keys. The two keys
Remember to update the keys with any Azure resources and apps
Storage account name
cookbookadlsgen2storage
key1
Rotate key
Key
3


cookbookadlsgen2storage | Shared access signature
Storage account
Search (Ctrl+/)
Tables
Security + networking
Networking
Access keys
Shared access signature
1
Encryption
Security
Data management
Geo-replication
2
An account-level SAS can delegate access to multiple storage services (i.e. blob, file, queue, table).
account-level SAS.
Learn more
Allowed services
Blob
File
Queue
Table
Allowed resource types
Service
Container
Object
Allowed permissions
Read
Write
Delete
List
Add
Create
Update
Process
Blob versioning permissions
Enables deletion of versions

Generate SAS and connection string

Connection string

BlobEndpoint=https://cookbookadlsgen2storage.blob.core.windows.net/;QueueEndpoint=https://cookbookadlsgen2storage.queue.core.windows.net/;FileEndpo...

SAS token

?sv=2020-08-[redacted]sDLt...

Blob service SAS URL

https://cookbookadlsgen2storage.blob.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-07-12T21:43:44Z&st=2021-06-03T13:43:...

File service SAS URL

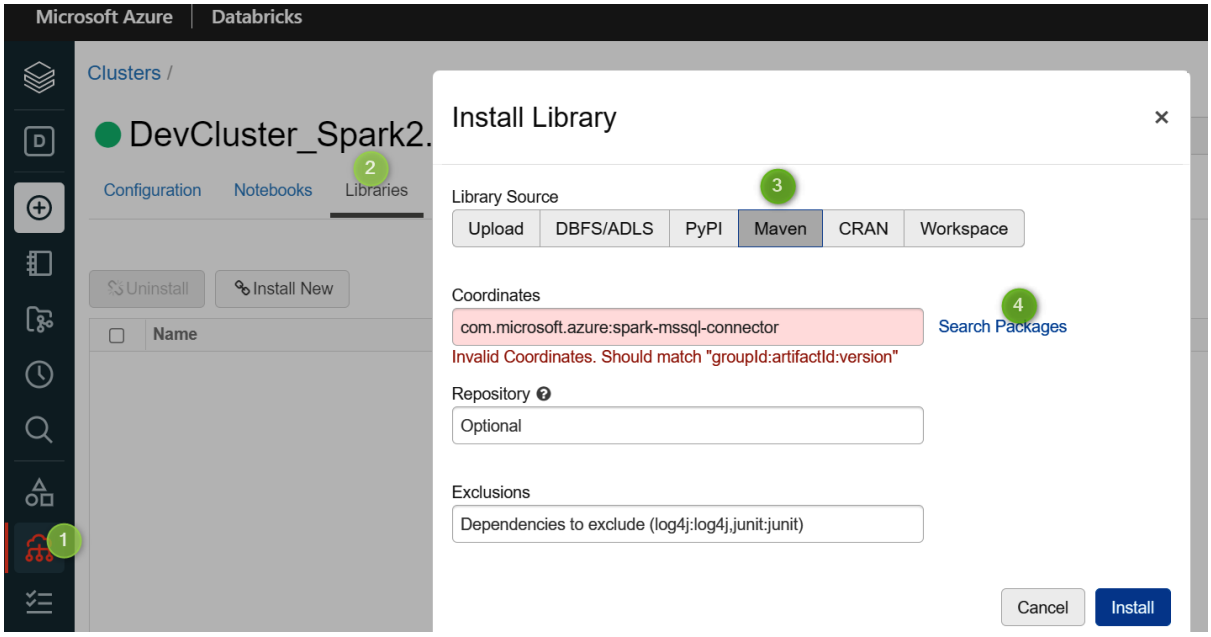
https://cookbookadlsgen2storage.file.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-07-12T21:43:44Z&st=2021-06-03T13:43:4...

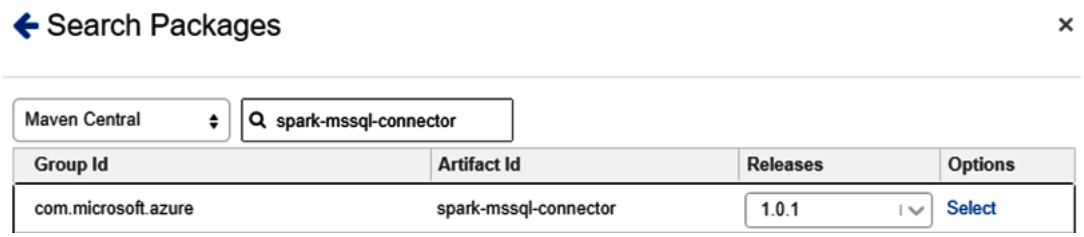
Queue service SAS URL

https://cookbookadlsgen2storage.queue.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-07-12T21:43:44Z&st=2021-06-03T13:4...

Table service SAS URL

https://cookbookadlsgen2storage.table.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-07-12T21:43:44Z&st=2021-06-03T13:43...





DevCluster_Spark2.x

Edit | Permissions | Clone | Restart

Configuration | Notebooks (0) | Libraries | Event Log | Spark UI | Driver Logs | Metrics | Apps | Spark

Uninstall | Install New

Name	Type	Status	Source
com.microsoft.azure:spark-mssql-connector:1.0.1	Maven	Installed	

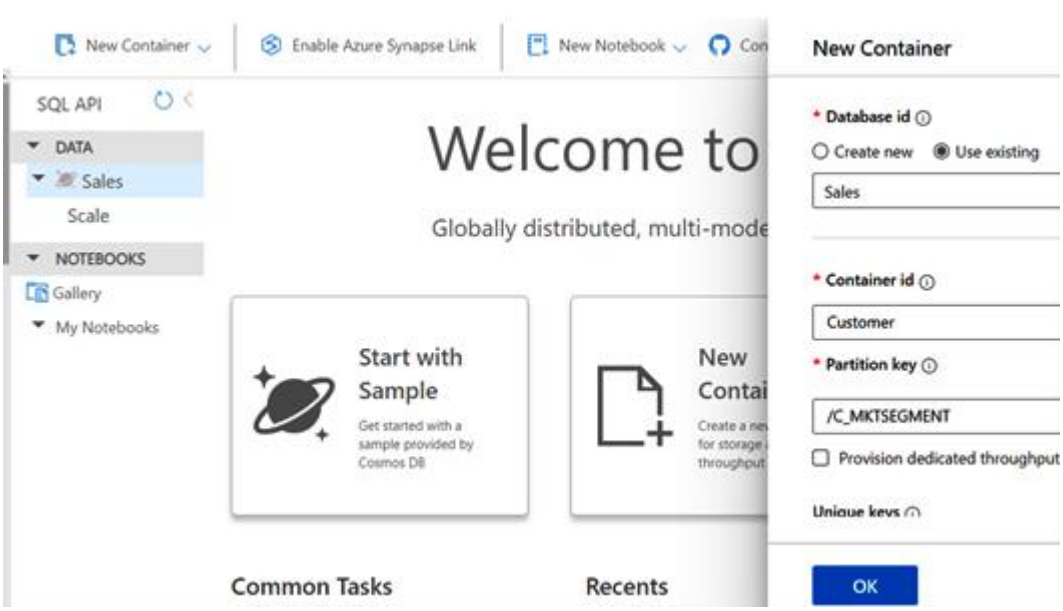
SQLQuery1.sql - az...icrosoft.com (126))*

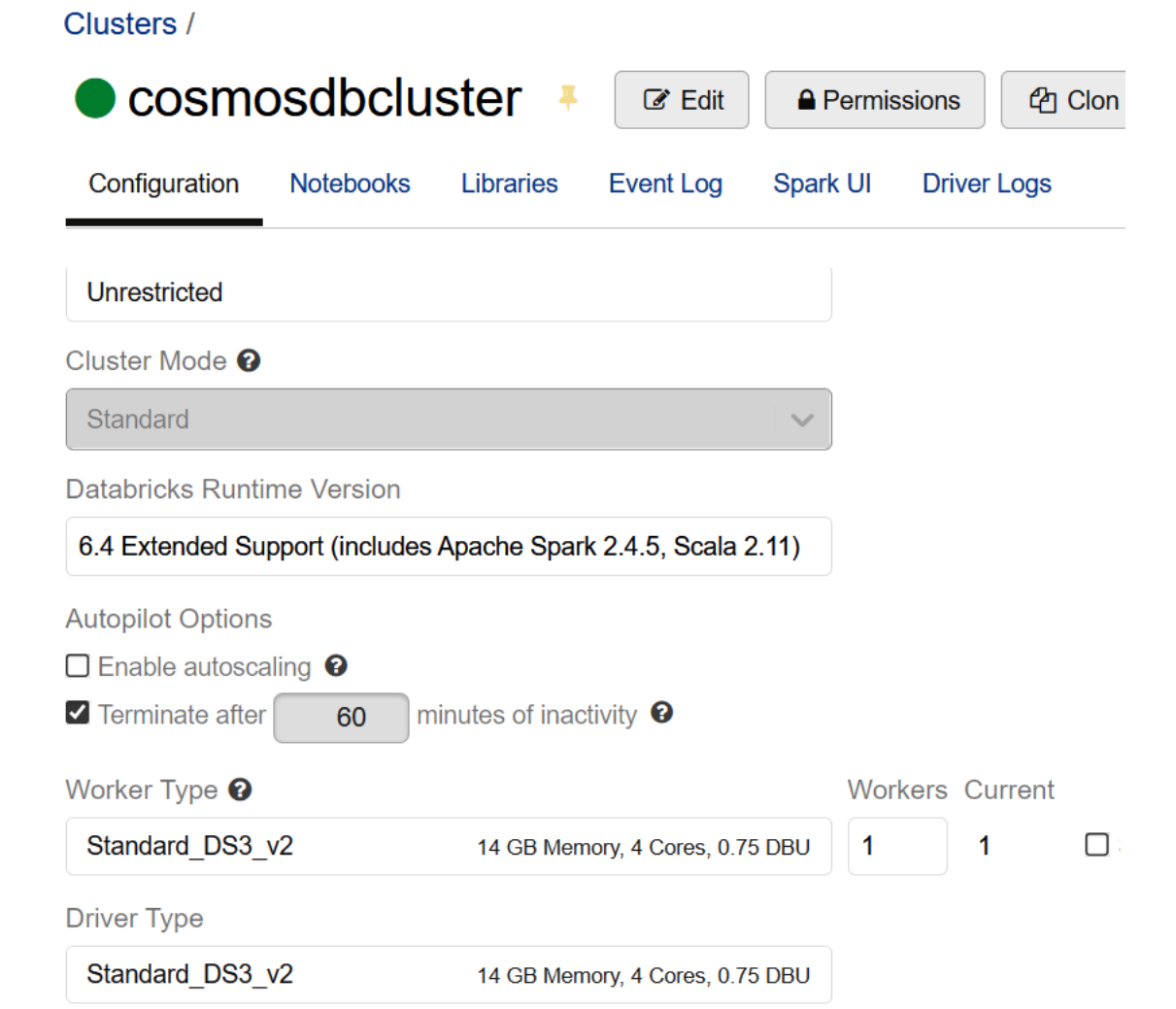
```
SELECT TOP 10 * FROM [dbo].[CustomerTable]
```

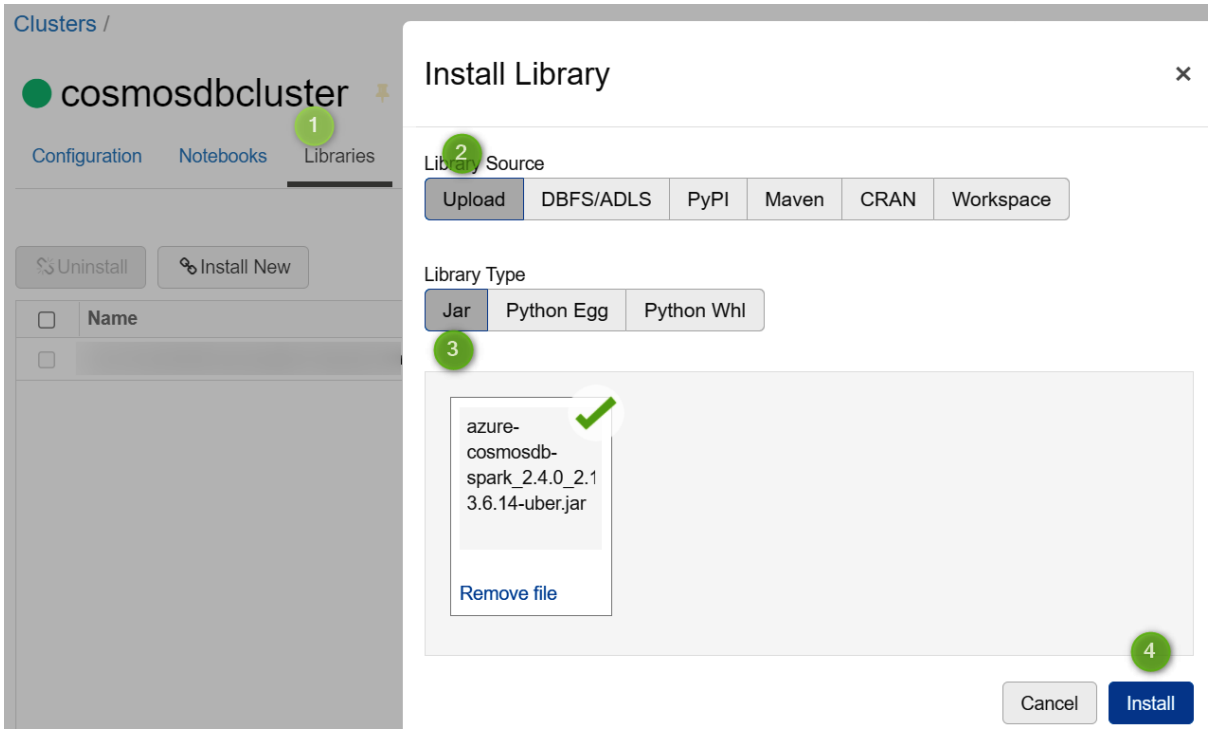
121 %

Results | Messages

	C_CUSTKEY	C_NAME	C_ADDRESS	C_NATIONKEY
1	23172	Customer#000023172	Y,GqQajAoOed9GXqdHZAuMObNEt19WJH	0
2	39243	Customer#000039243	20JFX538mRg2wrHjcHUEg g4	0
3	66522	Customer#000066522	1bJiQmqcr7mjooRovYIX	0
4	35373	Customer#000035373	5u20XRFD8JhBaLtOL3Kjh625GsK	0
5	30207	Customer#000030207	RC,iLpNQ1cCo2kSYncZJXIzze82OTrOTMxwelSO	0
6	42862	Customer#000042862	IP26TV1sfWipR9jZECG,QkXVn5o	0
7	42916	Customer#000042916	4yqldhxAihy	0
8	27274	Customer#000027274	G2satKy0awkzwk qwoL5nk O WDr6yDwm47iLdO7	0
9	39478	Customer#000039478	Lxp2V2TlYbaCh	0
10	50677	Customer#000050677	,XzNNjEz,W4MKEKBHVZWW	0




Clusters /
cosmosdbcluster
Edit
Permissions
Clon
Configuration
Notebooks
Libraries
Event Log
Spark UI
Driver Logs
Unrestricted
Cluster Mode
Standard
Databricks Runtime Version
6.4 Extended Support (includes Apache Spark 2.4.5, Scala 2.11)
Autopilot Options
Enable autoscaling
Terminate after
60
minutes of inactivity
Worker Type
Workers
Current
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU
1
1
Driver Type
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU


Clusters /
cosmosdbcluster
1
Configuration
Notebooks
Libraries
Uninstall
Install New
Name
Install Library
2
Library Source
Upload
DBFS/ADLS
PyPI
Maven
CRAN
Workspace
Library Type
Jar
Python Egg
Python Whl
3
azure-cosmosdb-spark_2.4.0_2.1 3.6.14-uber.jar
Remove file
4
Cancel
Install

Clusters /

cosmosdbcluster

Edit | Permissions | Clone

Configuration | Notebooks | Libraries | Event Log | Spark UI | Driver Logs | Metrics

Uninstall | Install New

Name	Type	Status
azure_cosmosdb_spark_2_4_0_2_11_3_6_14_...	JAR	Installed
[redacted]		

```
1  df_json = spark.read.option("multiline","true")
2  .json("dbfs:/mnt/SensorData/JsonData/SimpleJsonData/")
3  display(df_json)
```

▸ (3) Spark Jobs

▾ df_json: pyspark.sql.dataframe.DataFrame

```
Fuel: string
Transmission: string
_id: string
about: string
address: string
color: string
cost: string
currentowner: string
eventtime: string
index: long
latitude: double
longitude: double
phone: string
seatingcapacity: string
sellingcompany: string
```

```
1  df_json = spark.read.option("multiline","true")
2  .json("dbfs:/mnt/SensorData/JsonData/JsonData/")
3  display(df_json)
```

▸ (3) Spark Jobs

▸ df_json: pyspark.sql.dataframe.DataFrame = [Fuel: string, Transmission: string ...

	owners
1	▾ array ▾ 0: name: "Burton Chase" phone: "+1 (896) 412-2343" ▾ 1: name: "Williams Decker" phone: "+1 (857) 575-3797"
	▸ [{"name": "Galloway Woods", "phone": "+1 (868) 450-3041"}, {"name": "Cha

```
1  from pyspark.sql.functions import explode
2  data_df = df_json.select("_id",
3     explode("owners").alias("vehicleOwnersExplode")
4  ).select("_id","vehicleOwnersExplode.*")
5  display(data_df)
```

▸ (2) Spark Jobs

▾ data_df: pyspark.sql.dataframe.DataFrame

```
_id: string
name: string
phone: string
```

	_id	name	phone
1	60f623d0f2861584c56e2660	Burton Chase	+1 (896) 412-2343
2	60f623d0f2861584c56e2660	Williams Decker	+1 (857) 575-3797

```
jsonDF = data_df.withColumn("jsonCol",
                            to_json(struct([data_df[x] for x in data_df.columns])))
                            .select("jsonCol")
display(jsonDF)
```

▸ (2) Spark Jobs

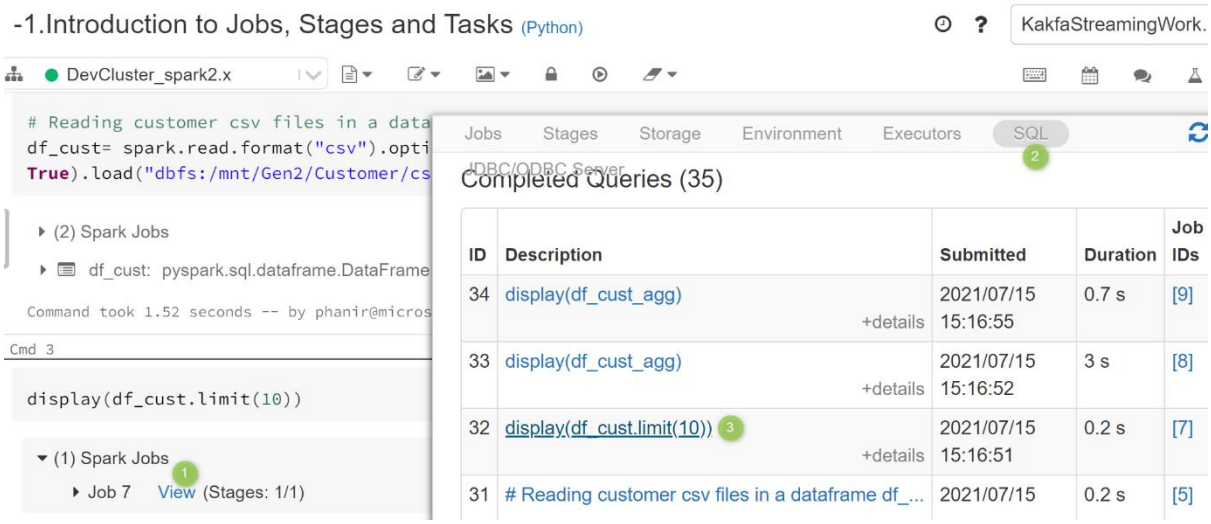
▸ jsonDF: pyspark.sql.dataframe.DataFrame = [jsonCol: string]

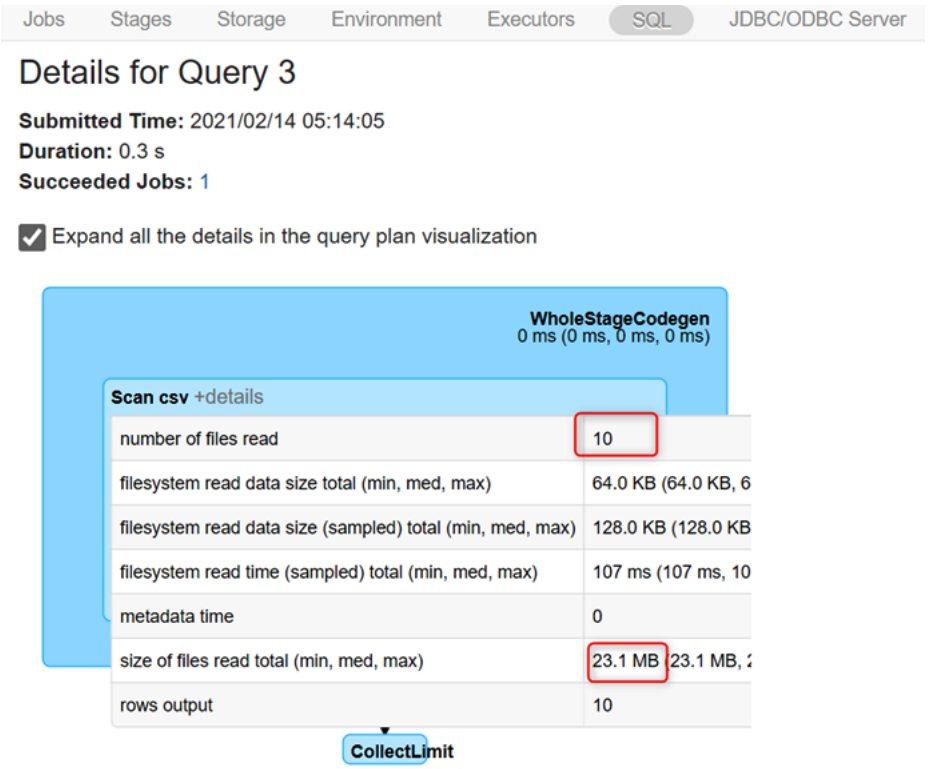
	jsonCol
1	{"_id":"60f623d0f2861584c56e2660","name":"Burton Chase","phone":"+1 (896) 412-2343"}
2	{"_id":"60f623d0f2861584c56e2660","name":"Williams Decker","phone":"+1 (857) 575-3797"}
3	{"_id":"60f623d0b164af3c012566e9","name":"Galloway Woods","phone":"+1 (868) 450-3041"}
4	{"_id":"60f623d0b164af3c012566e9","name":"Chaney Martin","phone":"+1 (996) 430-3875"}
5	{"_id":"60f623d0722a2004debafe2a","name":"Susanna Weber","phone":"+1 (919) 431-2833"}

Component	Version
Apache Spark	2.4.x, 2.3.x, 2.2.x, and 2.1.x
Scala	2.11
Azure Databricks runtime version	> 3.4

Connector	Maven Coordinate	Scala Version
Spark 2.4.x compatible connnector	com.microsoft.azure:spark-mssql-connector:1.0.2	2.11
Spark 3.0.x compatible connnector	com.microsoft.azure:spark-mssql-connector_2.12:1.1.0	2.12

Chapter 3: Understanding Spark Query Execution





```
1 display(df_cust.limit(10))
```

▾ (1) Spark Jobs
▾ Job 1 View (Stages: 1/1)
Stage 1: 1/1

Summary Metrics for 1 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.2 s	0.2 s	0.2 s	0.2 s	0.2 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Input Size / Records	64.0 KB / 10	64.0 KB / 10	64.0 KB / 10	64.0 KB / 10	64.0 KB / 10

▾Aggregated Metrics by Executor

Executor ID ▲	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Input Size / Records	Blacklisted
0 stdout stderr	10.139.64.5:42133	0.3 s	1	0	0	1	64.0 KB / 10	false

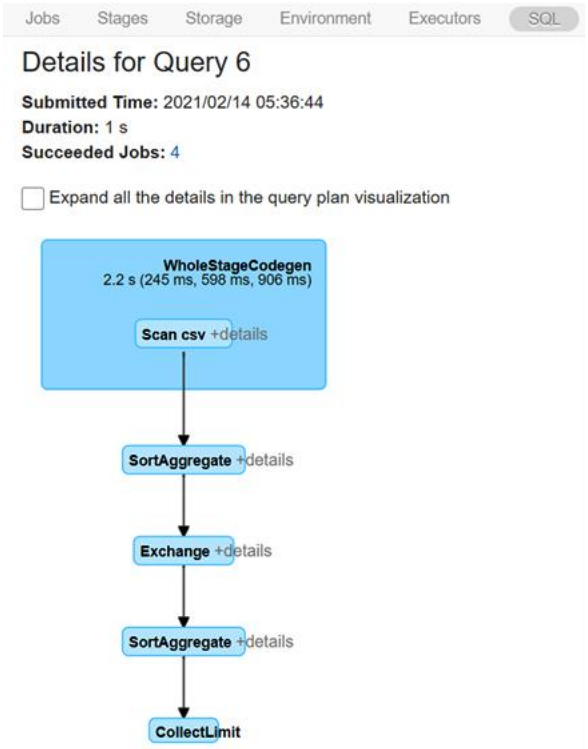
▾Tasks (1)

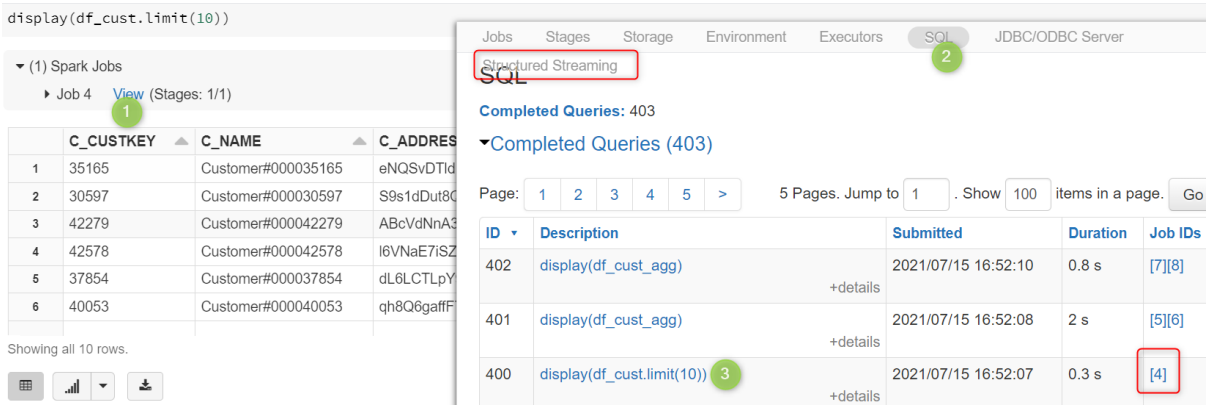
Index ▲	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Input Size / Records	Errors
0	1	0	SUCCESS	PROCESS_LOCAL	0	10.139.64.5 stdout stderr	2021/02/14 05:14:05	0.2 s		64.0 KB / 10	

```
1 display(df_cust_agg)
```

▾ (1) Spark Jobs
▾ Job 2 View (Stages: 2/2)
Stage 2: 4/4
Stage 3: 200/200

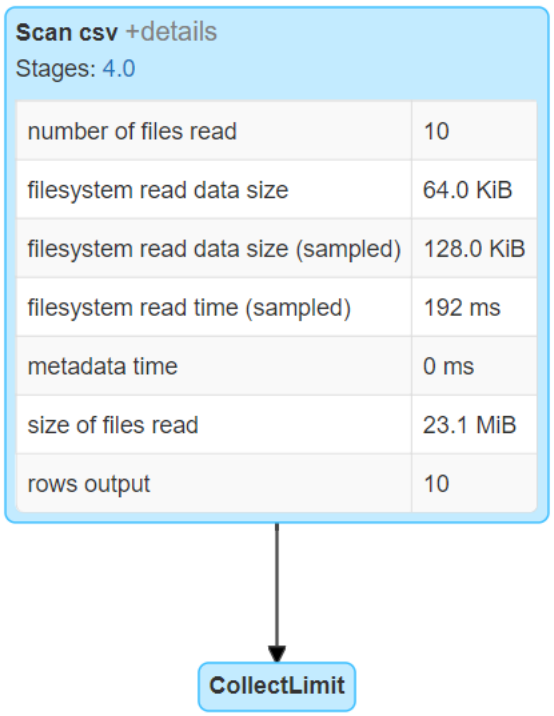
	C_MKTSEGMENT ▲	sum_acctbal ▲	avg_acctbal ▲	max_bonus ▲
1	BUILDING	135888621.94	4508.28153208148	9999.99
2	HOUSEHOLD	135873341.17	4500.756605717316	9999.23
3	AUTOMOBILE	133866847.09	4499.423470354938	9999.96
4	MACHINERY	134438861.67	4488.92656415906	9999.64

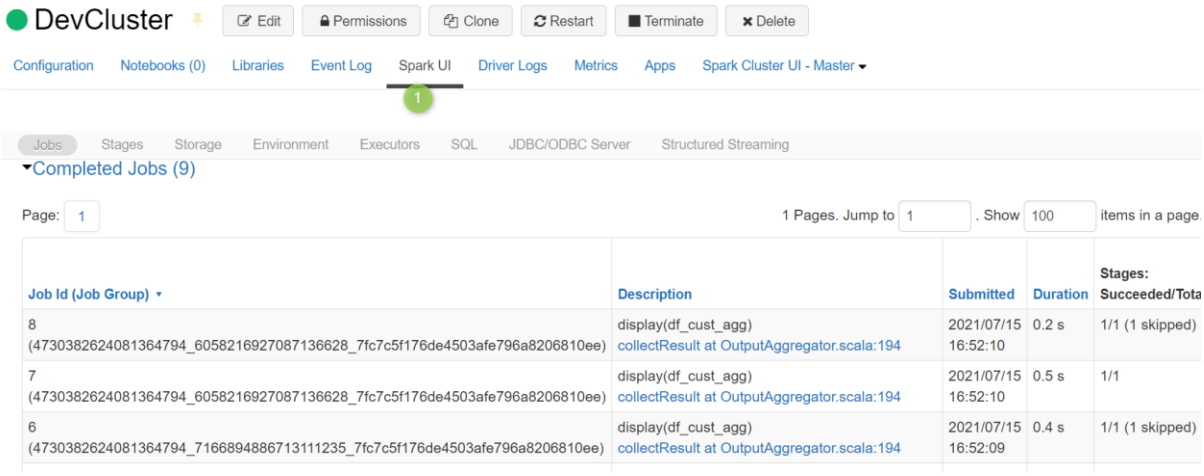

Jobs
Stages
Storage
Environment
Executors
SQL
Details for Query 6
Submitted Time: 2021/02/14 05:36:44
Duration: 1 s
Succeeded Jobs: 4
Expand all the details in the query plan visualization
WholeStageCodegen
2.2 s (245 ms, 598 ms, 906 ms)
Scan csv +details
SortAggregate +details
Exchange +details
SortAggregate +details
CollectLimit

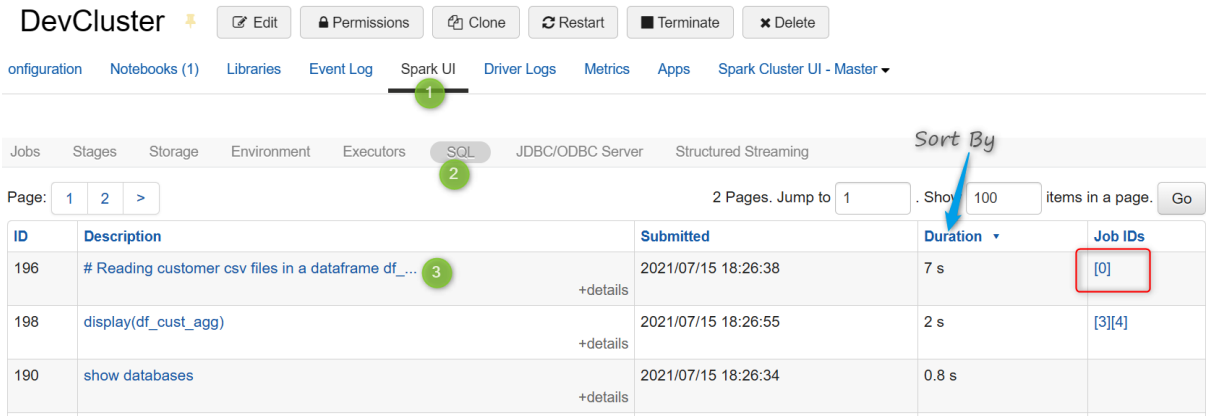

display(df_cust.limit(10))
(1) Spark Jobs
Job 4 View (Stages: 1/1)
1
C_CUSTKEY
C_NAME
C_ADDRES
1 35165 Customer#000035165 eNQSvDTld
2 30597 Customer#000030597 S9s1dDut8
3 42279 Customer#000042279 ABcVdNnA3
4 42578 Customer#000042578 l6VNaE7iSZ
5 37854 Customer#000037854 dL6LCTLpY
6 40053 Customer#000040053 qh8Q6gaffF
Showing all 10 rows.
Jobs
Stages
Storage
Environment
Executors
SQL
JDBC/ODBC Server
2
Structured Streaming
SQL
Completed Queries: 403
Completed Queries (403)
Page: 1 2 3 4 5 >
5 Pages. Jump to 1 . Show 100 items in a page. Go
ID
Description
Submitted
Duration
Job IDs
402 display(df_cust_agg) +details 2021/07/15 16:52:10 0.8 s [7][8]
401 display(df_cust_agg) +details 2021/07/15 16:52:08 2 s [5][6]
400 display(df_cust.limit(10)) 3 +details 2021/07/15 16:52:07 0.3 s [4]

Succeeded Jobs:

☑ Expand all the details in the query plan visualization







Associated SQL Query: 196
Job Group: 8287089703055764537_9154657314170381234_d0ddf8fc1a4c456f932860dfd7316f42
Completed Stages: 1

▸ Event Timeline
▸ DAG Visualization

▾ Completed Stages (1)

Page: 1 1 Pages. Jump to 1 . Show

Stage Id ▾	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input
0	8287089703055764537	# Reading customer csv files in a dataframe df_... load at NativeMethodAccessorImpl.java:0 +details	2021/07/15 18:26:38	7 s	1/1	64.0 KiB

Jobs Stages Storage Environment Executors SQL JDBC/ODBC Server Structured Streaming

▸ Aggregated Metrics by Executor

Tasks (10)

Show 20 entries Search:

Index ▲	Task ID	Attempt	Status	Locality level	Executor ID	Host	Logs	Launch Time	Duration	GC Time	Input Size / Records	Write Time	Shuffle Write Size / Records
0	23	0	SUCCESS	PROCESS_LOCAL	1	10.1.128.5	stdout stderr	2021-07-15 23:56:57	0.3 s		2.3 MiB / 15000	0.0 ms	75 B / 1
1	24	0	SUCCESS	PROCESS_LOCAL	0	10.1.128.4	stdout stderr	2021-07-15 23:56:57	0.2 s		2.3 MiB / 14999	0.0 ms	75 B / 1
2	25	0	SUCCESS	PROCESS_LOCAL	1	10.1.128.5	stdout stderr	2021-07-15 23:56:57	0.3 s		2.3 MiB / 15000	0.0 ms	75 B / 1

DevCluster Edit Permissions Clone Restart Terminate Delete

Configuration Notebooks (1) Libraries Event Log Spark UI Driver Logs Metrics Apps Spark Cluster UI - Master

Jobs Stages Storage Environment Executors SQL JDBC/ODBC Server Structured Streaming

Executor ID ▲	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read
0	10.1.128.4:41935	Active	0	22.1 KiB / 3.6 GiB	0.0 B	4	0	0	17	17	39 s (2 s)	37 MiB	750 B
driver	10.1.128.6:42891	Active	0	22.1 KiB / 3.3 GiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B
1	10.1.128.5:38627	Active	0	22.1 KiB / 3.6 GiB	0.0 B	4	0	0	17	17	28 s (2 s)	32.5 MiB	4.8 KiB

```
# Reading customer csf files in a dataframe with specify
df_cust= spark.read.format("csv").option("header",True).
True).load("dbfs:/mnt/Gen2/Customer/csvFiles")
```

▾ (2) Spark Jobs
▸ Job 7 View (Stages: 1/1)
▸ Job 8 View (Stages: 1/1)

▸ df_cust: pyspark.sql.dataframe.DataFrame = [C_CUSTKEY: intege

Cmd 6

```
#Creating the dataframe by specifying the schema using .schem() option. We don;t see any DAG getting
created with schema is specified
df_cust_sch=
spark.read.format("csv").option("header",True).schema(cust_schema).load("/mnt/Gen2/Customer/csvFiles/")
```

▸ df_cust_sch: pyspark.sql.dataframe.DataFrame = [C_CUSTKEY: integer, C_NAME: string ... 6 more fields]

Command took 0.25 seconds -- by [redacted] at [redacted]/2021, 12:23:25 AM on DevCluster

Cmd 4

```
1 # Getting the number of partitions
2 print(f" Number of partitions with all files read = {df_cust.rdd.getNumPartitions()}")
```

```
Number of partitions with all files read = 10
```

Cmd 8

```
1 #Getting the number of partitions by specifying only one csv file.
2 print(f" Number of partitions with 1 file read  = {df_cust_sch.rdd.getNumPartitions()}")
```

```
Number of partitions with 1 file read  = 1
```

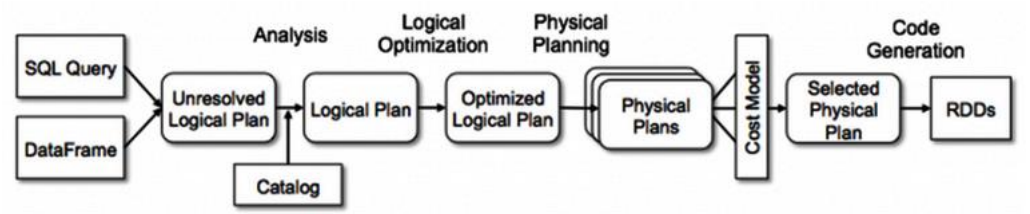
Jobs Stages Storage Environment Executors SQL JDBC/ODBC Server Structured Streaming

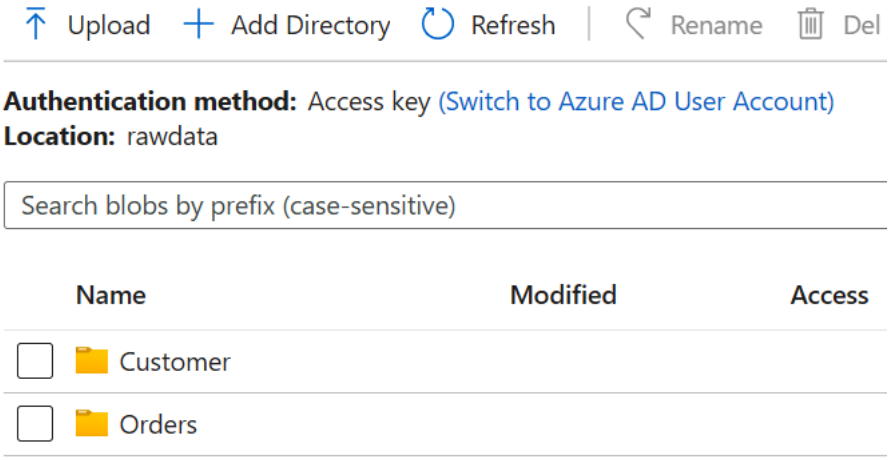
▾Details

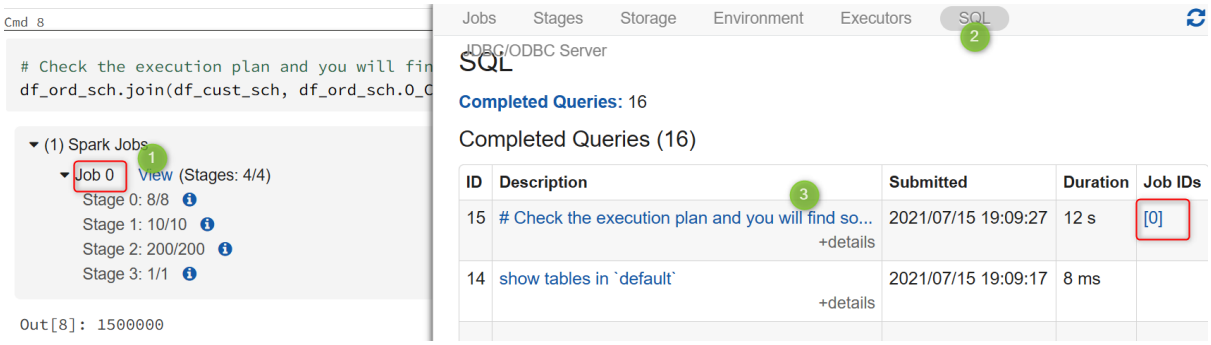
```
== Parsed Logical Plan ==
GlobalLimit 21
+- LocalLimit 21
   +- Project [cast(C_MKTSEGMENT#2240 as string) AS C_MKTSEGMENT#2317, cast(AvgAcctBal#2284 as string) AS AvgAcctBa
      +- Filter (AvgAcctBal#2284 > cast(4500 as double))
         +- Filter (C_MKTSEGMENT#2240 = MACHINERY)
            +- Aggregate [C_MKTSEGMENT#2240], [C_MKTSEGMENT#2240, avg(C_ACCTBAL#2239) AS AvgAcctBal#2284]
               +- Relation[C_CUSTKEY#2234,C_NAME#2235,C_ADDRESS#2236,C_NATIONKEY#2237,C_PHONE#2238,C_ACCTBAL#2239,C

== Analyzed Logical Plan ==
C_MKTSEGMENT: string, AvgAcctBal: string
GlobalLimit 21
+- LocalLimit 21
   +- Project [cast(C_MKTSEGMENT#2240 as string) AS C_MKTSEGMENT#2317, cast(AvgAcctBal#2284 as string) AS AvgAcctBa
      +- Filter (AvgAcctBal#2284 > cast(4500 as double))
         +- Filter (C_MKTSEGMENT#2240 = MACHINERY)
```

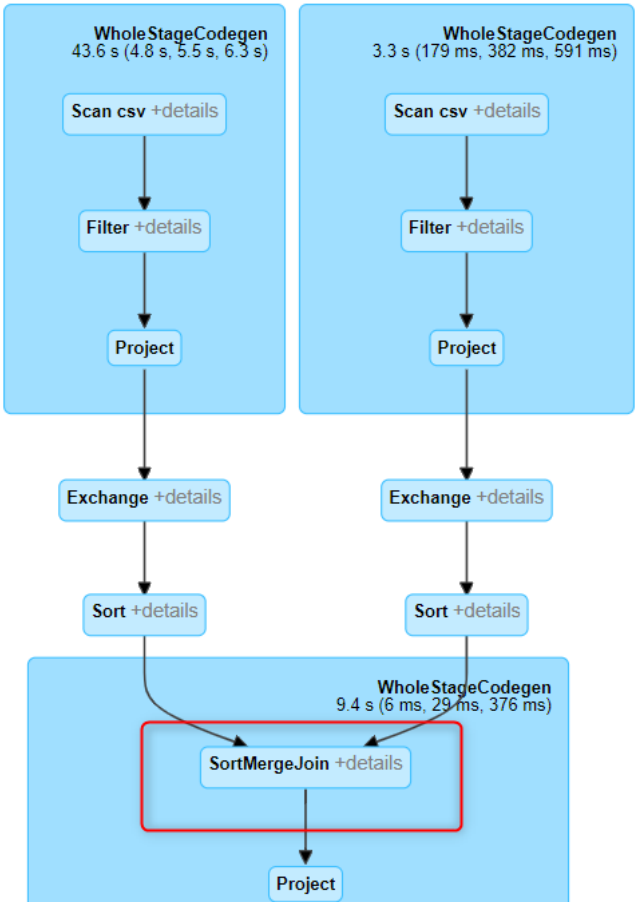
```
== Physical Plan ==
AdaptiveSparkPlan isFinalPlan=false
+- == Current Plan ==
   HashAggregate(keys=[C_MKTSEGMENT#2240], functions=[finalmerge_count(merge count#2357L) AS count(1)#2340L]
   +- Exchange hashpartitioning(C_MKTSEGMENT#2240, 200), true, [id=#2016]
      +- HashAggregate(keys=[C_MKTSEGMENT#2240], functions=[partial_count(1) AS count#2357L])
         +- FileScan csv [C_MKTSEGMENT#2240] Batched: false, DataFilters: [], Format: CSV, Location: InMemor
iles], PartitionFilters: [], PushedFilters: [], ReadSchema: struct<C_MKTSEGMENT:string>
+- == Initial Plan ==
   HashAggregate(keys=[C_MKTSEGMENT#2240], functions=[finalmerge_count(merge count#2357L) AS count(1)#2340L]
   +- Exchange hashpartitioning(C_MKTSEGMENT#2240, 200), true, [id=#2016]
      +- HashAggregate(keys=[C_MKTSEGMENT#2240], functions=[partial_count(1) AS count#2357L])
         +- FileScan csv [C_MKTSEGMENT#2240] Batched: false, DataFilters: [], Format: CSV, Location: InMemor
iles], PartitionFilters: [], PushedFilters: [], ReadSchema: struct<C_MKTSEGMENT:string>


== Physical Plan ==
AdaptiveSparkPlan isFinalPlan=false
+- == Current Plan ==
   HashAggregate(keys=[C_MKTSEGMENT#2240], functions=[finalmerge_count(merge count#2362L) AS count(1)#2352L]
   +- Exchange hashpartitioning(C_MKTSEGMENT#2240, 200), true, [id=#2062]
```

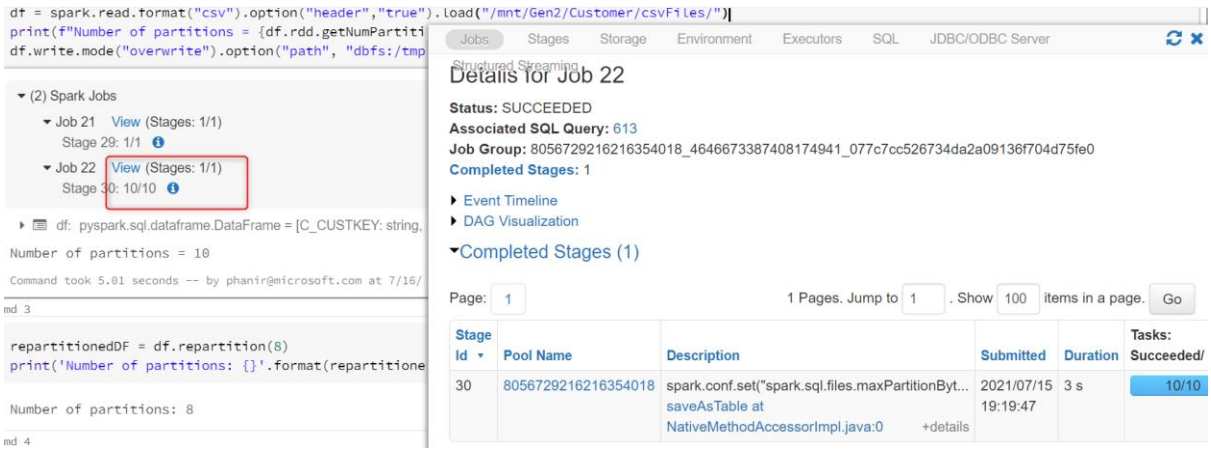
```
DataFilters: [isnotnull(C_MKTSEGMENT#370), (C_MKTSEGMENT#370 = MACHINERY)], Format: CSV, Location:
InMemoryFileIndex[dbfs:/mnt/Gen2/Customer/csvFiles], PartitionFilters: [], PushedFilters:
[IsNotNull(C_MKTSEGMENT), EqualTo(C_MKTSEGMENT,MACHINERY)], ReadSchema:
struct<C_ACCTBAL:double,C_MKTSEGMENT:string>
```






Cmd 8
Check the execution plan and you will fin
df_ord_sch.join(df_cust_sch, df_ord_sch.O_C
(1) Spark Jobs
Job 0 View (Stages: 4/4)
Stage 0: 8/8
Stage 1: 10/10
Stage 2: 200/200
Stage 3: 1/1
Out[8]: 1500000
1
Jobs Stages Storage Environment Executors SQL
2
JDBC/ODBC Server
SQL
Completed Queries: 16
Completed Queries (16)
ID Description Submitted Duration Job IDs
3
15 # Check the execution plan and you will find so... +details 2021/07/15 19:09:27 12 s [0]
14 show tables in `default` +details 2021/07/15 19:09:17 8 ms


WholeStageCodegen
43.6 s (4.8 s, 5.5 s, 6.3 s)
Scan csv +details
Filter +details
Project
WholeStageCodegen
3.3 s (179 ms, 382 ms, 591 ms)
Scan csv +details
Filter +details
Project
Exchange +details
Exchange +details
Sort +details
Sort +details
WholeStageCodegen
9.4 s (6 ms, 29 ms, 376 ms)
SortMergeJoin +details
Project


df = spark.read.format("csv").option("header","true").load("/mnt/Gen2/Customer/csvFiles/")
print(f"Number of partitions = {df.rdd.getNumPartiti
df.write.mode("overwrite").option("path", "dbfs:/tmp
(2) Spark Jobs
Job 21 View (Stages: 1/1)
Stage 29: 1/1
Job 22 View (Stages: 1/1)
Stage 30: 10/10
df: pyspark.sql.dataframe.DataFrame = [C_CUSTKEY: string,
Number of partitions = 10
Command took 5.01 seconds -- by phanir@microsoft.com at 7/16/
md 3
repartitionedDF = df.repartition(8)
print('Number of partitions: {}'.format(repartitione
Number of partitions: 8
md 4
Jobs Stages Storage Environment Executors SQL JDBC/ODBC Server
Structured Streaming
Details for Job 22
Status: SUCCEEDED
Associated SQL Query: 613
Job Group: 805672921621635401 8_4646673387408174941_077c7cc526734da2a09136f704d75fe0
Completed Stages: 1
Event Timeline
DAG Visualization
Completed Stages (1)
Page: 1
1 Pages. Jump to 1 . Show 100 items in a page. Go
Stage Id Pool Name Description Submitted Duration Tasks: Succeeded/
30 805672921621635401 spark.conf.set("spark.sql.files.maxPartitionByt... saveAsTable at NativeMethodAccessorImpl.java:0 +details 2021/07/15 19:19:47 3 s 10/10

```
# Check the duration of execution with default 200 partitions
spark.conf.set("spark.sql.shuffle.partitions", 200)
mktSegmentDF = df.groupBy("C_MKTSEGMENT").count().collect()
```

▸ (2) Spark Jobs

Command took 1.43 seconds -- by [redacted] on CookBookCluster

```
# Check the duration of execution with default 30 partitions
spark.conf.set("spark.sql.shuffle.partitions", 30)
mktSegmentDF = df.groupBy("C_MKTSEGMENT").count().collect()
```

▸ (2) Spark Jobs

Command took 0.61 seconds -- [redacted] on CookBookCluster

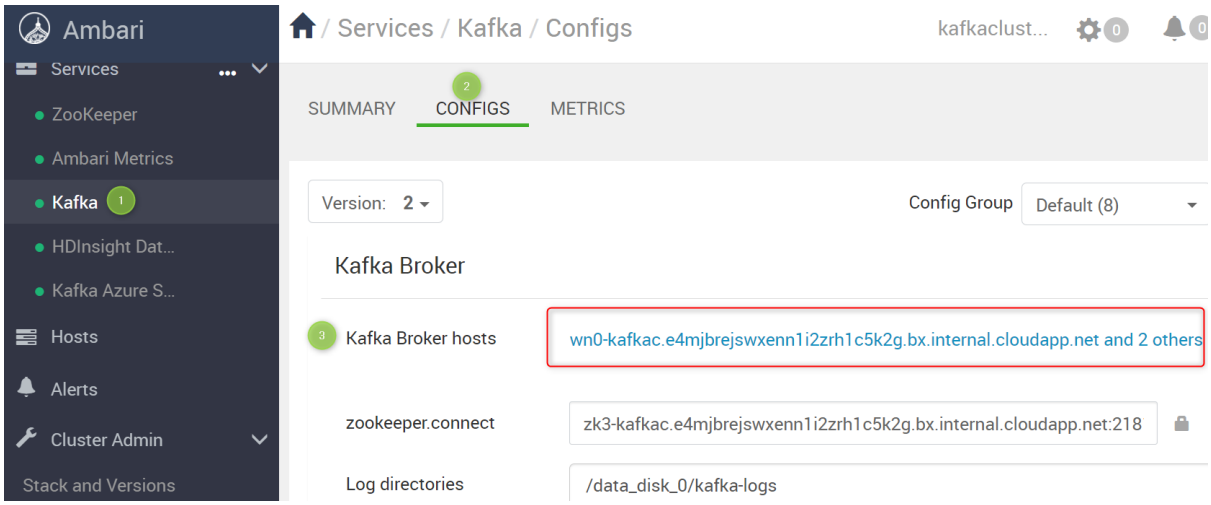
```
sortDF = df.sort('C_CUSTKEY')
display(sortDF)
```

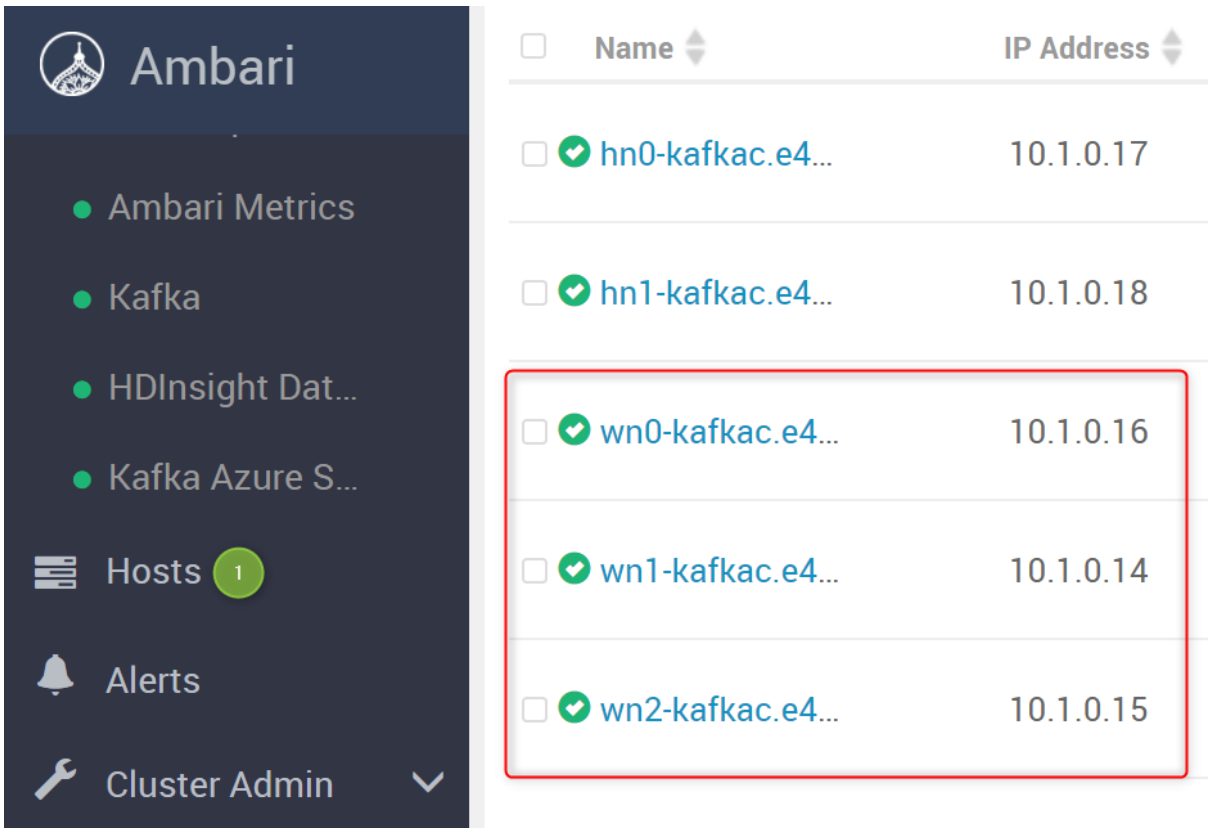
▸ (1) Spark Jobs

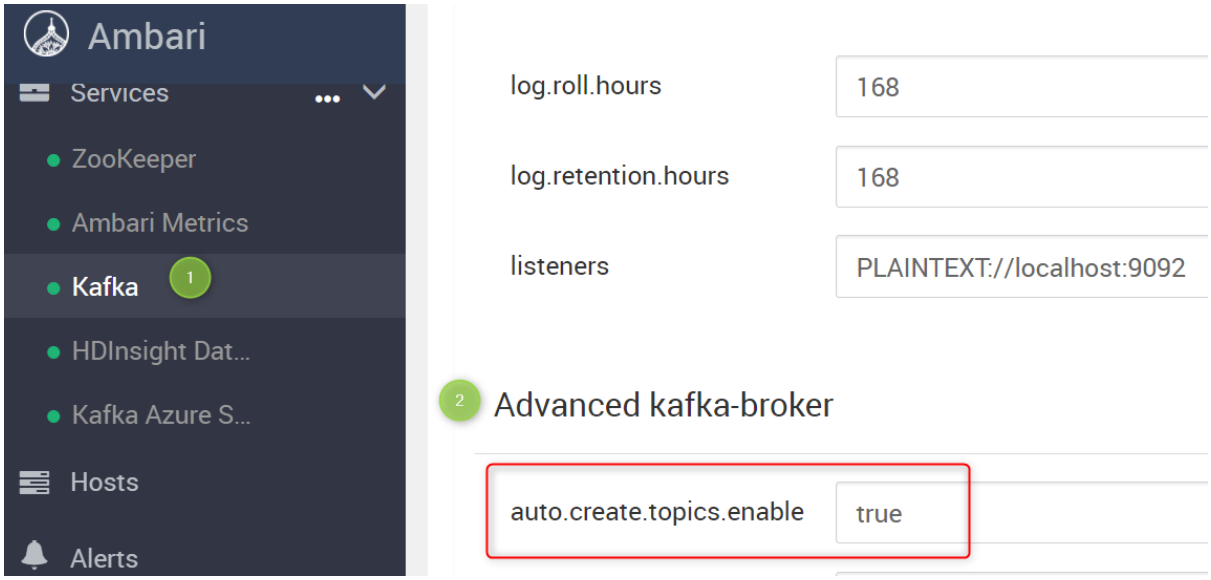
▸ sortDF: pyspark.sql.dataframe.DataFrame = [C_CUSTKEY: string, C_NAME: string ... 6 more fields]

	C_CUSTKEY	C_NAME	C_ADDRESS	C_NATIONKEY	C_PHONE
1	1	Customer#000000001	IVhzIApeRb ot,c,E	0	10-989-741-2988
2	10	Customer#000000010	Lbrg3ElDieI0B10bB0Aymm	0	10-741-346-9870
3	100	Customer#000000100	mFowIuhnHjp2GjCiYYavkW kUwOjIaTCQ	0	10-749-445-4907
4	1000	Customer#000001000	Ne,bJD40ae8ZXo XFrOkEmACe	0	10-730-275-2976
5	10000	Customer#000010000	OI7fOBe9O7SEIX7xSxqR	0	10-923-910-8193
6	100000	Customer#000100000	GIw,RrxSCw6UDA5	0	10-488-516-6598

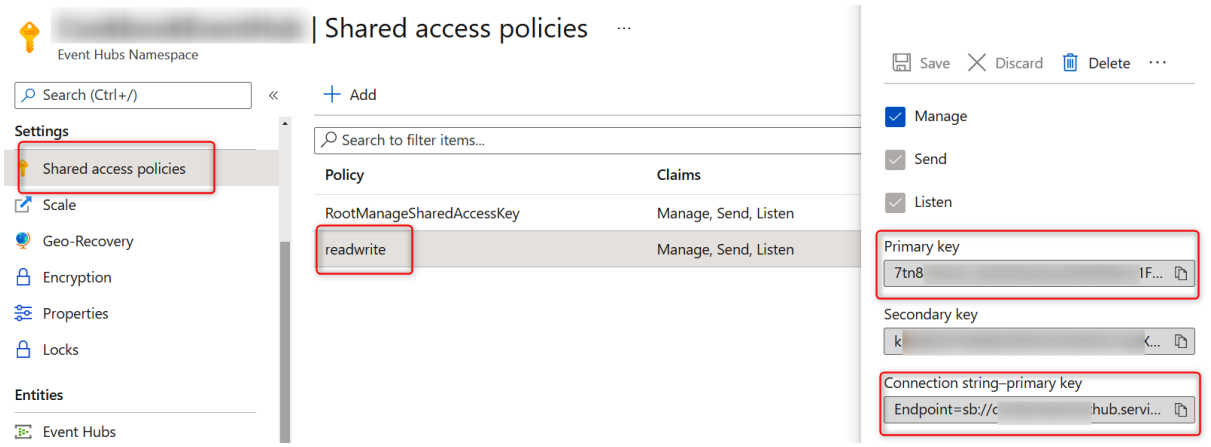
Chapter 4: Working with Streaming Data

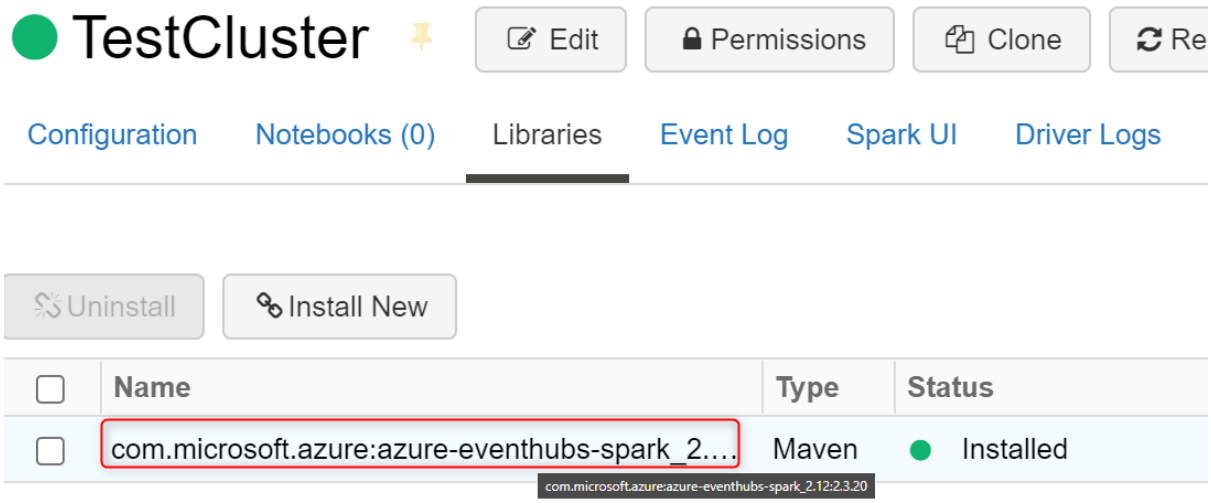






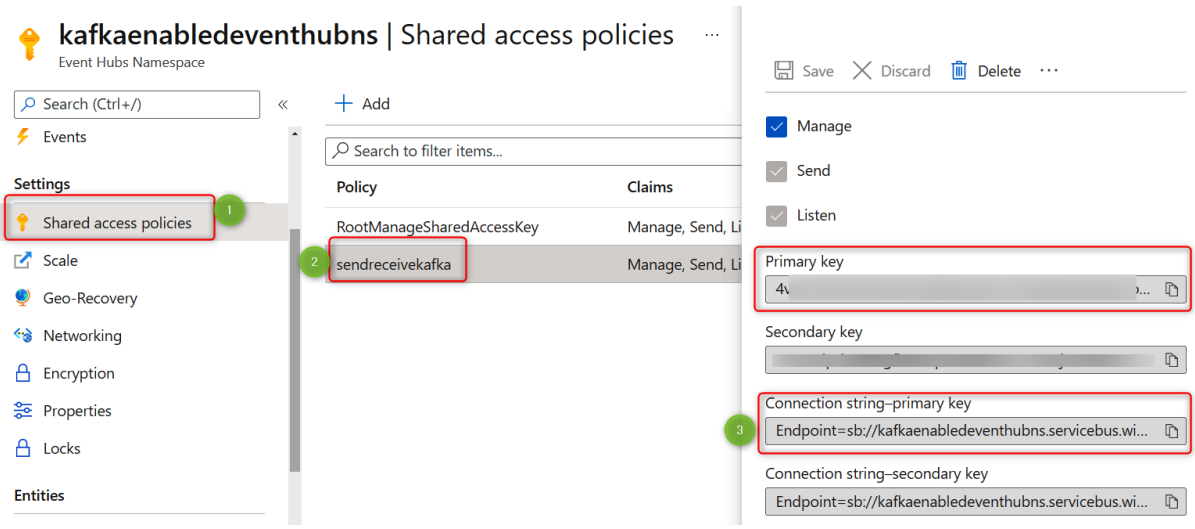

```
producer = KafkaProducer(bootstrap_servers=['10.0.0.16:9092','10.1.0.14:9092',
                                            '10.1.0.15:9092'],
                         value_serializer=lambda x:
                         dumps(x).encode('utf-8'))
```





```
print( EventHubdf.isStreaming)
print( EventHubdf.printSchema())
```

```
True
root
 |-- body: binary (nullable = true)
 |-- partition: string (nullable = true)
 |-- offset: string (nullable = true)
 |-- sequenceNumber: long (nullable = true)
 |-- enqueuedTime: timestamp (nullable = true)
 |-- publisher: string (nullable = true)
 |-- partitionKey: string (nullable = true)
 |-- properties: map (nullable = true)
 |    |-- key: string
 |    |-- value: string (valueContainsNull = true)
 |-- systemProperties: map (nullable = true)
 |    |-- key: string
 |    |-- value: string (valueContainsNull = true)
```



```
display(dbutils.fs.ls("/mnt/Gen2/LogFiles"))
```

▸ (2) Spark Jobs

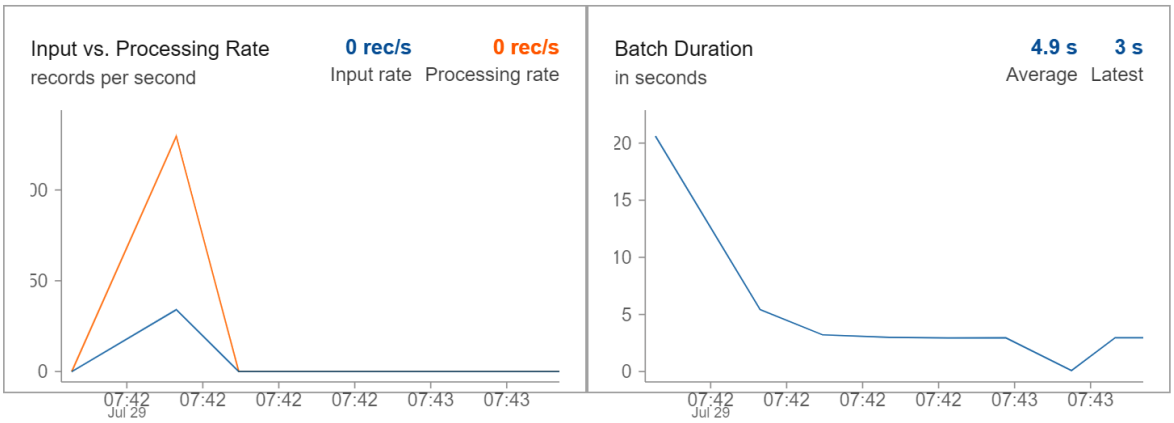
	path	name	size
1	dbfs:/mnt/Gen2/LogFiles/Log1.json	Log1.json	211
2	dbfs:/mnt/Gen2/LogFiles/Log2.json	Log2.json	209
3	dbfs:/mnt/Gen2/LogFiles/Log3.json	Log3.json	209
4	dbfs:/mnt/Gen2/LogFiles/Log4 - Copy.json	Log4 - Copy.json	209
5	dbfs:/mnt/Gen2/LogFiles/Log4.json	Log4.json	209

```
%fs head dbfs:/mnt/Gen2/LogFiles/Log3.json
```


```
{
  "source": "Application 1",
  "data": "Log data example 1",
  "IP": "10.22.135.7",
  "log_level": "DEBUG",
  "log_type": "Error",
  "log_app": "app",
  "log_timestamp": "2020-08-25T17:45:13+07:00"
}
```


▾ 0a223bb6-9302-435d-981b-55b7b4baf3da *Last updated: 15 seconds ago*

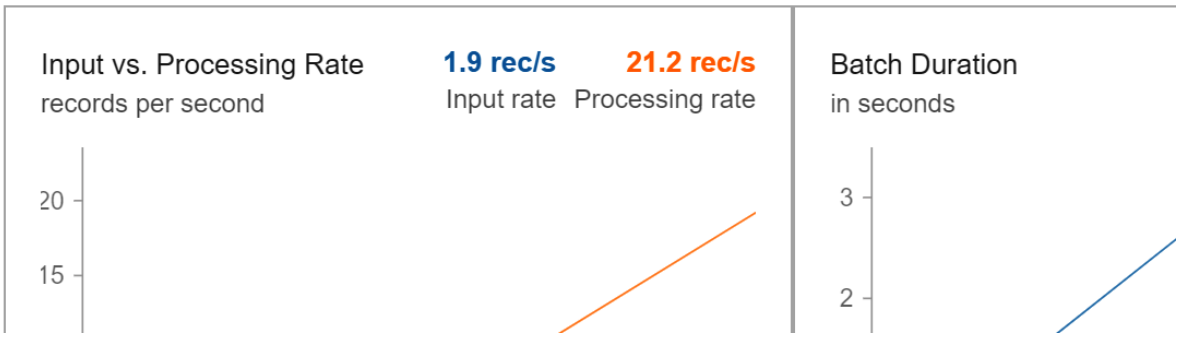
Dashboard | Raw Data



▸ (1) Spark Jobs

▾ 0a223bb6-9302-435d-981b-55b7b4baf3da *Last updated: 30 seconds ago*

Dashboard Raw Data




```
%sql
SELECT * FROM VehiclechkpointKafkaEventHub_Delta_Agg_Tumbling ORDER BY Window desc
```

▸ (1) Spark Jobs

	window	id	count
1	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	19eeda6f-0489-4c68-937a-f91ba866643b	6
2	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	afa56211-c75c-4d76-9f17-2e550cd51fa8	5
3	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	66ce2ce6-3b23-4f92-a3b7-a7dfd3fe7caf	5
4	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	a529fb2a-5f7c-4aa1-976c-e4b683b919f2	5
5	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	c7bb7aa9-cc1c-453f-a86e-ca171c710e85	5
6	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	c91ff0d5-bffc-458f-9a15-25fac0fc6cc8	6

Showing all 10 rows.

```
%sql
SELECT * FROM VehiclechkpointKafkaEventHub_Delta_Agg_Tumbling ORDER BY Window desc
```

▸ (1) Spark Jobs

	window	id	count
1	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	a1a4b4ba-7f9c-4058-aa15-b79f3fd5dd7d	2
2	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	c7bb7aa9-cc1c-453f-a86e-ca171c710e85	1
3	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	67f5187f-d1c0-431b-844a-2f99f49d59e3	1
4	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	45e71b4f-f46e-4e58-959d-e2afa7830875	2
5	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	66ce2ce6-3b23-4f92-a3b7-a7dfd3fe7caf	1
6	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	a529fb2a-5f7c-4aa1-976c-e4b683b919f2	1

Showing all 30 rows.

```
%sql
SELECT * FROM VehiclechkpointKafkaEventHub_Delta_Agg_Tumbling where id='c7bb7aa9-cc1c-453f-a86e-ca171c710e85'
```

▸ (2) Spark Jobs

	window	id
1	{"start": "2021-07-30T01:54:00.000+0000", "end": "2021-07-30T01:55:00.000+0000"}	c7bb7aa9-cc1c-453f-a86e-ca171c710e85
2	{"start": "2021-07-30T01:55:00.000+0000", "end": "2021-07-30T01:56:00.000+0000"}	c7bb7aa9-cc1c-453f-a86e-ca171c710e85
3	{"start": "2021-07-30T01:53:00.000+0000", "end": "2021-07-30T01:54:00.000+0000"}	c7bb7aa9-cc1c-453f-a86e-ca171c710e85

```sql
%sql
select * from VehiclechkpointKafkaEventHub_Delta_Agg_Overlapping
where id ='c2c7cb35-2f97-4fab-ab23-62fe24eca7af'
order by window desc
```

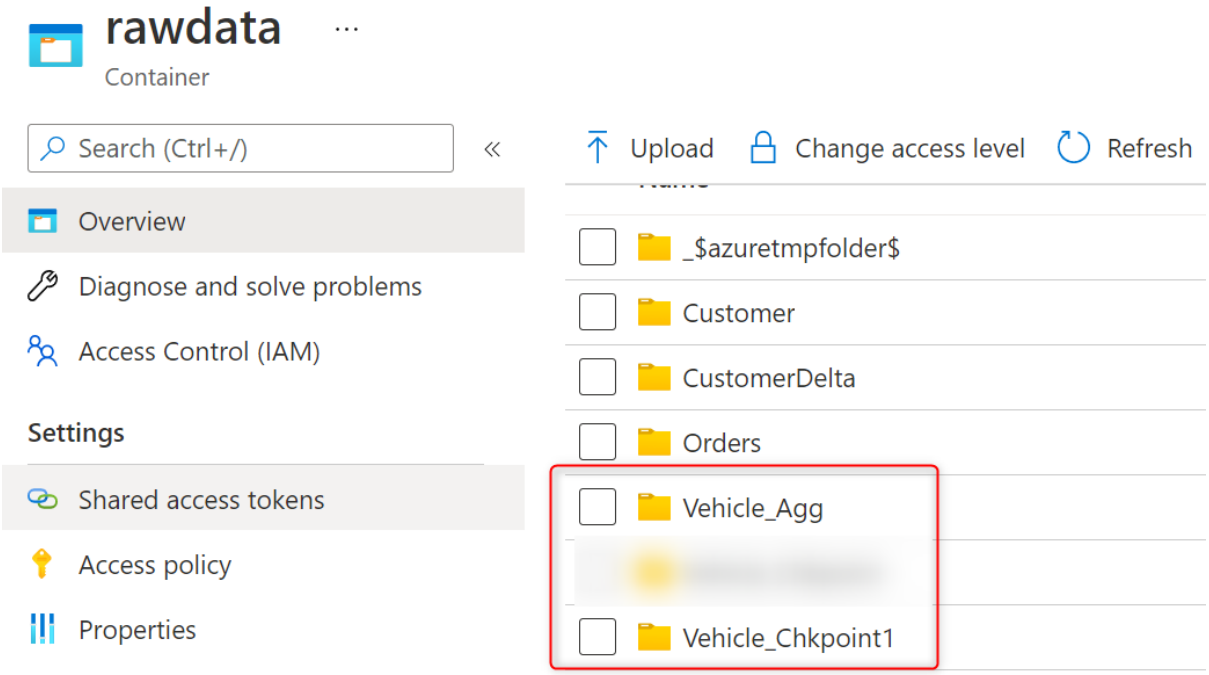
▸ (1) Spark Jobs

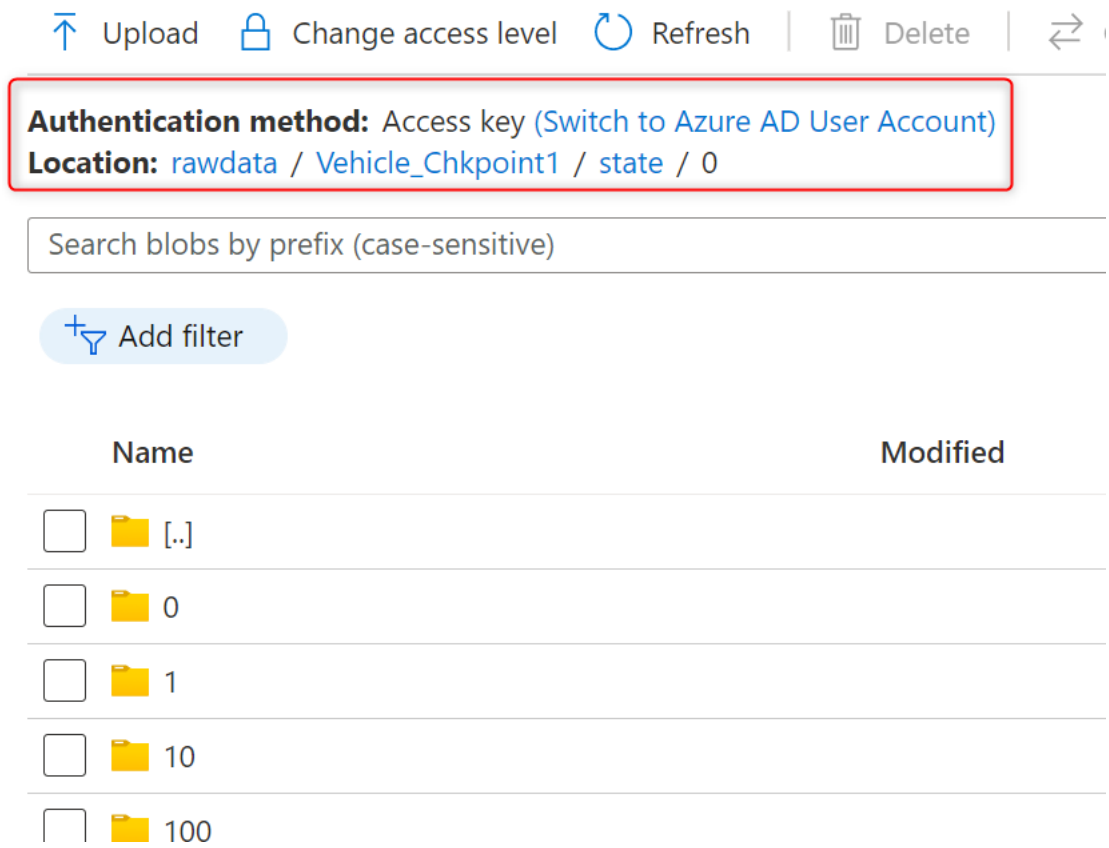
	window	id	count
1	{"start": "2021-07-30T03:08:00.000+0000", "end": "2021-07-30T03:10:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	1
2	{"start": "2021-07-30T03:07:00.000+0000", "end": "2021-07-30T03:09:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	13
3	{"start": "2021-07-30T03:06:00.000+0000", "end": "2021-07-30T03:08:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	34
4	{"start": "2021-07-30T03:05:00.000+0000", "end": "2021-07-30T03:07:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	28
5	{"start": "2021-07-30T03:04:00.000+0000", "end": "2021-07-30T03:06:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	6

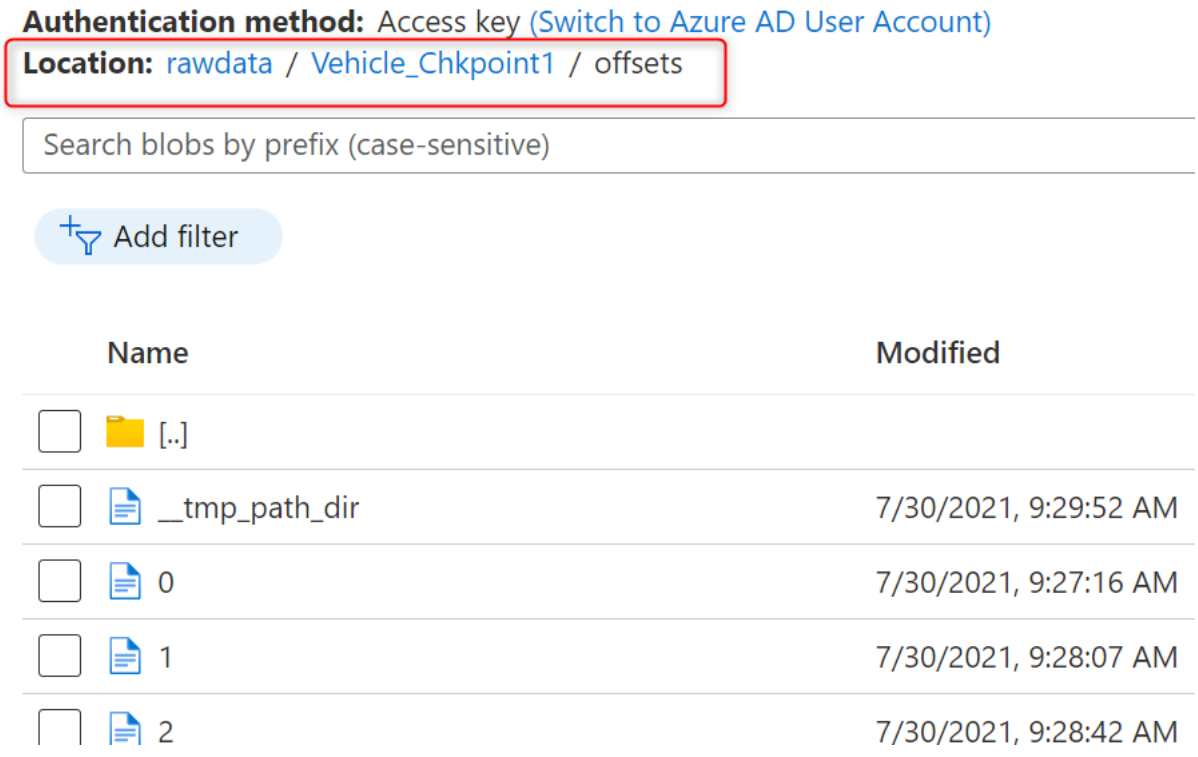
```sql
%sql
select * from VehiclechkpointKafkaEventHub_Delta_Agg_Overlapping
where id ='c2c7cb35-2f97-4fab-ab23-62fe24eca7af'
order by window desc
```

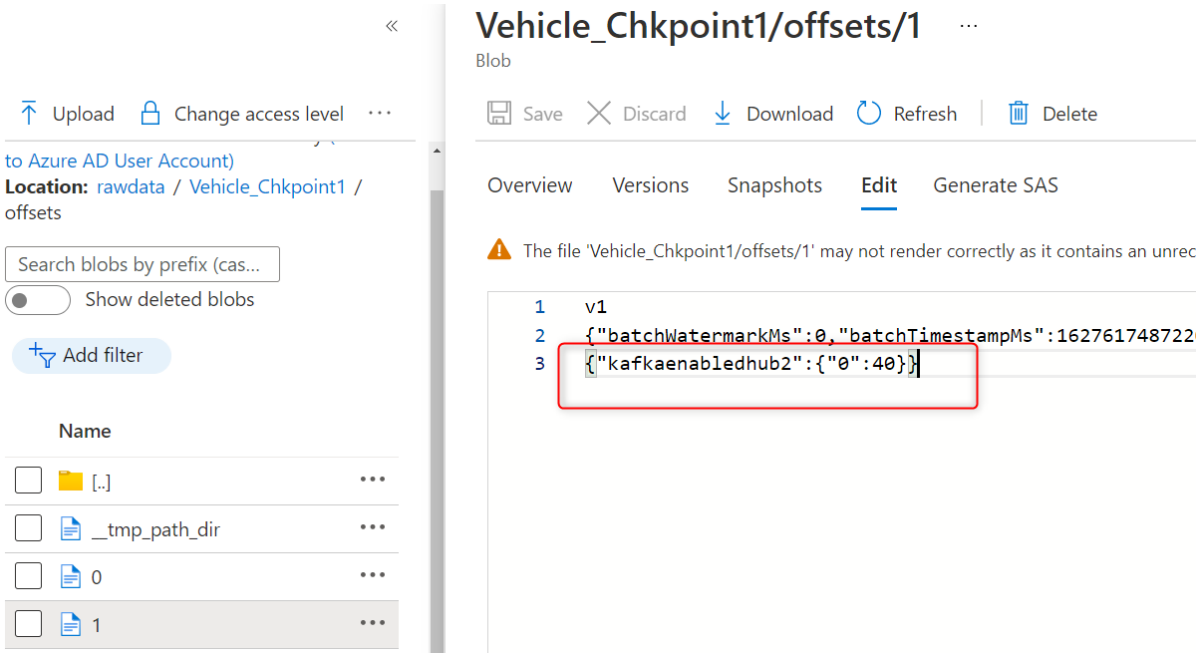
▸ (1) Spark Jobs

	window	id	count
1	{"start": "2021-07-30T03:08:00.000+0000", "end": "2021-07-30T03:10:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	2
2	{"start": "2021-07-30T03:07:00.000+0000", "end": "2021-07-30T03:09:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	14
3	{"start": "2021-07-30T03:06:00.000+0000", "end": "2021-07-30T03:08:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	34
4	{"start": "2021-07-30T03:05:00.000+0000", "end": "2021-07-30T03:07:00.000+0000"}	c2c7cb35-2f97-4fab-ab23-62fe24eca7af	28

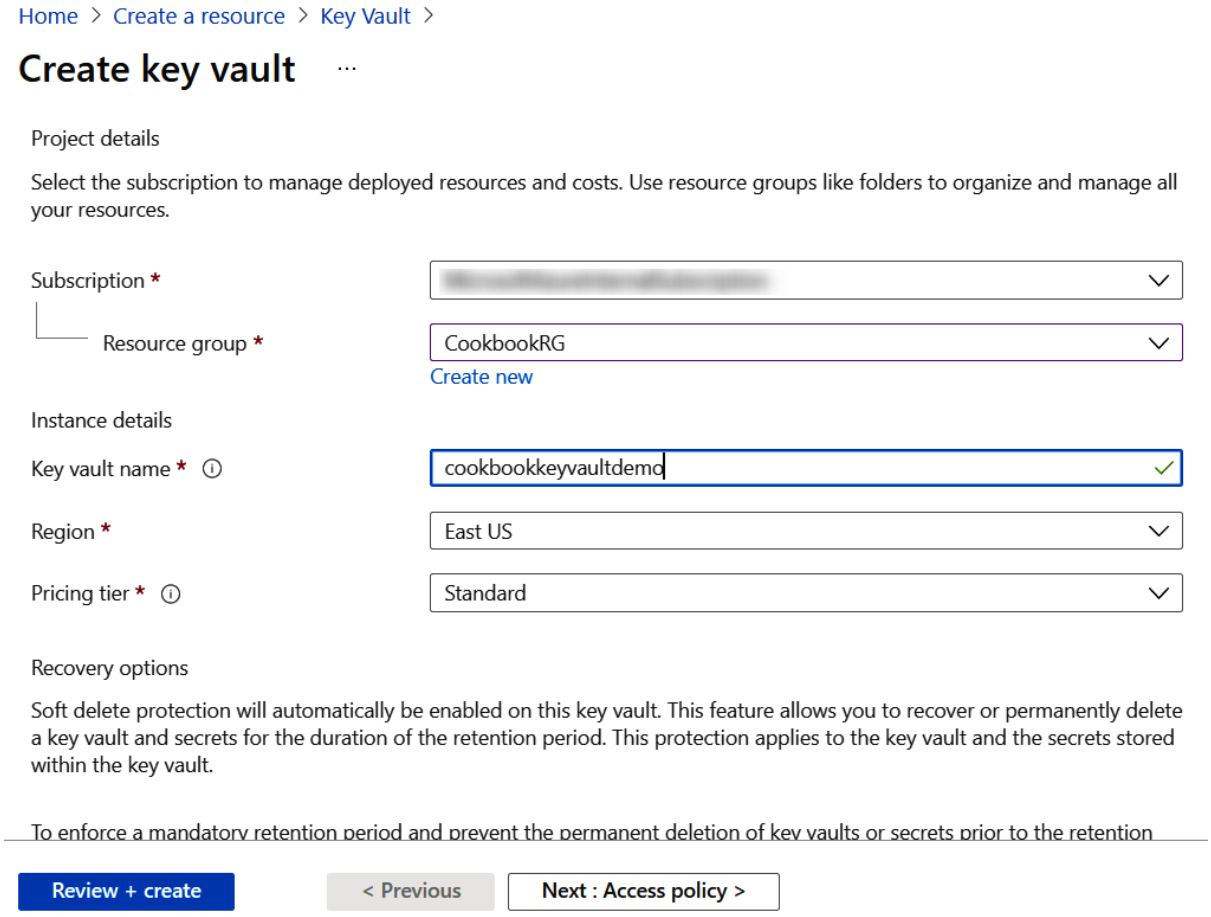



Upload
Change access level
Refresh
Delete
Authentication method: Access key (Switch to Azure AD User Account)
Location: rawdata / Vehicle_Chkpoint1 / state / 0
Search blobs by prefix (case-sensitive)
Add filter
Name
Modified
[..]
0
1
10
100


Authentication method: Access key (Switch to Azure AD User Account)
Location: rawdata / Vehicle_Chkpoint1 / offsets
Search blobs by prefix (case-sensitive)
Add filter
Name
Modified
[..]
__tmp_path_dir
7/30/2021, 9:29:52 AM
0
7/30/2021, 9:27:16 AM
1
7/30/2021, 9:28:07 AM
2
7/30/2021, 9:28:42 AM


Upload
Change access level
to Azure AD User Account)
Location: rawdata / Vehicle_Chkpoint1 / offsets
Search blobs by prefix (cas...
Show deleted blobs
Add filter
Name
[..]
_tmp_path_dir
0
1
Vehicle_Chkpoint1/offsets/1
Blob
Save
Discard
Download
Refresh
Delete
Overview
Versions
Snapshots
Edit
Generate SAS
The file 'Vehicle_Chkpoint1/offsets/1' may not render correctly as it contains an unrec
1 v1
2 {"batchWatermarkMs":0,"batchTimestampMs":162761748722
3 {"kafkaenabledhub2":{"0":40}}

Chapter 5: Integrating with Azure Key Vault, App Configuration, and Log Analytics



Home > Create a resource > Key Vault >

Create key vault ...

Basics | **Access policy** | Networking | Tags | Review + create

Enable Access to:

- [ ] Azure Virtual Machines for deployment
- [ ] Azure Resource Manager for template deployment
- [ ] Azure Disk Encryption for volume encryption

Permission model

- (•) Vault access policy
- ( ) Azure role-based access control

+ Add Access Policy

Current Access Policies

Name	Email	Key Permissions	Secret Permissions	Certificate Permissions	Action
USER					
Vinod Jaiswal	Vinod.Jaiswal	9 selected	7 selected	15 selected	Delete

Review + create | < Previous | Next : Networking >

Create key vault ...

Basics | Access policy | **Networking** | Tags | Review + create

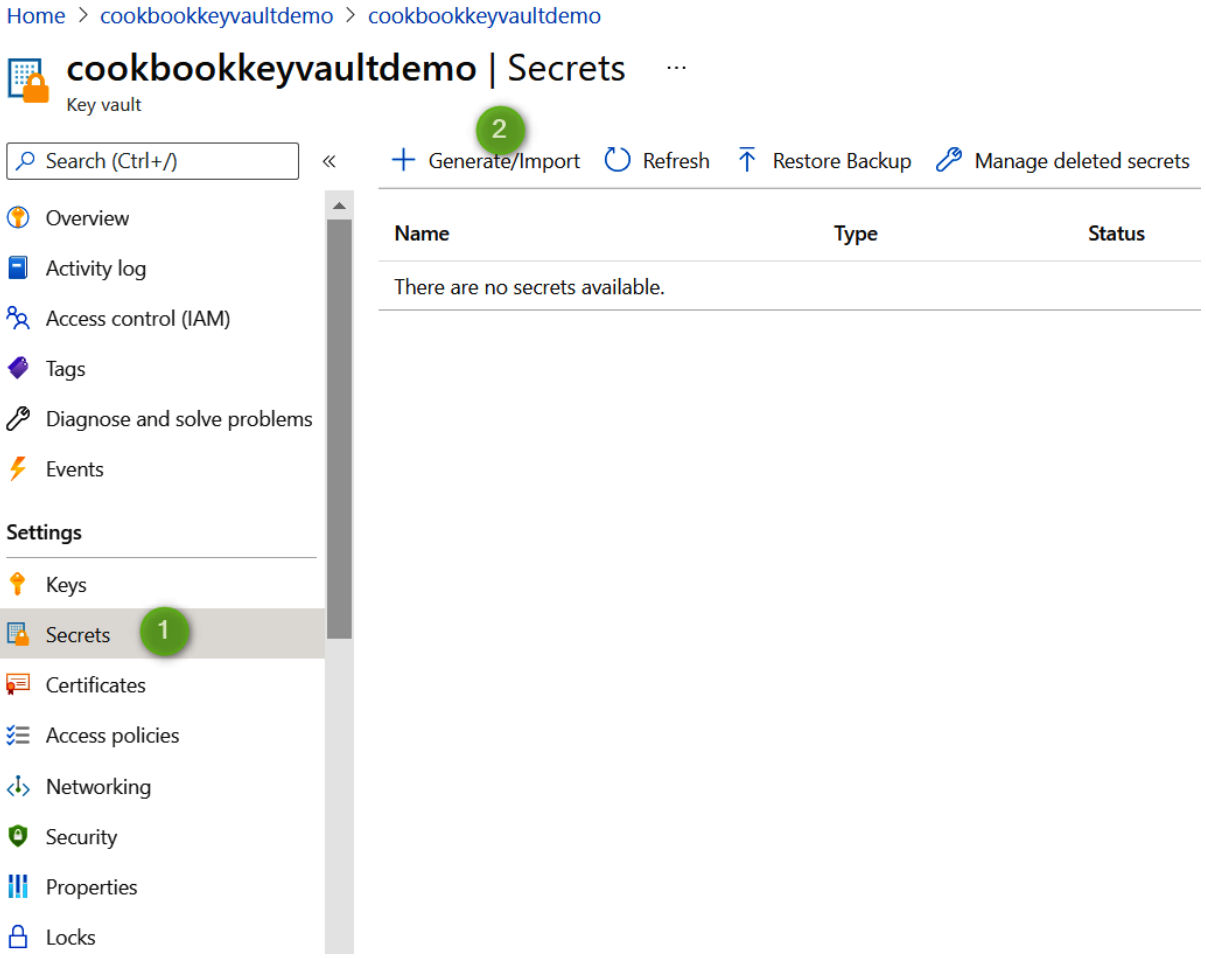
Network connectivity

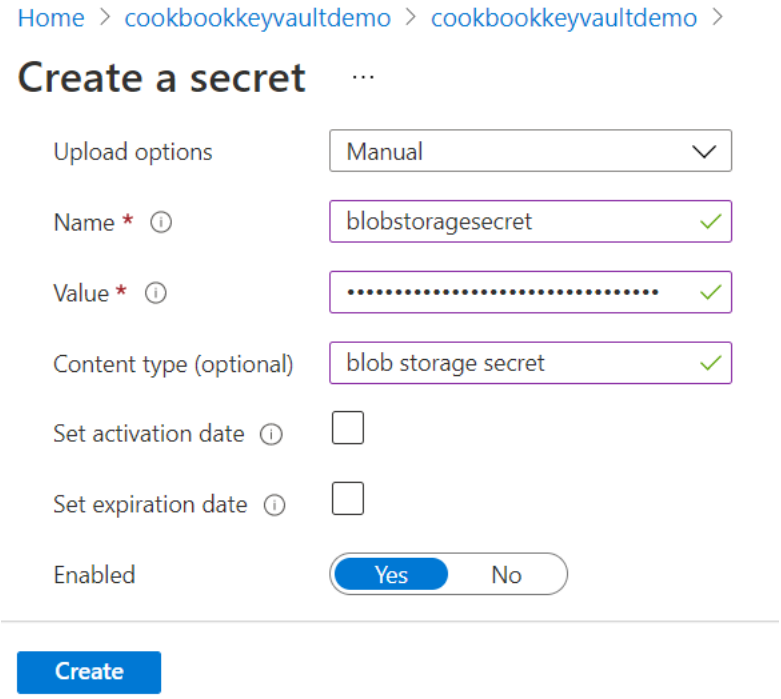
You can connect to this key vault either publicly, via public IP addresses or service endpoints, or privately, using a private endpoint.

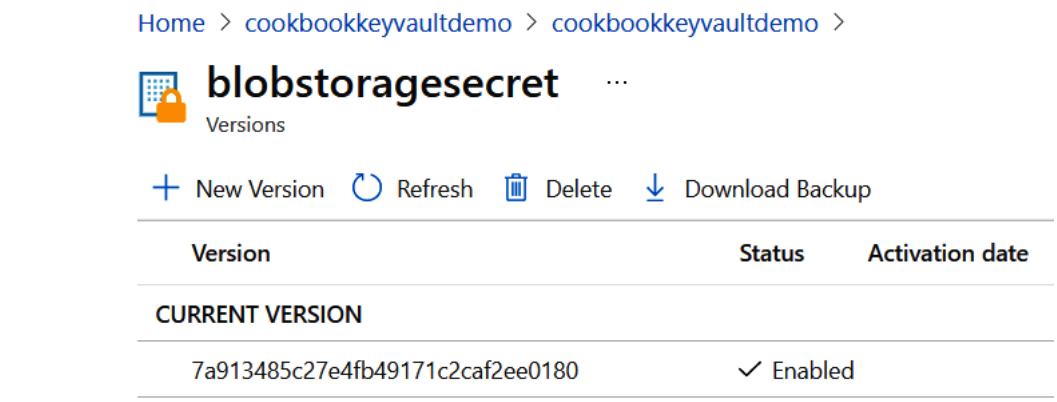
Connectivity method

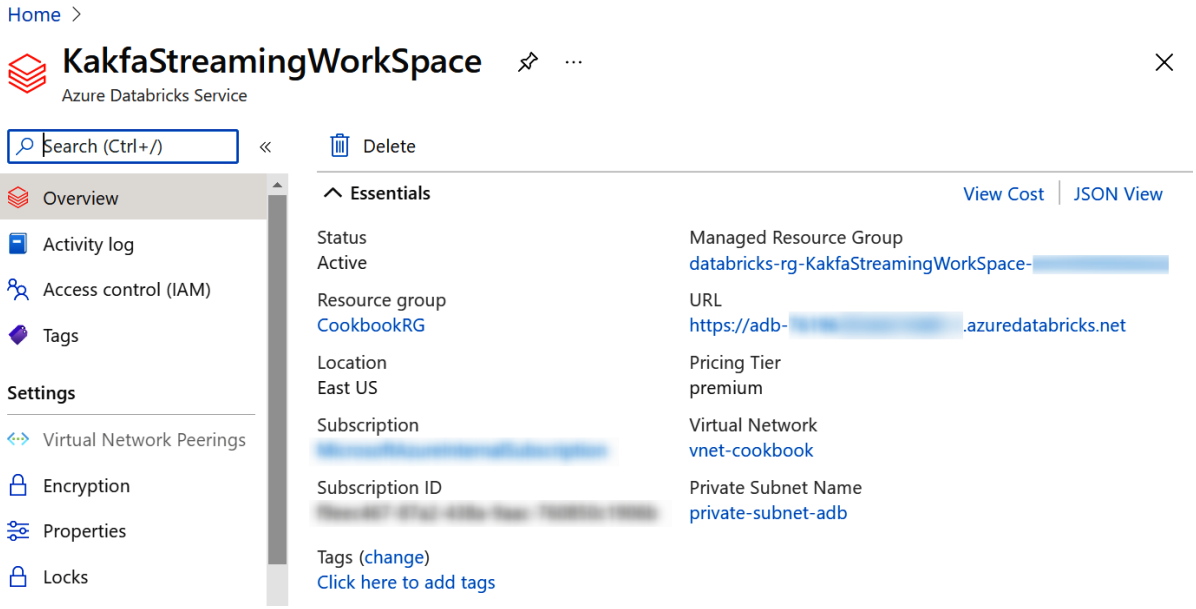
- (•) Public endpoint (all networks)
- ( ) Public endpoint (selected networks)
- ( ) Private endpoint

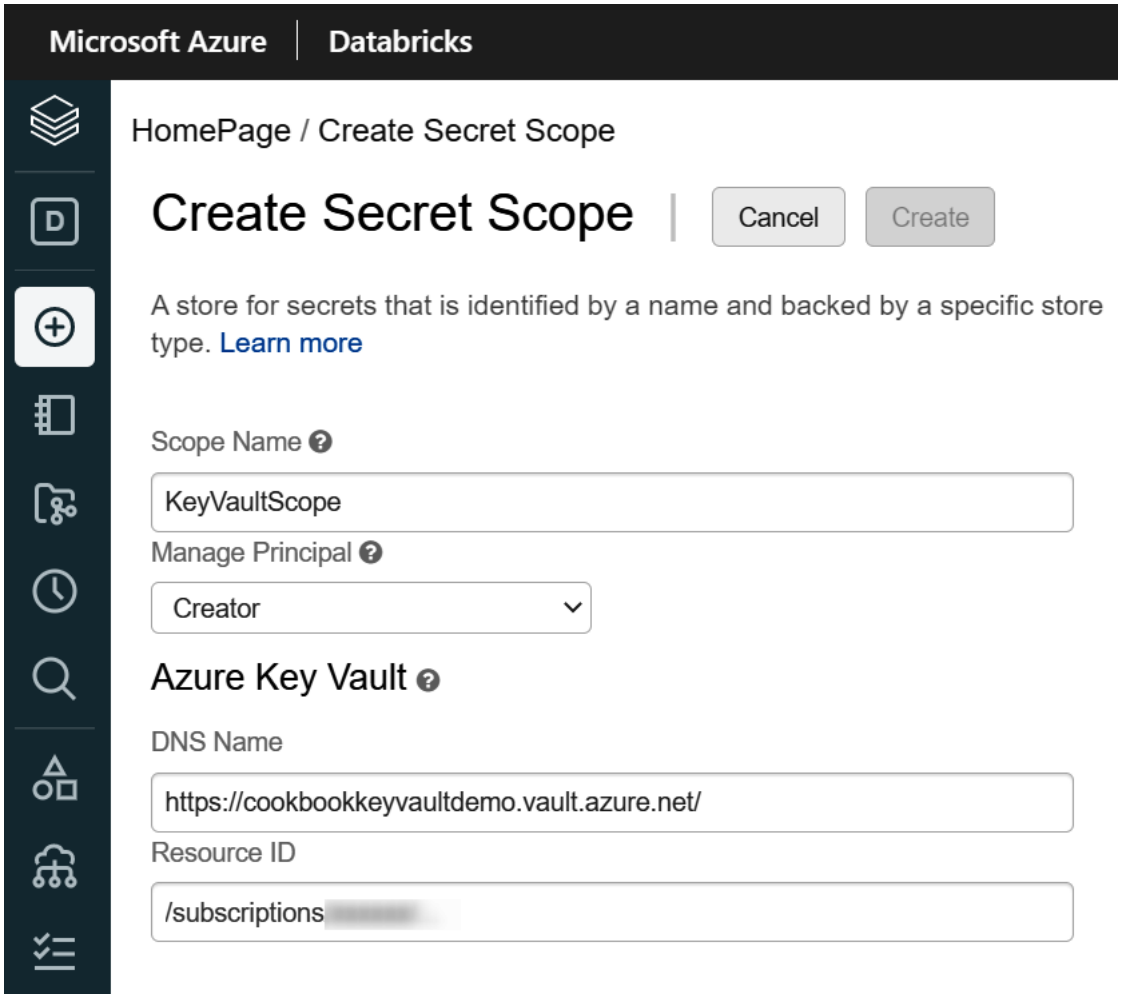
Review + create | < Previous | Next : Tags >

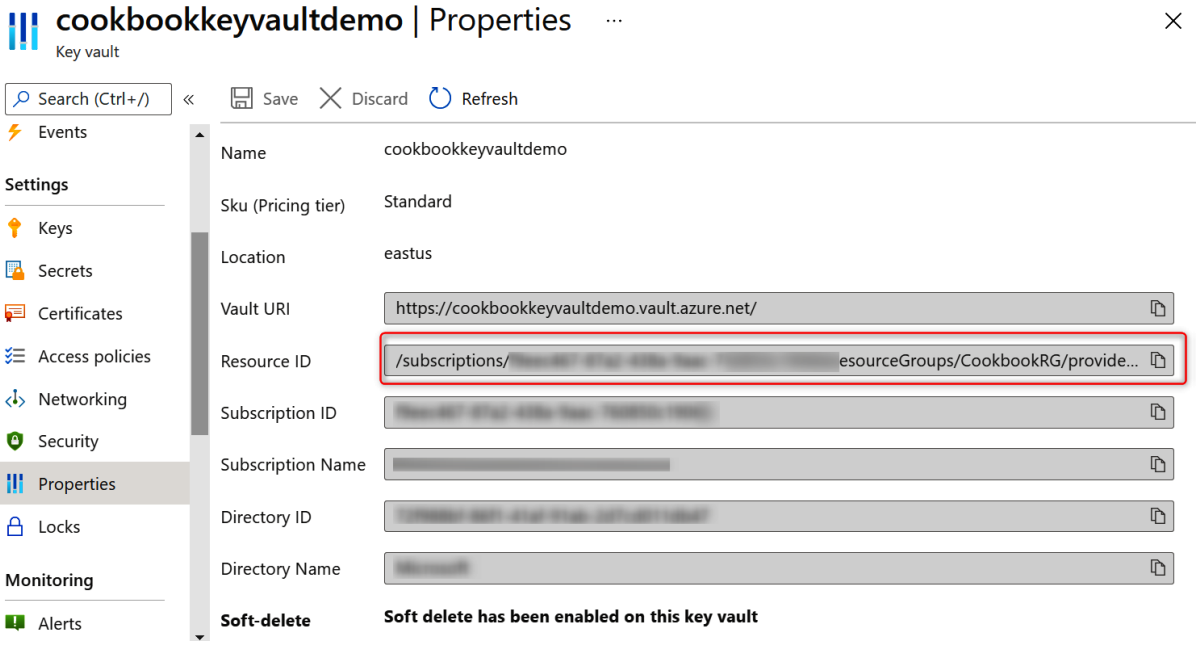

Home > cookbookkeyvaultdemo > cookbookkeyvaultdemo
cookbookkeyvaultdemo | Secrets
Key vault
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Events
Settings
Keys
Secrets
1
Certificates
Access policies
Networking
Security
Properties
Locks
2
Generate/Import
Refresh
Restore Backup
Manage deleted secrets
Name
Type
Status
There are no secrets available.


Home > cookbookkeyvaultdemo > cookbookkeyvaultdemo >
Create a secret
Upload options
Manual
Name *
blobstoragesecret
Value *
Content type (optional)
blob storage secret
Set activation date
Set expiration date
Enabled
Yes
No
Create


Home > cookbookkeyvaultdemo > cookbookkeyvaultdemo >
blobstoragesecret
Versions
New Version
Refresh
Delete
Download Backup
Version
Status
Activation date
CURRENT VERSION
7a913485c27e4fb49171c2caf2ee0180
Enabled


Home >
KakfaStreamingWorkSpace
Azure Databricks Service
Search (Ctrl+/)
Delete
Overview
Activity log
Access control (IAM)
Tags
Settings
Virtual Network Peerings
Encryption
Properties
Locks
Essentials
View Cost
JSON View
Status
Active
Managed Resource Group
databricks-rg-KakfaStreamingWorkSpace-
Resource group
CookbookRG
URL
https://adb-
.azuredatabricks.net
Location
East US
Pricing Tier
premium
Subscription
Virtual Network
vnet-cookbook
Subscription ID
Private Subnet Name
private-subnet-adb
Tags (change)
Click here to add tags


Microsoft Azure | Databricks
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
KeyVaultScope
Manage Principal
Creator
Azure Key Vault
DNS Name
https://cookbookkeyvaultdemo.vault.azure.net/
Resource ID
/subscriptions


cookbookkeyvaultdemo | Properties
Key vault
Search (Ctrl+/)
Save
Discard
Refresh
Events
Settings
Keys
Secrets
Certificates
Access policies
Networking
Security
Properties
Locks
Monitoring
Alerts
Name
cookbookkeyvaultdemo
Sku (Pricing tier)
Standard
Location
eastus
Vault URI
https://cookbookkeyvaultdemo.vault.azure.net/
Resource ID
/subscriptions/
esourceGroups/CookbookRG/provide...
Subscription ID
Subscription Name
Directory ID
Directory Name
Soft-delete
Soft delete has been enabled on this key vault

Home > Create a resource >

App Configuration

Microsoft

App Configuration

Add to Favorites

Microsoft

★★★★½ 4.6 (18 ratings)

Azure benefit eligible

Create

Overview Plans Usage Information + Support Reviews

Azure App Configuration allows developers to store, retrieve and manage access to application settings all in one place. It is easy to set up and simple to use from any application. It gives developers the ability to modify an application's behavior on demand without having to redeploy the application. App Configuration offers the following benefits:

Home > Create a resource > Marketplace > App Configuration >

Create App Configuration

Azure App Configuration provides a service to centrally manage application settings and feature flags. Modern programs, especially programs running in a cloud, generally have many components that are distributed in nature. Spreading configuration settings across these components can lead to hard-to-troubleshoot errors during an application deployment. Use App Configuration to store all the settings for your application and secure their accesses in one place. Learn more

Project Details

Subscription *

Resource group * CookbookRG

Create new

Instance Details

Resource name * Enter resource name

Location * East US

Pricing tier * Standard

View full pricing details

Review + create < Previous Next: Tags >

Home > App Configuration > DevAppconfigurationRes

DevAppconfigurationRes | Configuration explorer

App Configuration

Search (Ctrl+/)

Access keys
Scale up
Identity
Encryption
Private endpoint connections
Properties
Locks

Operations

Configuration explorer 1

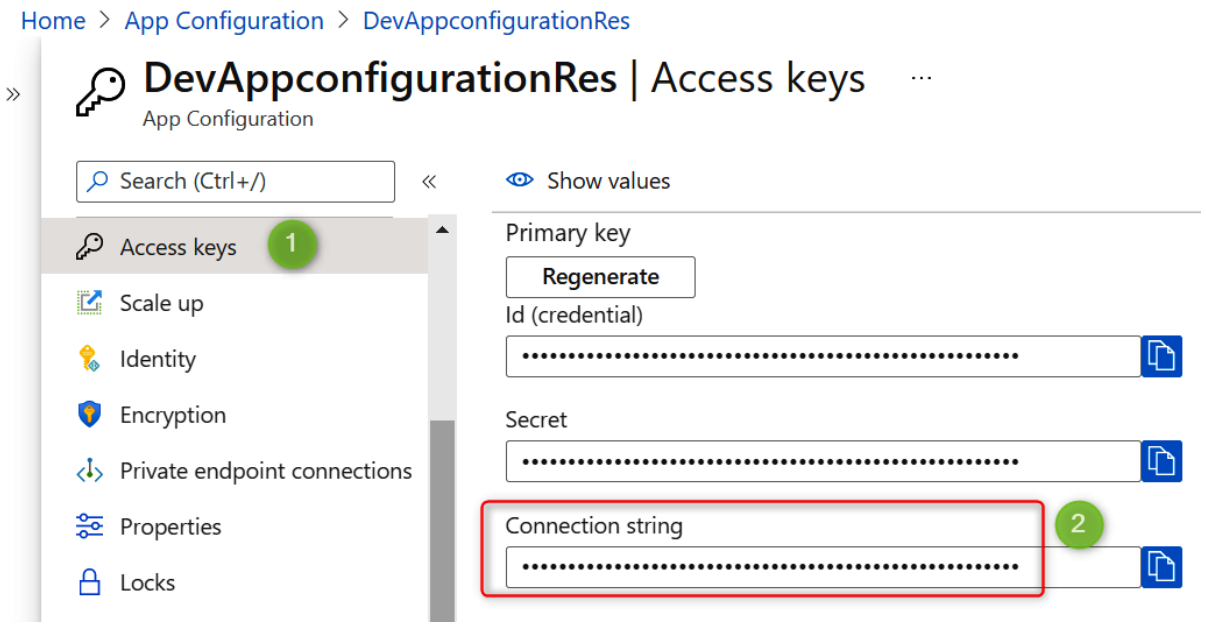
2 Create Refresh

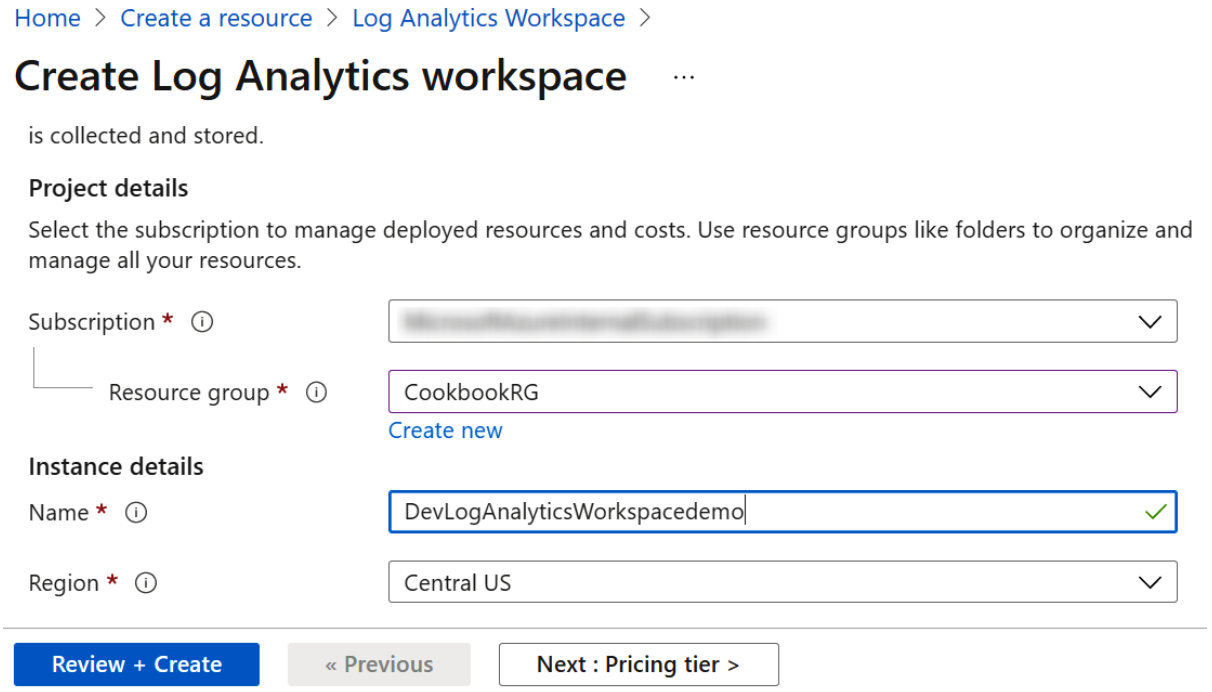
Authentication method: Access keys (Switch to Azure AD)

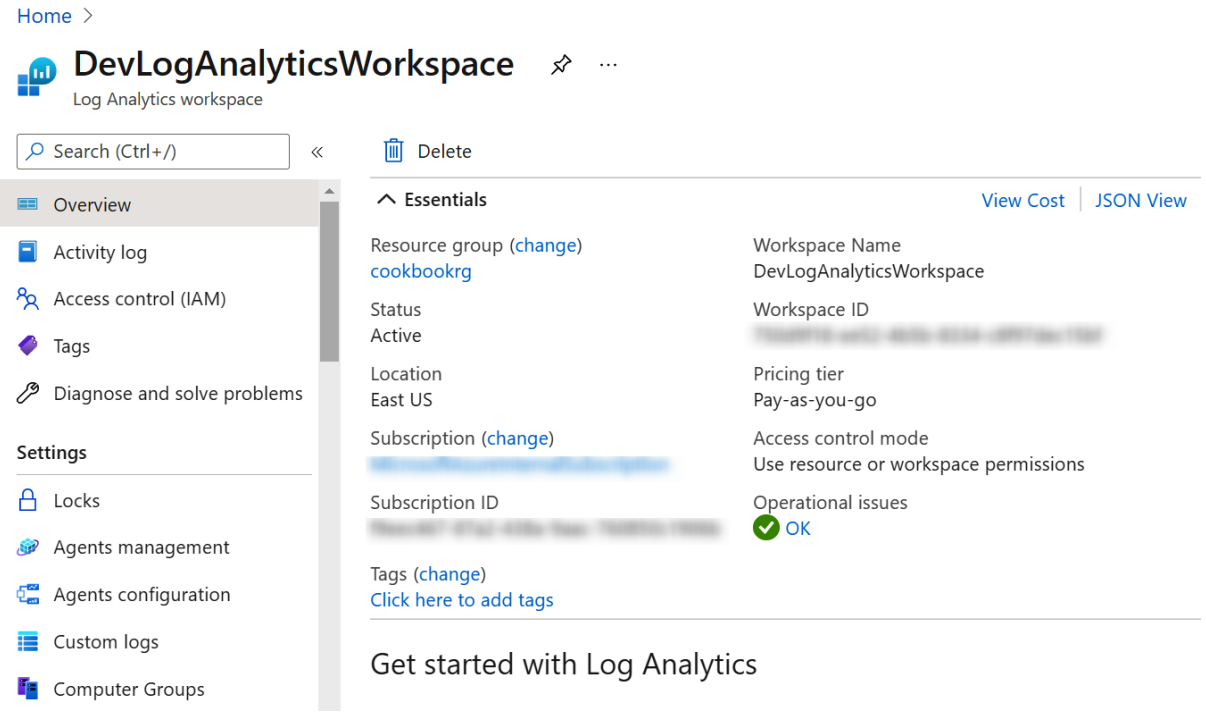
Date : Select date Keys : Select key Labels : Select label

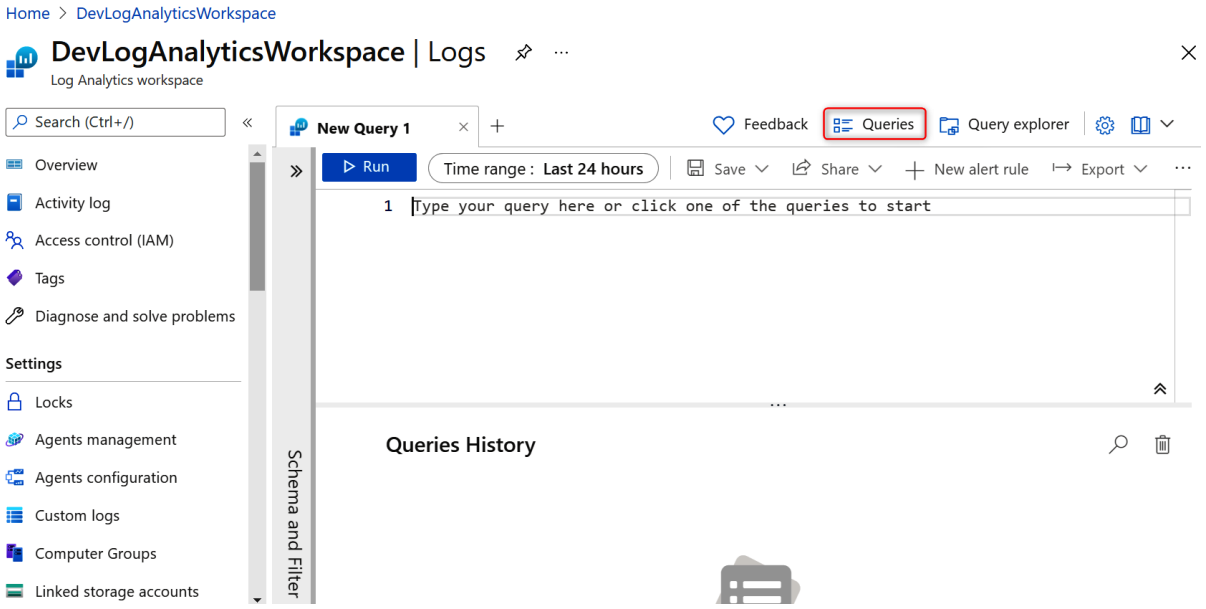
Loaded 1 key-values with 1 unique keys. Show values Expand all Collapse all

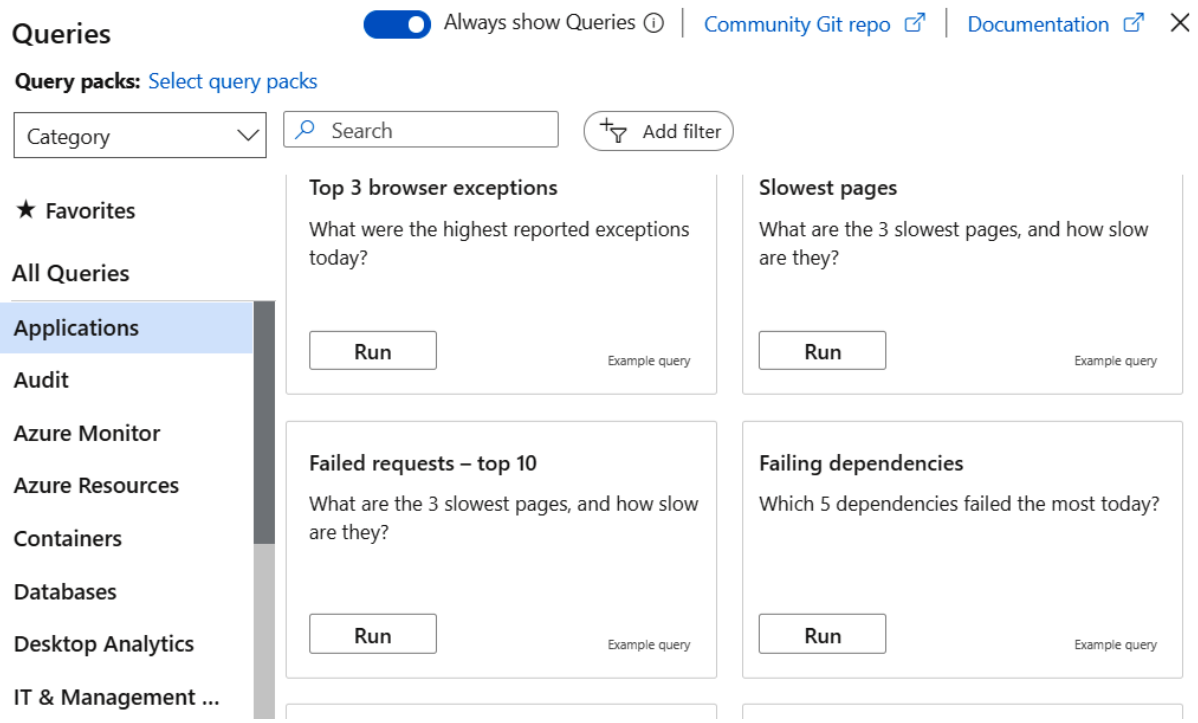
Key	Value	Label
StorageKey	*(Hidden value)*	*(No label)*


Home > App Configuration > DevAppconfigurationRes
DevAppconfigurationRes | Access keys
App Configuration
Search (Ctrl+/)
Access keys
1
Scale up
Identity
Encryption
Private endpoint connections
Properties
Locks
Show values
Primary key
Regenerate
Id (credential)
Secret
Connection string
2


Home > Create a resource > Log Analytics Workspace >
Create Log Analytics workspace
is collected and stored.
Project details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription *
Resource group *
CookbookRG
Create new
Instance details
Name *
DevLogAnalyticsWorkspacedemo
Region *
Central US
Review + Create
« Previous
Next : Pricing tier >


Home
DevLogAnalyticsWorkspace
Log Analytics workspace
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Locks
Agents management
Agents configuration
Custom logs
Computer Groups
Delete
Essentials
View Cost
JSON View
Resource group (change)
cookbookrg
Workspace Name
DevLogAnalyticsWorkspace
Status
Active
Workspace ID
Location
East US
Pricing tier
Pay-as-you-go
Subscription (change)
Access control mode
Use resource or workspace permissions
Subscription ID
Operational issues
OK
Tags (change)
Click here to add tags
Get started with Log Analytics


Home
DevLogAnalyticsWorkspace
DevLogAnalyticsWorkspace | Logs
Log Analytics workspace
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Locks
Agents management
Agents configuration
Custom logs
Computer Groups
Linked storage accounts
New Query 1
Feedback
Queries
Query explorer
Run
Time range : Last 24 hours
Save
Share
New alert rule
Export
1 Type your query here or click one of the queries to start
Schema and Filter
Queries History



Diagnostic setting

Save | Discard | Delete | Feedback

A diagnostic setting specifies a list of categories of platform logs and/or metrics that you want to collect from a resource, and one or more destinations that you would stream them to. Normal usage charges for the destination will occur. Learn more about the different log categories and contents of those logs

Diagnostic setting name * Collect Databricks Metrices

Category details

log

- [ ] Audit

metric

- [ ] AllMetrics

Destination details

- [ ] Send to Log Analytics workspace
- [ ] Archive to a storage account
- [ ] Stream to an event hub
- [ ] Send to partner solution

Diagnostic setting

Save | Discard | Delete | Feedback

A diagnostic setting specifies a list of categories of platform logs and/or metrics that you want to collect from a resource, and one or more destinations that you would stream them to. Normal usage charges for the destination will occur. Learn more about the different log categories and contents of those logs

Diagnostic setting name * Collect Databricks Metrices

Category details

log

- [x] Audit

metric

- [x] AllMetrics

Destination details

- [x] Send to Log Analytics workspace

Subscription

Log Analytics workspace

DevLogAnalyticsWorkspace (eastus)

- [ ] Archive to a storage account
- [ ] Stream to an event hub
- [ ] Send to partner solution

Home > Log Analytics workspaces >

DevLogAnalyticsWorkspace

Log Analytics workspace

Search (Ctrl+/)

Delete

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

Settings

- Locks
- Agents management
- Agents configuration
- Custom logs
- Computer Groups

Select one or more data sources to connect to the workspace

Azure virtual machines (VMs)
Windows and Linux Agents management
Azure Activity logs
Storage account log
System Center Operations Manager

Add monitoring solutions that provide insights for applications and services in your environment

View solutions

Create alerts to proactivley detect any issue that arise in your workspace

Learn more

Maximize your Log Analytics experience

Search and analyze logs

Use Log Analytics rich query language to analyze logs

View logs

Manage alert rules

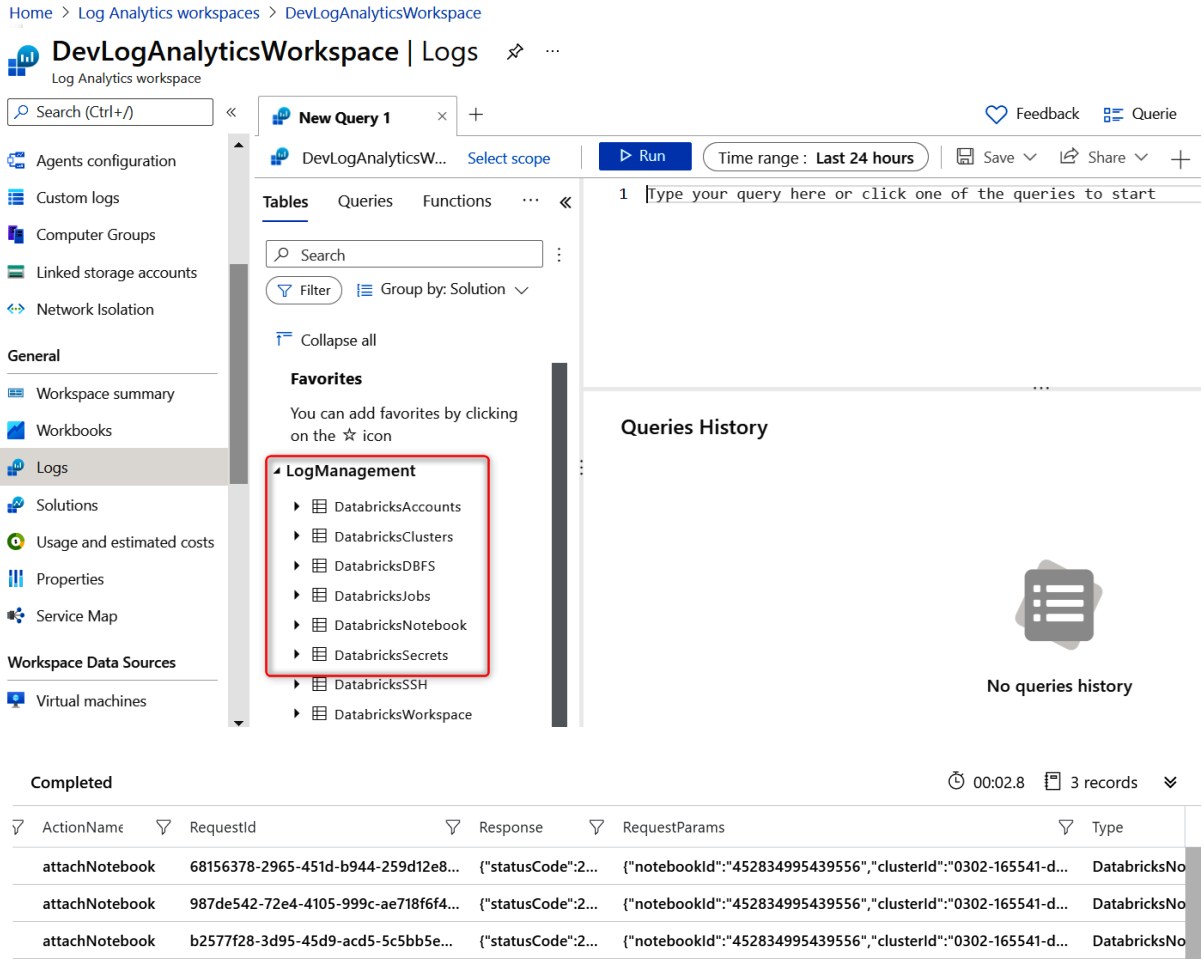
Notify or take action in response to important information in your data

Set alerts

Manage usage and costs

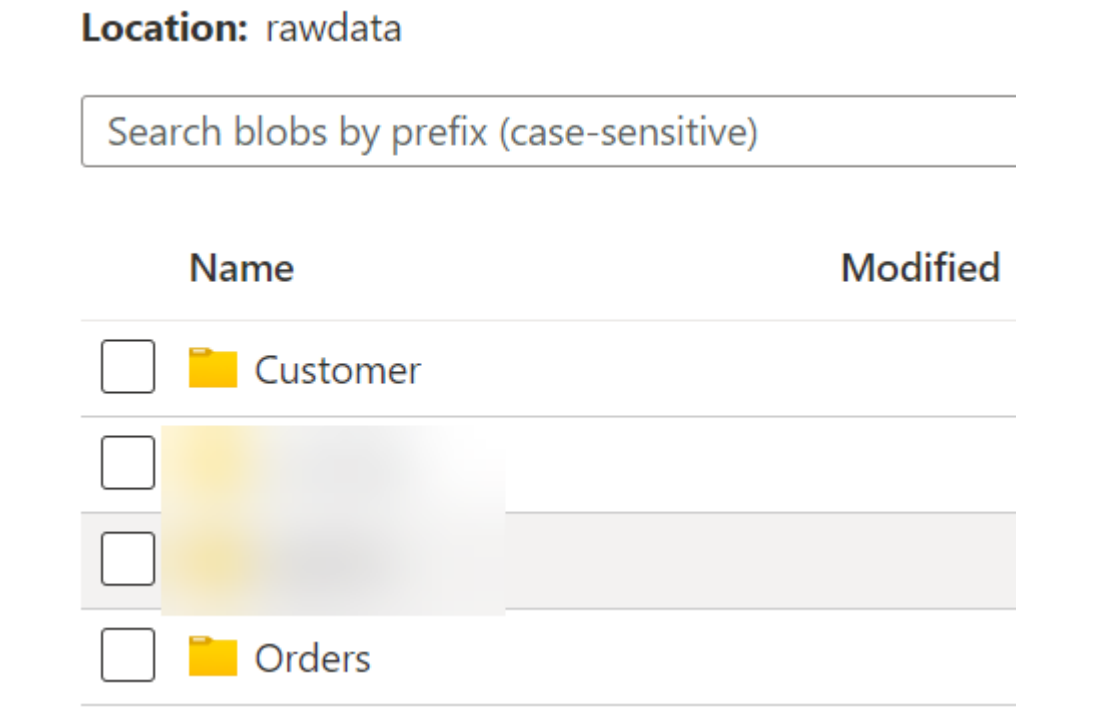
Understand your usage of Log Analytics and estimate your costs for each month

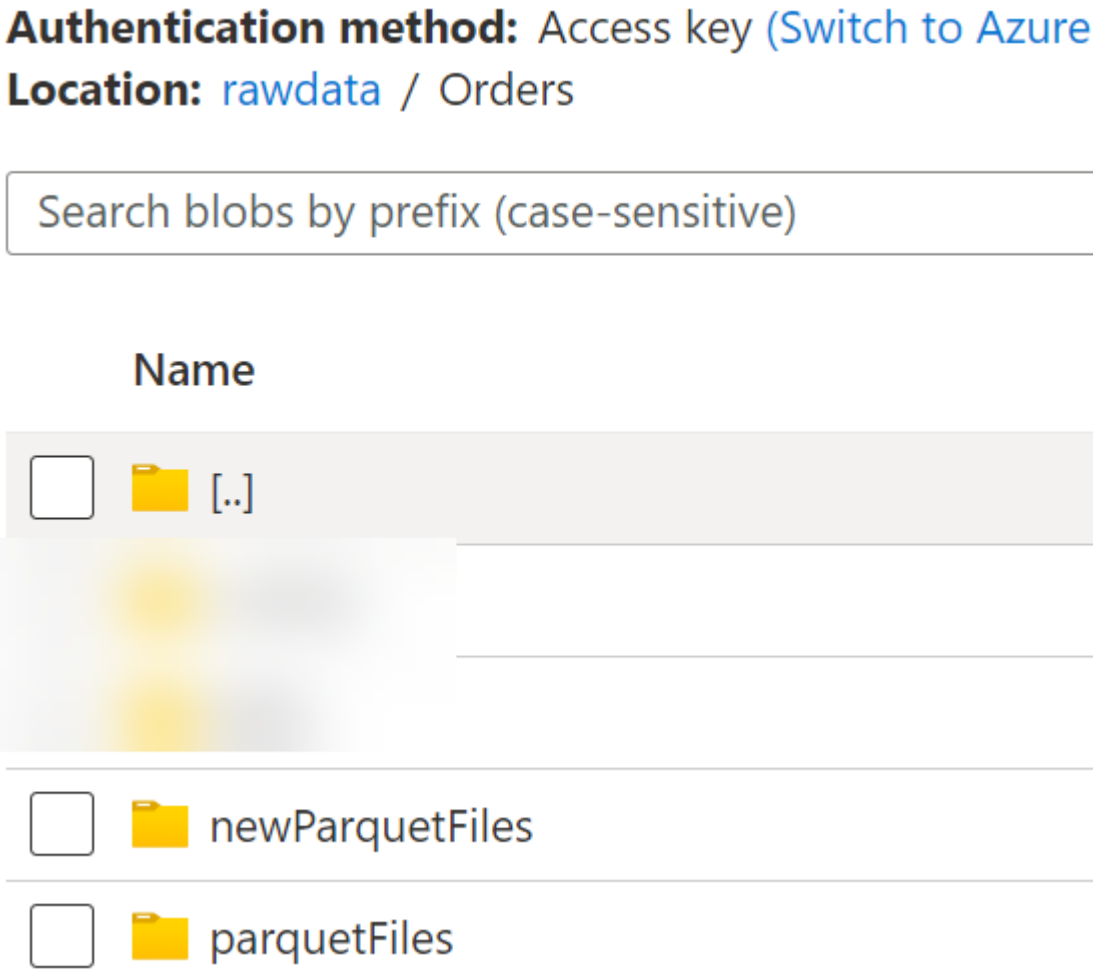
Manage costs



Completed 00:02.8 3 records

ActionName	RequestId	Response	RequestParams	Type
attachNotebook	68156378-2965-451d-b944-259d12e8...	{"statusCode":2...	{"notebookId":"452834995439556","clusterId":"0302-165541-d...	DatabricksNo
attachNotebook	987de542-72e4-4105-999c-ae718f6f4...	{"statusCode":2...	{"notebookId":"452834995439556","clusterId":"0302-165541-d...	DatabricksNo
attachNotebook	b2577f28-3d95-45d9-acd5-5c5bb5e...	{"statusCode":2...	{"notebookId":"452834995439556","clusterId":"0302-165541-d...	DatabricksNo


Location: rawdata
Search blobs by prefix (case-sensitive)
Name
Modified
Customer
Orders


Authentication method: Access key (Switch to Azure
Location: rawdata / Orders
Search blobs by prefix (case-sensitive)
Name
[..]
newParquetFiles
parquetFiles

```
query_source1 = append_kafkadata_stream(topic='eventhubsource1')
```

Cancel

(1) Spark Jobs

edd4228e-e858-4f6e-a1c1-5f5acd0f62f4 *Last updated: 15 seconds ago*

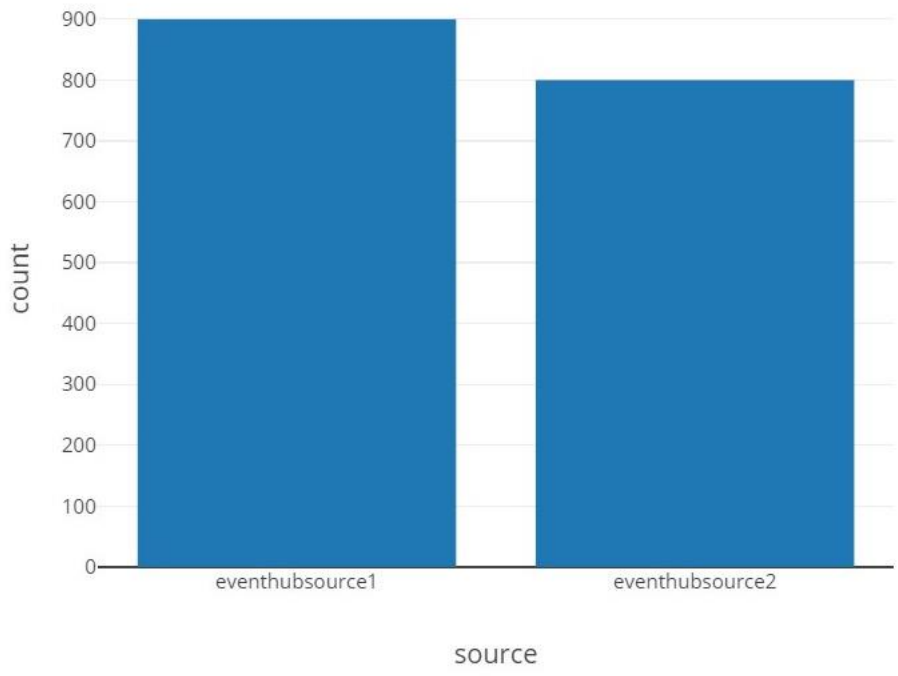
Cmd 9

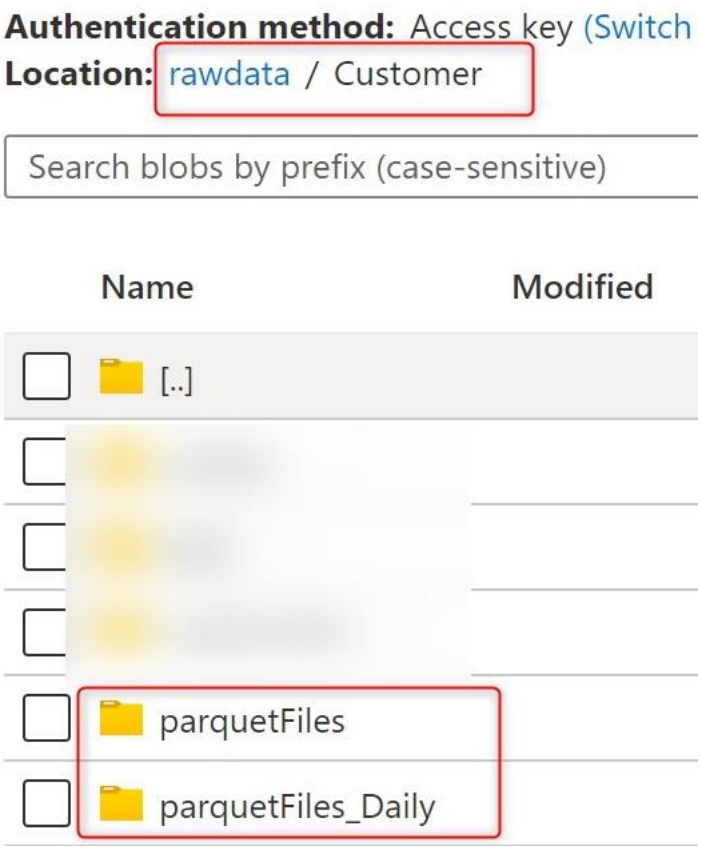
```
query_source2 = append_kafkadata_stream(topic='eventhubsource2')
```

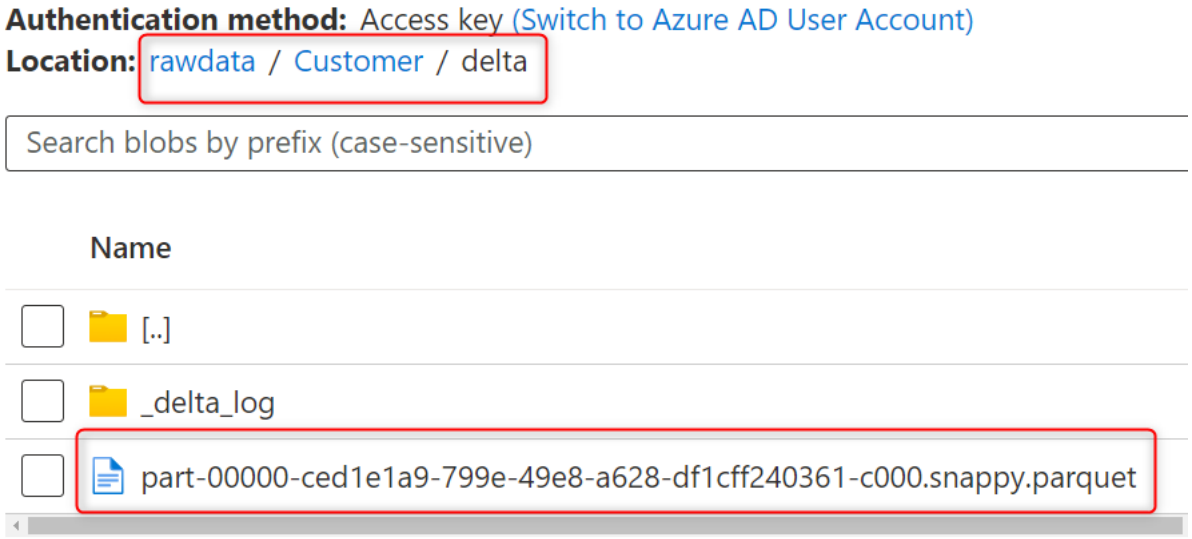
Cancel

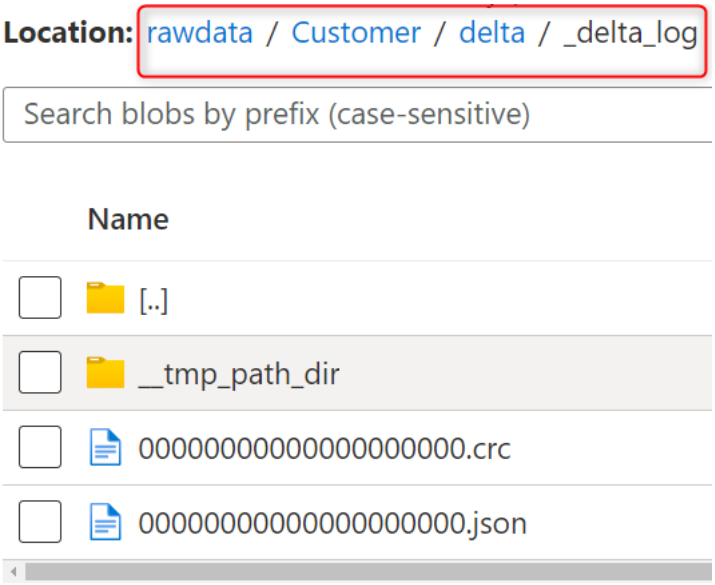
(1) Spark Jobs

76343cb9-5960-48d2-9c1a-d1f49ff27749 *Last updated: 15 seconds ago*




Authentication method: Access key (Switch
Location: rawdata / Customer
Search blobs by prefix (case-sensitive)
Name
Modified
[..]
parquetFiles
parquetFiles_Daily


Authentication method: Access key (Switch to Azure AD User Account)
Location: rawdata / Customer / delta
Search blobs by prefix (case-sensitive)
Name
[..]
_delta_log
part-00000-ced1e1a9-799e-49e8-a628-df1cff240361-c000.snappy.parquet



Customer/delta/_delta_log/00000000000000000000.json

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

```
1  {"commitInfo":{"timestamp":1628417004020,"userId":"1231225035485455","userName":"                    ","operation":"WRITE","
2  {"protocol":{"minReaderVersion":1,"minWriterVersion":2}}
3  {"metaData":{"id":"67bfb2f9-75e0-41d9-a4b4-107750160f90","format":{"provider":"parquet","options":{}},"schemaString":"{\"typ
4  {"add":{"path":"part-00000-ced1e1a9-799e-49e8-a628-df1cff240361-c000.snappy.parquet","partitionValues":{},"size":8618802,"mo
5
```

Customer/delta/_delta_log/00000000000000000000.crc

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

The file 'Customer/delta/_delta_log/00000000000000000000.crc' may not render correctly as it contains an unrecognized extension.

```
1  {"tableSizeBytes":8618802,"numFiles":1,"numMetadata":1,"numProtocol":1,"numTransactions":0}
2
```

Customer/delta/_delta_log/00000000000000000001.json

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

```
1        ","operation":"WRITE","operationParameters":{"mode":"Append","partitionBy":"[]"},"notebook":{"notebook
2  ıValues":{},"size":3691463,"modificationTime":1628417370000,"dataChange":true,"stats":"{\"numRecords\":45000,
3
```

Customer/delta/_delta_log/00000000000000000002.json

Blob

Save | Discard | Download | Refresh | Delete

Overview | Versions | Edit | Generate SAS


```
1  {"commitInfo":{"timestamp":1628417520625,"userId":"1231225035485455","userName":[redacted],"operation":"DELETE"
2  {"remove":{"path":"part-00000-ced1e1a9-799e-49e8-a628-df1cff240361-c000.snappy.parquet","deletionTimestamp":1628417520623,"
3  {"add":{"path":"part-00000-527e3cd6-fbce-474e-be7c-00821ede3c5c-c000.snappy.parquet","partitionValues":{},"size":8620958,"m
4
```


Customer/delta/_delta_log/00000000000000000003.json

Blob

Save | Discard | Download | Refresh | Delete

Overview | Versions | Edit | Generate SAS


```
1  com","operation":"UPDATE","operationParameters":{"predicate":"(C_CUSTKEY#4112 = 101275)"},"notebook":{"notebookId":
2  imestamp":1628417846199,"dataChange":true,"extendedFileMetadata":true,"partitionValues":{},"size":8620958}}
3  ues":{},"size":8620802,"modificationTime":1628417847000,"dataChange":true,"stats":"{\"numRecords\":104999,\"minValu
4
```


Vehicle_Delta/_delta_log/00000000000000000000.json

Blob

Save | Discard | Download | Refresh | Delete

Overview | Versions | Edit | Generate SAS


```
1  ","operation":"STREAMING UPDATE","operationParameters":{"outputMode":"Append","queryId":"76[redacted]277
2
```


Overview | Versions | Edit | Generate SAS


```
1  tionMetrics":{"numRemovedFiles":"0","numOutputRows":"75","numOutputBytes":"3440","numAddedFiles":"1"}}}
2
3  d-4e91-ba08-1f5cd3dd\",\"timestamp\":\"2021-04-16T16:59:57.818Z\",\"rpm\":99,\"speed\":100,\"kms\":995,
4
```


```
%sql
DELETE FROM CustomerDelta WHERE C_MKTSEGMENT='FURNITURE'
```

▸ (3) Spark Jobs

Error in SQL statement: ConcurrentAppendException: Files were added to the root of the table by a concurrent update. Please try the operation again.

```
%sql
DELETE FROM CustomerDeltaPartition WHERE C_MKTSEGMENT='FURNITURE'
```

▸ (4) Spark Jobs

OK

```
%sql
DELETE FROM CustomerDeltaPartition WHERE C_MKTSEGMENT='MACHINERY' AND C_CUSTKEY=1377
```

▸ (4) Spark Jobs

Error in SQL statement: ConcurrentAppendException: Files were added to partition [C_MKTSEGMENT=MACHINERY] by a concurrent update. Please try the operation again.
Conflicting commit: {"timestamp":1628440339088,"userId":"[redacted]55","userName":"[redacted]","operation":"WRITE","operationParameters":{"mode":Append,"partitionBy":["C_MKTSEGMENT"]},"notebook":{"notebookId":"[redacted]2"},"clusterId":"[redacted]","readVersion":33,"isolationLevel":"Serializable","isBlindAppend":true,"operationMetrics":{"numFiles":"4","numOutputBytes":"2461672","numOutputRows":"29949"}}
Refer to https://docs.microsoft.com/azure/databricks/delta/concurrency-control for more details.

```
%sql
DELETE FROM CustomerDeltaPartition WHERE C_MKTSEGMENT='MACHINERY' AND C_CUSTKEY=1377
```

▸ (8) Spark Jobs

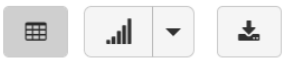
OK

```
display(ordersDF.filter("O_ORDERSTATUS =
'O'").groupBy("O_OrderYear","O_ORDERPRIORITY").agg(count("*").alias("TotalOrde
ascending=False).limit(20))
```

▸ (2) Spark Jobs

	O_OrderYear	O_ORDERPRIORITY	TotalOrders
1	1996	1-URGENT	46014
2	1996	4-NOT SPECIFIED	45776
3	1997	4-NOT SPECIFIED	45754
4	1997	2-HIGH	45689
5	1997	1-URGENT	45662
6	1996	3-MEDIUM	45657
7	1996	5-LOW	45639

Showing all 20 rows.



Command took 1.70 seconds -- [redacted]

```
display(spark.sql("DROP TABLE  IF EXISTS Orders"))

display(spark.sql("CREATE TABLE Orders USING DELTA LOCATION '/mnt/Gen2Source/Orders/OrdersDelta'"))

display(spark.sql("OPTIMIZE Orders ZORDER BY (O_ORDERPRIORITY)"))
```

▸ (17) Spark Jobs

OK
OK

	path	metrics
1	null	▸ {"numFilesAdded": 7, "numFilesRemoved": 28, "filesAdded": {"min": 2178106, "max": 3726077, "avg": 3452511, "totalFiles": 7, "totalSize": 24167579}, "filesRemoved": {"min": 443929, "max": 1024790, "avg": 887174, "totalFiles": 28, "totalSize": 24840888}, "partitionsOptimized": 7, "zOrderStats": {"strategyName": "minCubeSize(107374182400)", "inputCubeFiles": {"num": 0, "size": 0}, "inputOtherFiles": {"num": 28, "size": 24840888}, "inputNumCubes": 0, "mergedFiles": {"num": 28, "size": 24840888}, "numOutputCubes": 7, "mergedNumCubes": null}, "numBatches": 1}

▸ ordersDeltaDF: pyspark.sql.dataframe.DataFrame = [O_ORDERKEY: integer, O_C

	O_OrderYear	O_ORDERPRIORITY	TotalOrders
1	1996	1-URGENT	46014
2	1996	4-NOT SPECIFIED	45776
3	1997	4-NOT SPECIFIED	45754
4	1997	2-HIGH	45689
5	1997	1-URGENT	45662
6	1996	3-MEDIUM	45657
7	1996	5-LOW	45639

Showing all 20 rows.

Command took 0.72 seconds --

```
display(spark.sql("DROP TABLE  IF EXISTS Orders"))

display(spark.sql("CREATE TABLE Orders USING DELTA LOCATION '/mnt/Gen2Source/Orders/OrdersDelta'"))

display(spark.sql("OPTIMIZE Orders ZORDER BY (O_ORDERPRIORITY)"))
```

▸ (20) Spark Jobs

OK
OK

	path	metrics
1	null	{"numFilesAdded": 7, "numFilesRemoved": 28, "filesAdded": {"min": 2178106, "max": 3726077, "avg": 3452511, "totalFiles": 7, "totalSize": 24167579}, "filesRemoved": {"min": 443929, "max": 1024790, "avg": 887174, "totalFiles": 28, "totalSize": 24840888}, "partitionsOptimized": 7, "zOrderStats": {"strategyName": "minCubeSize(107374182400)", "inputCubeFiles": {"num": 0, "size": 0}, "inputOtherFiles": {"num": 28, "size": 24840888}, "inputNumCubes": 0, "mergedFiles": {"num": 28, "size": 24840888}, "numOutputCubes": 7, "mergedNumCubes": null}, "numBatches": 1}

```
df_cust.write.format("delta").mode("append").saveAsTable("customer")
```

▸ (1) Spark Jobs

com.databricks.sql.transaction.tahoe.schema.InvariantViolationException: NOT NULL constraint violated for column: C_NAME.

```
%sql
-- Properties column of DESCRIBE output has the details about constraints
DESCRIBE DETAIL customer;
```

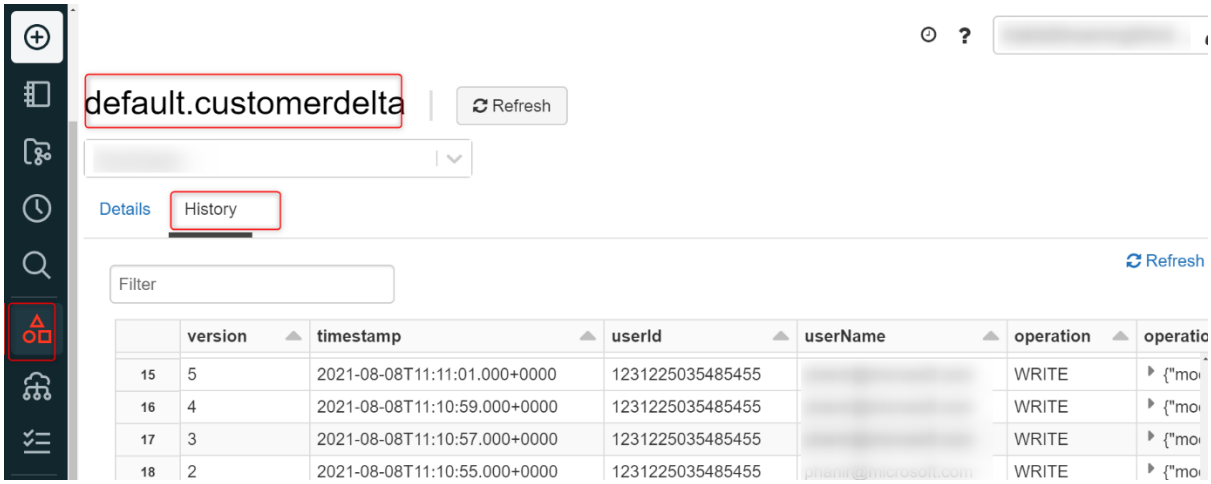
		partitionColumns	numFiles	sizeInBytes	properties
1	00	[]	0	0	{"delta.constraints.mktsegment": "C_MKTSEGMENT = 'BUILDING'"}

Showing all 1 rows.

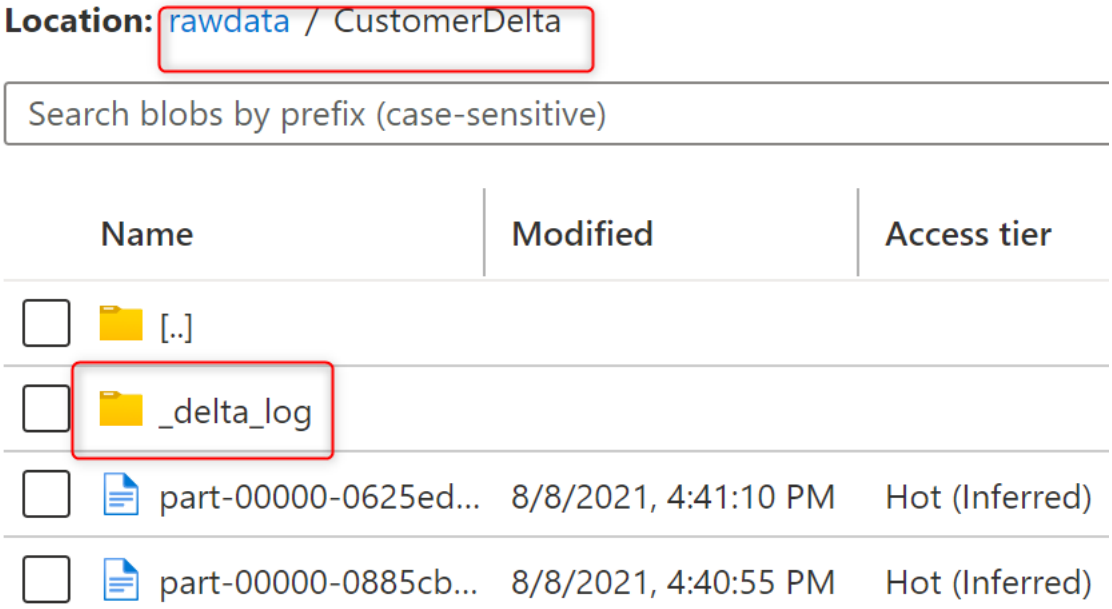
```
display(spark.sql("DESCRIBE HISTORY customerdelta"))
```

▸ (2) Spark Jobs

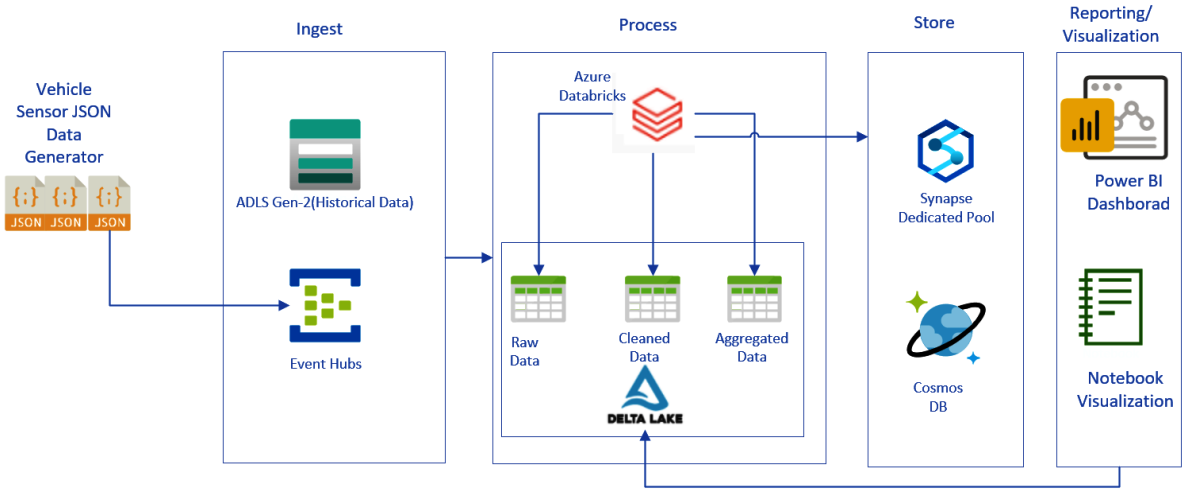
	version	timestamp	userId	userName	operation	operationParameters
15	5	2021-08-08T11:11:01.000+0000	1231225035485455		WRITE	{"mode": "Append", "partitio
16	4	2021-08-08T11:10:59.000+0000	1231225035485455		WRITE	{"mode": "Append", "partitio
17	3	2021-08-08T11:10:57.000+0000	1231225035485455		WRITE	{"mode": "Append", "partitio
18	2	2021-08-08T11:10:55.000+0000	1231225035485455		WRITE	{"mode": "Append", "partitio
19	1	2021-08-08T11:10:52.000+0000	1231225035485455		WRITE	{"mode": "Append", "partitio
20	0	2021-08-08T11:10:50.000+0000	1231225035485455		WRITE	{"mode": "Append", "partitio

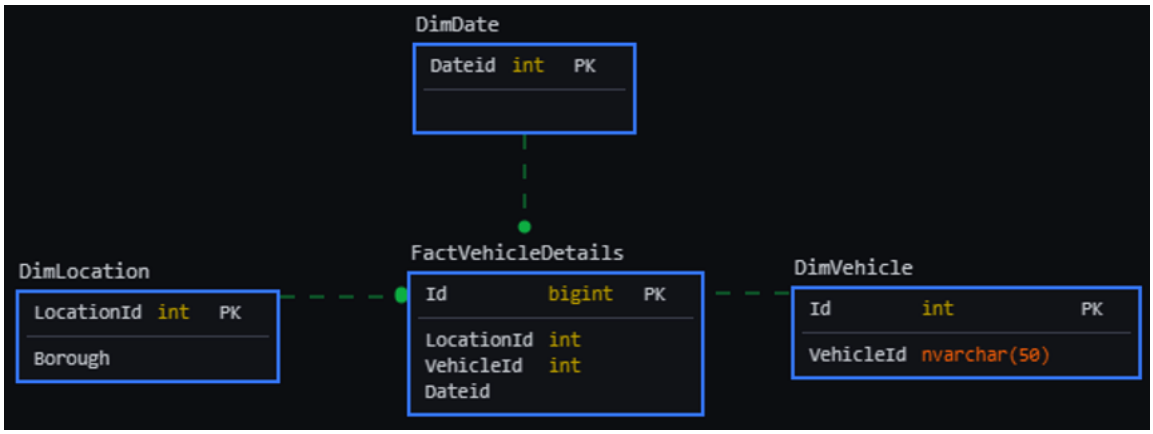


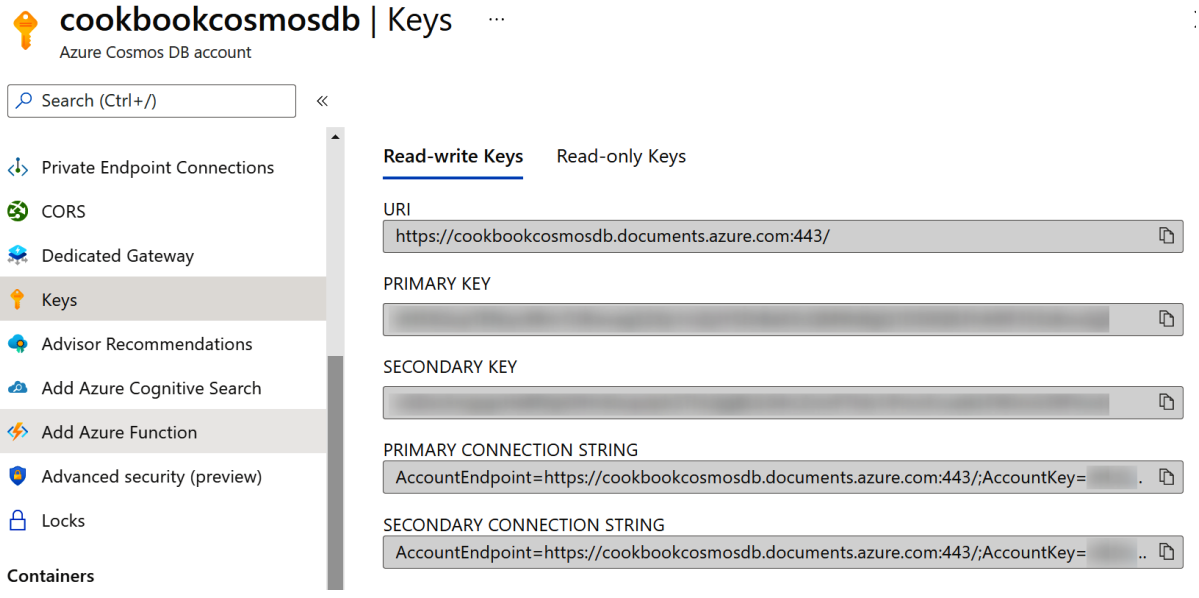
	version	timestamp	userId	userName	operation	operatio
15	5	2021-08-08T11:11:01.000+0000	1231225035485455		WRITE	{"mo
16	4	2021-08-08T11:10:59.000+0000	1231225035485455		WRITE	{"mo
17	3	2021-08-08T11:10:57.000+0000	1231225035485455		WRITE	{"mo
18	2	2021-08-08T11:10:55.000+0000	1231225035485455		WRITE	{"mo

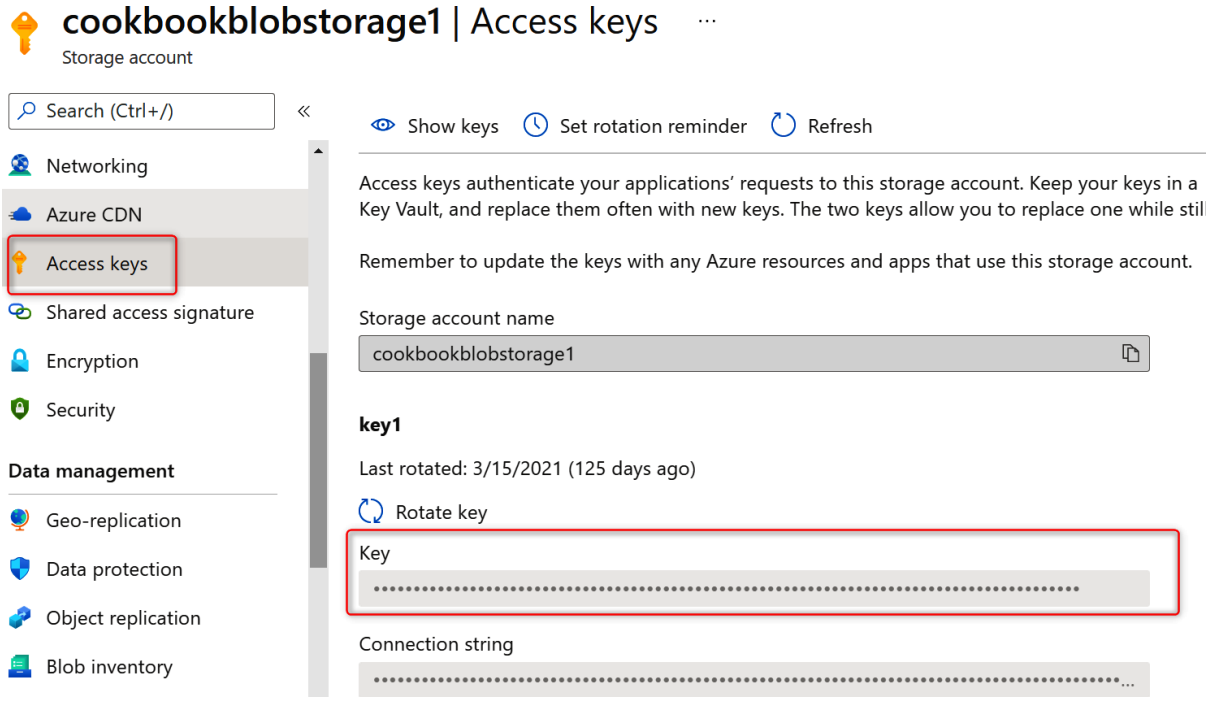


Chapter 7: Implementing Near-Real-Time Analytics and Building a Modern Data Warehouse

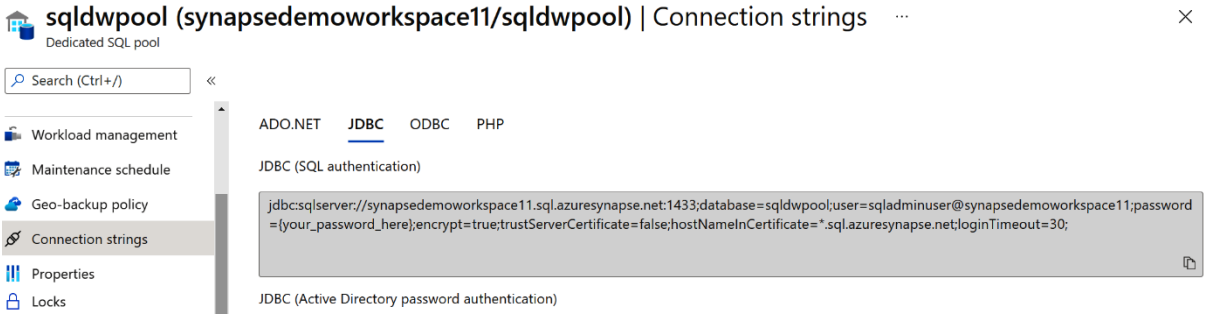


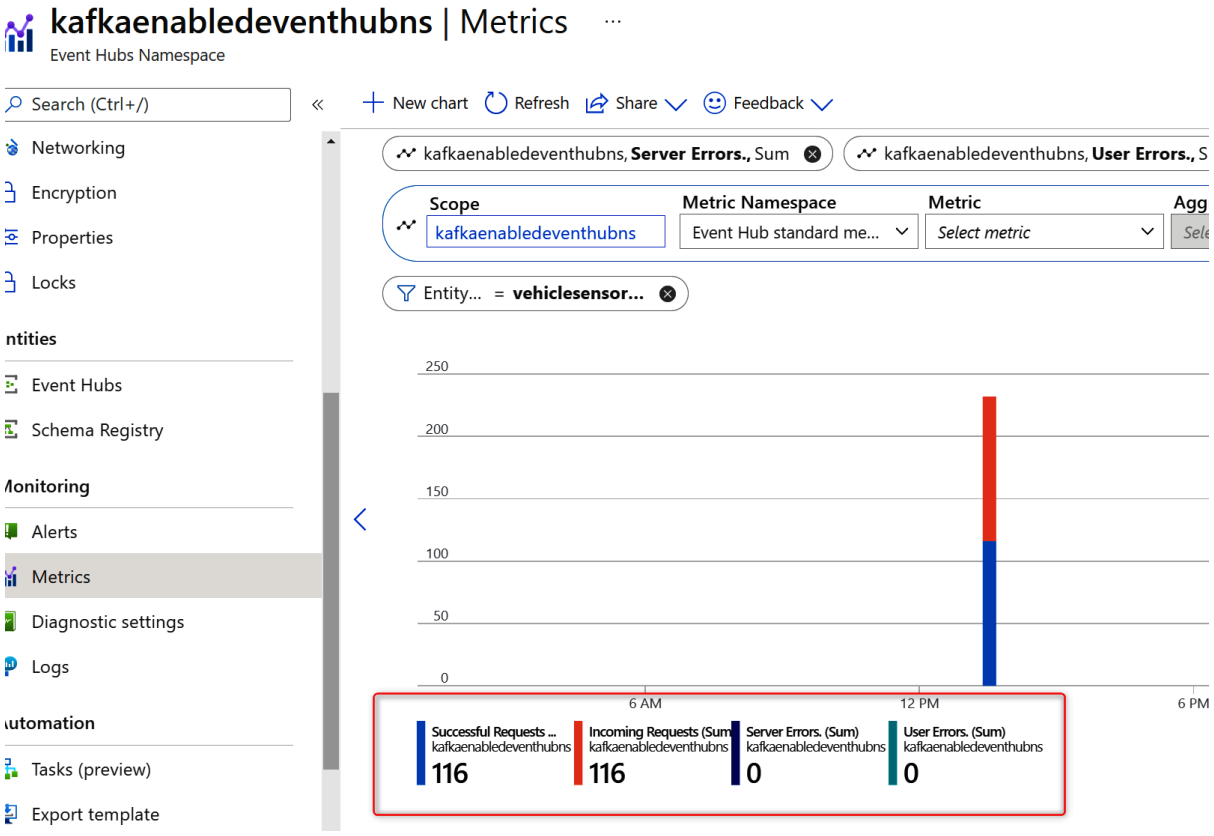


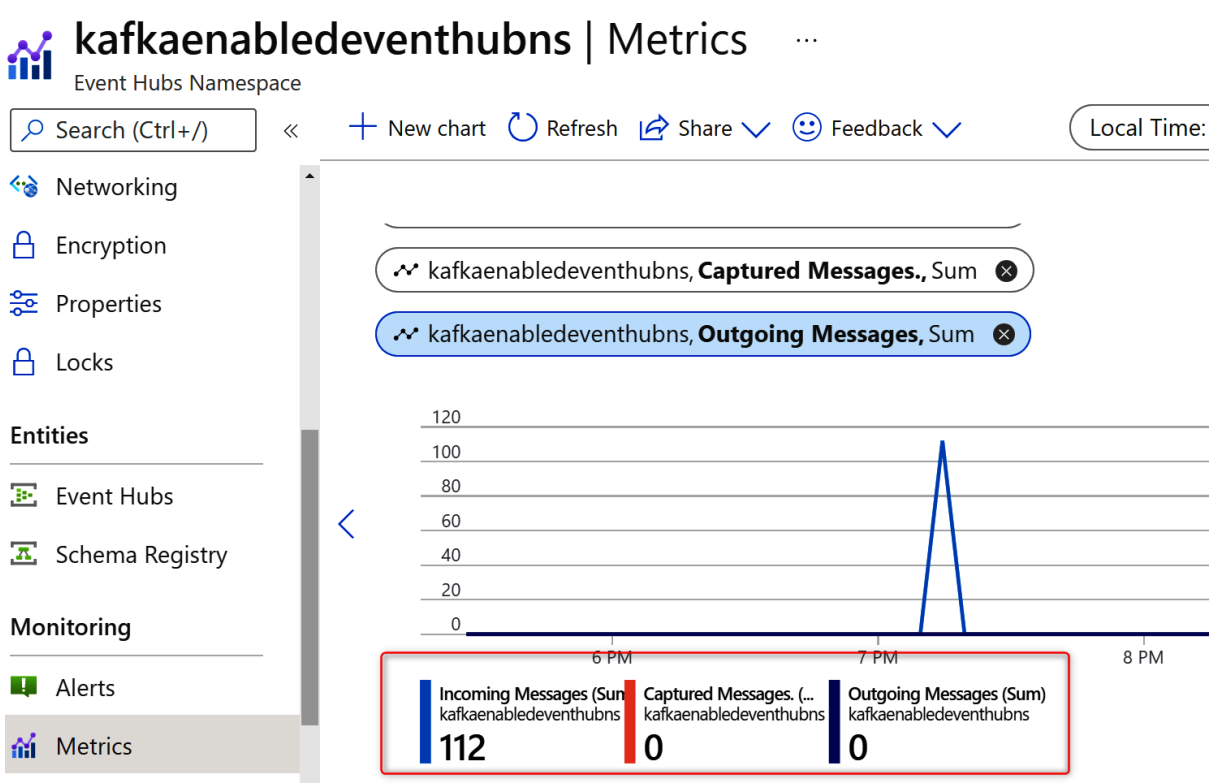


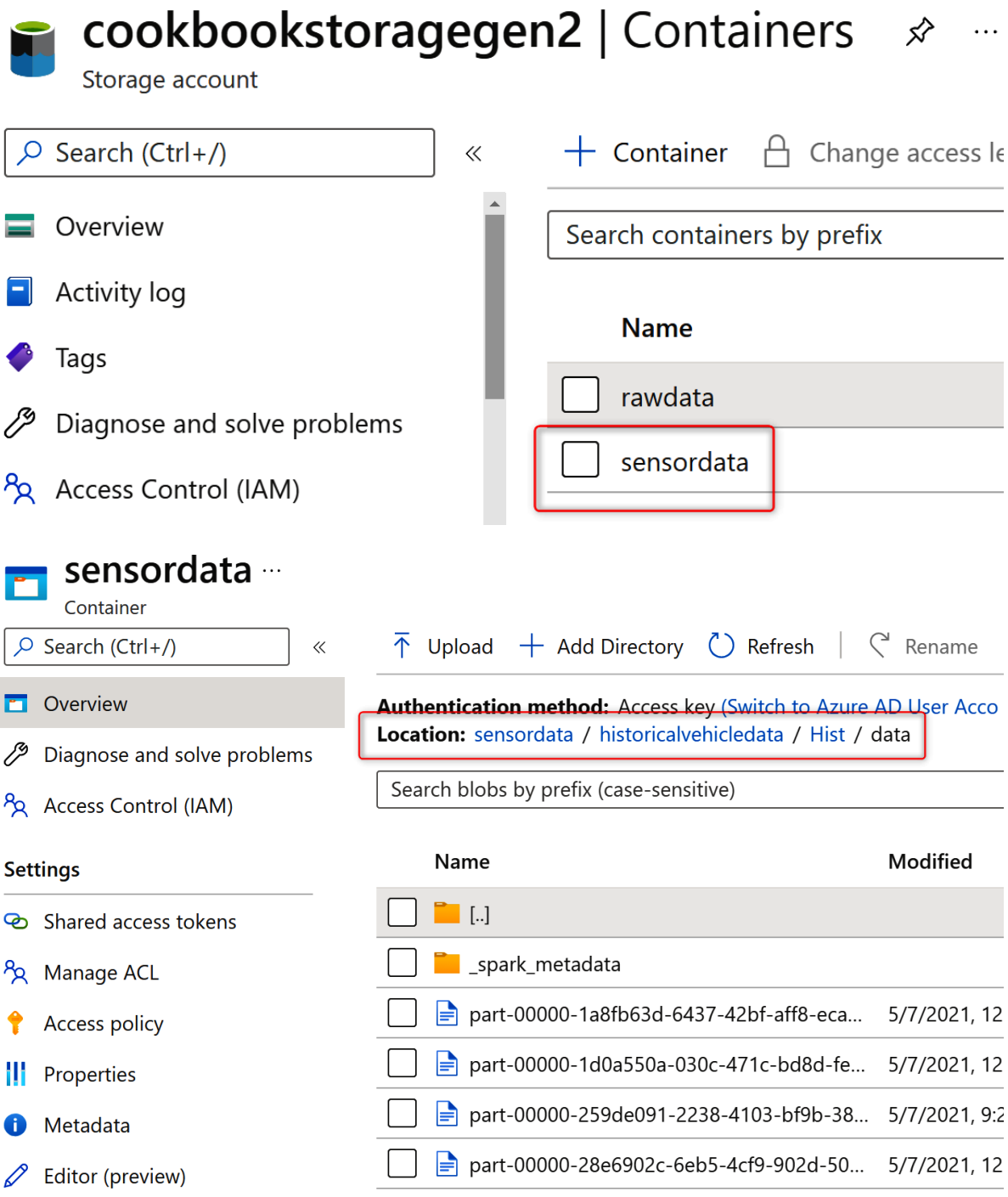

cookbookblobstorage1 | Access keys
Storage account
Search (Ctrl+/)
Networking
Azure CDN
Access keys
Shared access signature
Encryption
Security
Data management
Geo-replication
Data protection
Object replication
Blob inventory
Show keys
Set rotation reminder
Refresh
Access keys authenticate your applications' requests to this storage account. Keep your keys in a
Key Vault, and replace them often with new keys. The two keys allow you to replace one while stil
Remember to update the keys with any Azure resources and apps that use this storage account.
Storage account name
cookbookblobstorage1
key1
Last rotated: 3/15/2021 (125 days ago)
Rotate key
Key
Connection string

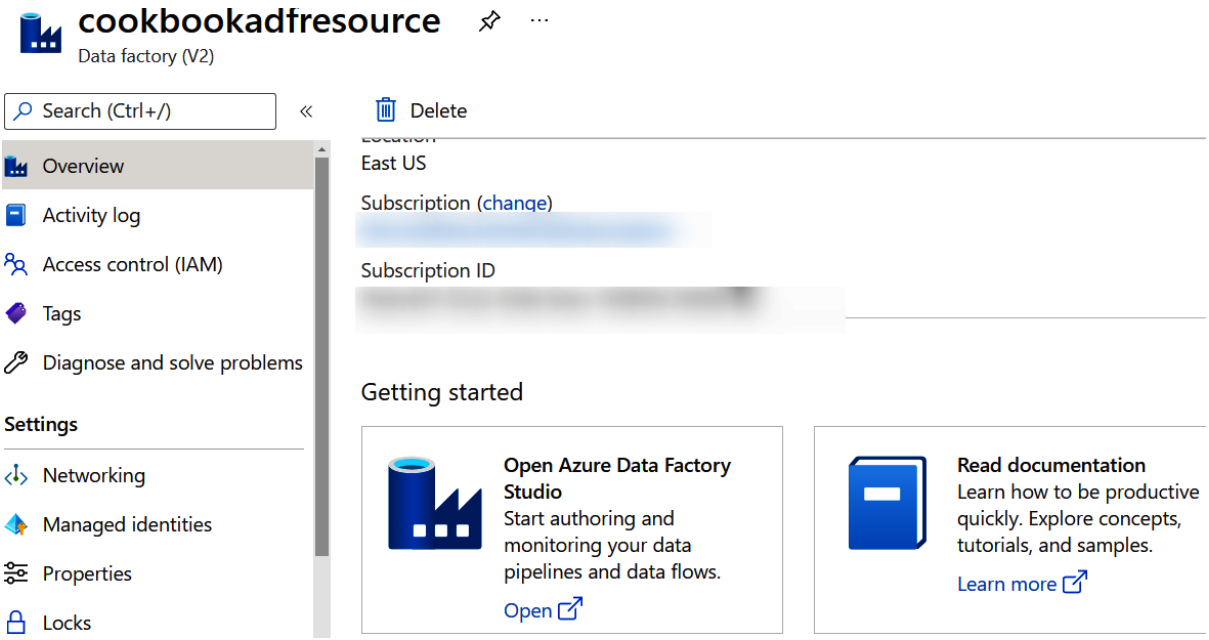

VehicleInformationDB (srvr/VehicleInformationDB) | Conn...
SQL database
Search (Ctrl+/)
Power Automate (preview)
Settings
Compute + storage
Connection strings
Maintenance
Properties
Locks
ADO.NET
JDBC
ODBC
PHP
Go
JDBC (SQL authentication)
jdbc:sqlserver://srvr.database.windows.net:1433;database=VehicleInformationDB;user=sqladmin in
@vehicleinformatiosrvr;password=
{your_password_here};encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.database.windows.net;lo lo
ginTimeout=30;
Download JDBC driver for SQL server

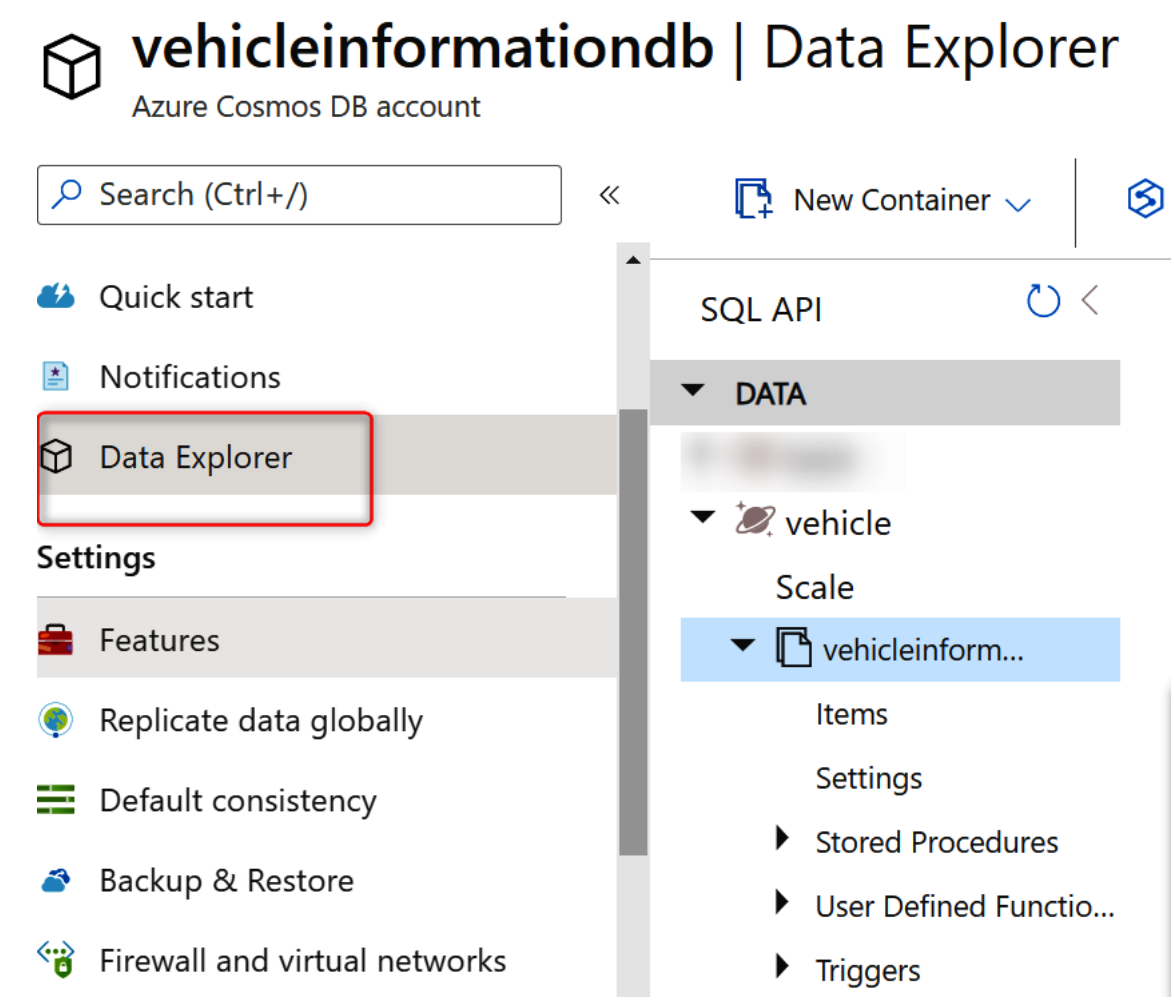

sqldwpool (synapsedemoworkspace11/sqldwpool) | Connection strings
Dedicated SQL pool
Search (Ctrl+/)
Workload management
Maintenance schedule
Geo-backup policy
Connection strings
Properties
Locks
ADO.NET
JDBC
ODBC
PHP
JDBC (SQL authentication)
jdbc:sqlserver://synapsedemoworkspace11.sql.azuresynapse.net:1433;database=sqldwpool;user=sqladminuser@synapsedemoworkspace11;password
={your_password_here};encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.sql.azuresynapse.net;loginTimeout=30;
JDBC (Active Directory password authentication)

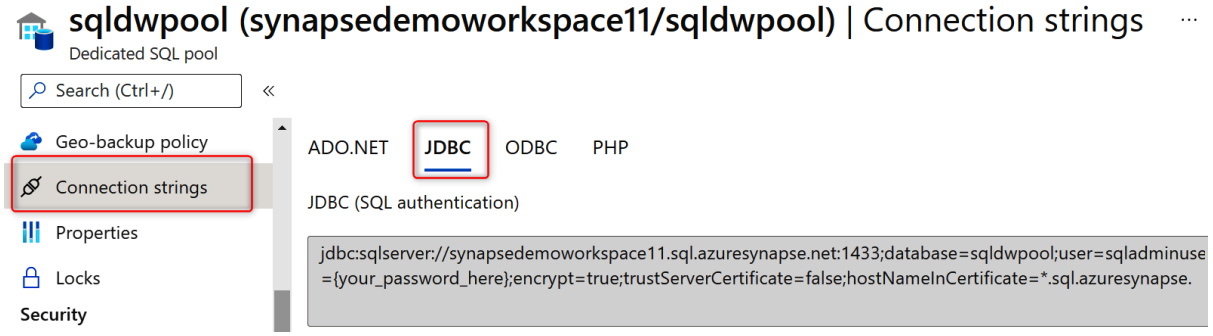

kafkaenabledeventhubns | Metrics
Event Hubs Namespace
Search (Ctrl+/)
New chart
Refresh
Share
Feedback
Networking
Encryption
Properties
Locks
ntities
Event Hubs
Schema Registry
Monitoring
Alerts
Metrics
Diagnostic settings
Logs
utomation
Tasks (preview)
Export template
kafkaenabledeventhubns, Server Errors., Sum
kafkaenabledeventhubns, User Errors., S
Scope
kafkaenabledeventhubns
Metric Namespace
Event Hub standard me...
Metric
Select metric
Agg
Entity... = vehiclesensor...
250
200
150
100
50
0
6 AM
12 PM
6 PM
Successful Requests ...
kafkaenabledeventhubns
116
Incoming Requests (Sum)
kafkaenabledeventhubns
116
Server Errors. (Sum)
kafkaenabledeventhubns
0
User Errors. (Sum)
kafkaenabledeventhubns
0

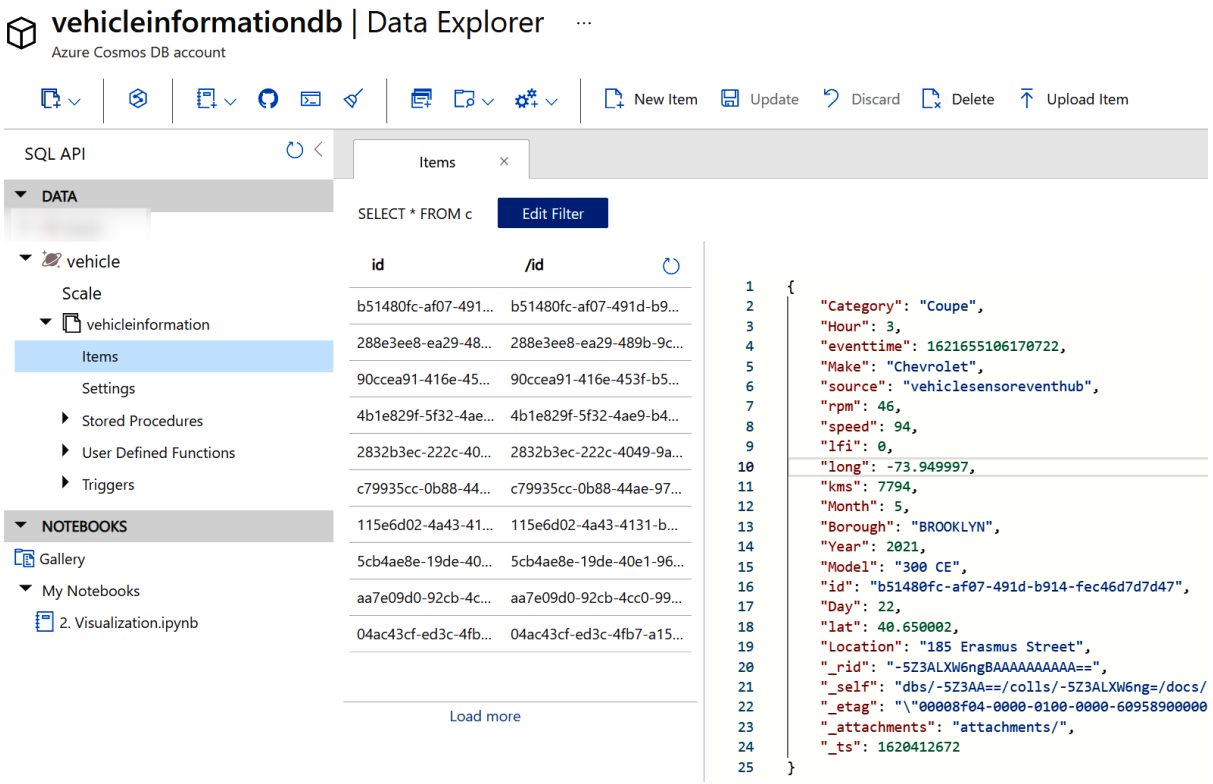

kafkaenabledeventhubns | Metrics
Event Hubs Namespace
Search (Ctrl+/)
New chart
Refresh
Share
Feedback
Local Time:
Networking
Encryption
Properties
Locks
Entities
Event Hubs
Schema Registry
Monitoring
Alerts
Metrics
kafkaenabledeventhubns, Captured Messages., Sum
kafkaenabledeventhubns, Outgoing Messages, Sum
120
100
80
60
40
20
0
6 PM
7 PM
8 PM
Incoming Messages (Sum)
kafkaenabledeventhubns
112
Captured Messages. (...
kafkaenabledeventhubns
0
Outgoing Messages (Sum)
kafkaenabledeventhubns
0

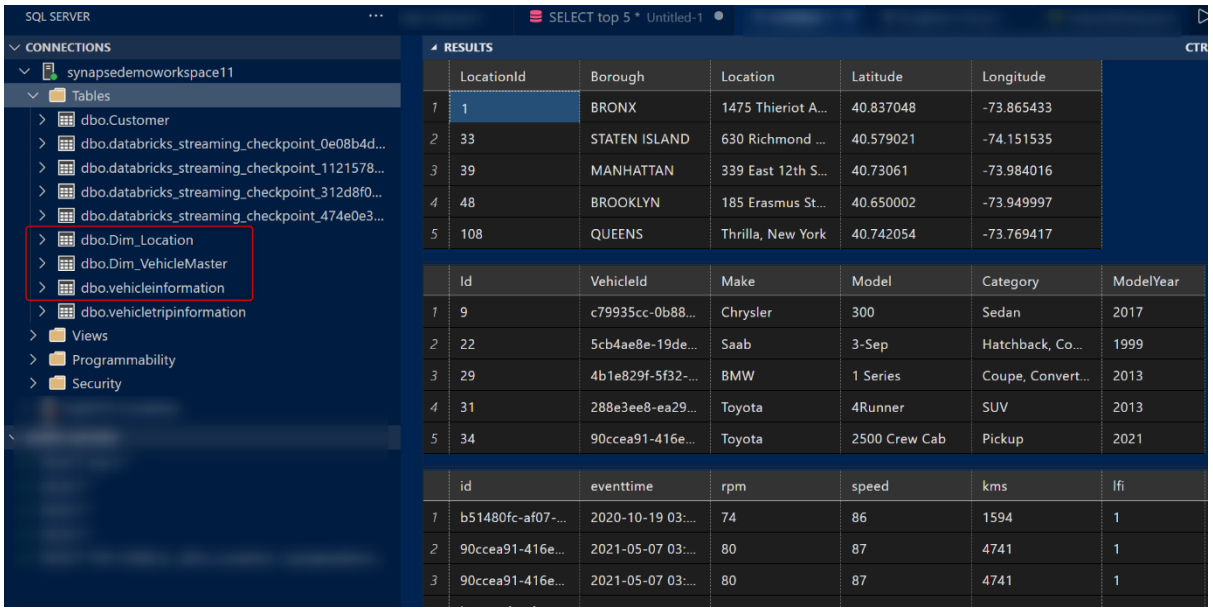

cookbookstoragegen2 | Containers
Storage account
Search (Ctrl+/)
Container
Change access le
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Search containers by prefix
Name
rawdata
sensordata
sensordata
Container
Search (Ctrl+/)
Upload
Add Directory
Refresh
Rename
Overview
Diagnose and solve problems
Access Control (IAM)
Authentication method: Access key (Switch to Azure AD User Acco
Location: sensordata / historicalvehicledata / Hist / data
Search blobs by prefix (case-sensitive)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Editor (preview)
Name
Modified
[..]
_spark_metadata
part-00000-1a8fb63d-6437-42bf-aff8-eca...
5/7/2021, 12
part-00000-1d0a550a-030c-471c-bd8d-fe...
5/7/2021, 12
part-00000-259de091-2238-4103-bf9b-38...
5/7/2021, 9:2
part-00000-28e6902c-6eb5-4cf9-902d-50...
5/7/2021, 12


cookbookadfresource
Data factory (V2)
Search (Ctrl+/)
Delete
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Networking
Managed identities
Properties
Locks
East US
Subscription (change)
Subscription ID
Getting started
Open Azure Data Factory Studio
Start authoring and monitoring your data pipelines and data flows.
Open
Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more


vehicleinformationdb | Data Explorer
Azure Cosmos DB account
Search (Ctrl+/)
New Container
Quick start
Notifications
Data Explorer
Settings
Features
Replicate data globally
Default consistency
Backup & Restore
Firewall and virtual networks
SQL API
DATA
vehicle
Scale
vehicleinform...
Items
Settings
Stored Procedures
User Defined Functio...
Triggers


sqldwpool (synapsedemoworkspace11/sqldwpool) | Connection strings
Dedicated SQL pool
Search (Ctrl+/)
Geo-backup policy
Connection strings
Properties
Locks
Security
ADO.NET
JDBC
ODBC
PHP
JDBC (SQL authentication)
jdbc:sqlserver://synapsedemoworkspace11.sql.azuresynapse.net:1433;database=sqldwpool;user=sqladminuse
={your_password_here};encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.sql.azuresynapse.


vehicleinformationdb | Data Explorer
Azure Cosmos DB account
New Item
Update
Discard
Delete
Upload Item
SQL API
DATA
vehicle
Scale
vehicleinformation
Items
Settings
Stored Procedures
User Defined Functions
Triggers
NOTEBOOKS
Gallery
My Notebooks
2. Visualization.ipynb
Items
SELECT * FROM c
Edit Filter
id
/id
b51480fc-af07-491...
b51480fc-af07-491d-b9...
288e3ee8-ea29-48...
288e3ee8-ea29-489b-9c...
90ccea91-416e-45...
90ccea91-416e-453f-b5...
4b1e829f-5f32-4ae...
4b1e829f-5f32-4ae9-b4...
2832b3ec-222c-40...
2832b3ec-222c-4049-9a...
c79935cc-0b88-44...
c79935cc-0b88-44ae-97...
115e6d02-4a43-41...
115e6d02-4a43-4131-b...
5cb4ae8e-19de-40...
5cb4ae8e-19de-40e1-96...
aa7e09d0-92cb-4c...
aa7e09d0-92cb-4cc0-99...
04ac43cf-ed3c-4fb...
04ac43cf-ed3c-4fb7-a15...
Load more
{
"Category": "Coupe",
"Hour": 3,
"eventtime": 1621655106170722,
"Make": "Chevrolet",
"source": "vehiclesensoreventhub",
"rpm": 46,
"speed": 94,
"lfi": 0,
"long": -73.949997,
"kms": 7794,
"Month": 5,
"Borough": "BROOKLYN",
"Year": 2021,
"Model": "300 CE",
"id": "b51480fc-af07-491d-b914-fec46d7d7d47",
"Day": 22,
"lat": 40.650002,
"Location": "185 Erasmus Street",
"_rid": "-5Z3ALXW6ngBAAAAAAAAAA==",
"_self": "dbs/-5Z3AA==/colls/-5Z3ALXW6ng=/docs/
"_etag": "\"00008f04-0000-0100-0000-60958900000
"_attachments": "attachments/",
"_ts": 1620412672
}

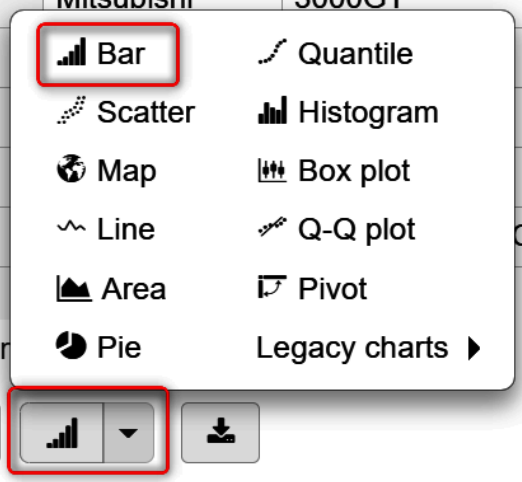

SQL SERVER
SELECT top 5 * Untitled-1
CONNECTIONS
synapsedemoworkspace11
Tables
dbo.Customer
dbo.databricks_streaming_checkpoint_0e08b4d...
dbo.databricks_streaming_checkpoint_1121578...
dbo.databricks_streaming_checkpoint_312d8f0...
dbo.databricks_streaming_checkpoint_474e0e3...
dbo.Dim_Location
dbo.Dim_VehicleMaster
dbo.vehicleinformation
dbo.vehicletripinformation
Views
Programmability
Security
RESULTS
LocationId Borough Location Latitude Longitude
1 BRONX 1475 Thieriot A... 40.837048 -73.865433
33 STATEN ISLAND 630 Richmond ... 40.579021 -74.151535
39 MANHATTAN 339 East 12th S... 40.73061 -73.984016
48 BROOKLYN 185 Erasmus St... 40.650002 -73.949997
108 QUEENS Thrilla, New York 40.742054 -73.769417
Id VehicleId Make Model Category ModelYear
9 c79935cc-0b88... Chrysler 300 Sedan 2017
22 5cb4ae8e-19de... Saab 3-Sep Hatchback, Co... 1999
29 4b1e829f-5f32-... BMW 1 Series Coupe, Convert... 2013
31 288e3ee8-ea29... Toyota 4Runner SUV 2013
34 90ccea91-416e... Toyota 2500 Crew Cab Pickup 2021
id eventtime rpm speed kms lfi
b51480fc-af07-... 2020-10-19 03:... 74 86 1594 1
90ccea91-416e... 2021-05-07 03:... 80 87 4741 1
90ccea91-416e... 2021-05-07 03:... 80 87 4741 1

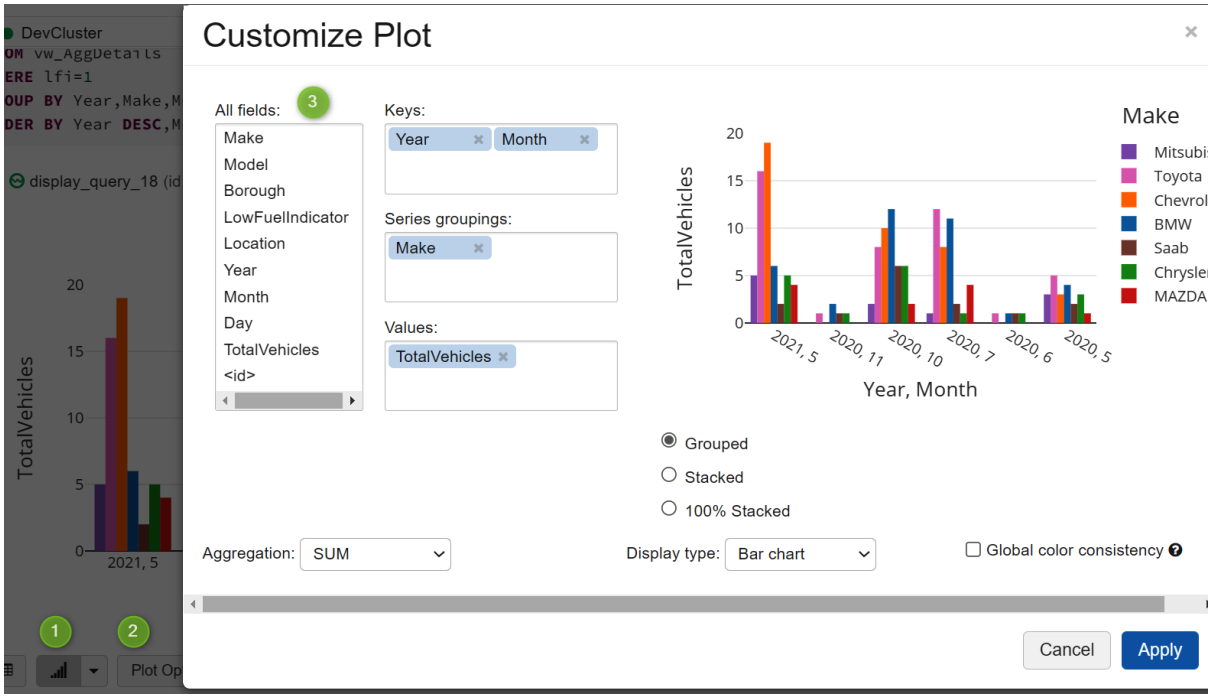
```
GROUP BY Year,Make,Model,coalesce(Borough,"UnKnown"),Location,
ORDER BY Year DESC,Month DESC,TotalVehicles DESC
```

▸ display_query_18 (id: fde636c1-8cfd-405a-a44a-7bebf9d64f6a) *Last update*

	Make	Model	Borough	LowFuelIn
1	Mitsubishi	3000GT	QUEENS	1
2	Toyota	2500 Crew Cab	QUEENS	1
3	Toyota	4Runner	STATEN ISLAND	1
4	Toyota	2500 Crew Cab	BRONX	1
5	Chevrolet	300 CE	MANHATTAN	1
6	BMW	M Series	STATEN ISLAND	1
7	Mitsubishi	3000GT	BROOKLYN	1
8			MANHATTAN	1
9			BRONX	1
10			BRONX	1
11		Cab	STATEN ISLAND	1
12			MANHATTAN	1

Showin



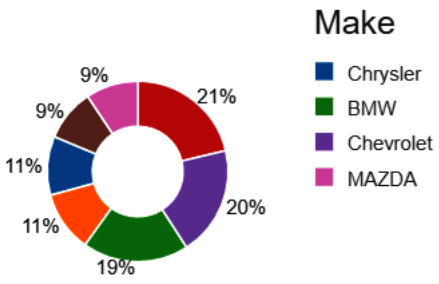


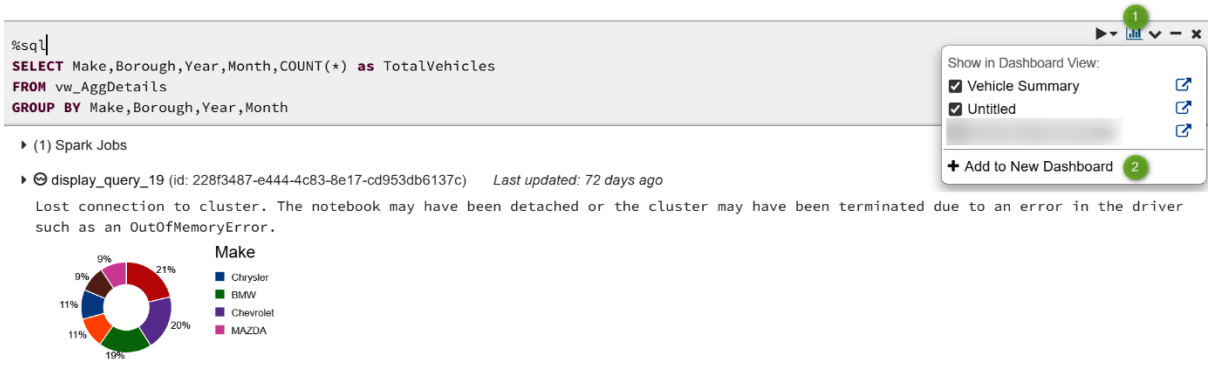

```
SELECT Make,Borough,Year,Month,COUNT(*) as TotalVehicles
FROM vw_AggDetails
GROUP BY Make,Borough,Year,Month
```

▸ (1) Spark Jobs

▸ display_query_19 (id: 228f3487-e444-4c83-8e17-cd953db6137c) *Last upd*

```
Lost connection to cluster. The notebook may have been detac
yError.
```

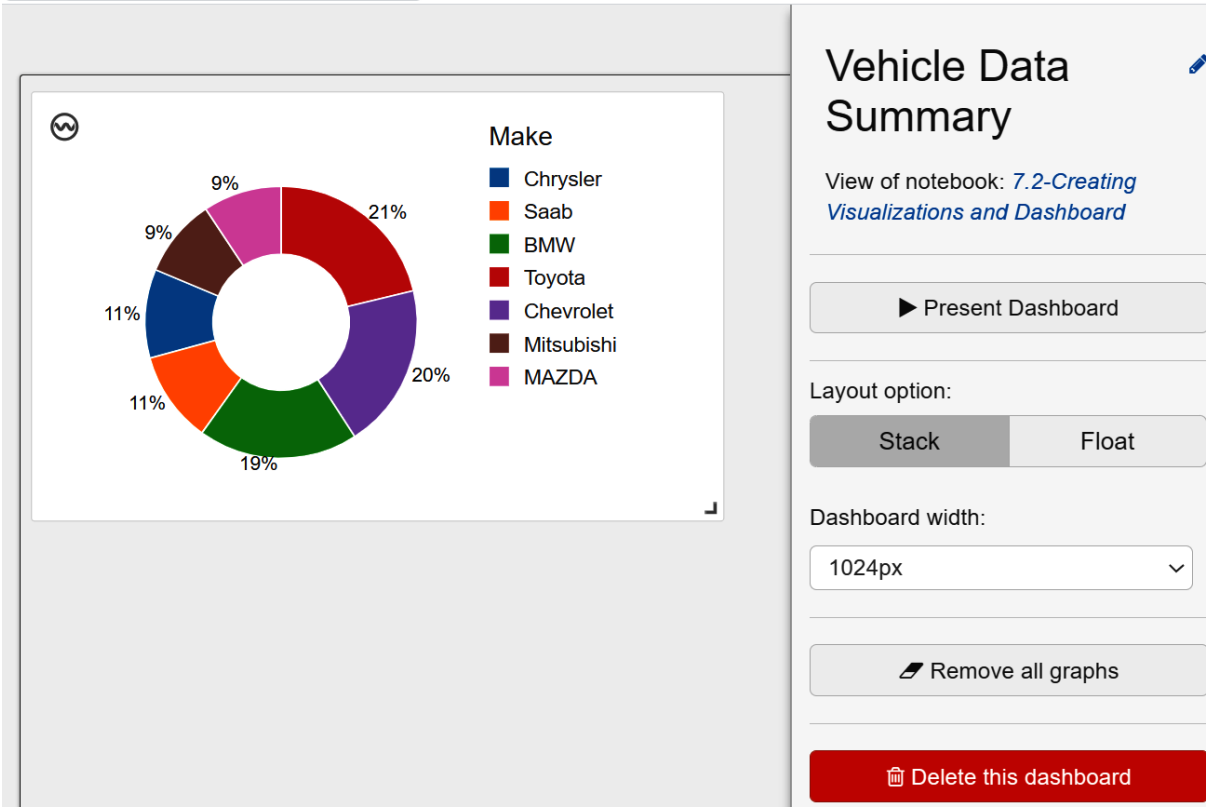




7.2-Creating Visualizations and Dashboard (Python)

KakfaStreamingWork...

DevCluster



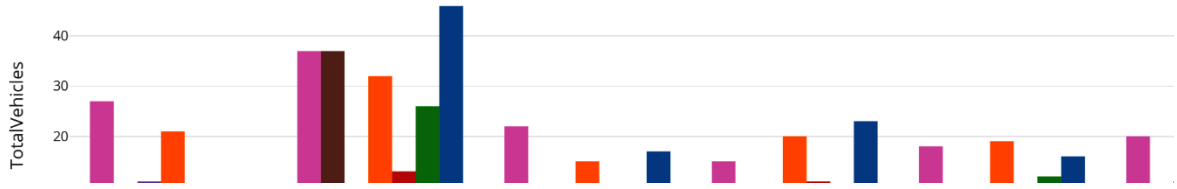

```
%sql
SELECT Id as VehicleId,Make,Model,coalesce(Borough,"Unknown") AS Borough,Location,Year,Month,Day,COUNT(*) as
FROM vw_AggDetails
GROUP BY Make,Model,coalesce(Borough,"Unknown"),Location,Year,Month,Day,Id
```

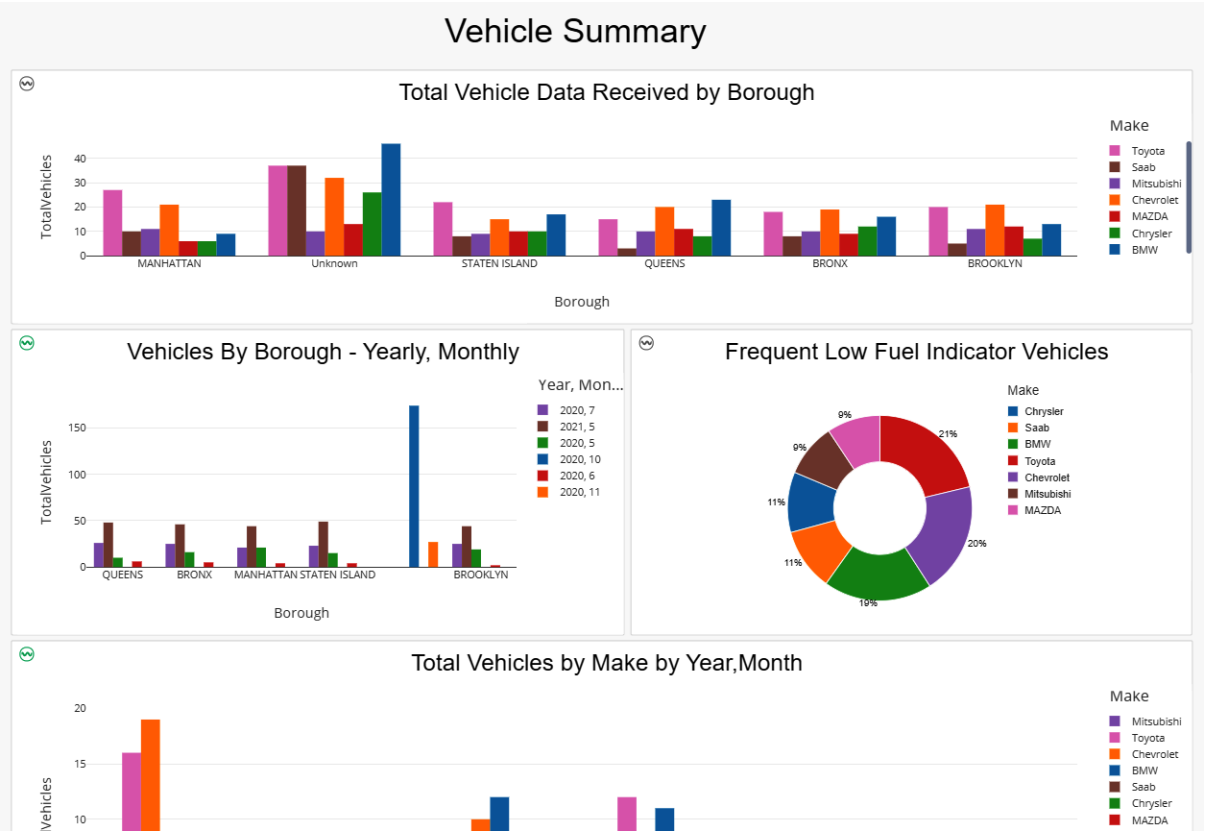
Show in Dashboard View:
- Vehicle Summary
- Untitled
- Vehicle Data Summary
- Vehicle Data Summary
- Add to New Dashboard

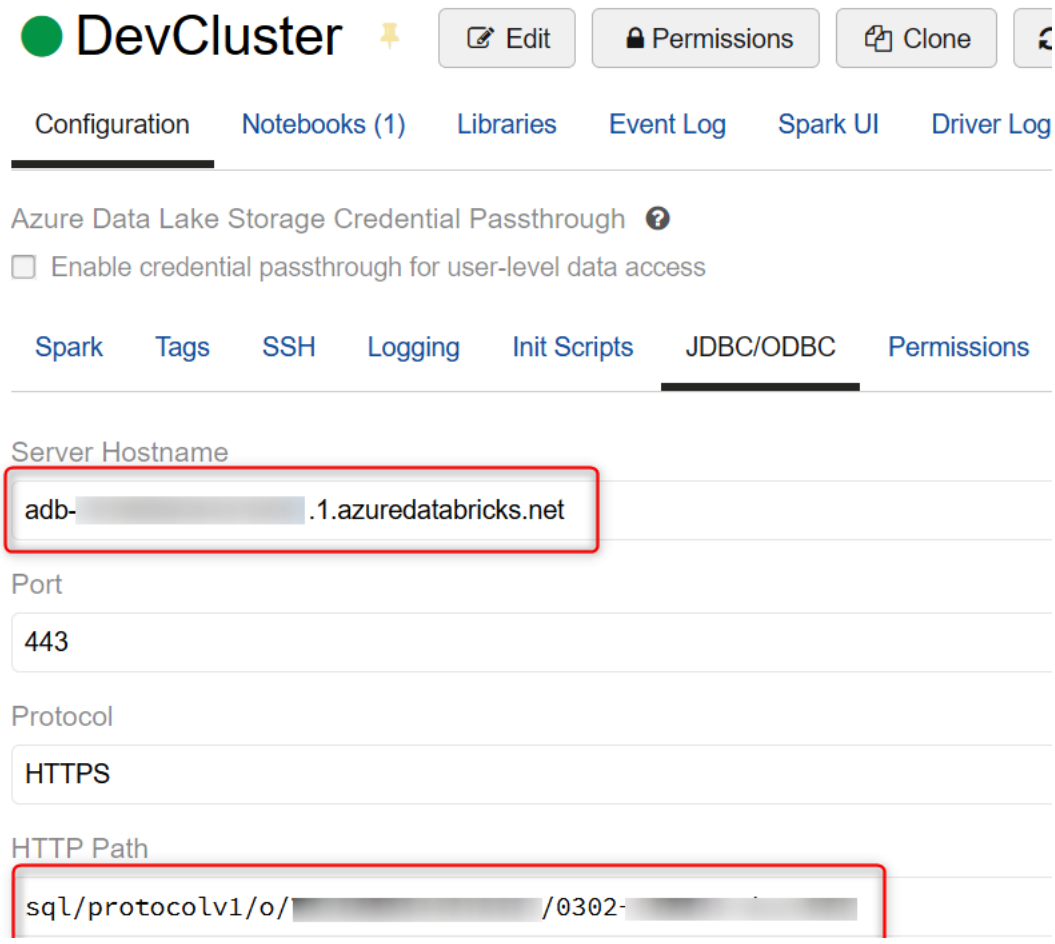
(1) Spark Jobs

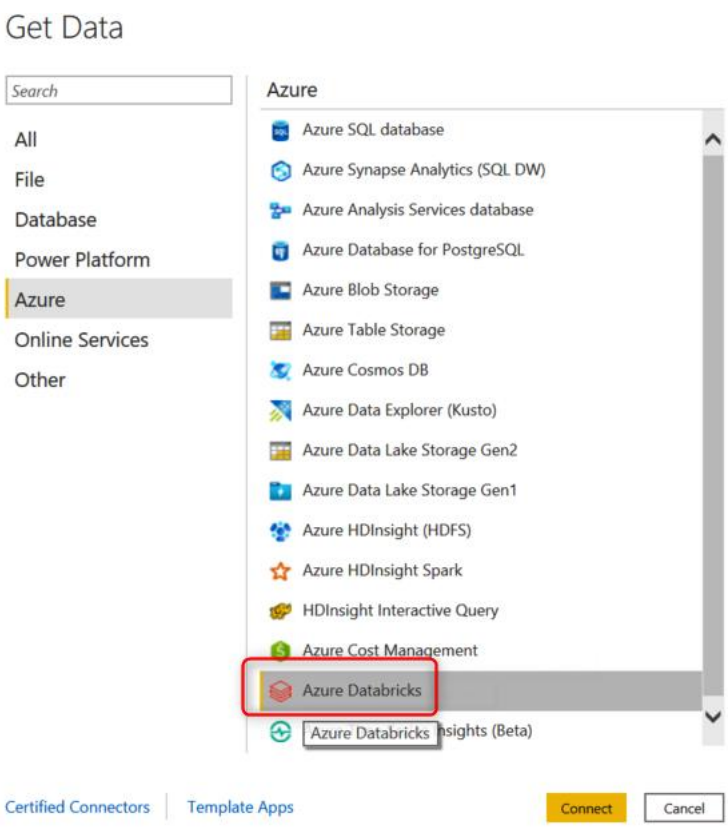
display_query_20 (id: 209f3957-ad3a-40f0-96da-9c1a20b64544) *Last updated: 72 days ago*

Lost connection to cluster. The notebook may have been detached or the cluster may have been terminated due to an error in the driver such as an OutOfMemoryError.




Vehicle Summary
Total Vehicle Data Received by Borough
TotalVehicles
40
30
20
10
0
MANHATTAN
Unknown
STATEN ISLAND
QUEENS
BRONX
BROOKLYN
Borough
Make
Toyota
Saab
Mitsubishi
Chevrolet
MAZDA
Chrysler
BMW
Vehicles By Borough - Yearly, Monthly
TotalVehicles
150
100
50
0
QUEENS
BRONX
MANHATTAN
STATEN ISLAND
BROOKLYN
Borough
Year, Mon...
2020, 7
2021, 5
2020, 5
2020, 10
2020, 6
2020, 11
Frequent Low Fuel Indicator Vehicles
Make
Chrysler
Saab
BMW
Toyota
Chevrolet
Mitsubishi
MAZDA
9%
21%
9%
11%
20%
11%
19%
Total Vehicles by Make by Year,Month
Vehicles
20
15
10
Make
Mitsubishi
Toyota
Chevrolet
BMW
Saab
Chrysler
MAZDA


DevCluster
Edit
Permissions
Clone
Configuration
Notebooks (1)
Libraries
Event Log
Spark UI
Driver Log
Azure Data Lake Storage Credential Passthrough
Enable credential passthrough for user-level data access
Spark
Tags
SSH
Logging
Init Scripts
JDBC/ODBC
Permissions
Server Hostname
adb-[redacted].1.azuredatabricks.net
Port
443
Protocol
HTTPS
HTTP Path
sql/protocolv1/o/[redacted]/0302-[redacted]



Azure Databricks

Server Hostname ⓘ

adb· 1.azuredatabricks.net

HTTP Path ⓘ

sql/protocolv1/o/ /03 1-dour999

▲ Advanced Options (optional)

Database (optional) ⓘ

Example: abc

Batch Size (rows) (optional) ⓘ

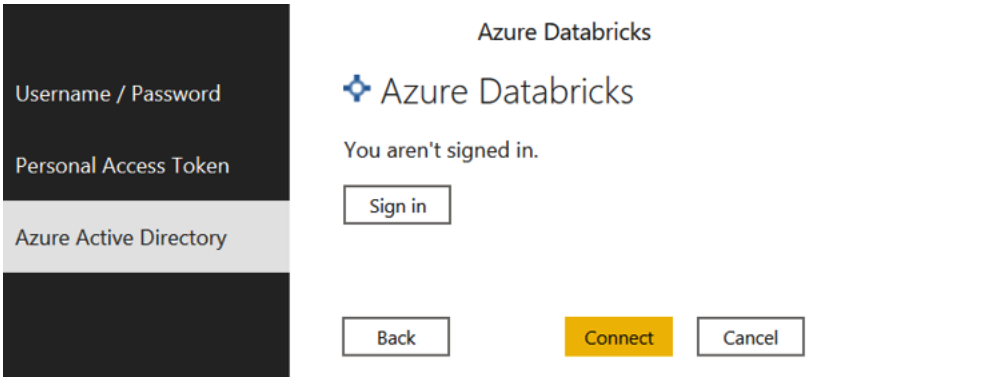
Example: 10000

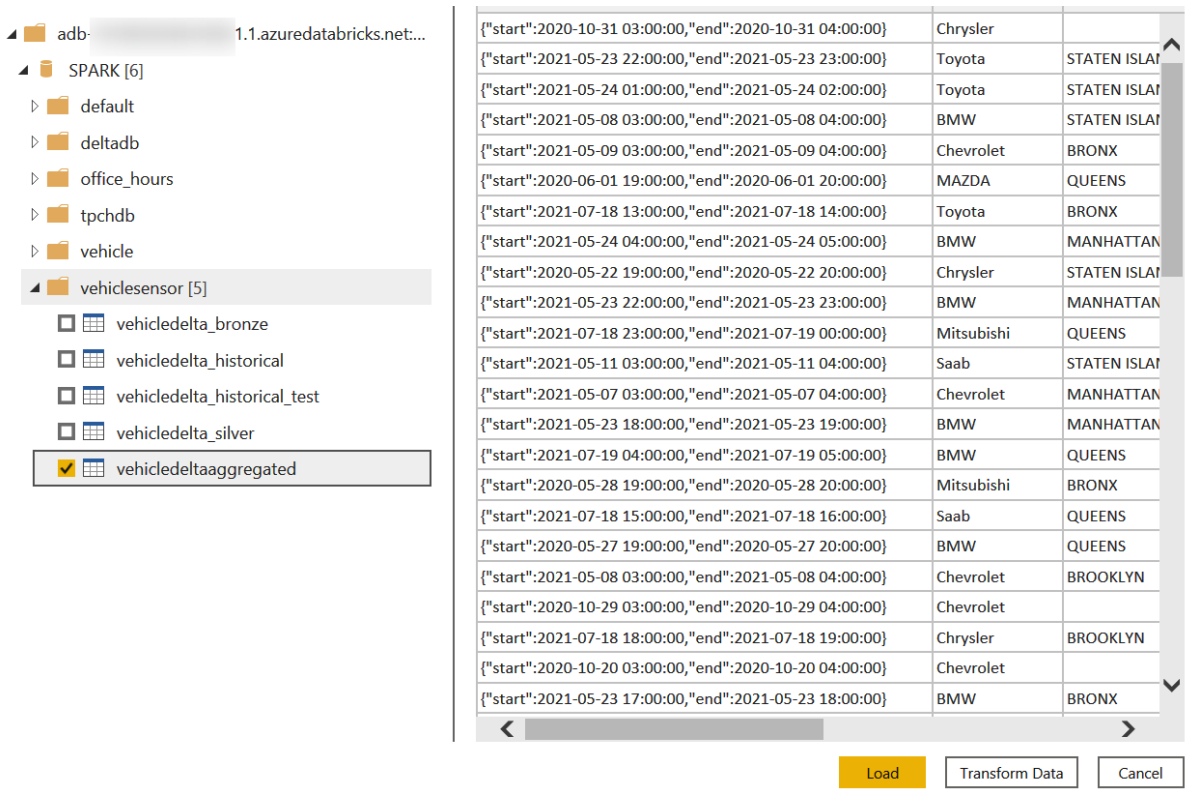
Data Connectivity mode ⓘ

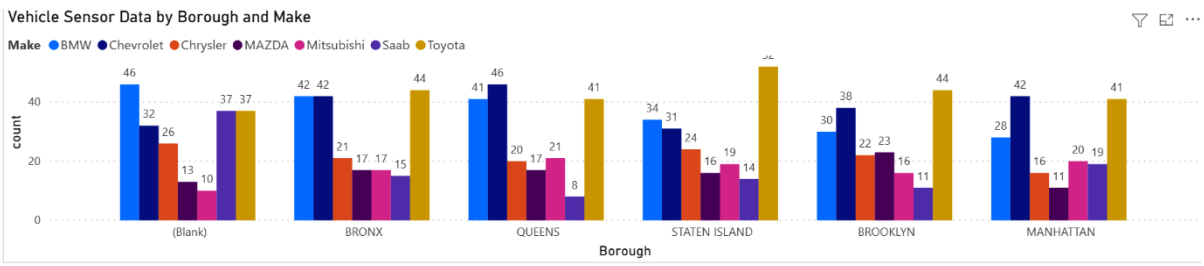
○ Import

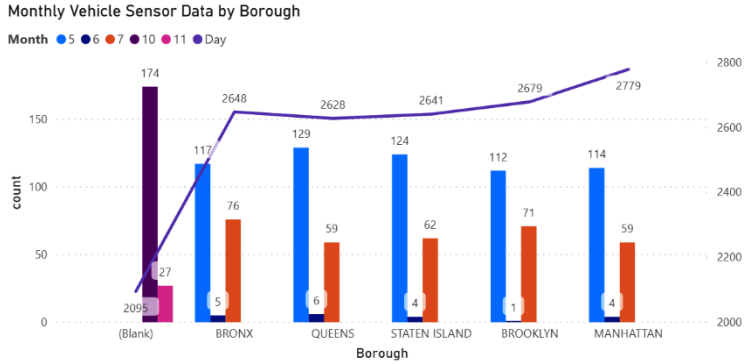
◉ DirectQuery

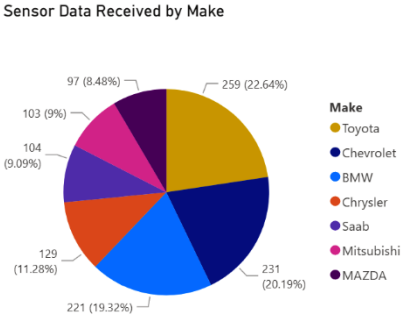
OK Cancel

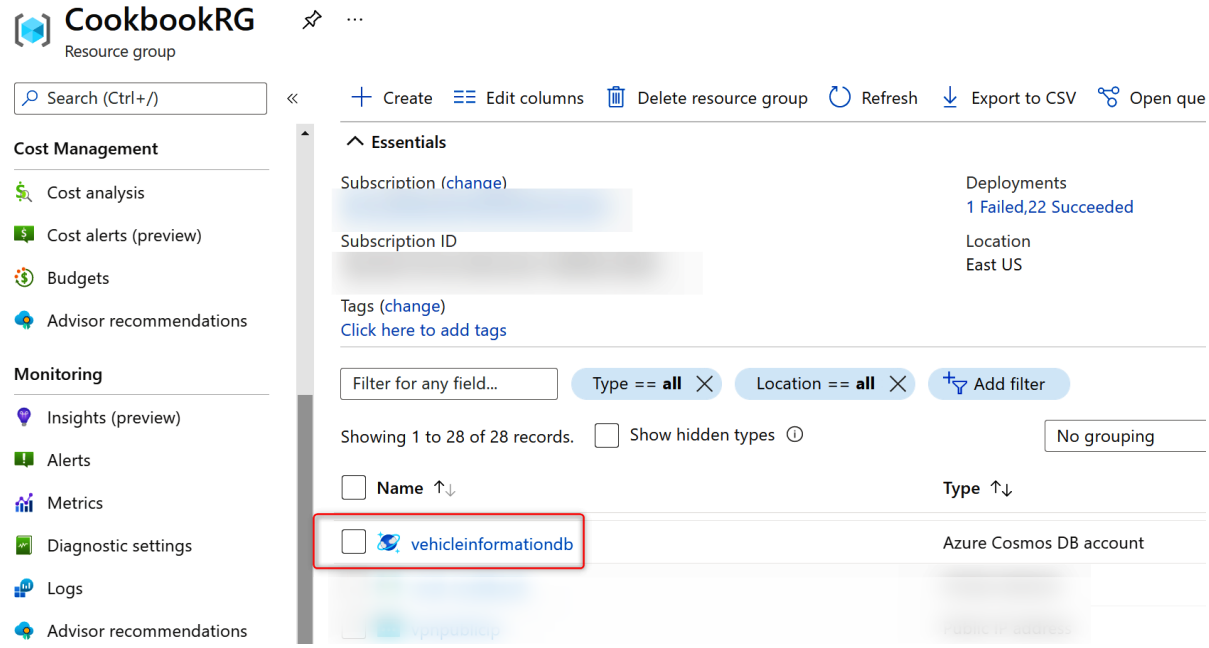

Username / Password
Personal Access Token
Azure Active Directory
Azure Databricks
Azure Databricks
You aren't signed in.
Sign in
Back
Connect
Cancel

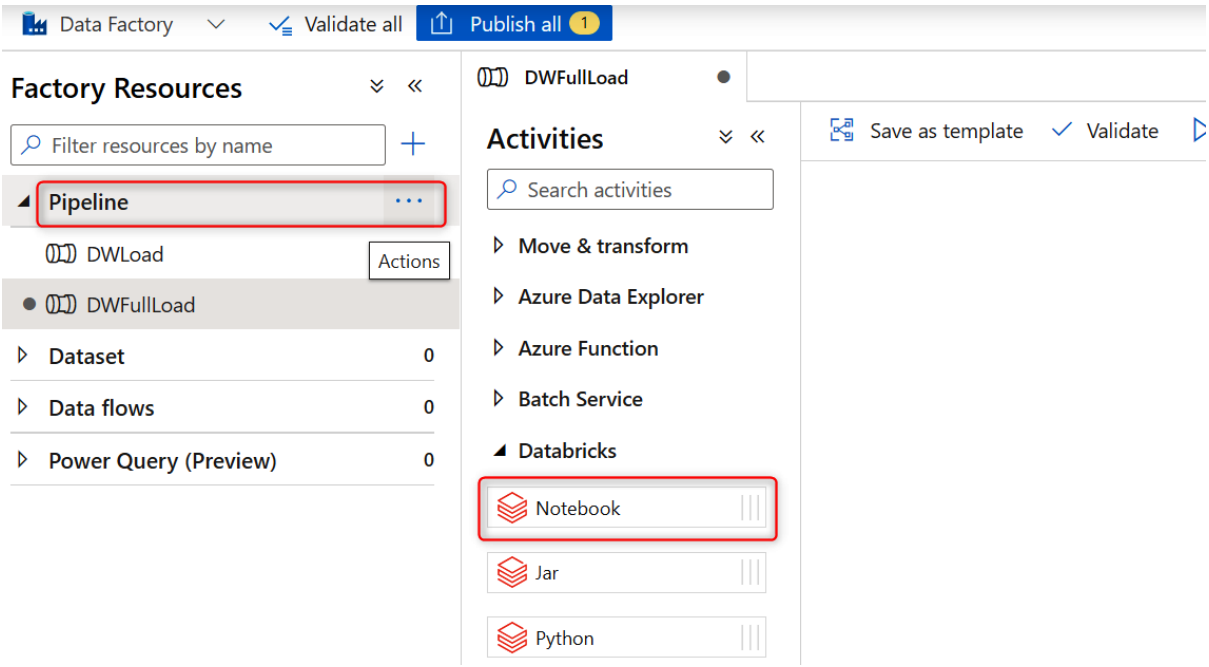

adb- 1.1.azuredatabricks.net:...
SPARK [6]
default
deltadb
office_hours
tpchdb
vehicle
vehiclesensor [5]
vehicledelta_bronze
vehicledelta_historical
vehicledelta_historical_test
vehicledelta_silver
vehicledeltaaggregated
{"start":2020-10-31 03:00:00,"end":2020-10-31 04:00:00} Chrysler
{"start":2021-05-23 22:00:00,"end":2021-05-23 23:00:00} Toyota STATEN ISLAI
{"start":2021-05-24 01:00:00,"end":2021-05-24 02:00:00} Toyota STATEN ISLAI
{"start":2021-05-08 03:00:00,"end":2021-05-08 04:00:00} BMW STATEN ISLAI
{"start":2021-05-09 03:00:00,"end":2021-05-09 04:00:00} Chevrolet BRONX
{"start":2020-06-01 19:00:00,"end":2020-06-01 20:00:00} MAZDA QUEENS
{"start":2021-07-18 13:00:00,"end":2021-07-18 14:00:00} Toyota BRONX
{"start":2021-05-24 04:00:00,"end":2021-05-24 05:00:00} BMW MANHATTAN
{"start":2020-05-22 19:00:00,"end":2020-05-22 20:00:00} Chrysler STATEN ISLAI
{"start":2021-05-23 22:00:00,"end":2021-05-23 23:00:00} BMW MANHATTAN
{"start":2021-07-18 23:00:00,"end":2021-07-19 00:00:00} Mitsubishi QUEENS
{"start":2021-05-11 03:00:00,"end":2021-05-11 04:00:00} Saab STATEN ISLAI
{"start":2021-05-07 03:00:00,"end":2021-05-07 04:00:00} Chevrolet MANHATTAN
{"start":2021-05-23 18:00:00,"end":2021-05-23 19:00:00} BMW MANHATTAN
{"start":2021-07-19 04:00:00,"end":2021-07-19 05:00:00} BMW QUEENS
{"start":2020-05-28 19:00:00,"end":2020-05-28 20:00:00} Mitsubishi BRONX
{"start":2021-07-18 15:00:00,"end":2021-07-18 16:00:00} Saab QUEENS
{"start":2020-05-27 19:00:00,"end":2020-05-27 20:00:00} BMW QUEENS
{"start":2021-05-08 03:00:00,"end":2021-05-08 04:00:00} Chevrolet BROOKLYN
{"start":2020-10-29 03:00:00,"end":2020-10-29 04:00:00} Chevrolet
{"start":2021-07-18 18:00:00,"end":2021-07-18 19:00:00} Chrysler BROOKLYN
{"start":2020-10-20 03:00:00,"end":2020-10-20 04:00:00} Chevrolet
{"start":2021-05-23 17:00:00,"end":2021-05-23 18:00:00} BMW BRONX
Load
Transform Data
Cancel

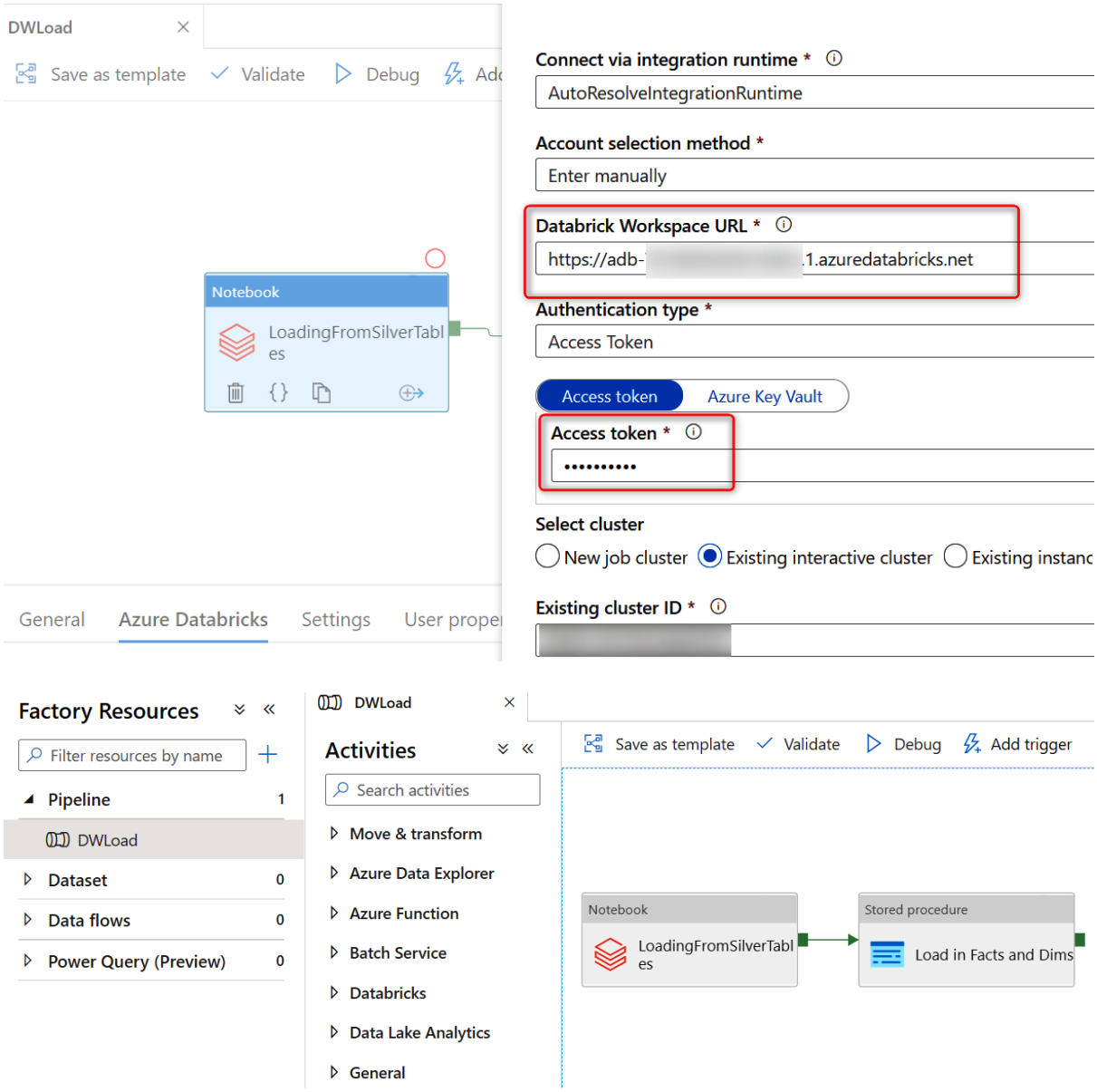

Vehicle Sensor Data by Borough and Make
Make BMW Chevrolet Chrysler MAZDA Mitsubishi Saab Toyota
count
Borough
(Blank)
BRONX
QUEENS
STATEN ISLAND
BROOKLYN
MANHATTAN


Monthly Vehicle Sensor Data by Borough
Month 5 6 7 10 11 Day
count
Borough
(Blank)
BRONX
QUEENS
STATEN ISLAND
BROOKLYN
MANHATTAN


Sensor Data Received by Make
97 (8.48%)
259 (22.64%)
103 (9%)
104 (9.09%)
129 (11.28%)
231 (20.19%)
221 (19.32%)
Make
Toyota
Chevrolet
BMW
Chrysler
Saab
Mitsubishi
MAZDA

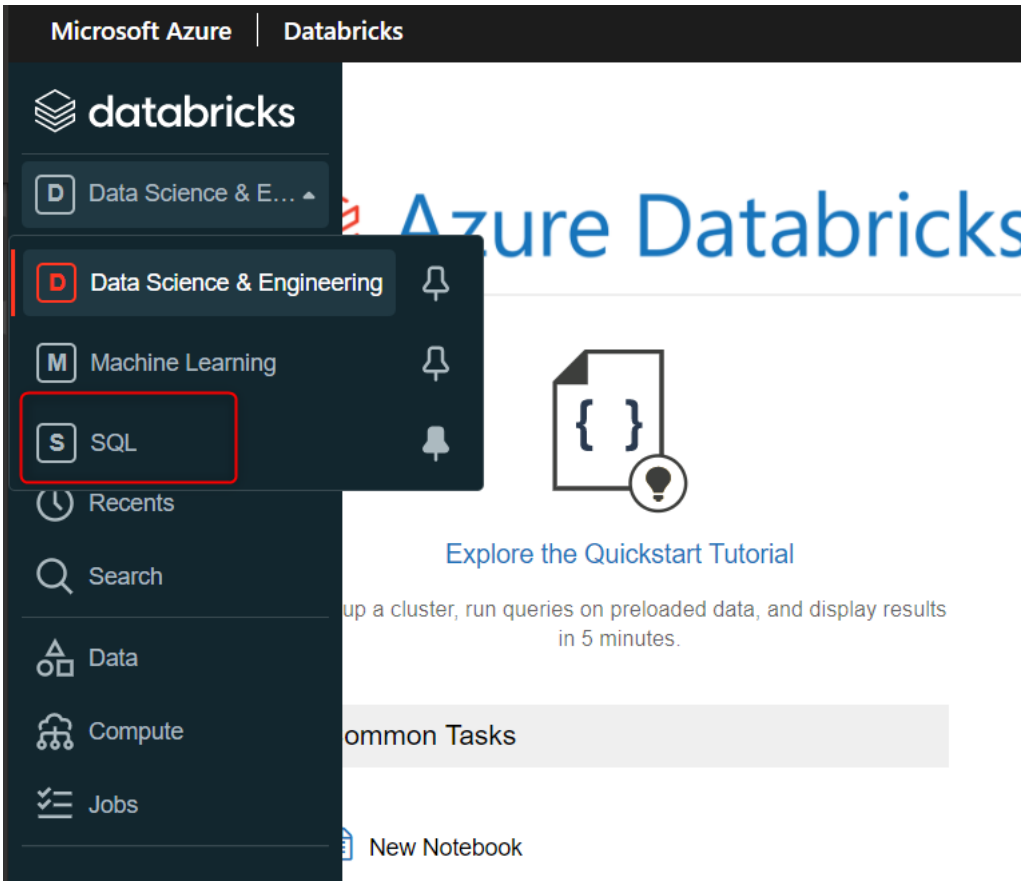

CookbookRG
Resource group
Search (Ctrl+/)
Create
Edit columns
Delete resource group
Refresh
Export to CSV
Open que
Cost Management
Cost analysis
Cost alerts (preview)
Budgets
Advisor recommendations
Monitoring
Insights (preview)
Alerts
Metrics
Diagnostic settings
Logs
Advisor recommendations
Essentials
Subscription (change)
Subscription ID
Deployments
1 Failed,22 Succeeded
Location
East US
Tags (change)
Click here to add tags
Filter for any field...
Type == all
Location == all
Add filter
Showing 1 to 28 of 28 records.
Show hidden types
No grouping
Name
Type
vehicleinformationdb
Azure Cosmos DB account

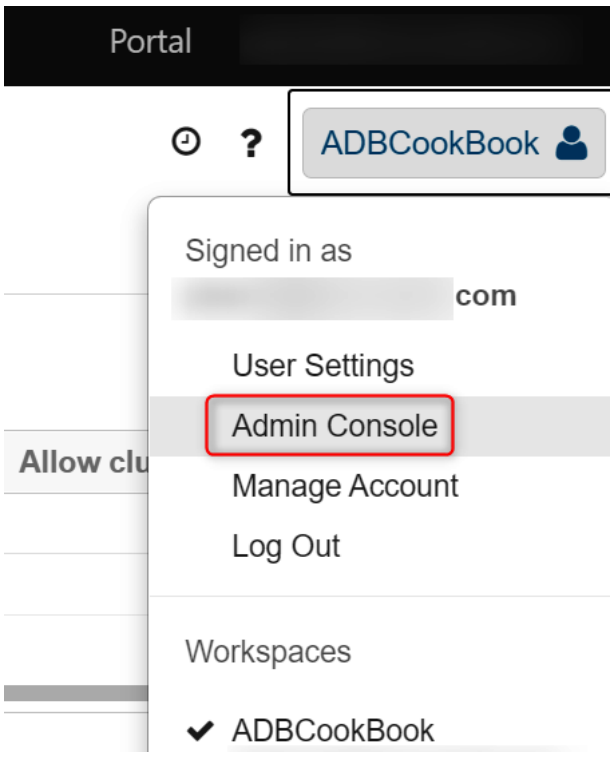

Data Factory
Validate all
Publish all 1
Factory Resources
Filter resources by name
Pipeline
DWLoad
Actions
DWFullLoad
Dataset 0
Data flows 0
Power Query (Preview) 0
DWFullLoad
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Notebook
Jar
Python
Save as template
Validate

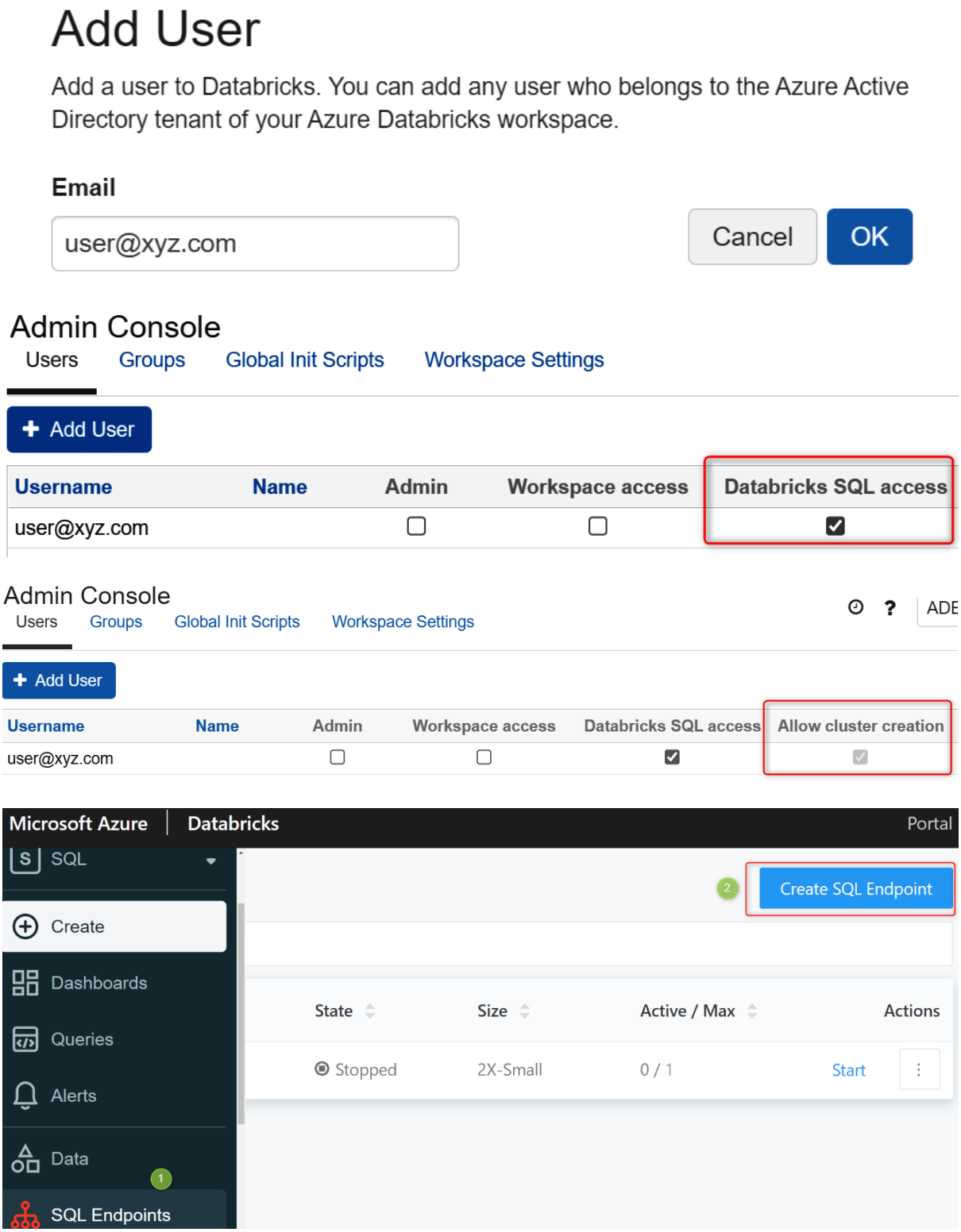

DWLoad
Save as template
Validate
Debug
Notebook
LoadingFromSilverTables
Connect via integration runtime *
AutoResolveIntegrationRuntime
Account selection method *
Enter manually
Databrick Workspace URL *
https://adb-
.1.azuredatabricks.net
Authentication type *
Access Token
Access token
Azure Key Vault
Access token *
Select cluster
New job cluster
Existing interactive cluster
Existing instanc
General
Azure Databricks
Settings
User proper
Existing cluster ID *

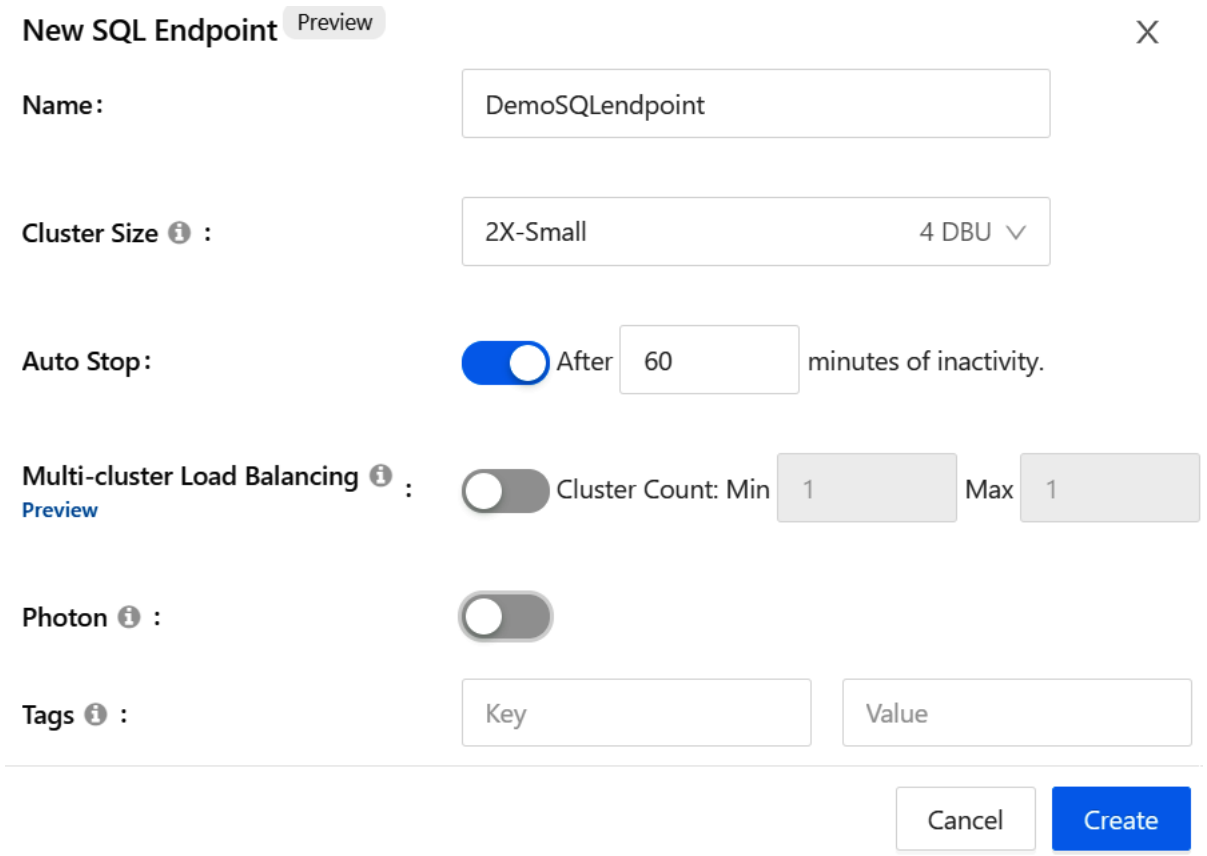
Factory Resources
Filter resources by name
Pipeline 1
DWLoad
Dataset 0
Data flows 0
Power Query (Preview) 0
DWLoad
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
Save as template
Validate
Debug
Add trigger
Notebook
LoadingFromSilverTables
Stored procedure
Load in Facts and Dims

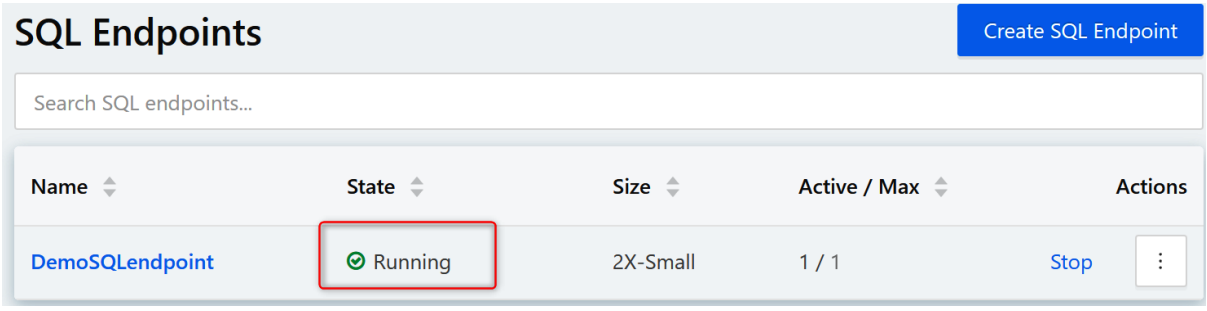
Chapter 8: Azure Databricks SQL Analytics

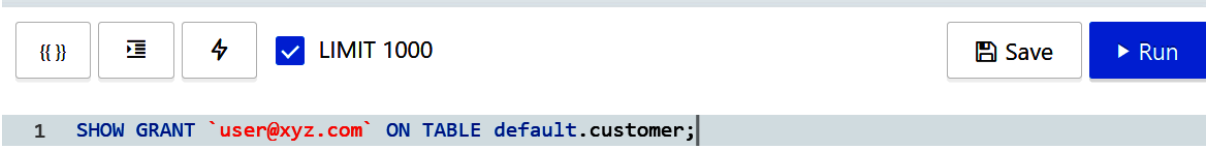





Add User
Add a user to Databricks. You can add any user who belongs to the Azure Active Directory tenant of your Azure Databricks workspace.
Email
user@xyz.com
Cancel
OK
Admin Console
Users
Groups
Global Init Scripts
Workspace Settings
Add User
Username
Name
Admin
Workspace access
Databricks SQL access
user@xyz.com
Admin Console
Users
Groups
Global Init Scripts
Workspace Settings
Add User
Username
Name
Admin
Workspace access
Databricks SQL access
Allow cluster creation
user@xyz.com
Microsoft Azure
Databricks
Portal
SQL
Create
Dashboards
Queries
Alerts
Data
SQL Endpoints
1
2
Create SQL Endpoint
State
Size
Active / Max
Actions
Stopped
2X-Small
0 / 1
Start



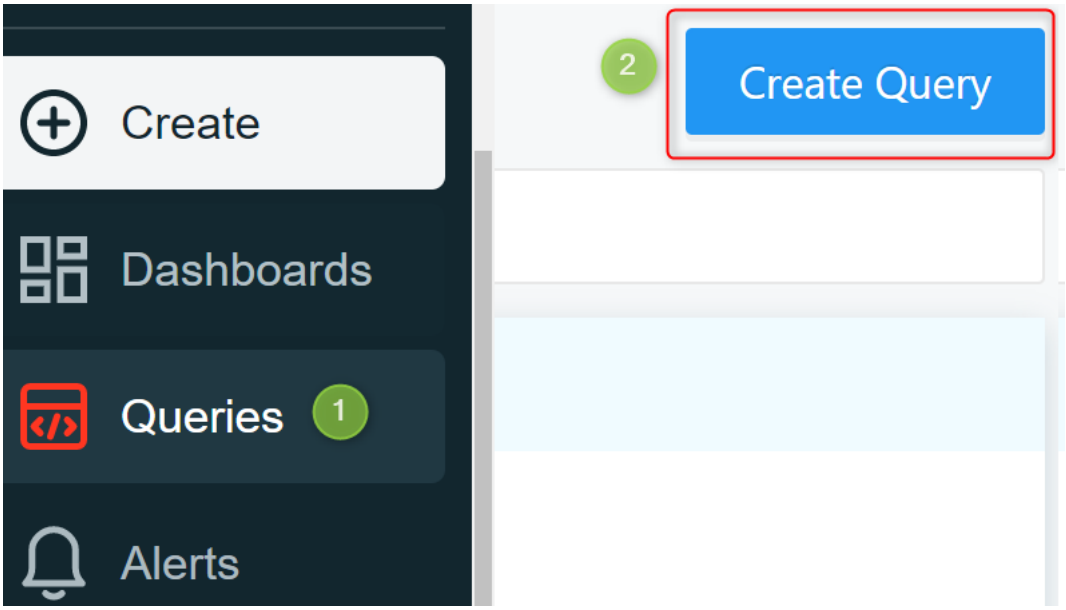




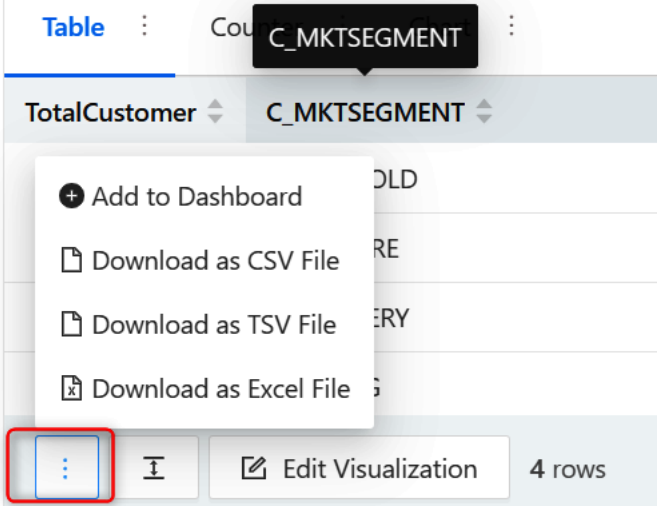
Table

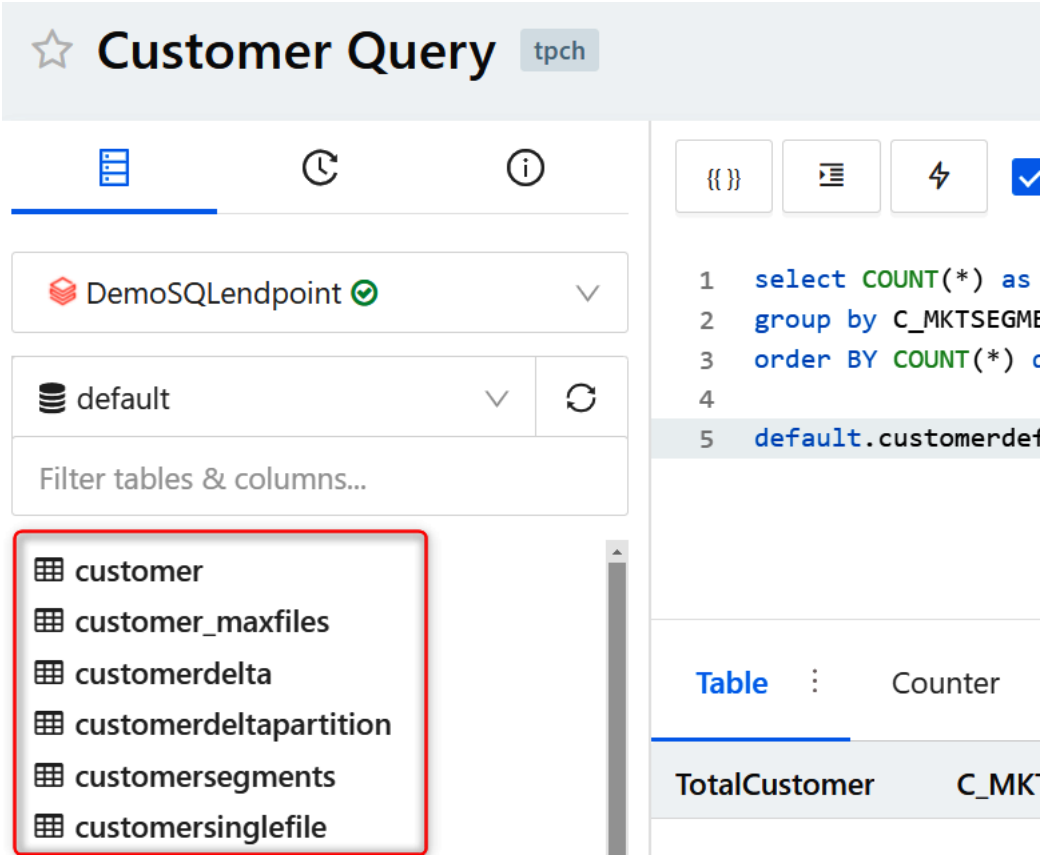
+ Add Visualization

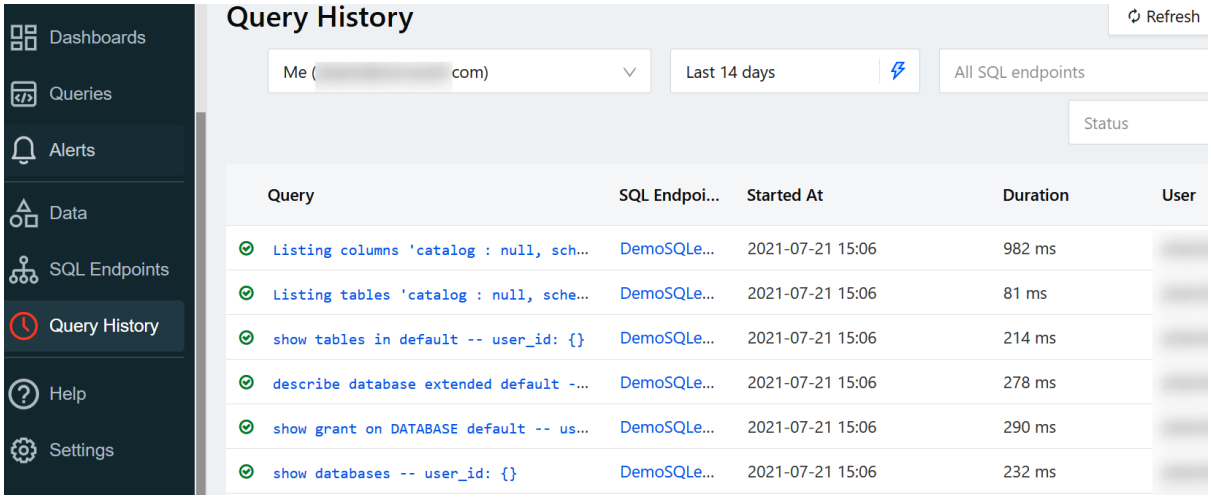
Principal	ActionType	ObjectType	ObjectKey
user@xyz.com	READ_METADATA	TABLE	`default`.`customer`
user@xyz.com	SELECT	TABLE	`default`.`customer`
user@xyz.com	USAGE	DATABASE	default

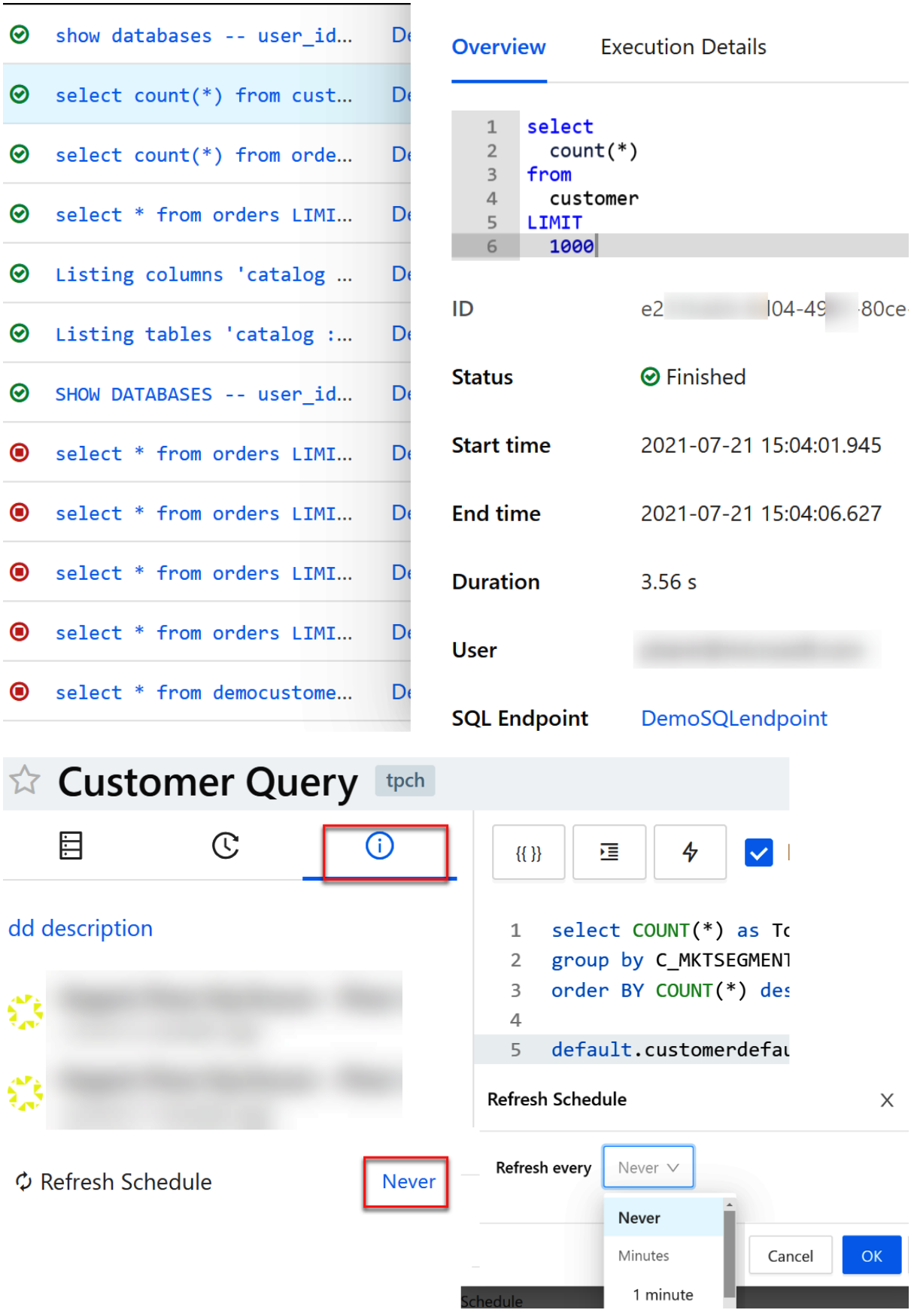


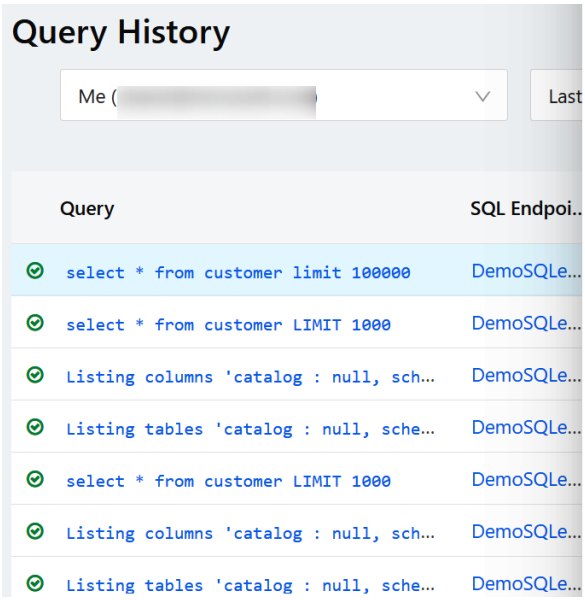

```
1  select COUNT(*) as TotalCustomer,C_MKTSEGME
2  group by C_MKTSEGMENT
3  order BY COUNT(*) desc
4
5
```

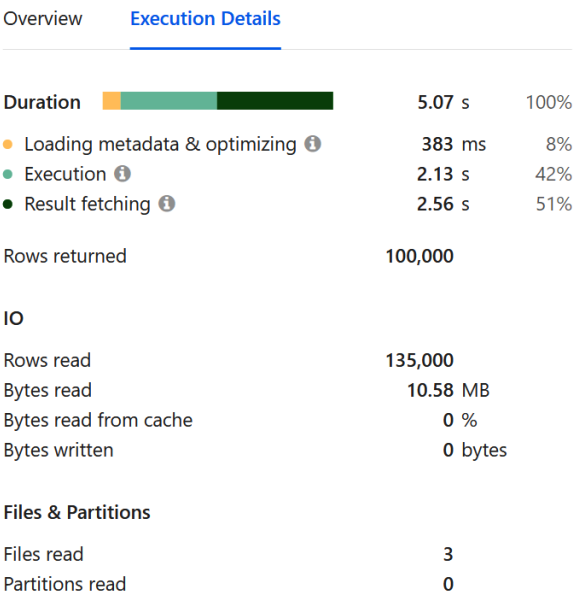



Customer Query
tpch
DemoSQLendpoint
default
Filter tables & columns...
customer
customer_maxfiles
customerdelta
customerdeltapartition
customersegments
customersinglefile
1 select COUNT(*) as
2 group by C_MKTSEGME
3 order BY COUNT(*) c
4
5 default.customerdet
Table
Counter
TotalCustomer
C_MK

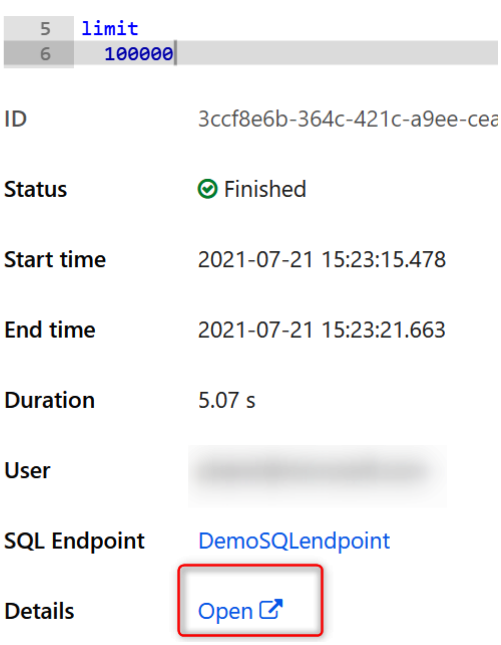

Dashboards
Queries
Alerts
Data
SQL Endpoints
Query History
Help
Settings
Query History
Refresh
Me (com)
Last 14 days
All SQL endpoints
Status
Query
SQL Endpoi...
Started At
Duration
User
Listing columns 'catalog : null, sch...
DemoSQLe...
2021-07-21 15:06
982 ms
Listing tables 'catalog : null, sche...
DemoSQLe...
2021-07-21 15:06
81 ms
show tables in default -- user_id: {}
DemoSQLe...
2021-07-21 15:06
214 ms
describe database extended default -...
DemoSQLe...
2021-07-21 15:06
278 ms
show grant on DATABASE default -- us...
DemoSQLe...
2021-07-21 15:06
290 ms
show databases -- user_id: {}
DemoSQLe...
2021-07-21 15:06
232 ms


show databases -- user_id...
select count(*) from cust...
select count(*) from orde...
select * from orders LIMI...
Listing columns 'catalog ...
Listing tables 'catalog :...
SHOW DATABASES -- user_id...
select * from democustome...
Overview
Execution Details
select
count(*)
from
customer
LIMIT
1000
ID
Status
Finished
Start time
2021-07-21 15:04:01.945
End time
2021-07-21 15:04:06.627
Duration
3.56 s
User
SQL Endpoint
DemoSQLendpoint
Customer Query
tpch
dd description
select COUNT(*) as Tc
group by C_MKTSEGMENT
order BY COUNT(*) des
default.customerdefau
Refresh Schedule
Refresh Schedule
Never
Refresh every
Never
Never
Minutes
1 minute
Cancel
OK
Schedule


Query History
Me (
Last
Query
SQL Endpoi..
select * from customer limit 100000
DemoSQLe...
select * from customer LIMIT 1000
DemoSQLe...
Listing columns 'catalog : null, sch...
DemoSQLe...
Listing tables 'catalog : null, sche...
DemoSQLe...
select * from customer LIMIT 1000
DemoSQLe...
Listing columns 'catalog : null, sch...
DemoSQLe...
Listing tables 'catalog : null, sche...
DemoSQLe...


Overview
Execution Details
Duration
5.07 s
100%
Loading metadata & optimizing
383 ms
8%
Execution
2.13 s
42%
Result fetching
2.56 s
51%
Rows returned
100,000
IO
Rows read
135,000
Bytes read
10.58 MB
Bytes read from cache
0 %
Bytes written
0 bytes
Files & Partitions
Files read
3
Partitions read
0


select * from customer LIMIT 1000
DemoSQLe...
Listing columns 'catalog : null, sch...
DemoSQLe...
Listing tables 'catalog : null, sche...
DemoSQLe...
select * from customer LIMIT 1000
DemoSQLe...
Listing columns 'catalog : null, sch...
DemoSQLe...
Listing tables 'catalog : null, sche...
DemoSQLe...
Listing columns 'catalog : null, sch...
DemoSQLe...
Listing tables 'catalog : null, sche...
DemoSQLe...
Listing columns 'catalog : null, sch...
DemoSQLe...
Listing tables 'catalog : null, sche...
DemoSQLe...
show tables in default -- user id: {}
DemoSQLe...


5 limit
6 100000
ID
3ccf8e6b-364c-421c-a9ee-cea
Status
Finished
Start time
2021-07-21 15:23:15.478
End time
2021-07-21 15:23:21.663
Duration
5.07 s
User
SQL Endpoint
DemoSQLendpoint
Details
Open

Details for Query 52

Download screen as png

Submitted Time: 2021/07/21 09:53:17
Duration: 2 s
Succeeded Jobs: 12 13

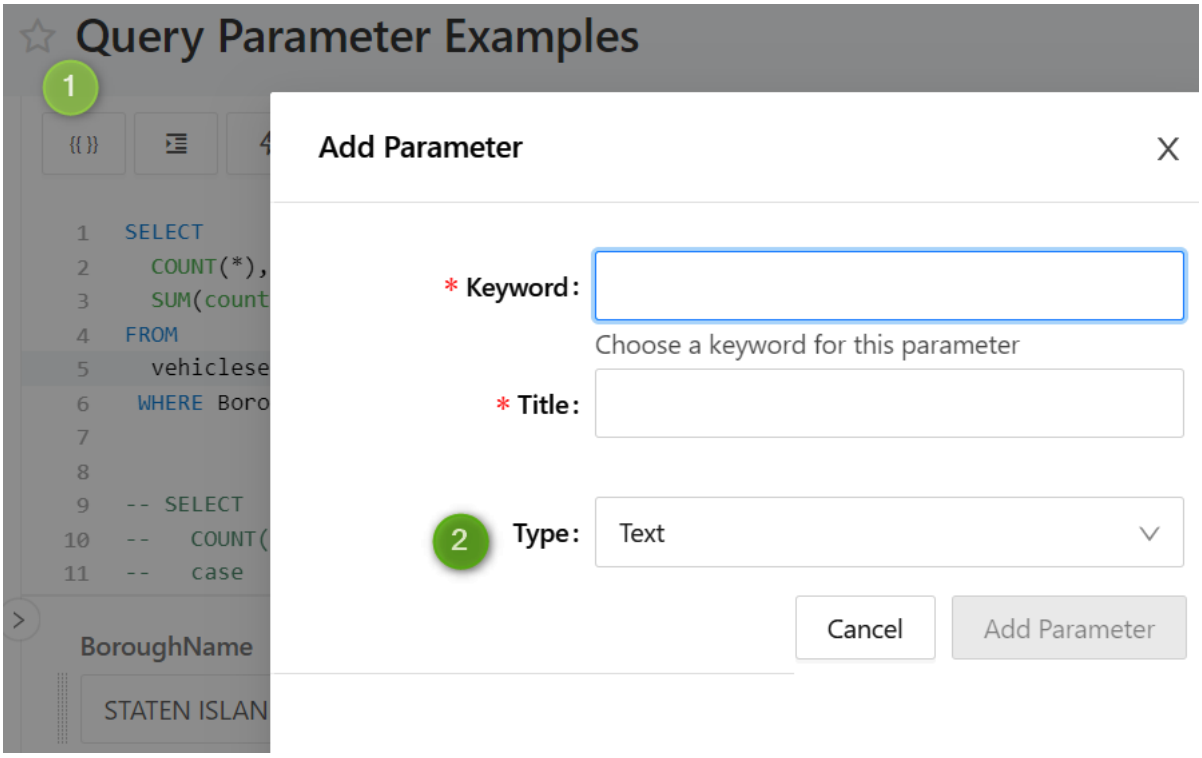
☑ Expand all the details in the query plan visualization

Scan parquet default.customer +details

Stages: 20.0 21.0

cache writes size (uncompressed) total (min, med, max)	20.5 MiB (6.8 MiB, 6.8 MiB, 6.8 MiB)
time spent waiting fetching data from cloud storage total (min, med, max)	592 ms (163 ms, 213 ms, 215 ms)
time spent in the cache locality manager in milliseconds total (min, med, max)	5 ms (0 ms, 0 ms, 5 ms)
number of files read	3
filesystem read data size total (min, med, max)	10.6 MiB (3.5 MiB, 3.5 MiB, 3.5 MiB)
cache async file status fetch waiting time total (min, med, max)	0 ms (0 ms, 0 ms, 0 ms)
scan time total (min, med, max)	763 ms (212 ms, 263 ms, 288 ms)
filesystem read data size (sampled) total (min, med, max)	21.2 MiB (7.0 MiB, 7.0 MiB, 7.1 MiB)
filesystem read time (sampled) total (min, med, max)	824 ms (268 ms, 276 ms, 280 ms)
metadata time	0 ms
size of files read	10.6 MiB
cache hits size total (min, med, max)	0.0 B (0.0 B, 0.0 B, 0.0 B)

```
1  SELECT
2    COUNT(*),
3    SUM(count) AS TotalCountofVehicles
4  FROM
5    vehiclesensor.vehicledeltaaggregated
6   WHERE Borough={{ BoroughName }} and Make ={{ MakeValue }}
7
8
```




```sql
SELECT
  COUNT(*) AS   TotalVehicles,
  case
    when lfi = 0 then "Low"
    else "High"
  end as LowFuelIndicator
FROM
  vehiclesensor.vehicledelta_silver
WHERE
  eventtime BETWEEN '{{ Date Range.start }}' AND '{{ Date Range.end }}'
  group BY LowFuelIndicator
```

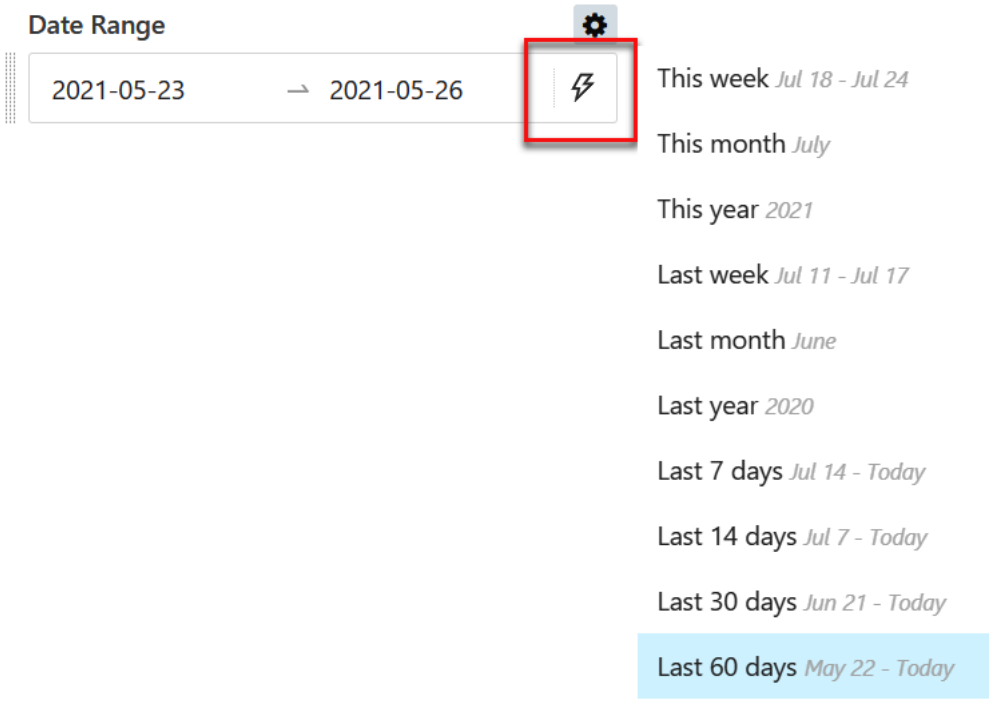
Date Range

2021-05-23 → 2021-05-26

Table

TotalVehicles	LowFuelIndicator
83	High
211	Low

```
8     vehiclesensor.vehicledelta_silver
9   WHERE
10    eventtime BETWEEN '{{ Date Range.start }}' AN
11    group BY LowFuelIndicator
```



Table

TotalVehicles	LowFuelIndicator
49	High
158	Low

Boroughs

* Title: Boroughs

Type: Query Based Dropdown List

Query: Boroughs

Select query to load dropdown values from

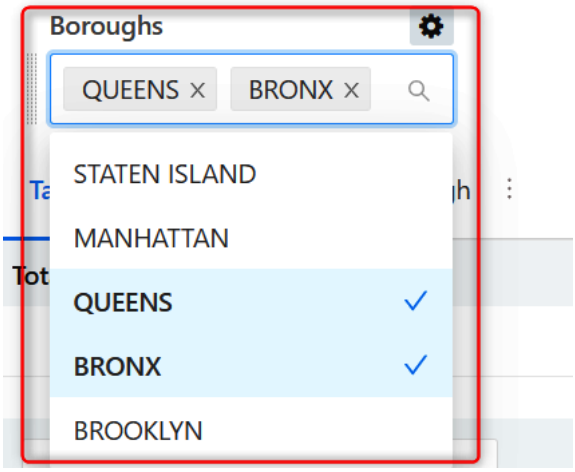
Allow multiple values

Quotation: Double Quotation Mark

Placed in query as: `"value1","value2","value3"`

Cancel OK

```
12    Borough  in ({{ Boroughs }})
13    group BY LowFuelIndicator,Borough
```



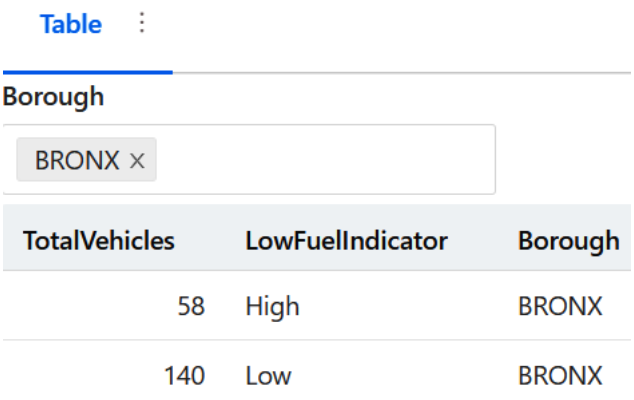
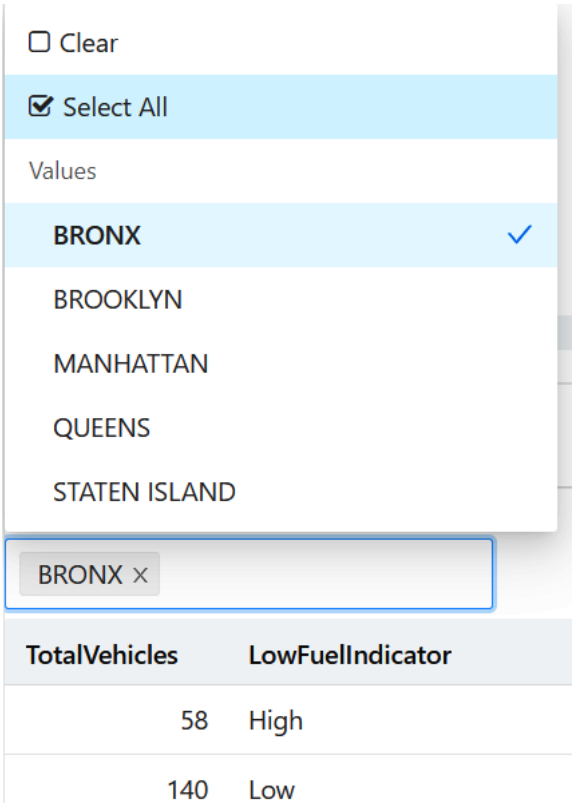
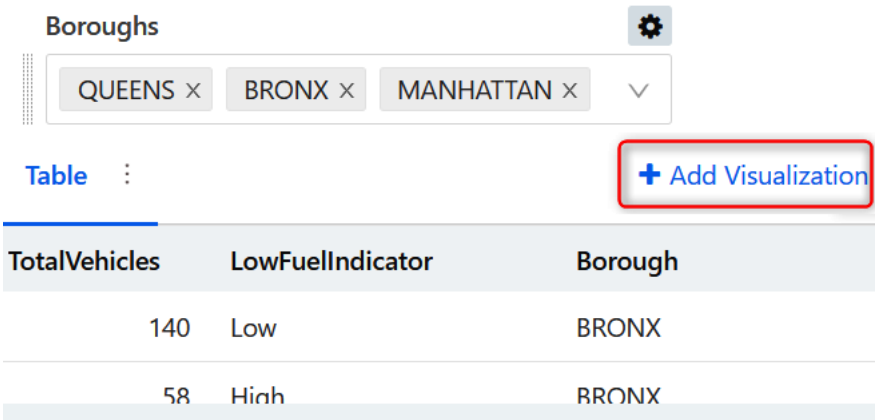


Table
Borough
BRONX ×
TotalVehicles
LowFuelIndicator
Borough
58
High
BRONX
140
Low
BRONX


Clear
Select All
Values
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
BRONX ×
TotalVehicles
LowFuelIndicator
58
High
140
Low


Boroughs
QUEENS ×
BRONX ×
MANHATTAN ×
Table
Add Visualization
TotalVehicles
LowFuelIndicator
Borough
140
Low
BRONX
58
High
BRONX

General | X Axis | Y Axis | Series | Colors | Data Labels

Chart Type

Bar

☐ Horizontal Chart

X Column

Borough

Y Columns

TotalVehicles ×

Group By

LowFuelIndicator

Visualization Name

LFI vehicles by Borough

General | X Axis | Y Axis | Series | Colors | Data Labels

☑ Show Data Labels

Number Values Format

0,0[.]00000

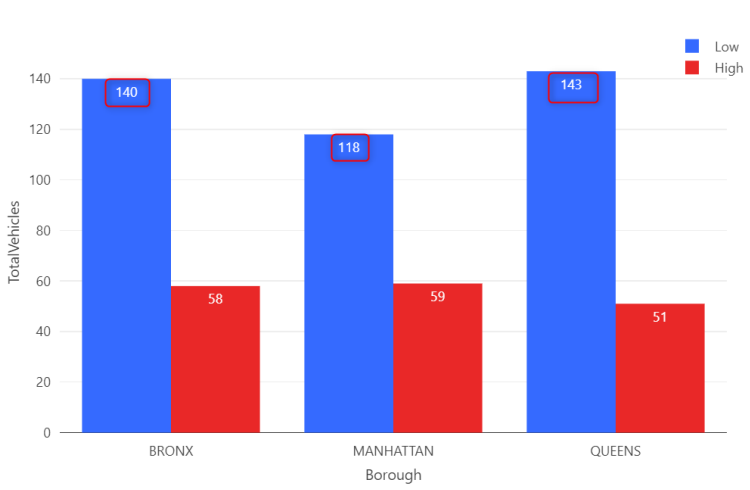
Percent Values Format

0[.]00%

Date/Time Values Format

YYYY-MM-DD HH:mm

Data Labels



Visualization Type

Funnel

Visualization Name

Funnel

General | Appearance

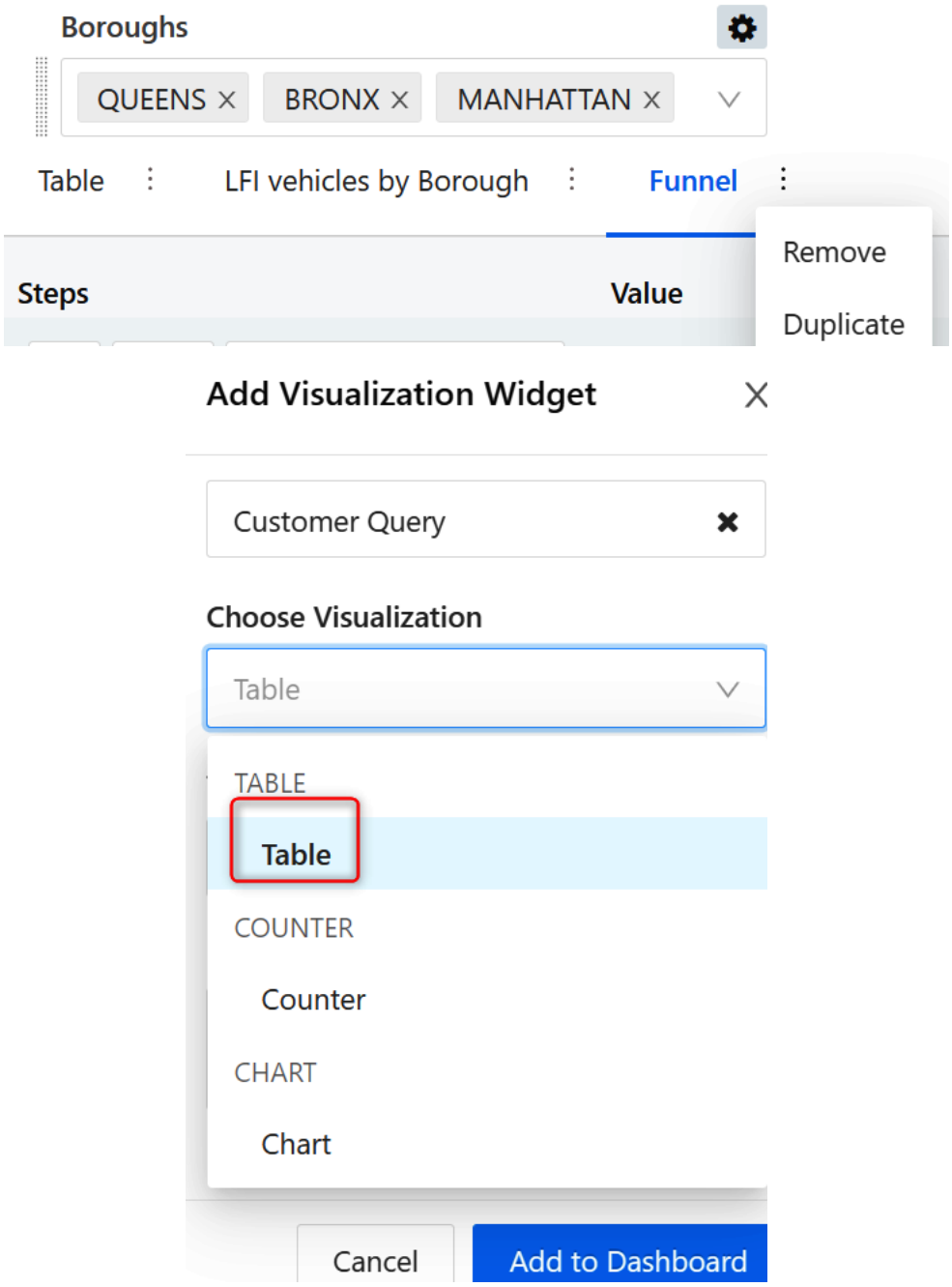
Step Column Borough

Step Column Title Steps

Value Column TotalVehicles

Steps	Value	% Max	% Previous
QUEENS	143	100%	100%
BRONX	140	97.90%	97.90%
MANHATTAN	118	82.52%	84.29%
MANHATTAN	59	41.26%	50%
BRONX	58	40.56%	98.31%
QUEENS	51	35.66%	87.93%

Cancel | Save


Boroughs
QUEENS
BRONX
MANHATTAN
Table
LFI vehicles by Borough
Funnel
Remove
Duplicate
Steps
Value
Add Visualization Widget
Customer Query
Choose Visualization
Table
TABLE
Table
COUNTER
Counter
CHART
Chart
Cancel
Add to Dashboard

Query Parameter Examples

Choose Visualization

LFI vehicles by Borough

Title

LFI vehicles by Borough - Query Parameter Examples

Description

Parameters

Title	Keyword	Default Value	Value Source
Boroughs	{{ Boroughs }}	QUEENS, BRONX	Dashboard Boroughs

Cancel | Add to Dashboard

Vehicle and Customer Stat

Share | Schedule | Refresh

Boroughs

QUEENS × BRONX × MANHATTAN ×

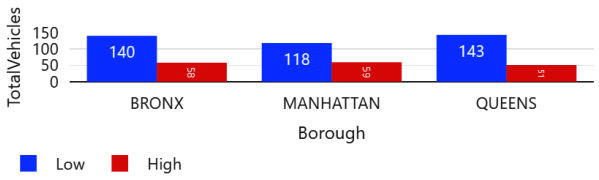
Counter - Vehicle Agg

652

Total Count

2 months ago

LFI vehicles by Borough - Query Parameter Examples

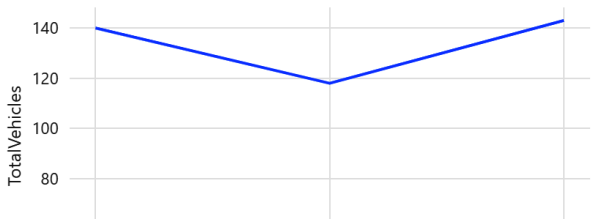


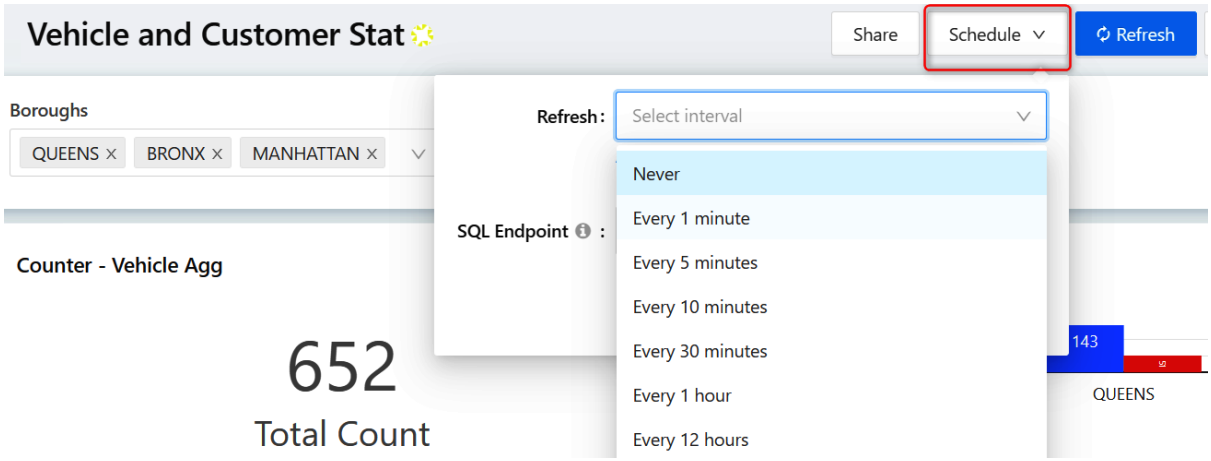
just now

Table - Customer Query

TotalCustomer	C_MKTSEGMENT
30,189	HOUSEHOLD
30,142	BUILDING
29,968	FURNITURE
29,949	MACHINERY

Line Chart - Query Parameter Examples





Overview | **Connection Details** | Monitoring

Server Hostname adb-[redacted]43.3.azuredatabricks.net

Port 443

Protocol https

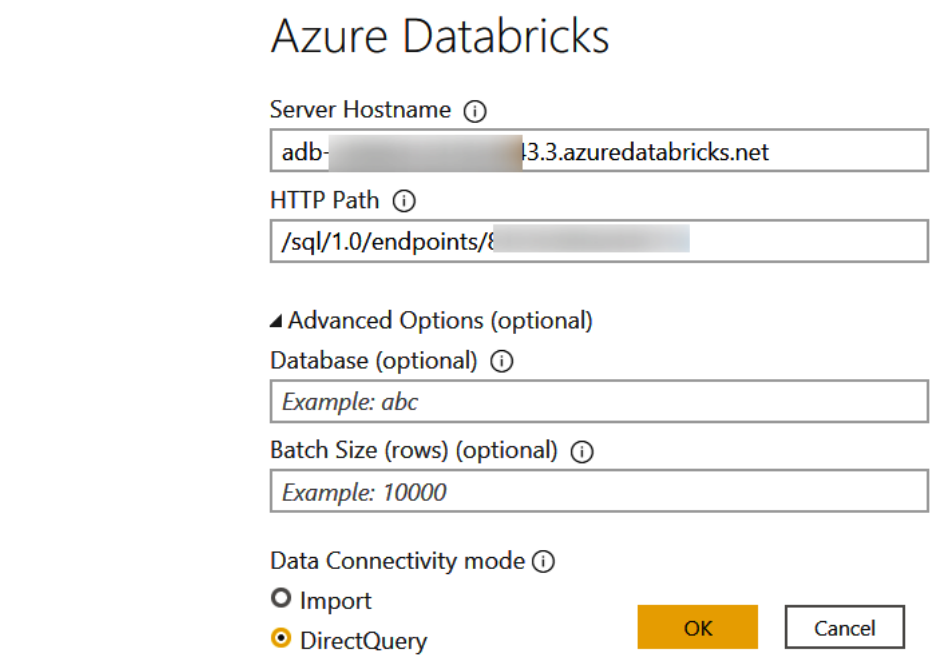
HTTP Path /sql/1.0/endpoints/[redacted]

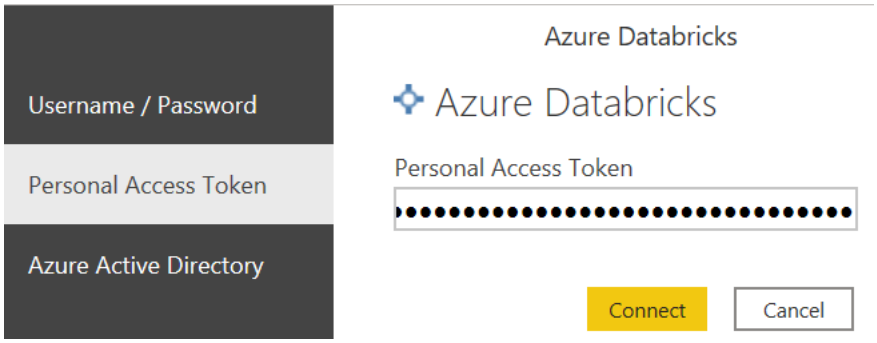
JDBC URL jdbc:spark://adb-[redacted].azuredatabricks.net:443/default;transportMode=http;ssl=1;AuthMech=3;httpPath=/sql/1.0/endpoints/84[redacted];

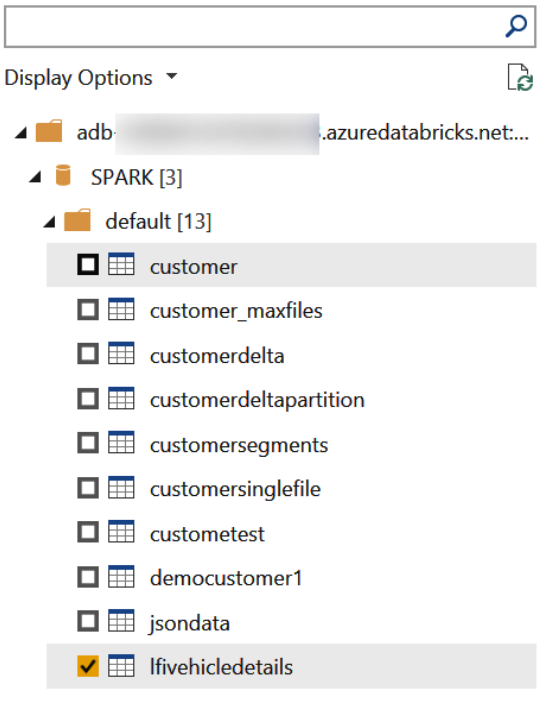
OAuth URL https://adb-[redacted].3.azuredatabricks.net/oidc

Use these details to connect to this endpoint from BI tools

Create a personal access token

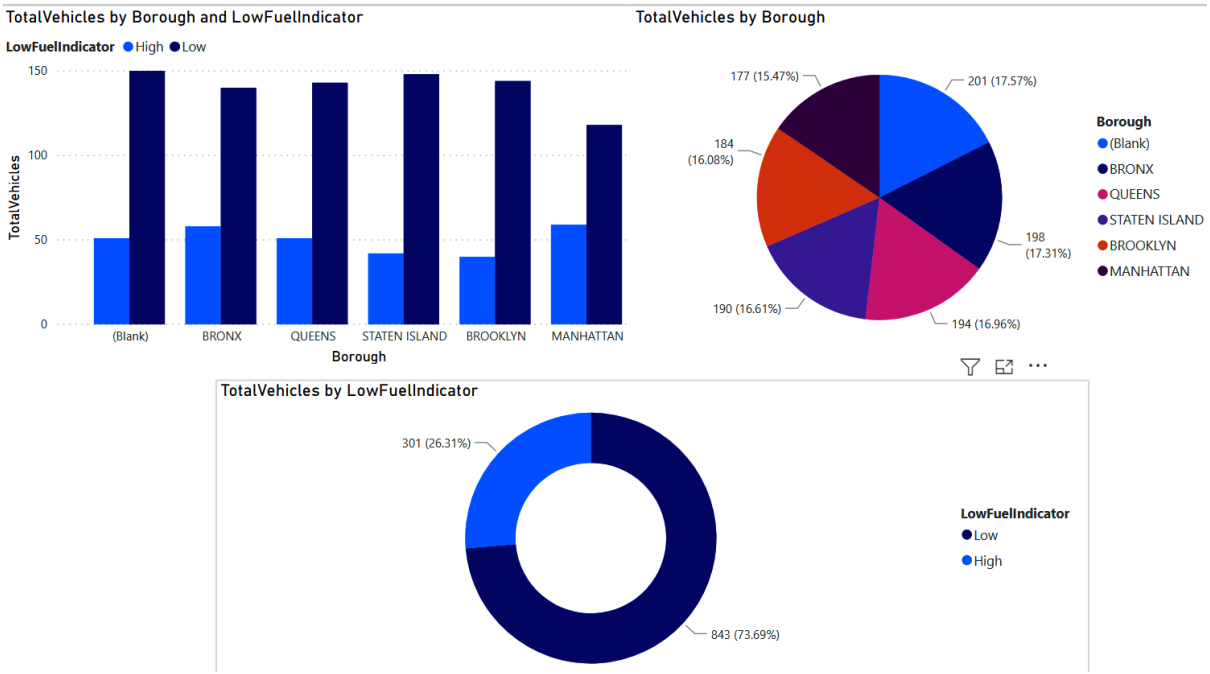




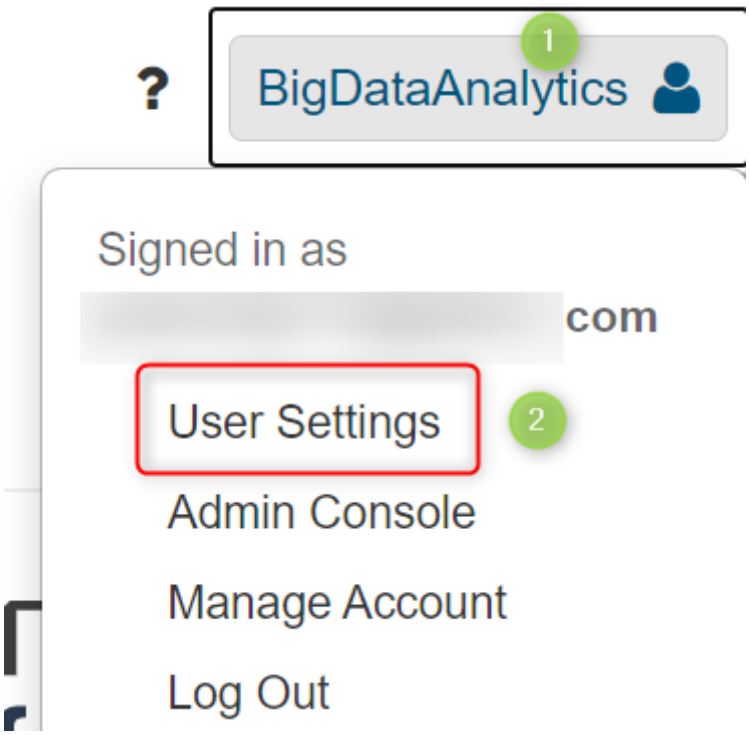


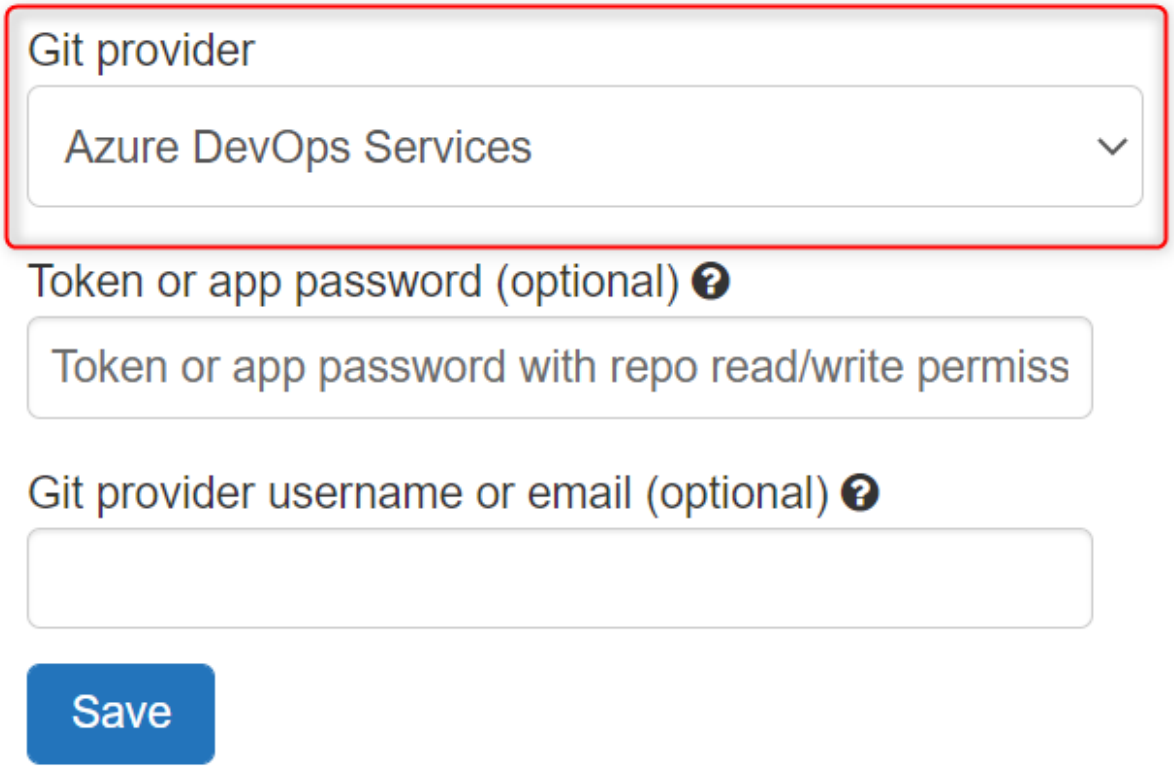
lfivehicledetails

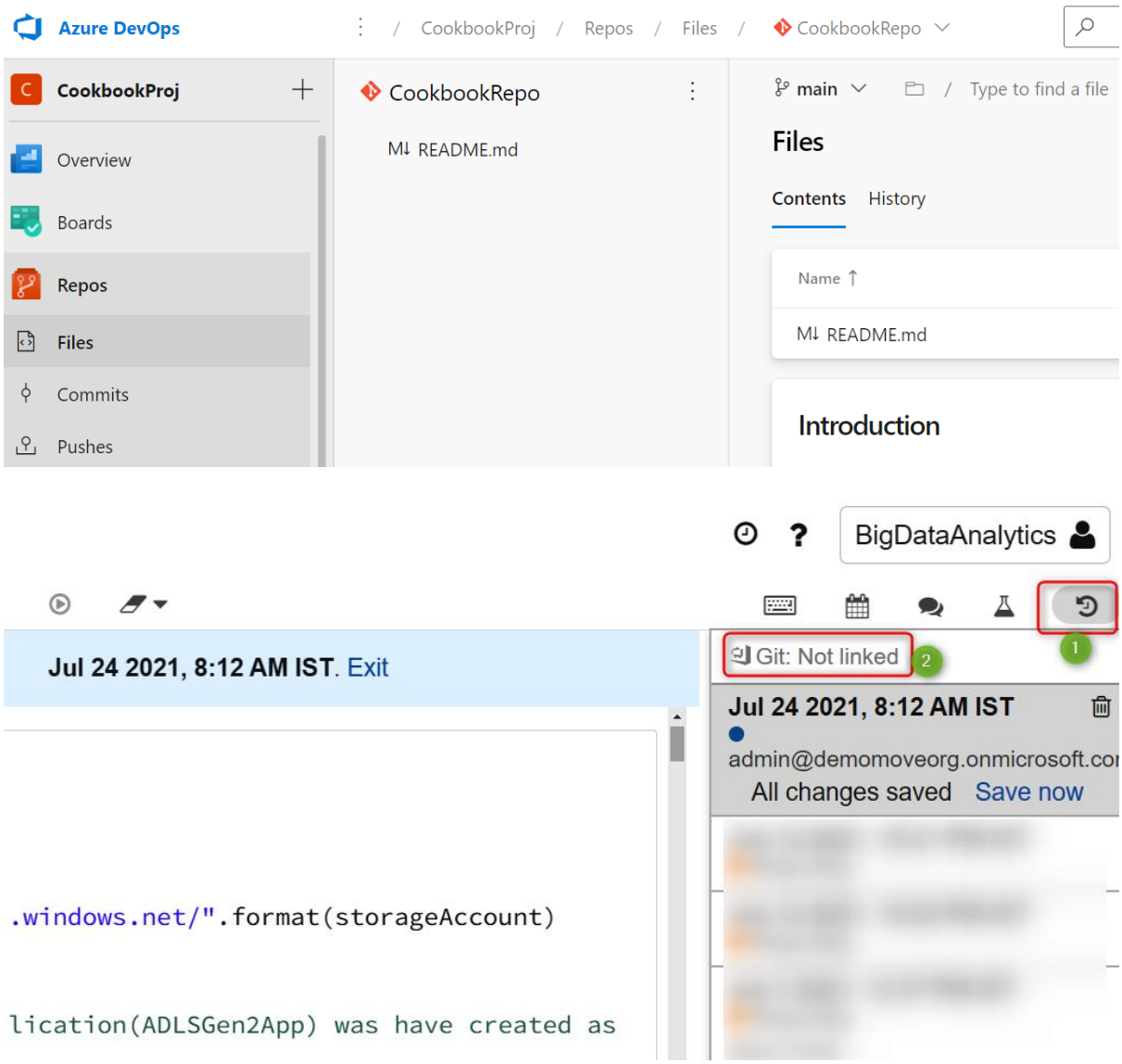
TotalVehicles	LowFuelIndicator	Borough
148	Low	STATEN ISLAND
51	High	QUEENS
118	Low	MANHATTAN
42	High	STATEN ISLAND
58	High	BRONX
40	High	BROOKLYN
140	Low	BRONX
59	High	MANHATTAN
143	Low	QUEENS
144	Low	BROOKLYN
51	High	*null*
150	Low	*null*

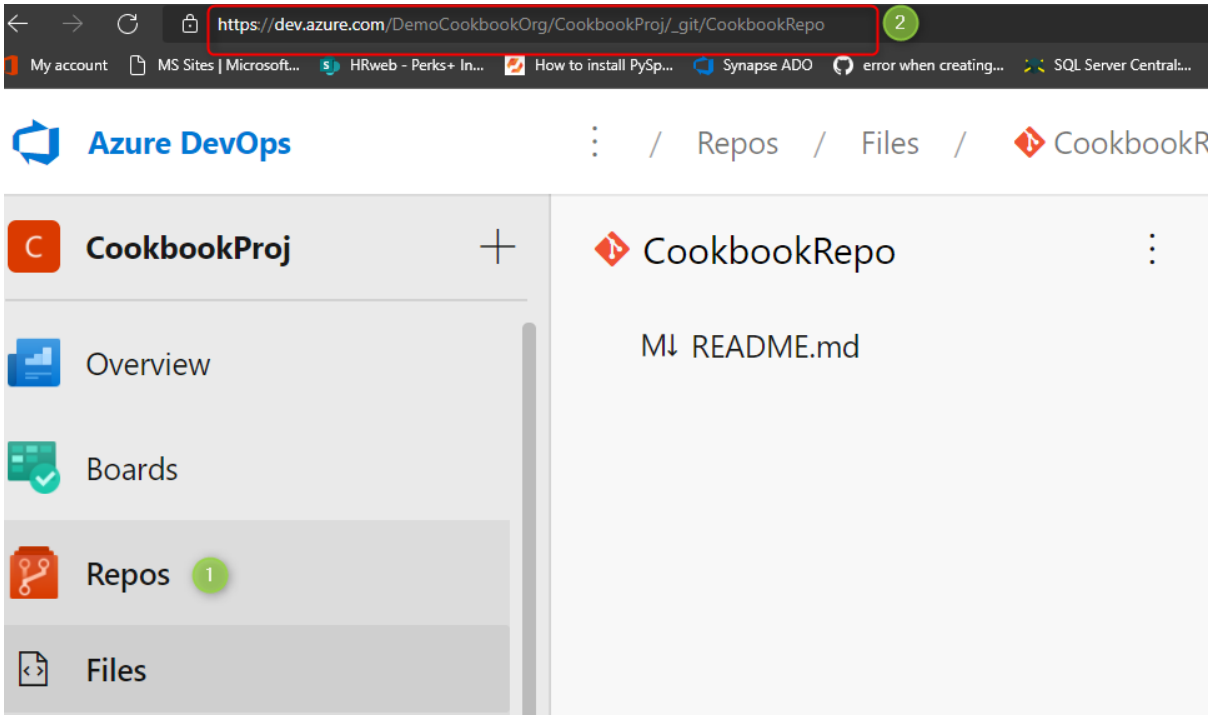

TotalVehicles by Borough and LowFuelIndicator
LowFuelIndicator High Low
150
100
50
0
TotalVehicles
(Blank)
BRONX
QUEENS
STATEN ISLAND
BROOKLYN
MANHATTAN
Borough
TotalVehicles by Borough
177 (15.47%)
201 (17.57%)
184 (16.08%)
198 (17.31%)
190 (16.61%)
194 (16.96%)
Borough
(Blank)
BRONX
QUEENS
STATEN ISLAND
BROOKLYN
MANHATTAN
TotalVehicles by LowFuelIndicator
301 (26.31%)
843 (73.69%)
LowFuelIndicator
Low
High

Chapter 9: DevOps Integrations and Implementing CI/CD for Azure Databricks






Azure DevOps
CookbookProj / Repos / Files / CookbookRepo
CookbookProj
Overview
Boards
Repos
Files
Commits
Pushes
CookbookRepo
README.md
main
Type to find a file
Files
Contents
History
Name
README.md
Introduction
BigDataAnalytics
Jul 24 2021, 8:12 AM IST. Exit
Git: Not linked
2
1
Jul 24 2021, 8:12 AM IST
admin@demomoveorg.onmicrosoft.com
All changes saved
Save now
.windows.net/".format(storageAccount)
lication(ADLSGen2App) was have created as



Git Preferences

Status ◉ Link ○ Unlink

Link https://dev.azure.com/DemoCookbookOrg/CookbookProj/_

Branch Dev Create PR

Path in Git Repo notebooks/6.1-Reading Writing to Delta Tables.py

Tip: You can also import/export multiple notebooks or an entire folder through Workspace API to your computer and check-in to your favorite version control system.

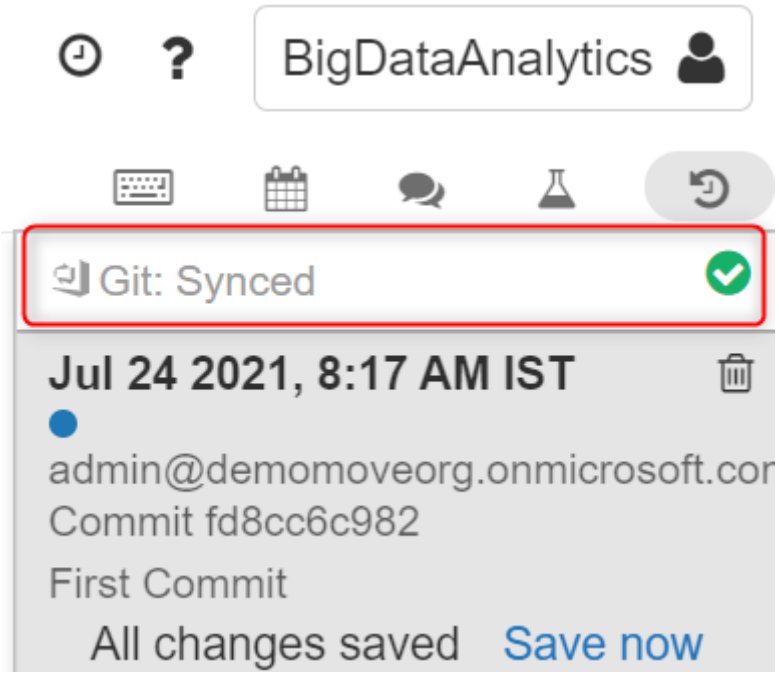
Close Save

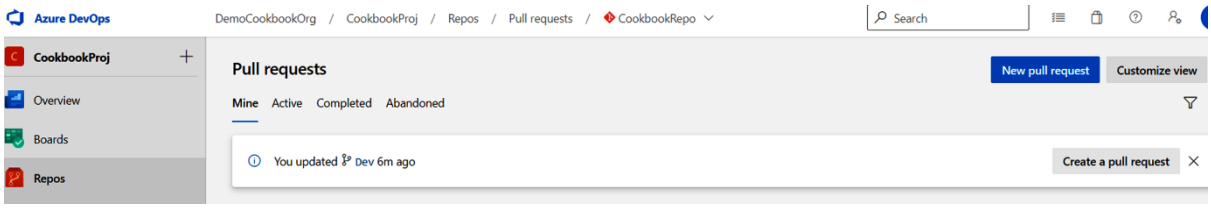
Save Notebook Revision

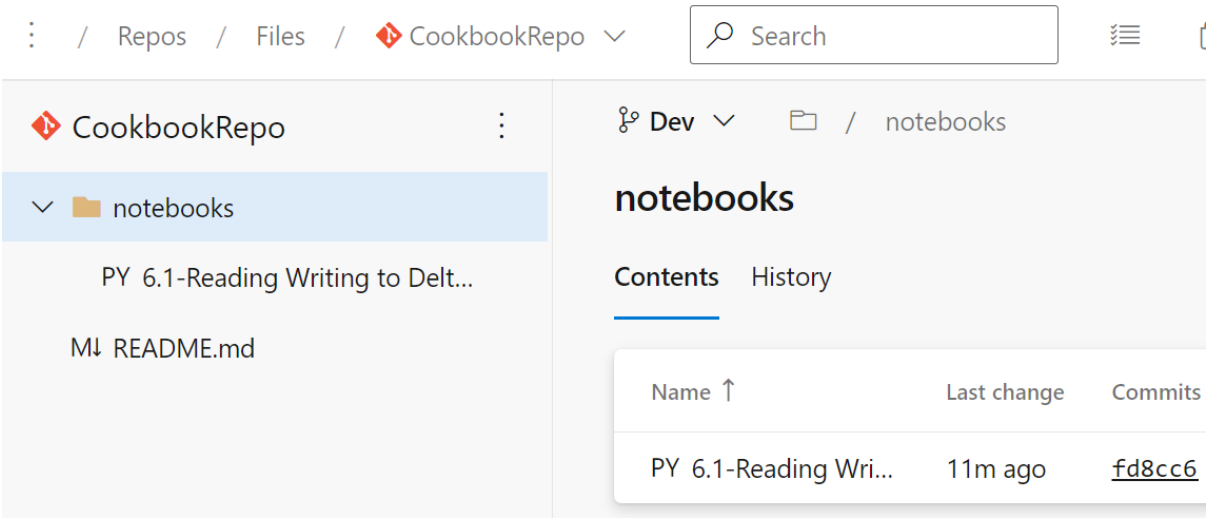
The file you linked to does not exist on Git yet. Would you like to make a commit and save the current version?

First Commit

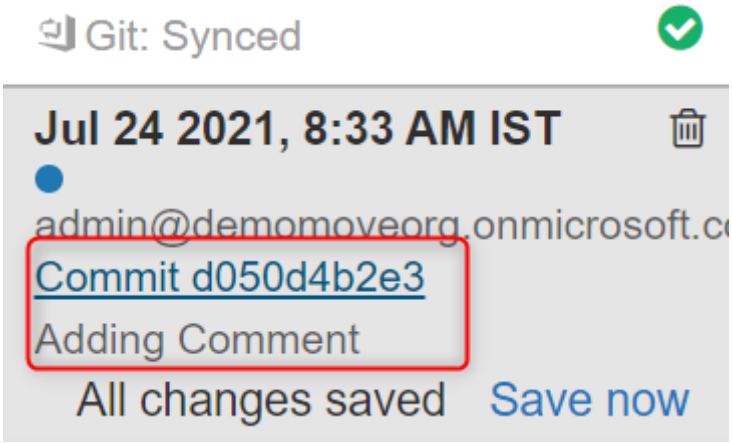
Close Save

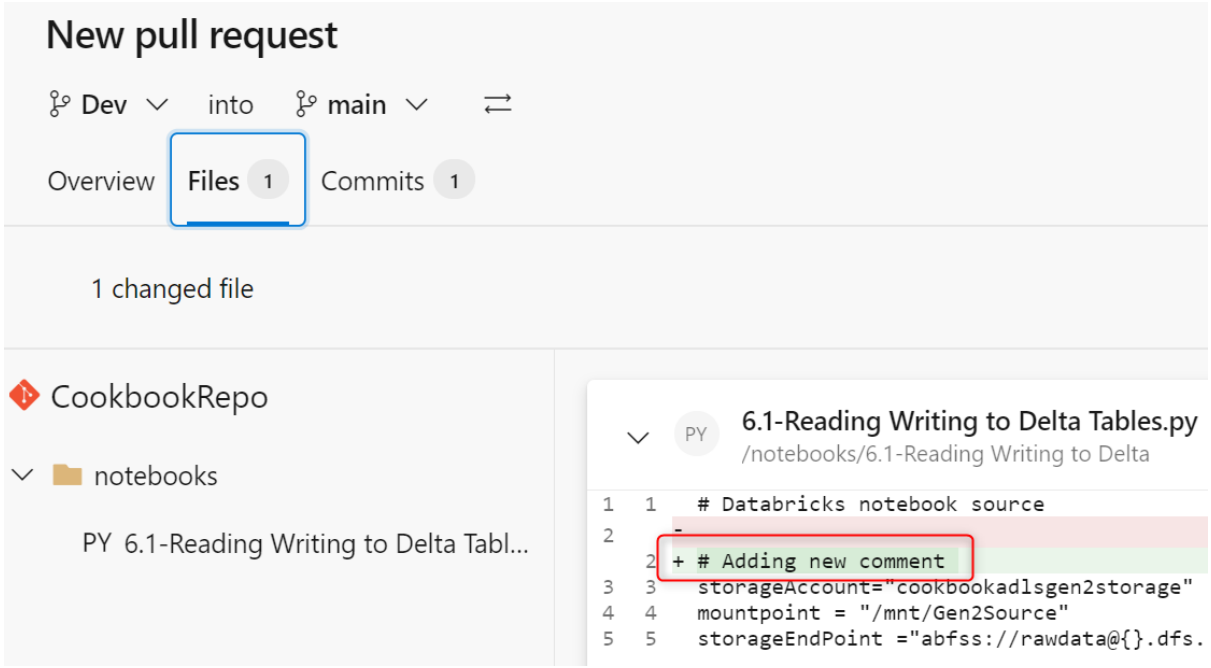


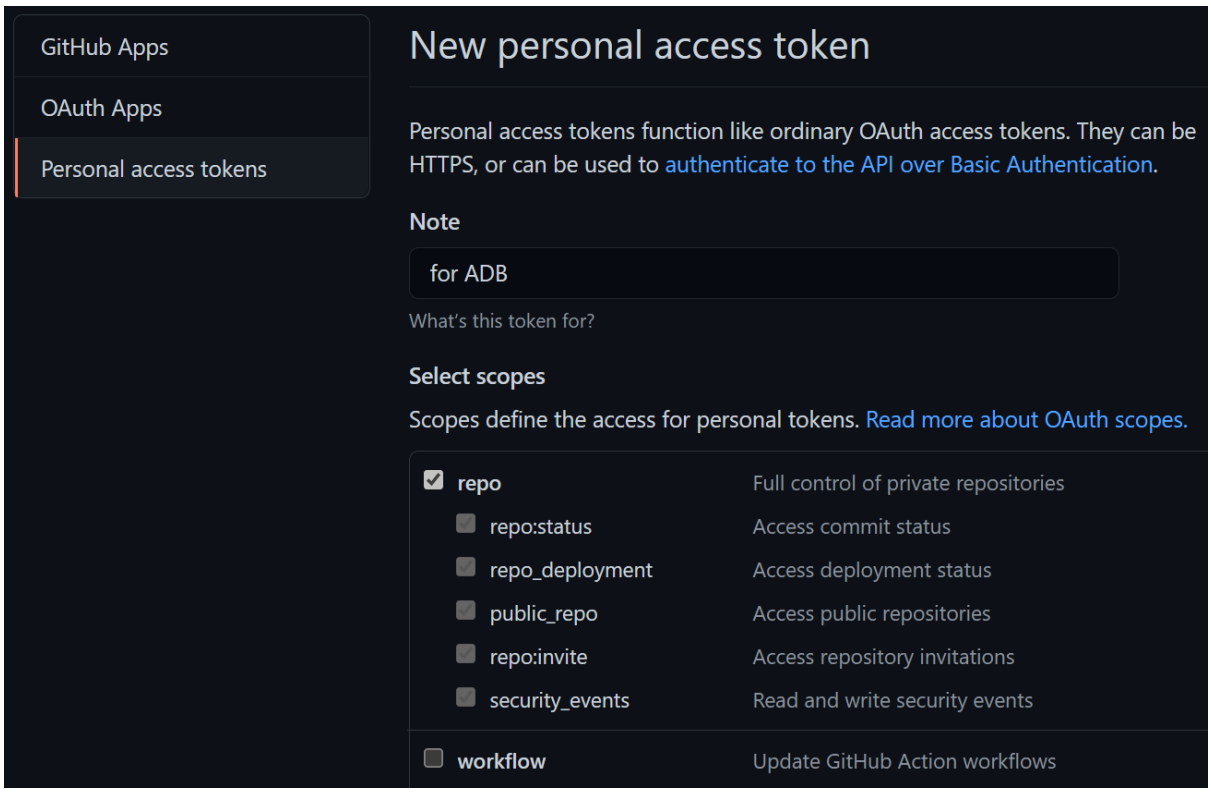





```
# Adding new comment
storageAccount="cookbookadlsgen2storage"
mountpoint = "/mnt/Gen2Source"
storageEndPoint ="abfss://rawdata@{}.dfs.cor
print ('Mount Point ='+mountpoint)
```




New pull request
Dev
into
main
Overview
Files 1
Commits 1
1 changed file
CookbookRepo
notebooks
PY 6.1-Reading Writing to Delta Tabl...
PY
6.1-Reading Writing to Delta Tables.py
/notebooks/6.1-Reading Writing to Delta
1 1 # Databricks notebook source
2 -
2 + # Adding new comment
3 3 storageAccount="cookbookadlsgen2storage"
4 4 mountpoint = "/mnt/Gen2Source"
5 5 storageEndPoint ="abfss://rawdata@{}.dfs.


GitHub Apps
OAuth Apps
Personal access tokens
New personal access token
Personal access tokens function like ordinary OAuth access tokens. They can be HTTPS, or can be used to authenticate to the API over Basic Authentication.
Note
for ADB
What's this token for?
Select scopes
Scopes define the access for personal tokens. Read more about OAuth scopes.
repo
Full control of private repositories
repo:status
Access commit status
repo_deployment
Access deployment status
public_repo
Access public repositories
repo:invite
Access repository invitations
security_events
Read and write security events
workflow
Update GitHub Action workflows

Generating tokens

To generate a GitHub personal access token, follow the GitHub permission.

To generate a Bitbucket Cloud app password, follow the Bitbuck must have "read" and "write" permission under repository.

To generate a GitLab personal access token, follow the GitLab "write_repository" permission.

To generate a Azure DevOps personal access token, follow the have "Full access" scope.

Git provider

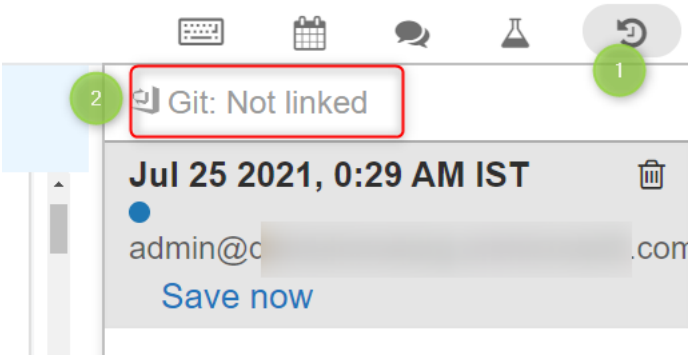
GitHub

Token or app password

••

Git provider username or email

Save



Status (•) Link ( ) Unlink

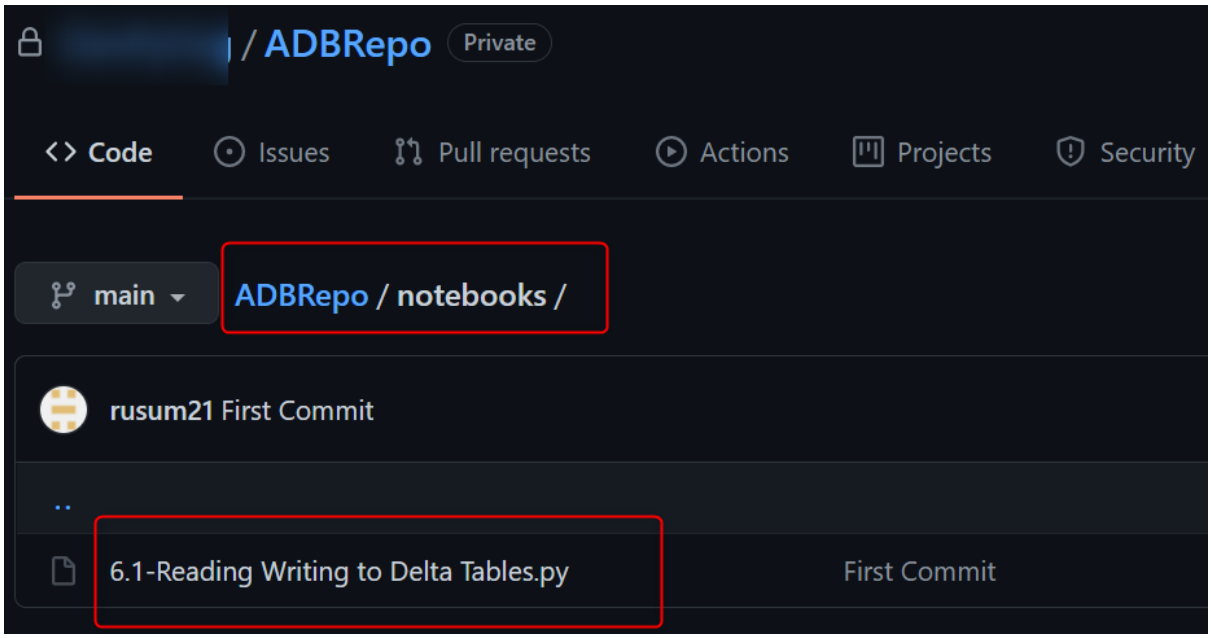
Link https://github.com/ /ADBRepo

Branch main Create PR

Path in Git Repo notebooks/6.1-Reading Writing to Delta Tables.py

Tip: You can also import/export multiple notebooks or an entire folder through Workspace API to your computer and check-in to your favorite version control system.

Close Save




```
#Adding comment for GitHub
display(dbutils.fs.ls("/mnt/Gen2Source/Customer/parquetFiles"))
```

Git Preferences

Status (•) Link ( ) Unlink

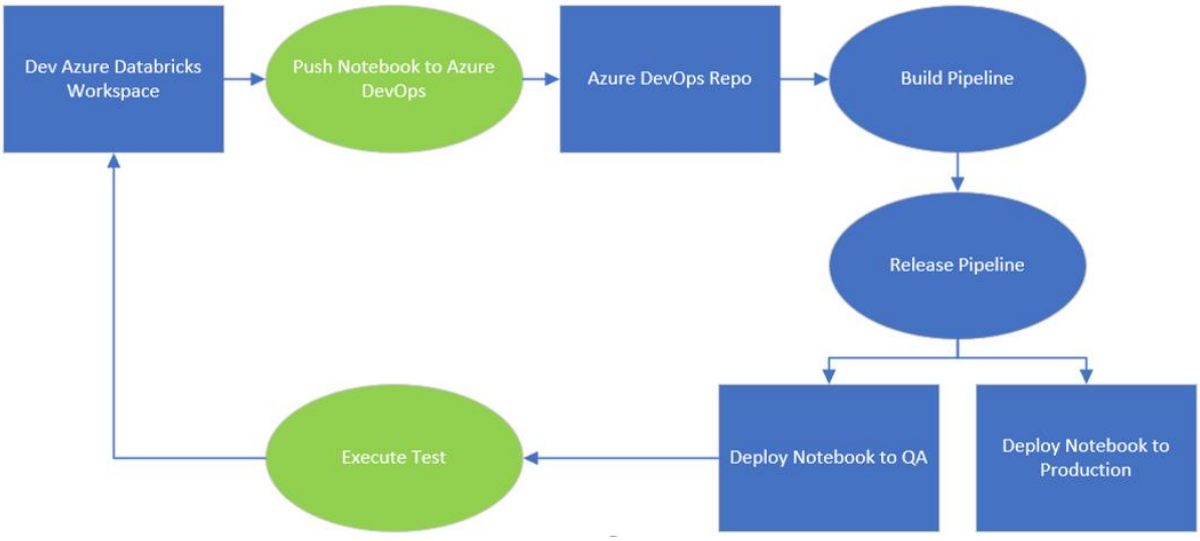
Link https://github.com/ /ADBRepo

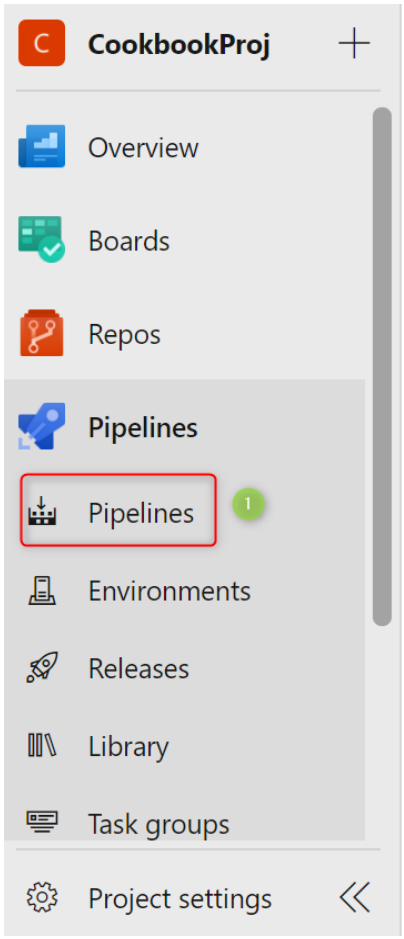
Branch dev-sprint1 Create PR

Path in Git Repo notebooks/6.1-Reading Writing to Delta Tables.py

Tip: You can also import/export multiple notebooks or an entire folder through Workspace API to your computer and check-in to your favorite version control system.

Close Save





New pipeline

Where is your code?

Azure Repos Git YAML
Free private Git repositories, pull requests, and code search

Bitbucket Cloud YAML
Hosted by Atlassian

GitHub YAML
Home to the world's largest community of developers

GitHub Enterprise Server YAML
The self-hosted version of GitHub Enterprise

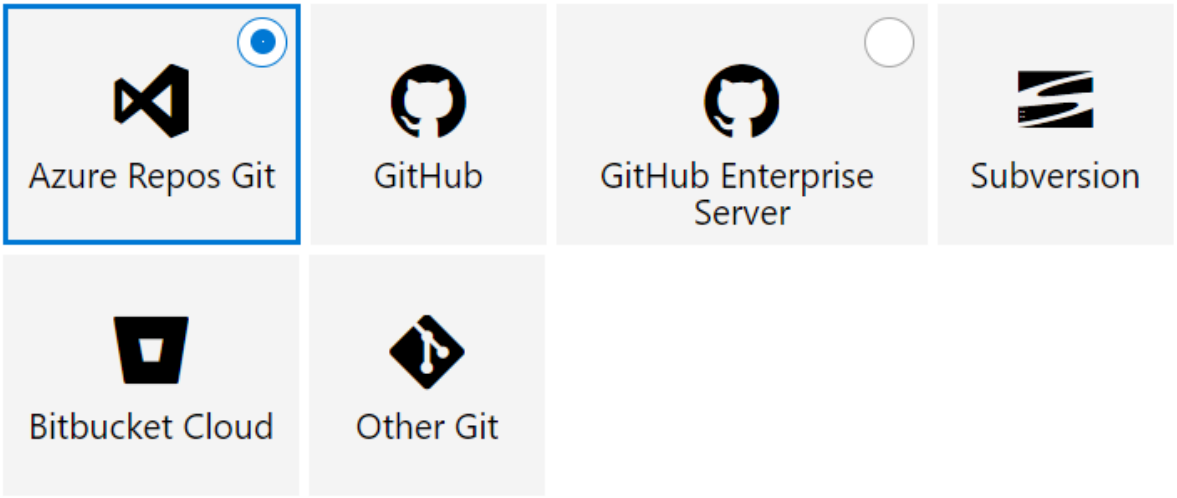
Other Git
Any generic Git repository

Subversion
Centralized version control by Apache

2

Use the classic editor to create a pipeline without YAML.

Select a source



Team project

CookbookProj

Repository

CookbookRepo

Default branch for manual and scheduled builds

main

Select a template

Or start with an **Empty job**

Configuration as code

YAML

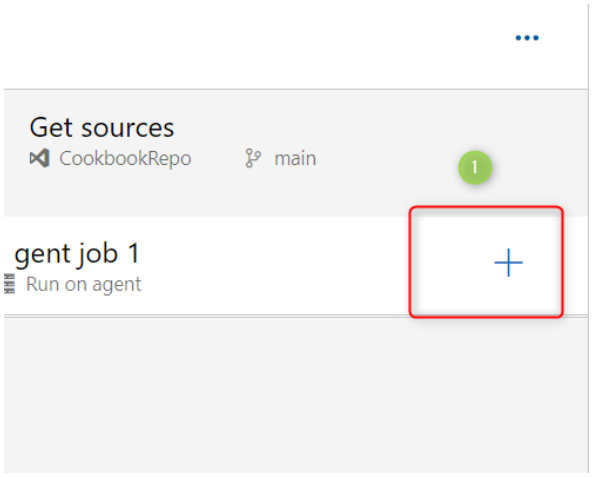
Looking for a better experience to configure your pipelines using YAML files? Try the new YAML pipeline creation experience. Learn more

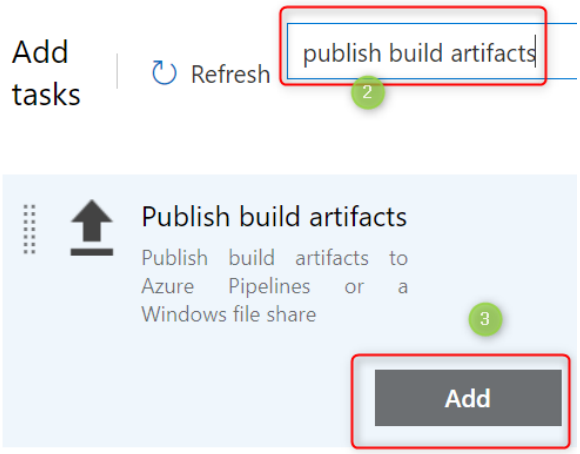
Featured

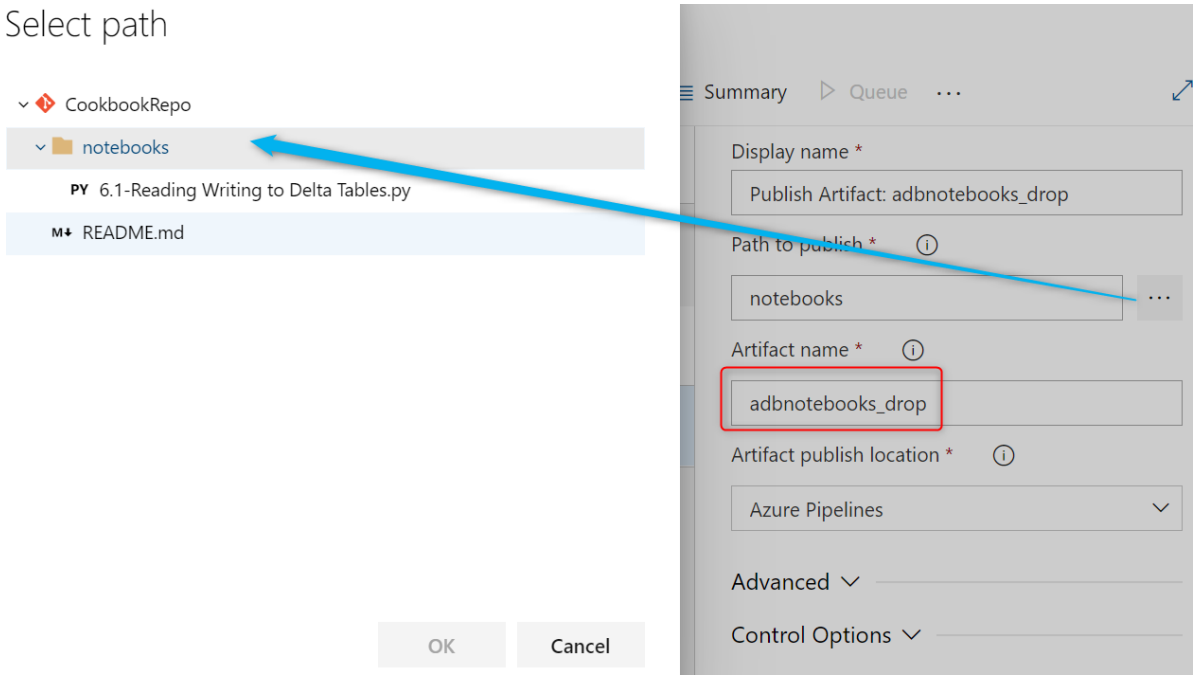
.NET Desktop

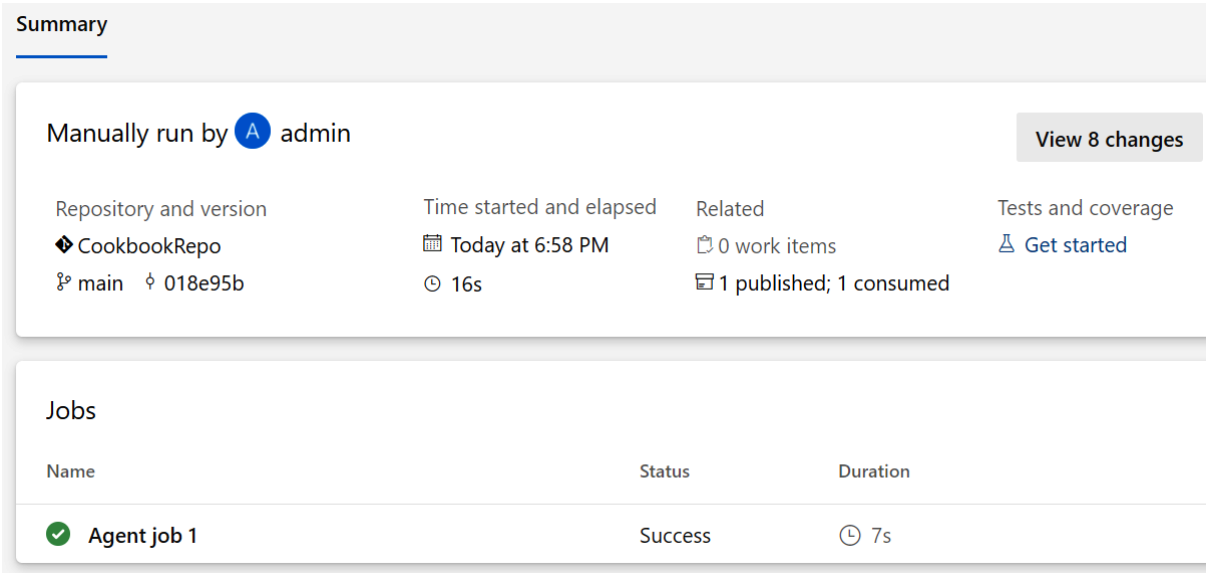
Build and test a .NET or Windows classic desktop solution.

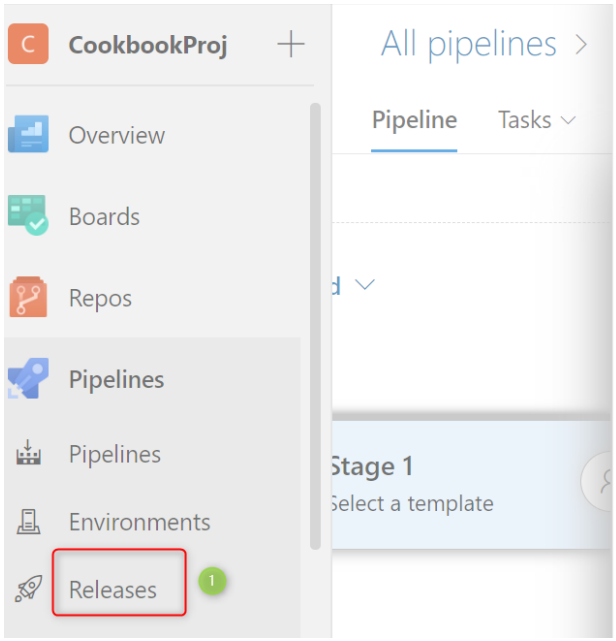
Android

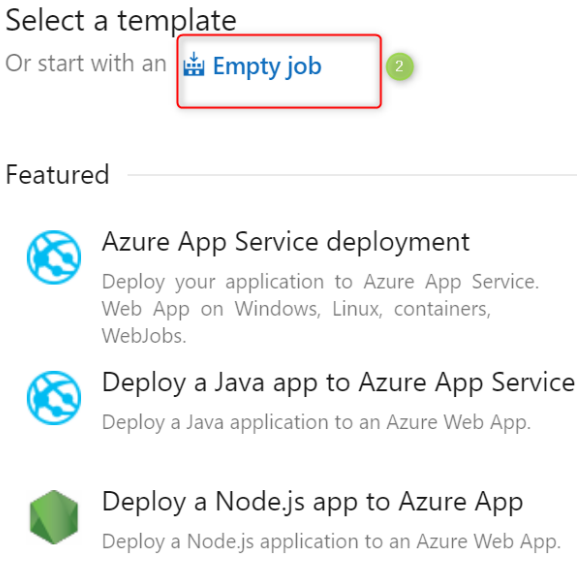


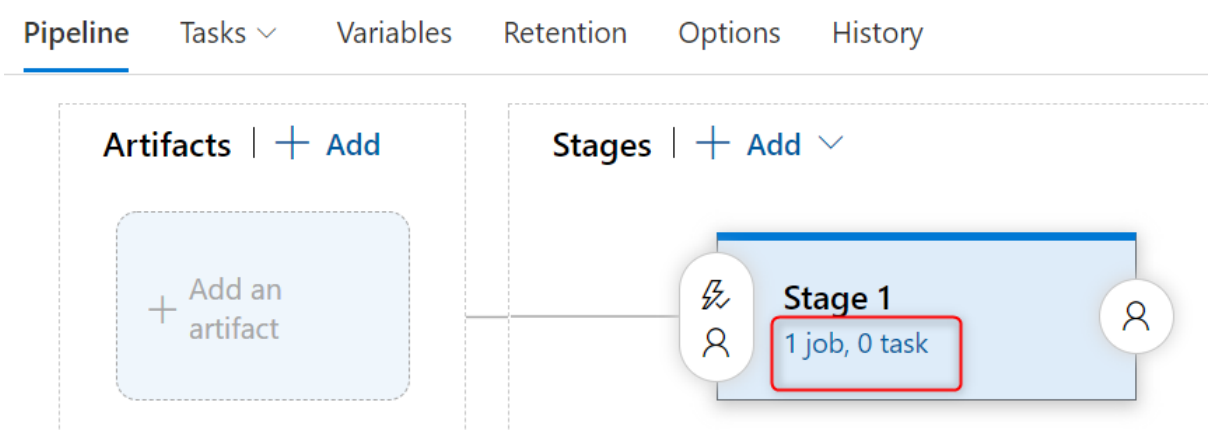


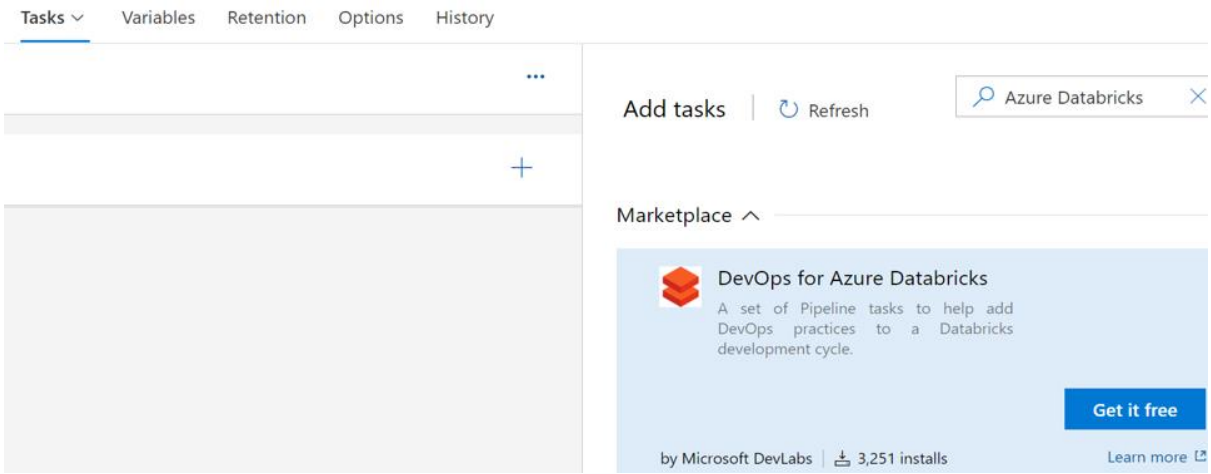

Select path
CookbookRepo
notebooks
6.1-Reading Writing to Delta Tables.py
README.md
OK
Cancel
Summary
Queue
Display name *
Publish Artifact: adbnotebooks_drop
Path to publish *
notebooks
Artifact name *
adbnotebooks_drop
Artifact publish location *
Azure Pipelines
Advanced
Control Options

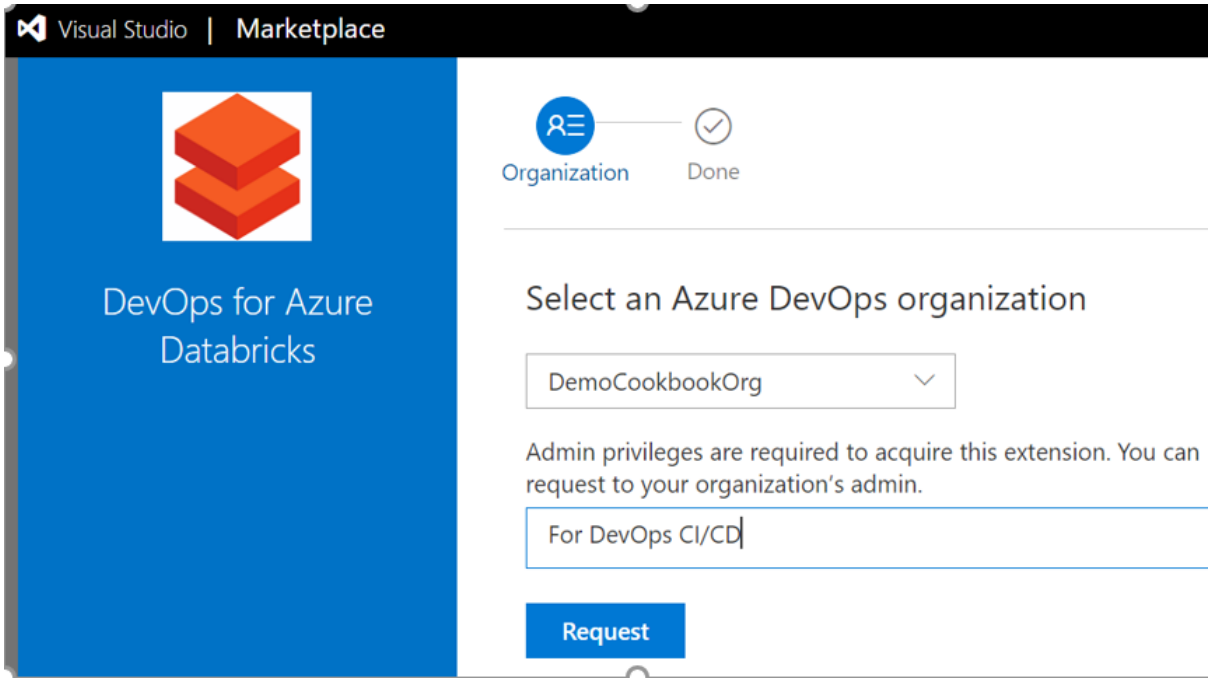

Summary
Manually run by admin
View 8 changes
Repository and version
CookbookRepo
main 018e95b
Time started and elapsed
Today at 6:58 PM
16s
Related
0 work items
1 published; 1 consumed
Tests and coverage
Get started
Jobs
Name
Status
Duration
Agent job 1
Success
7s

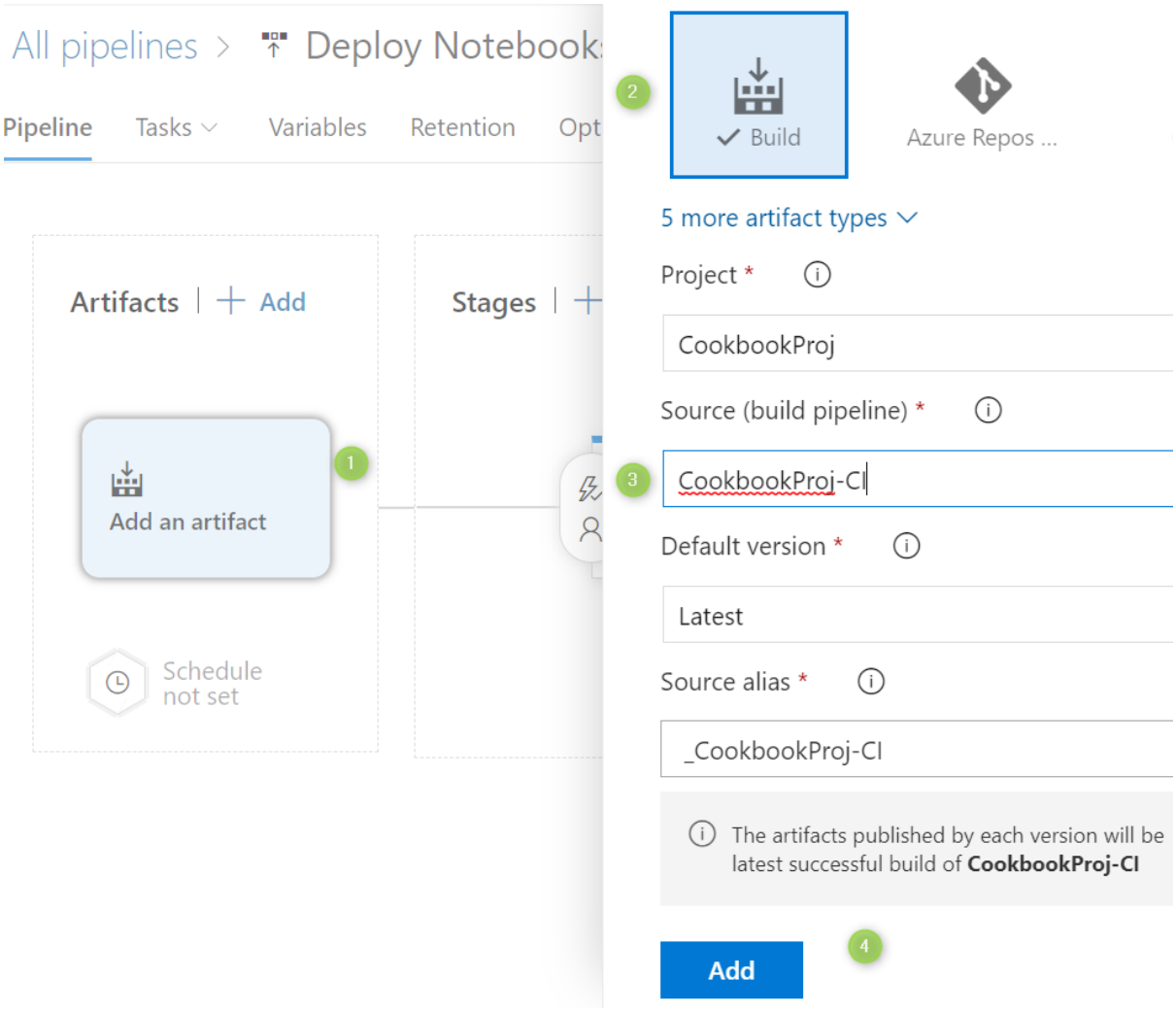

CookbookProj
Overview
Boards
Repos
Pipelines
Pipelines
Environments
Releases
1
All pipelines
Pipeline
Tasks
Stage 1
Select a template

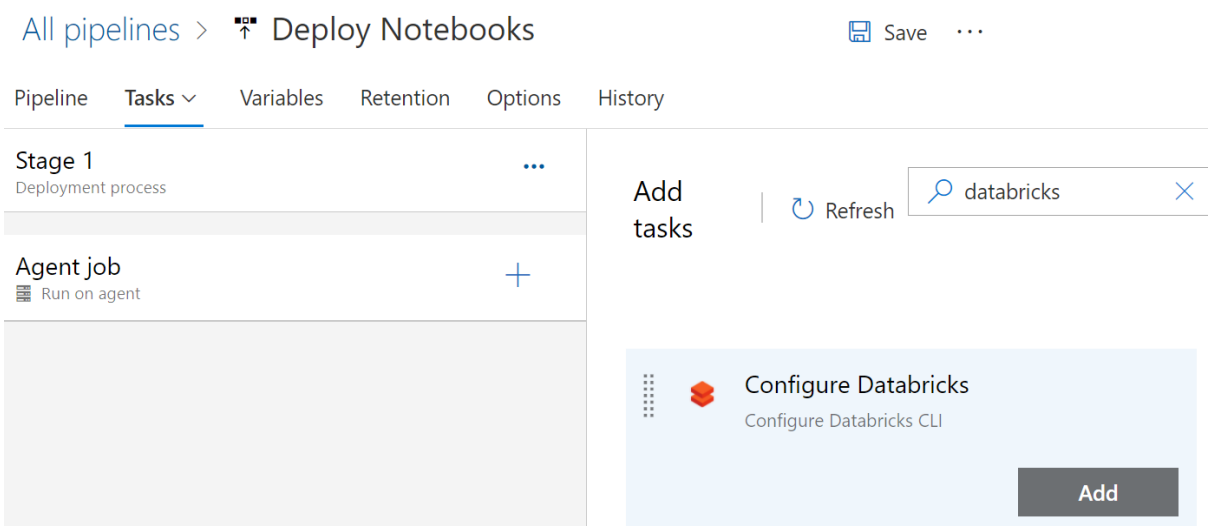

Select a template
Or start with an
Empty job
2
Featured
Azure App Service deployment
Deploy your application to Azure App Service. Web App on Windows, Linux, containers, WebJobs.
Deploy a Java app to Azure App Service
Deploy a Java application to an Azure Web App.
Deploy a Node.js app to Azure App
Deploy a Node.js application to an Azure Web App.

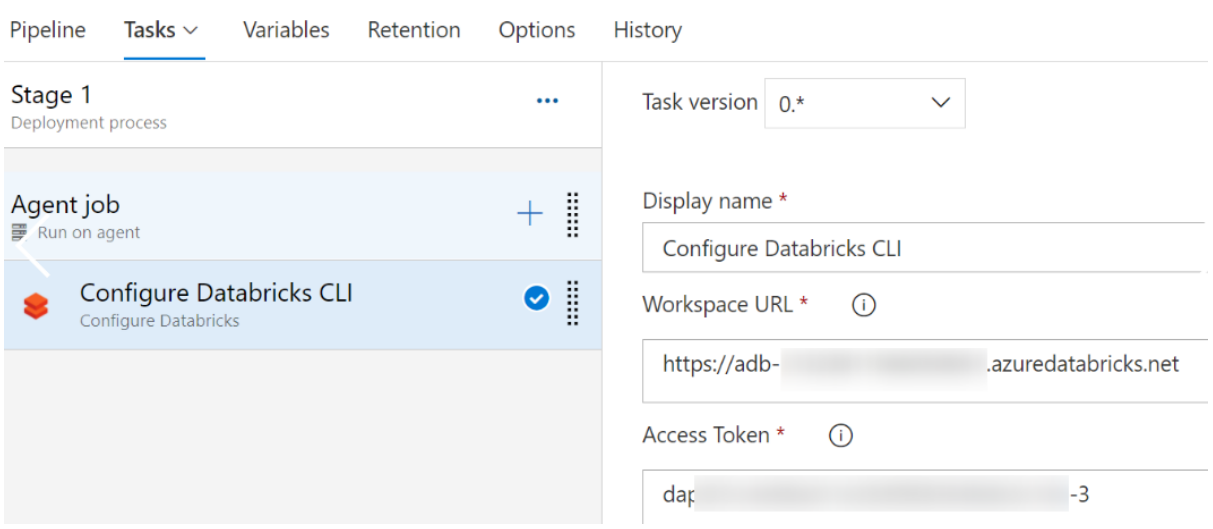

Pipeline
Tasks
Variables
Retention
Options
History
Artifacts
Add
Add an artifact
Stages
Add
Stage 1
1 job, 0 task

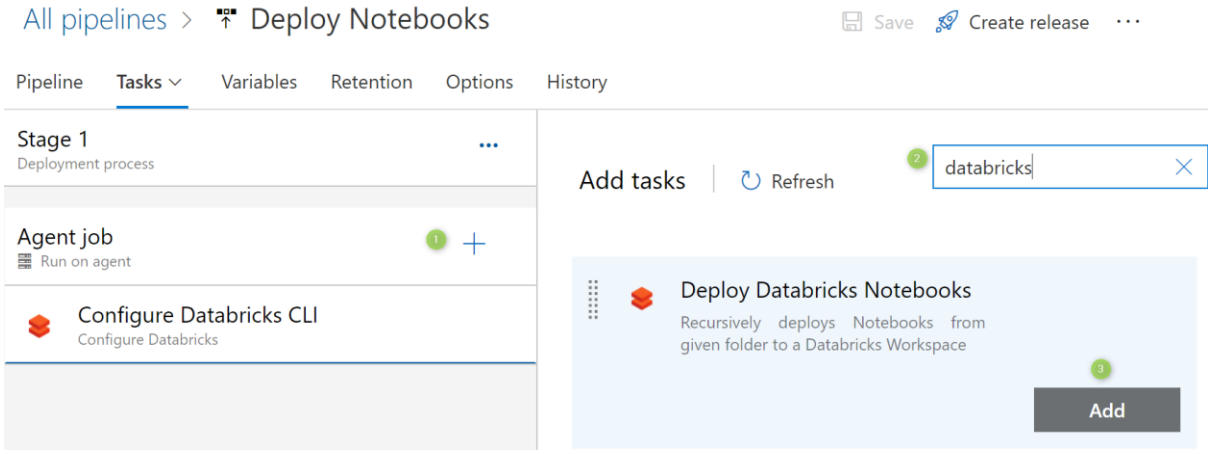

Tasks
Variables
Retention
Options
History
Add tasks
Refresh
Azure Databricks
Marketplace
DevOps for Azure Databricks
A set of Pipeline tasks to help add DevOps practices to a Databricks development cycle.
Get it free
by Microsoft DevLabs
3,251 installs
Learn more

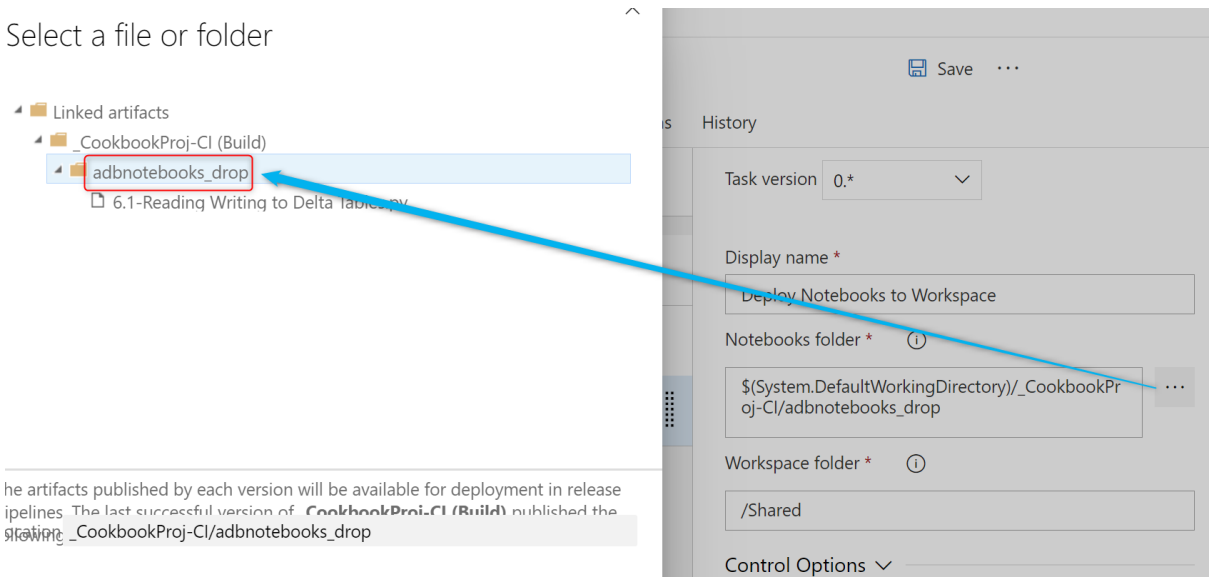

Visual Studio | Marketplace
DevOps for Azure Databricks
Organization
Done
Select an Azure DevOps organization
DemoCookbookOrg
Admin privileges are required to acquire this extension. You can request to your organization's admin.
For DevOps CI/CD
Request

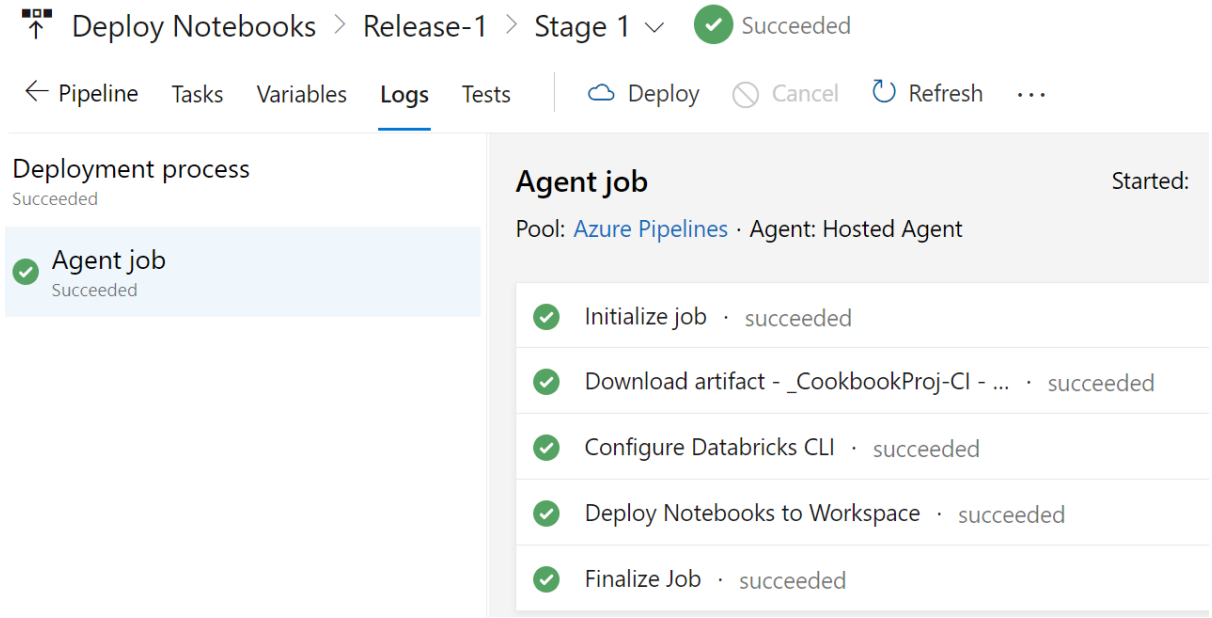

All pipelines > Deploy Notebook
Pipeline
Tasks
Variables
Retention
Opt
Artifacts | Add
Add an artifact
1
Schedule not set
Stages
2
Build
Azure Repos ...
5 more artifact types
Project *
CookbookProj
Source (build pipeline) *
3
CookbookProj-CI
Default version *
Latest
Source alias *
_CookbookProj-CI
The artifacts published by each version will be latest successful build of CookbookProj-CI
4
Add

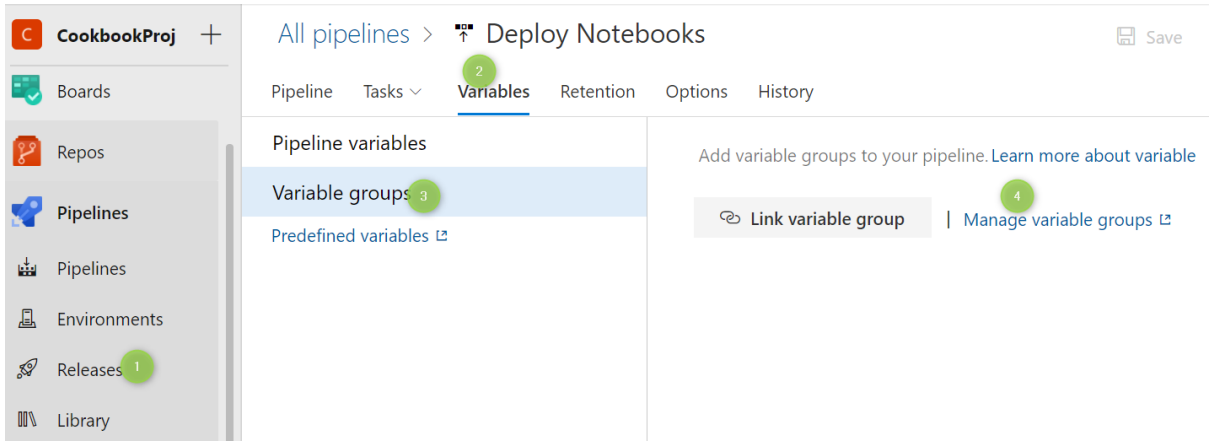

All pipelines
Deploy Notebooks
Save
Pipeline
Tasks
Variables
Retention
Options
History
Stage 1
Deployment process
Agent job
Run on agent
Add tasks
Refresh
databricks
Configure Databricks
Configure Databricks CLI
Add

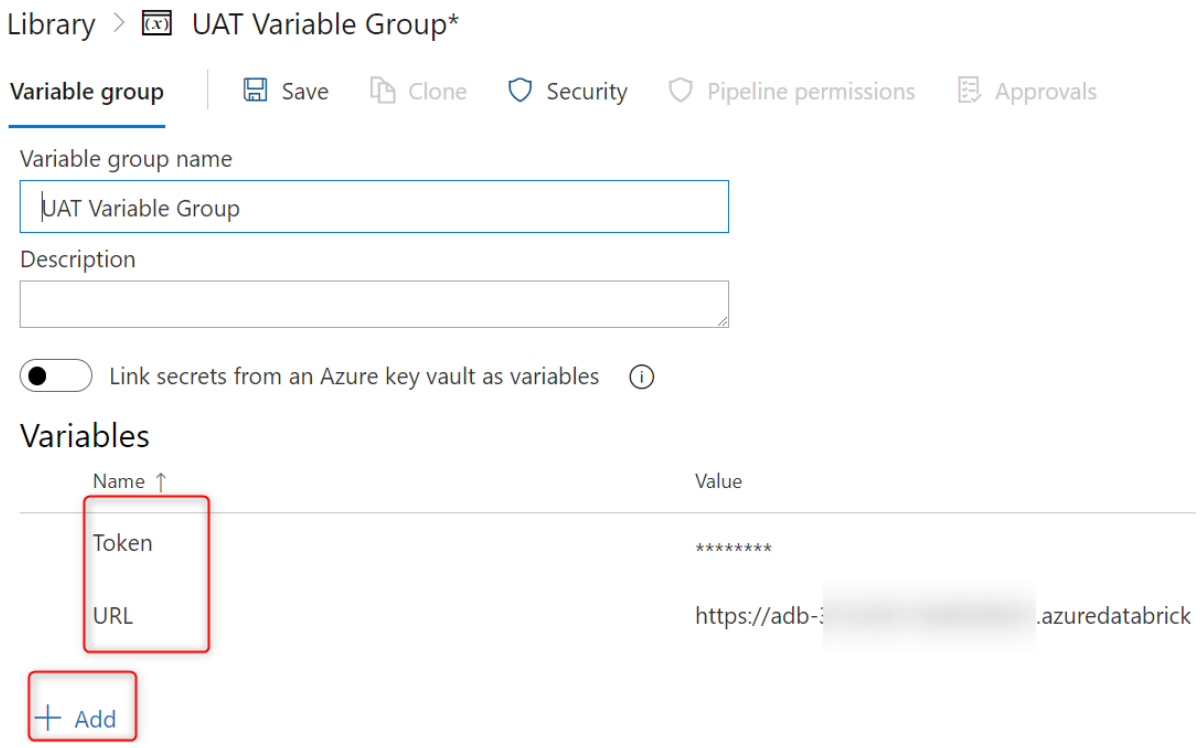

Pipeline
Tasks
Variables
Retention
Options
History
Stage 1
Deployment process
Agent job
Run on agent
Configure Databricks CLI
Configure Databricks
Task version
0.*
Display name *
Configure Databricks CLI
Workspace URL *
https://adb-
.azuredatabricks.net
Access Token *
-3

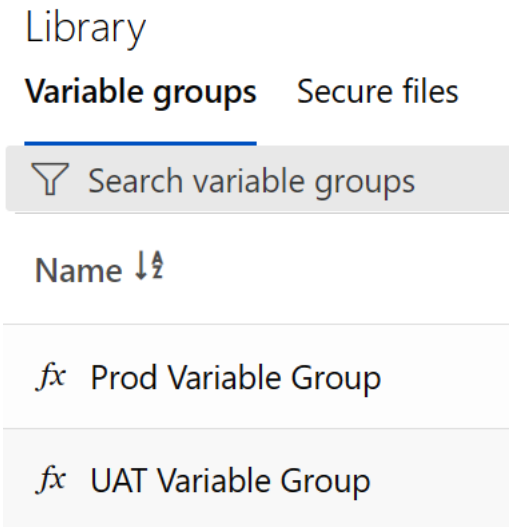

All pipelines
Deploy Notebooks
Save
Create release
Pipeline
Tasks
Variables
Retention
Options
History
Stage 1
Deployment process
Agent job
Run on agent
Configure Databricks CLI
Configure Databricks
Add tasks
Refresh
databricks
Deploy Databricks Notebooks
Recursively deploys Notebooks from given folder to a Databricks Workspace
Add

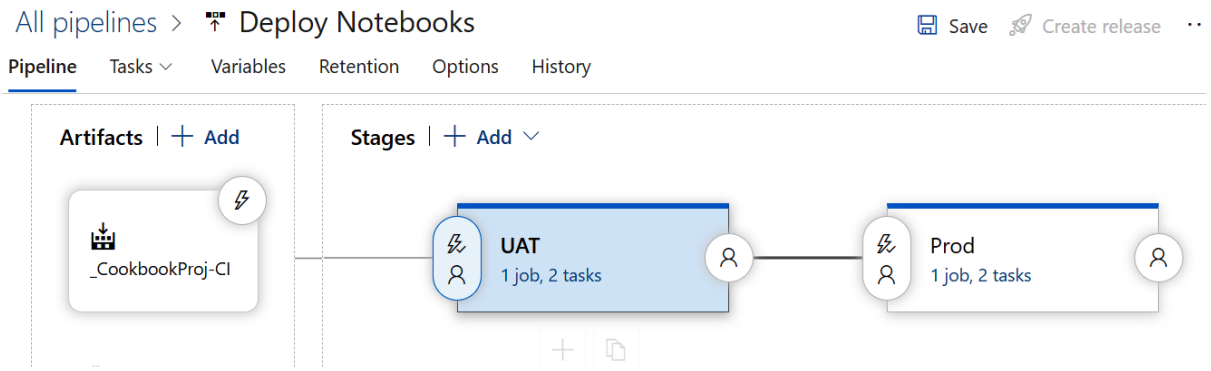

Select a file or folder
Linked artifacts
_CookbookProj-CI (Build)
adbnotebooks_drop
6.1-Reading Writing to Delta Tables.py
he artifacts published by each version will be available for deployment in release
ipelines. The last successful version of _CookbookProj-CI (Build) published the
ration _CookbookProj-CI/adbnotebooks_drop
Save
History
Task version 0.*
Display name *
Deploy Notebooks to Workspace
Notebooks folder *
$(System.DefaultWorkingDirectory)/_CookbookProj-CI/adbnotebooks_drop
Workspace folder *
/Shared
Control Options

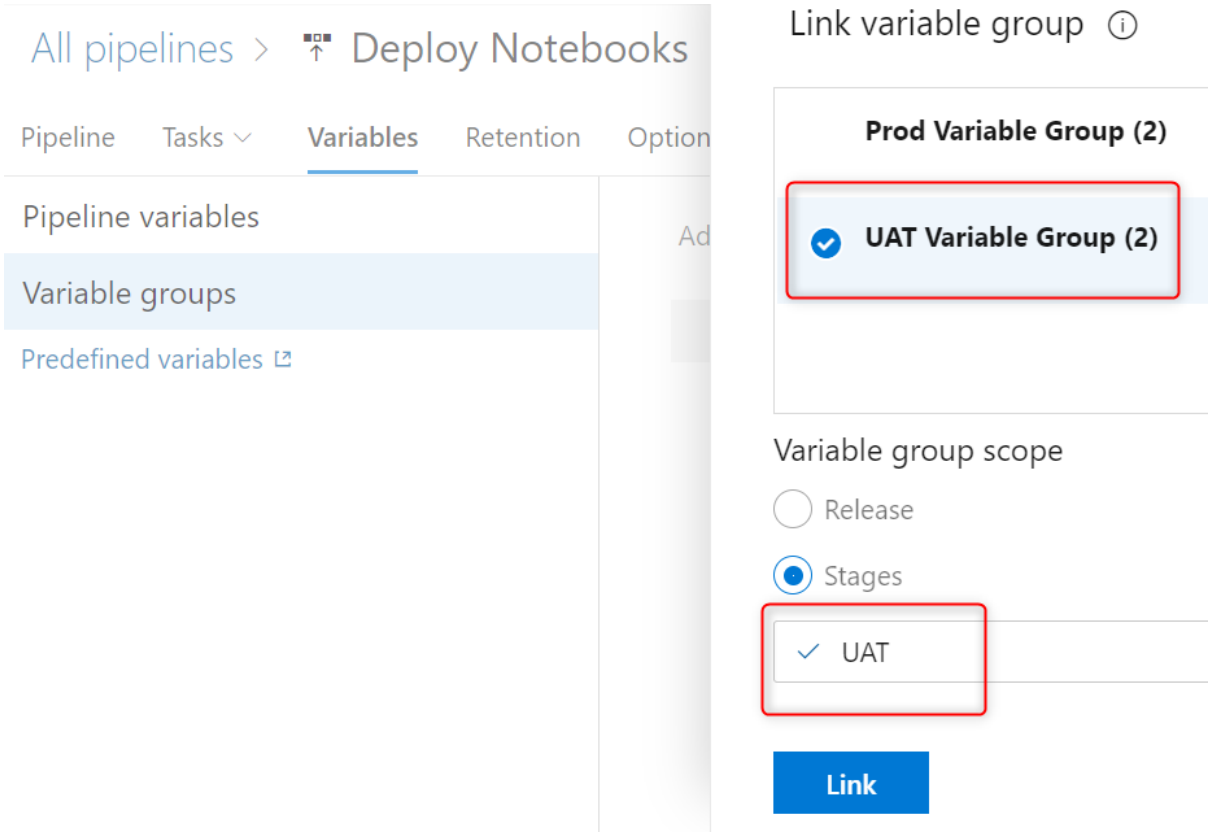

Deploy Notebooks
Release-1
Stage 1
Succeeded
Pipeline
Tasks
Variables
Logs
Tests
Deploy
Cancel
Refresh
Deployment process
Succeeded
Agent job
Succeeded
Agent job
Started:
Pool: Azure Pipelines · Agent: Hosted Agent
Initialize job · succeeded
Download artifact - _CookbookProj-CI - ... · succeeded
Configure Databricks CLI · succeeded
Deploy Notebooks to Workspace · succeeded
Finalize Job · succeeded

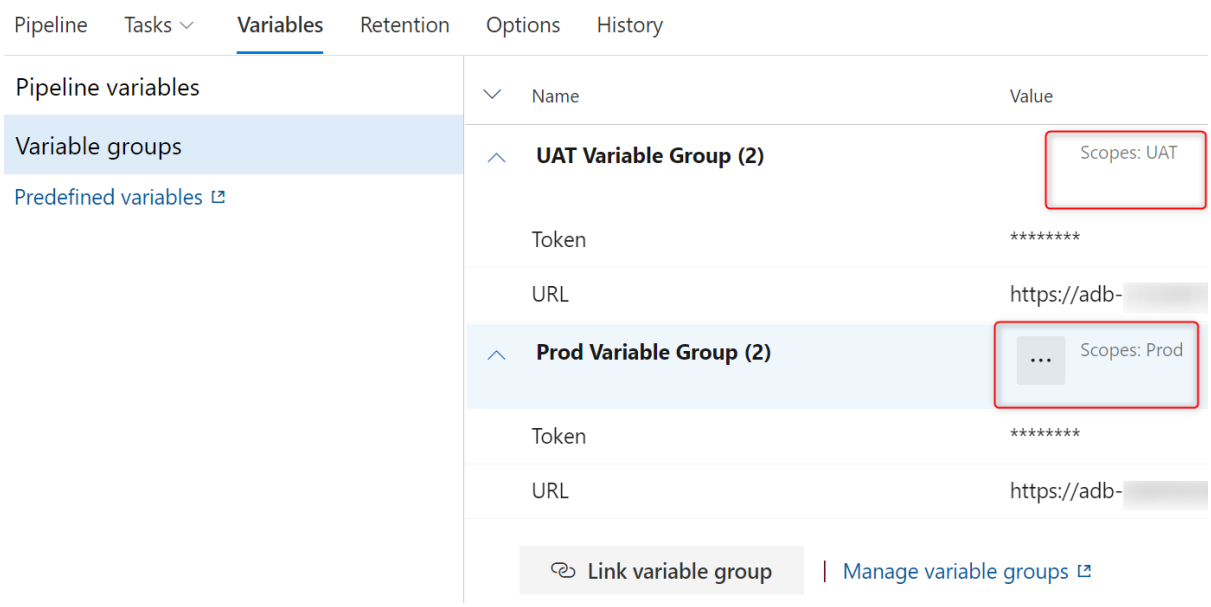

CookbookProj
Boards
Repos
Pipelines
Pipelines
Environments
Releases
Library
All pipelines
Deploy Notebooks
Save
Pipeline
Tasks
Variables
Retention
Options
History
Pipeline variables
Variable groups
Predefined variables
Add variable groups to your pipeline. Learn more about variable
Link variable group
Manage variable groups
1
2
3
4

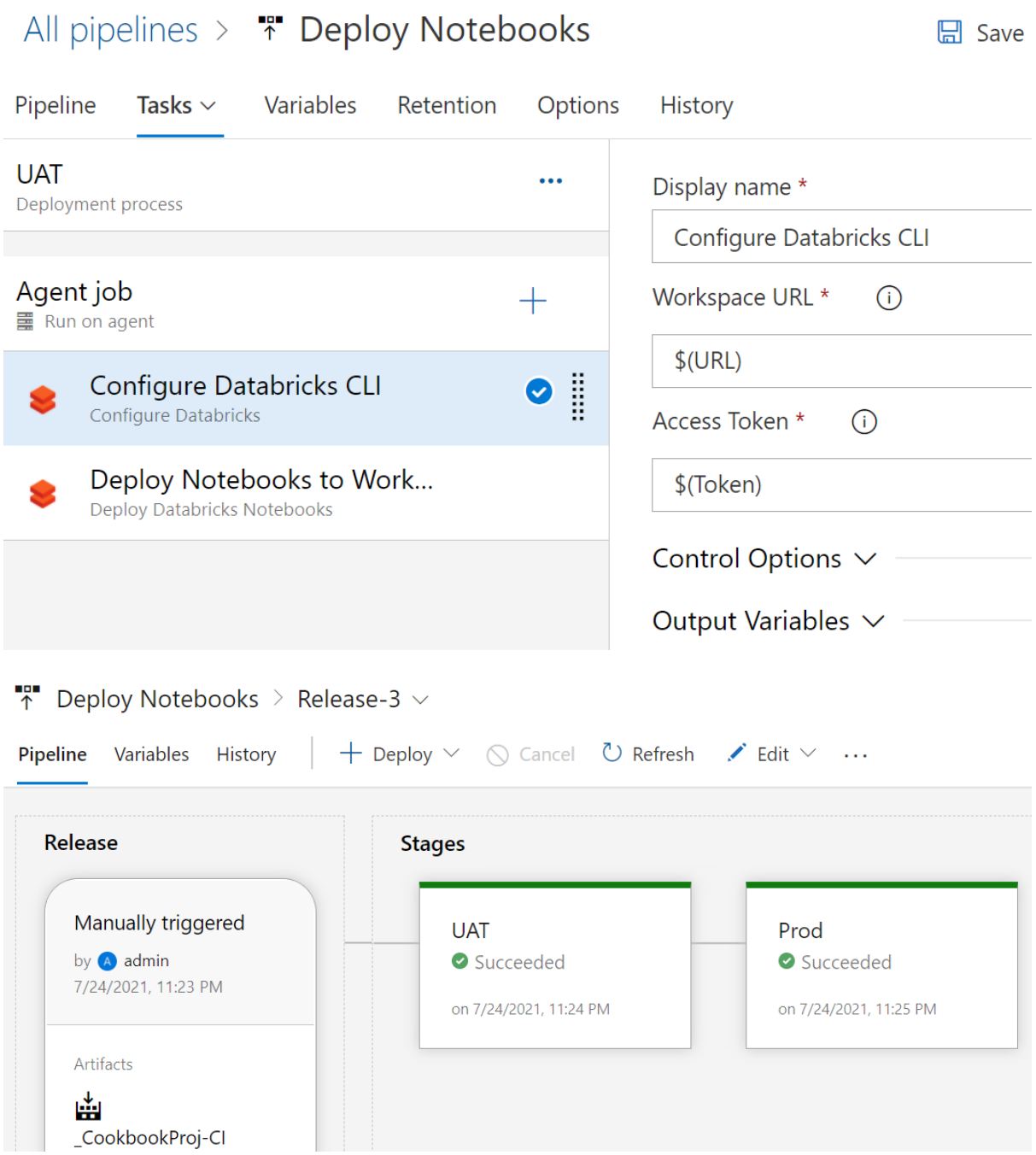

Library
UAT Variable Group*
Variable group
Save
Clone
Security
Pipeline permissions
Approvals
Variable group name
UAT Variable Group
Description
Link secrets from an Azure key vault as variables
Variables
Name
Value
Token
****
URL
https://adb-
.azuredatabrick
Add


Library
Variable groups
Secure files
Search variable groups
Name
Prod Variable Group
UAT Variable Group


All pipelines
Deploy Notebooks
Save
Create release
Pipeline
Tasks
Variables
Retention
Options
History
Artifacts
Add
_CookbookProj-CI
Stages
Add
UAT
1 job, 2 tasks
Prod
1 job, 2 tasks


All pipelines
Deploy Notebooks
Pipeline
Tasks
Variables
Retention
Option
Pipeline variables
Variable groups
Predefined variables
Link variable group
Prod Variable Group (2)
UAT Variable Group (2)
Variable group scope
Release
Stages
UAT
Link


Pipeline
Tasks
Variables
Retention
Options
History
Pipeline variables
Variable groups
Predefined variables
Name
Value
UAT Variable Group (2)
Scopes: UAT
Token
****
URL
https://adb-
Prod Variable Group (2)
...
Scopes: Prod
Token
****
URL
https://adb-
Link variable group
Manage variable groups


All pipelines
Deploy Notebooks
Save
Pipeline
Tasks
Variables
Retention
Options
History
UAT
Deployment process
Agent job
Run on agent
Configure Databricks CLI
Configure Databricks
Deploy Notebooks to Work...
Deploy Databricks Notebooks
Display name *
Configure Databricks CLI
Workspace URL *
$(URL)
Access Token *
$(Token)
Control Options
Output Variables
Deploy Notebooks
Release-3
Pipeline
Variables
History
Deploy
Cancel
Refresh
Edit
Release
Manually triggered
by admin
7/24/2021, 11:23 PM
Artifacts
_CookbookProj-CI
Stages
UAT
Succeeded
on 7/24/2021, 11:24 PM
Prod
Succeeded
on 7/24/2021, 11:25 PM

Library > Prod Variable Group*

Variable group | Save | Clone | Security | Pipeline permissions

Prod Variable Group

Description

Link secrets from an Azure key vault as variables

Azure subscription * | Manage

[redacted]

Scoped to subscription 'Visual Studio Enterprise'

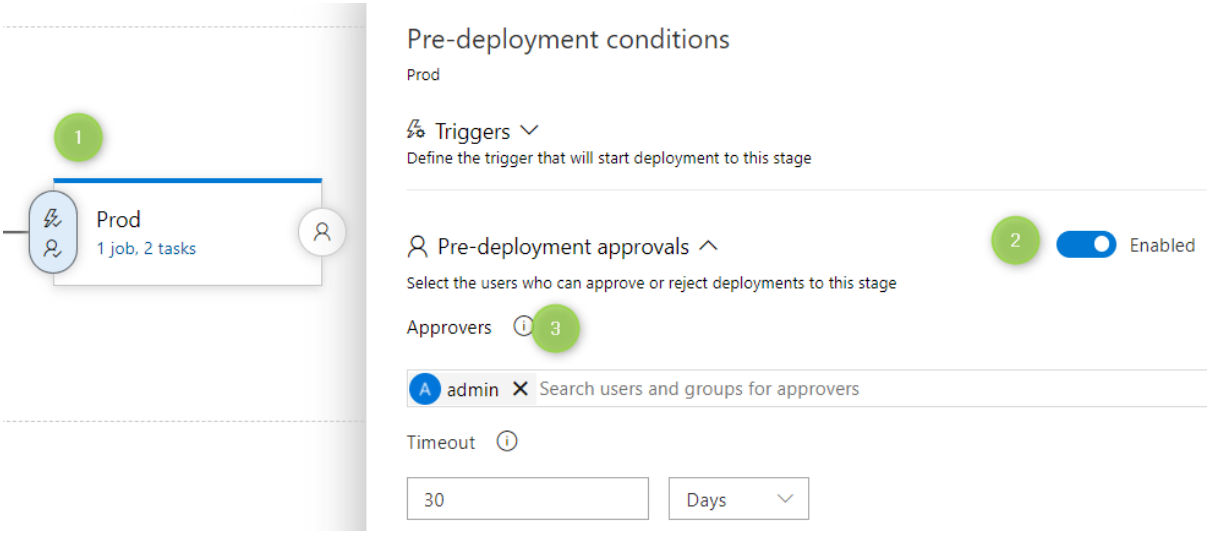
Key vault name * Manage

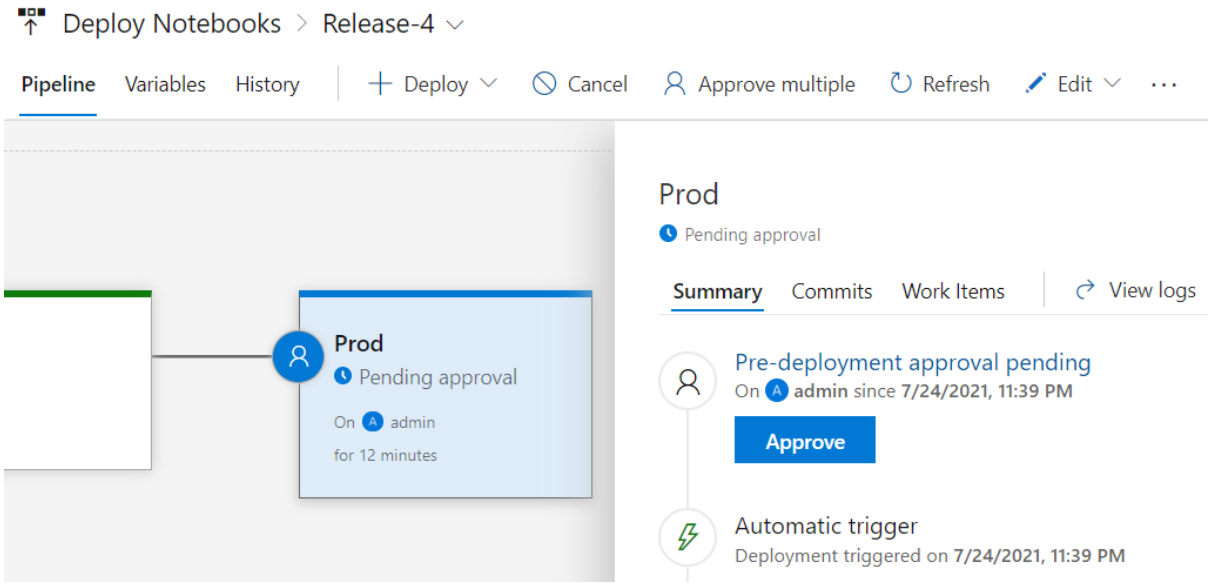
CookbookKeyVault

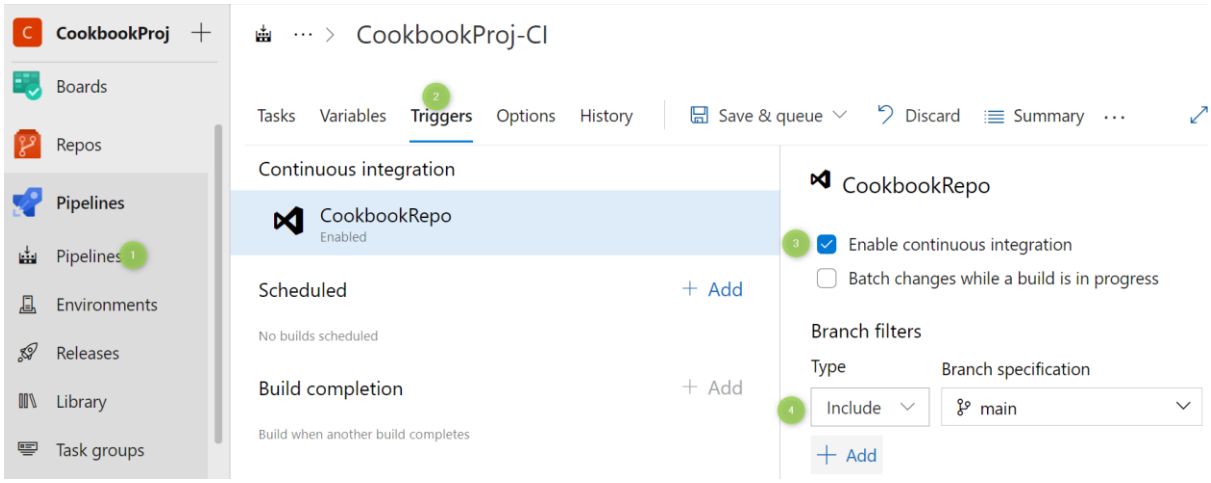
Variables

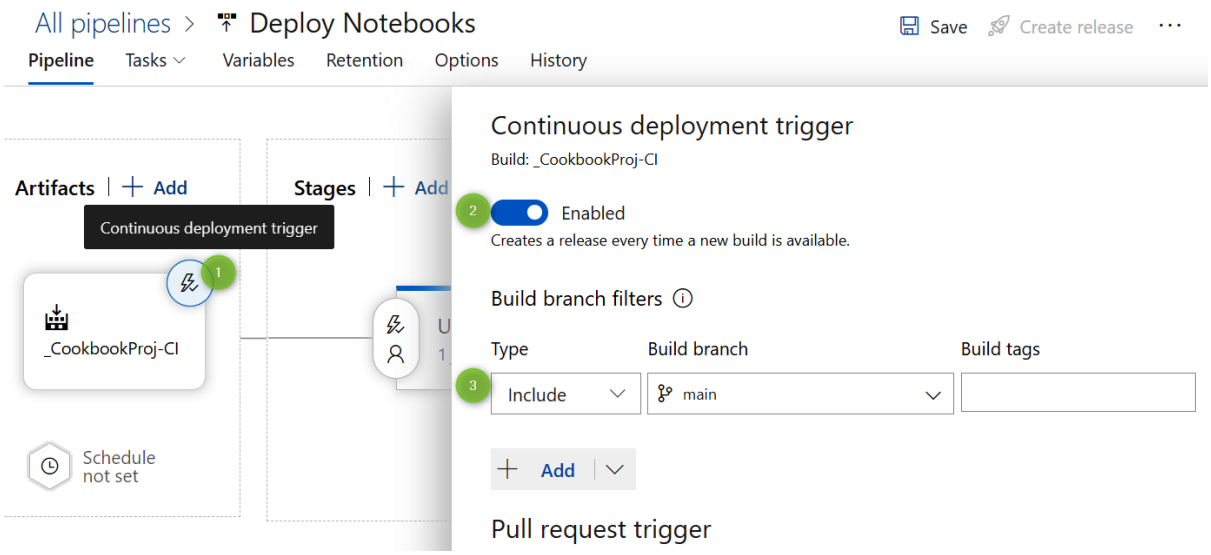
Last refreshed: 6 hours ago

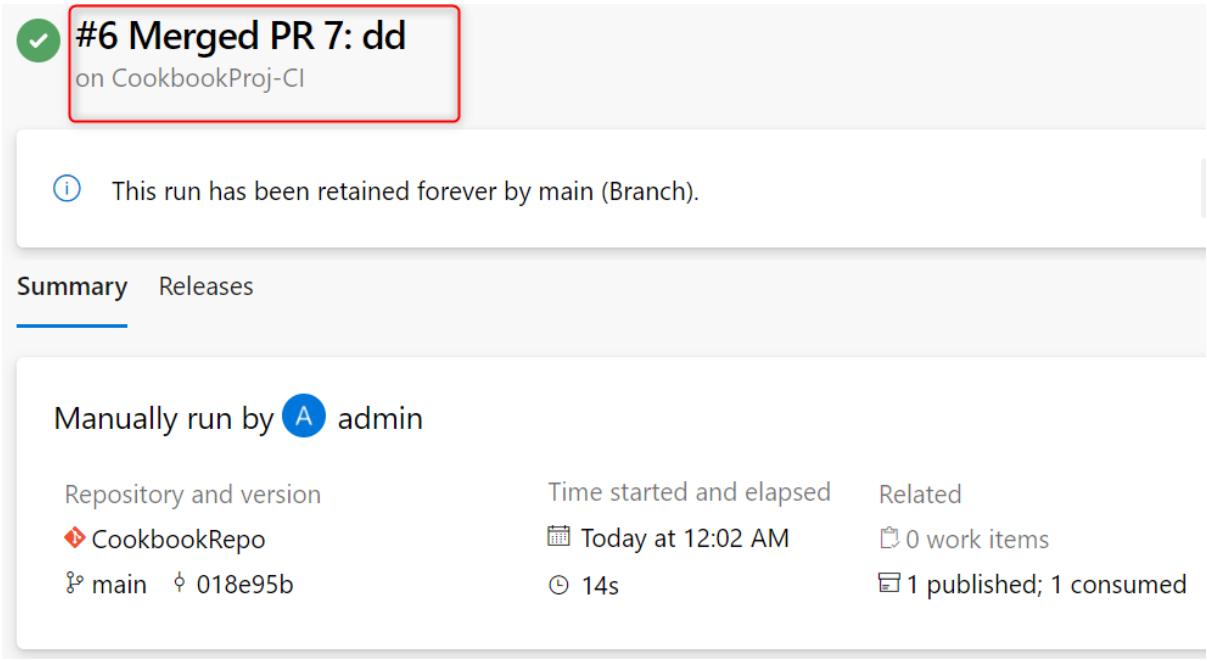
Delete	Secret name	Content type	Status	Expiration date
	ADBToken		Enabled	Never

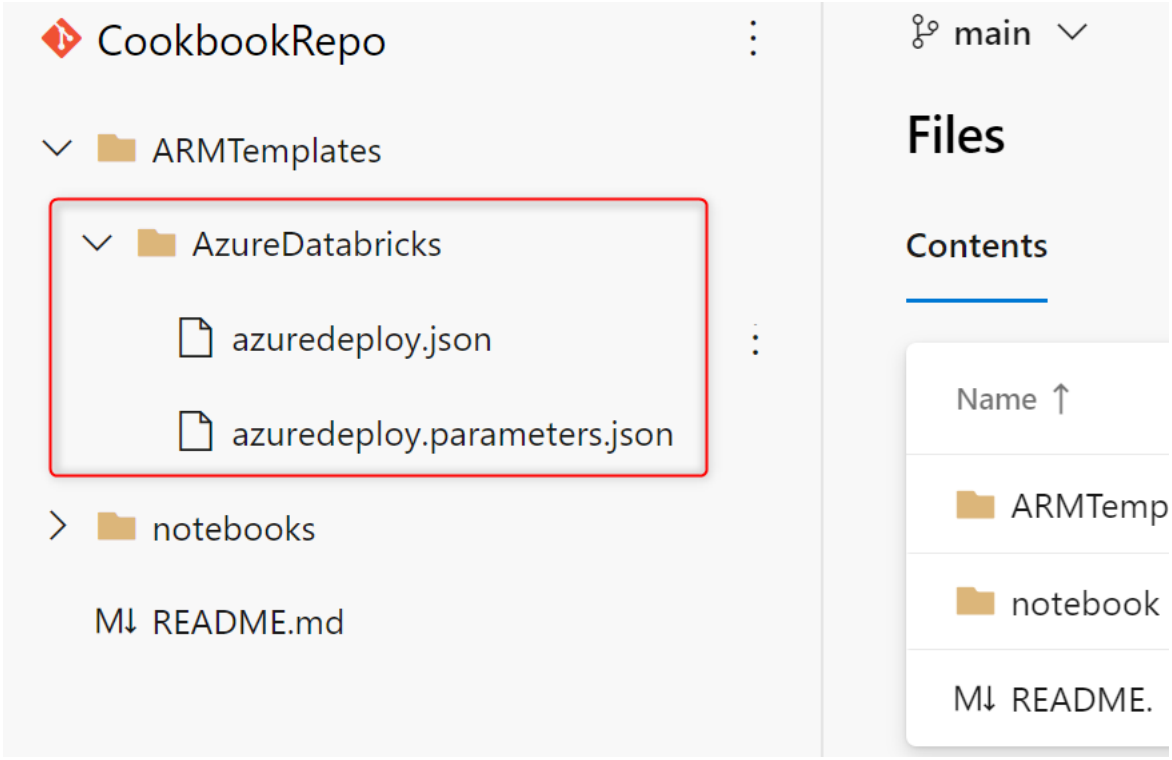


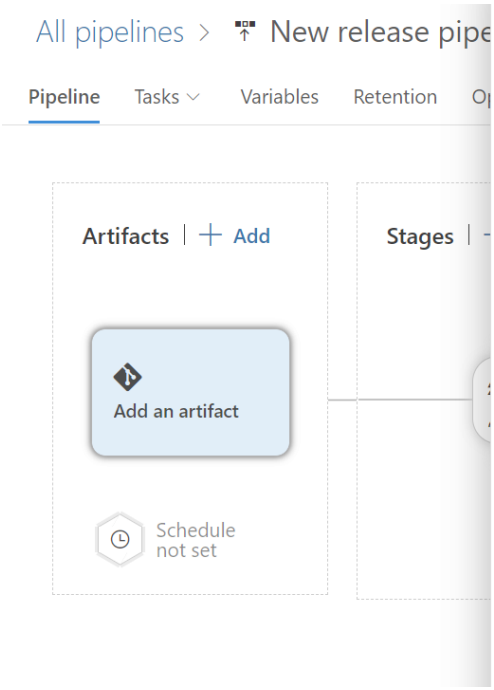

Deploy Notebooks
Release-4
Pipeline
Variables
History
Deploy
Cancel
Approve multiple
Refresh
Edit
Prod
Pending approval
On admin
for 12 minutes
Prod
Pending approval
Summary
Commits
Work Items
View logs
Pre-deployment approval pending
On admin since 7/24/2021, 11:39 PM
Approve
Automatic trigger
Deployment triggered on 7/24/2021, 11:39 PM

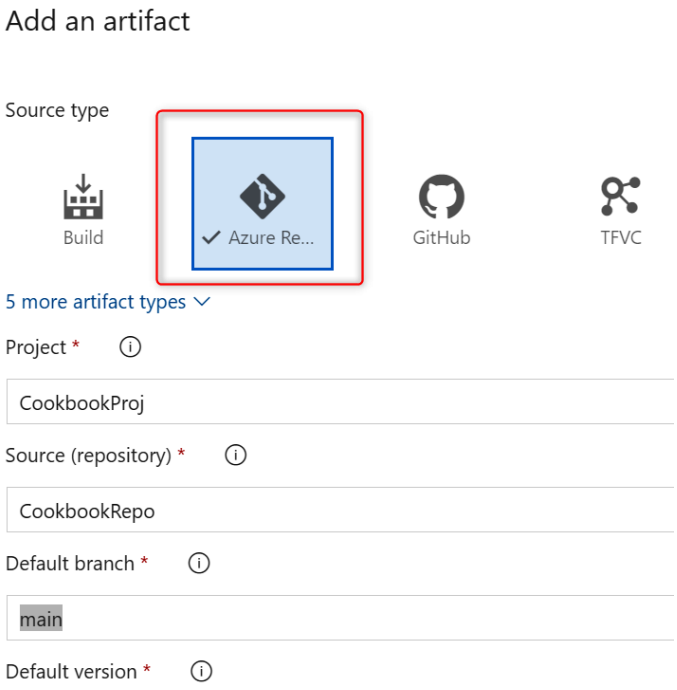

CookbookProj
Boards
Repos
Pipelines
Pipelines
Environments
Releases
Library
Task groups
CookbookProj-CI
Tasks
Variables
Triggers
Options
History
Save & queue
Discard
Summary
Continuous integration
CookbookRepo
Enabled
Scheduled
Add
No builds scheduled
Build completion
Add
Build when another build completes
CookbookRepo
Enable continuous integration
Batch changes while a build is in progress
Branch filters
Type
Branch specification
Include
main
Add

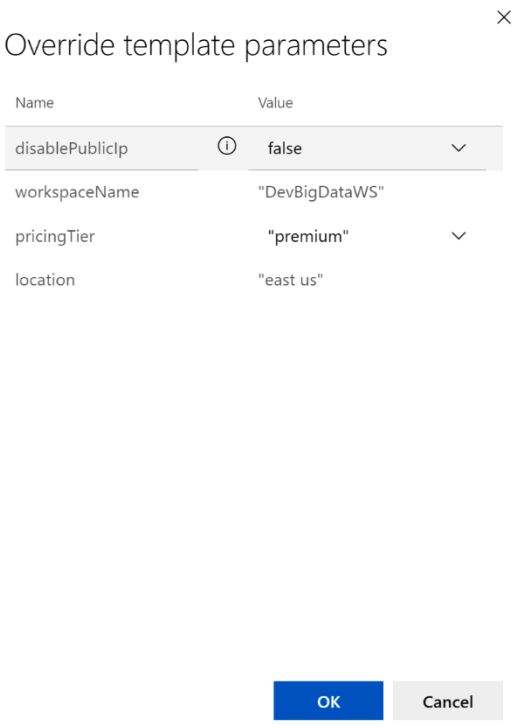

All pipelines
Deploy Notebooks
Save
Create release
Pipeline
Tasks
Variables
Retention
Options
History
Artifacts
Add
Stages
Add
Continuous deployment trigger
_CookbookProj-CI
Schedule not set
Continuous deployment trigger
Build: _CookbookProj-CI
Enabled
Creates a release every time a new build is available.
Build branch filters
Type
Build branch
Build tags
Include
main
Add
Pull request trigger

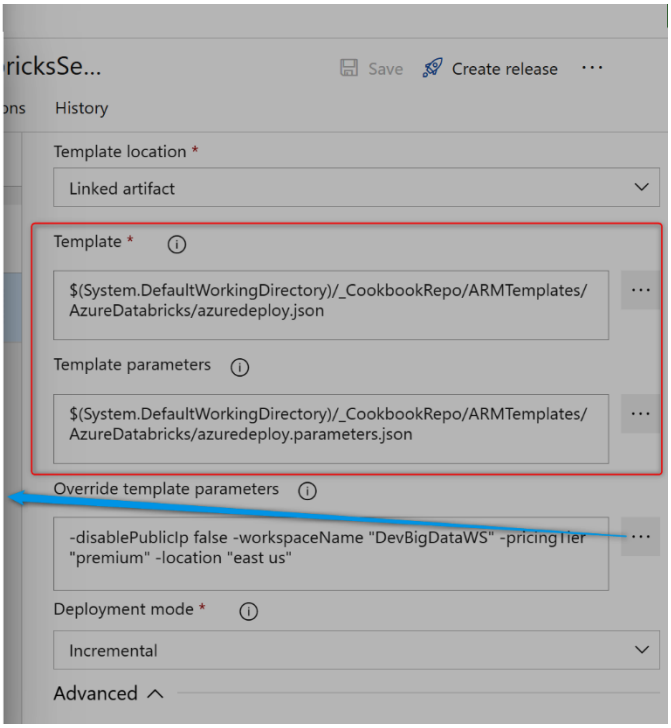

#6 Merged PR 7: dd
on CookbookProj-CI
This run has been retained forever by main (Branch).
Summary
Releases
Manually run by admin
Repository and version
CookbookRepo
main
018e95b
Time started and elapsed
Today at 12:02 AM
14s
Related
0 work items
1 published; 1 consumed

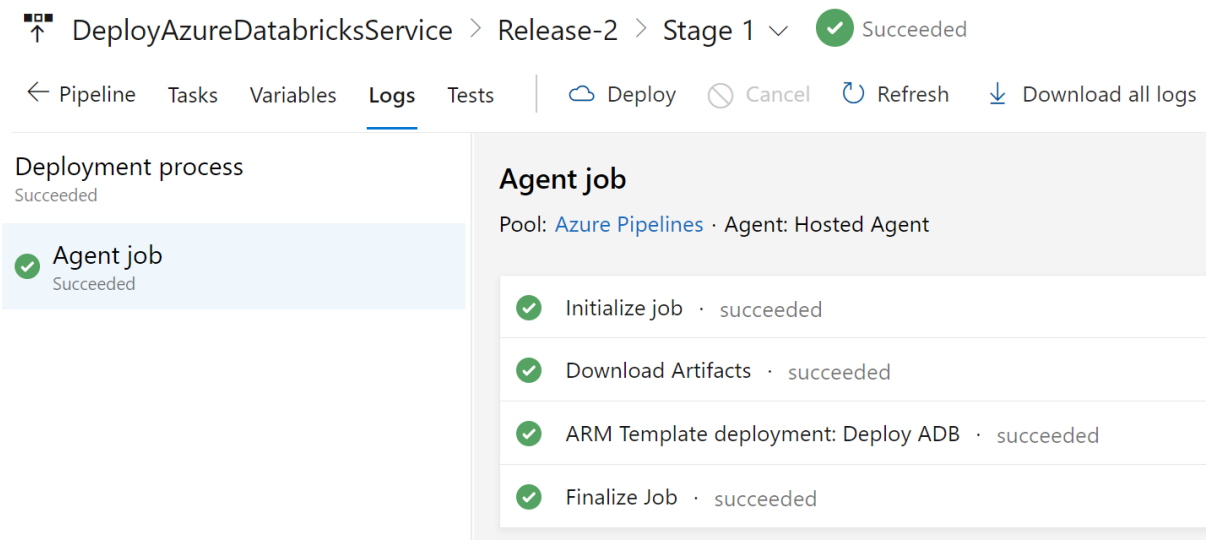

CookbookRepo
ARMTemplates
AzureDatabricks
azuredeploy.json
azuredeploy.parameters.json
notebooks
README.md
main
Files
Contents
Name
ARMTemp
notebook
README.


All pipelines
New release pipe
Pipeline
Tasks
Variables
Retention
Artifacts
Add
Add an artifact
Stages
Schedule
not set


Add an artifact
Source type
Build
Azure Re...
GitHub
TFVC
5 more artifact types
Project *
CookbookProj
Source (repository) *
CookbookRepo
Default branch *
main
Default version *

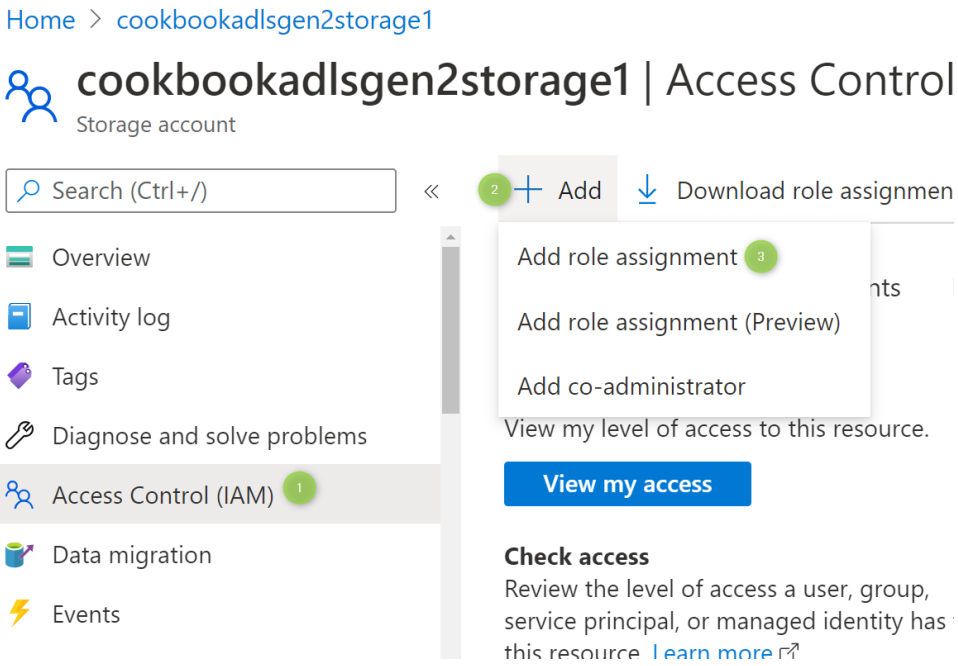

Override template parameters
Name
Value
disablePublicIp
false
workspaceName
"DevBigDataWS"
pricingTier
"premium"
location
"east us"
OK
Cancel

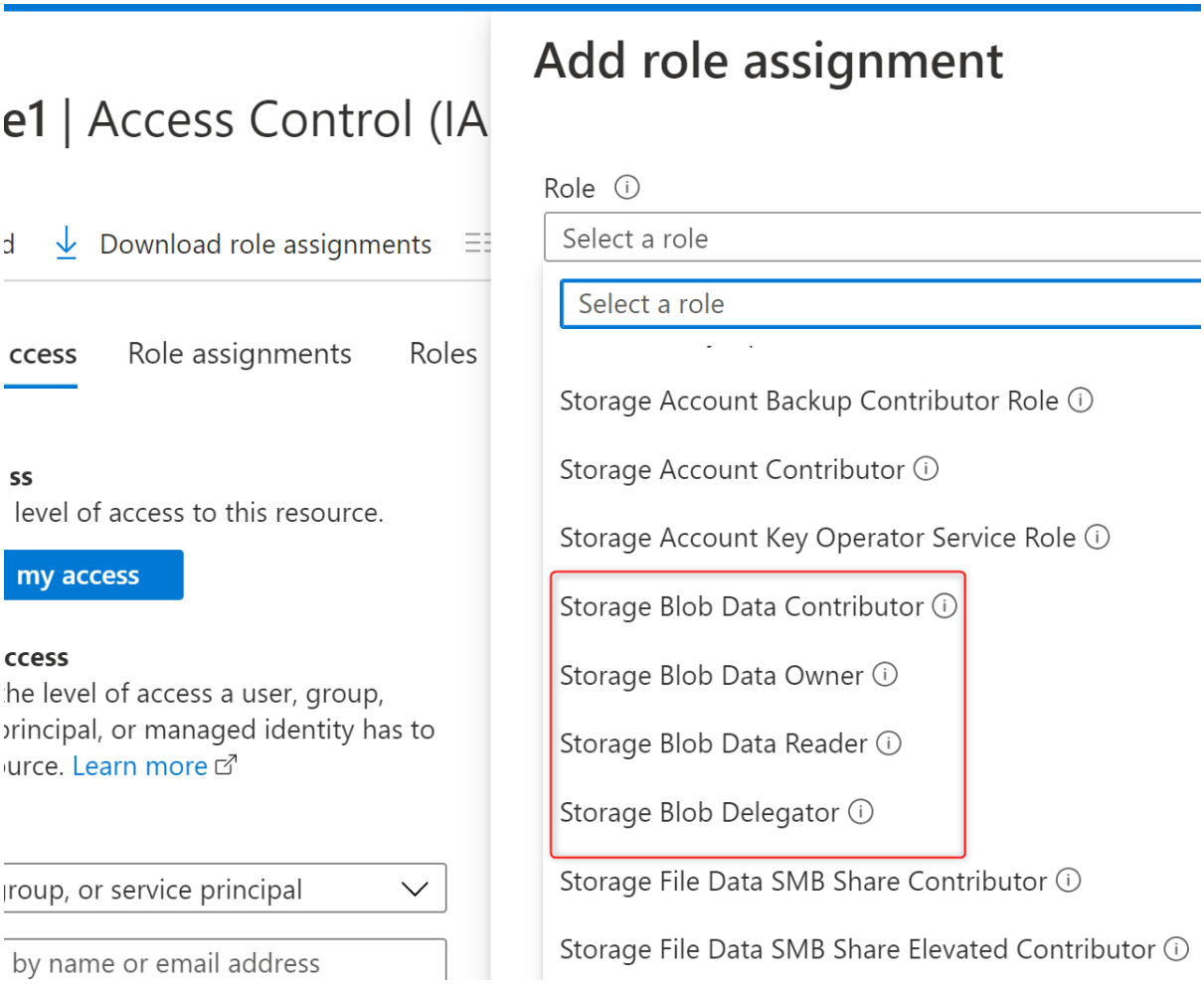

Save
Create release
History
Template location *
Linked artifact
Template *
$(System.DefaultWorkingDirectory)/_CookbookRepo/ARMTemplates/AzureDatabricks/azuredeploy.json
Template parameters
$(System.DefaultWorkingDirectory)/_CookbookRepo/ARMTemplates/AzureDatabricks/azuredeploy.parameters.json
Override template parameters
-disablePublicIp false -workspaceName "DevBigDataWS" -pricingTier "premium" -location "east us"
Deployment mode *
Incremental
Advanced



Environment	Resource Group	Workspace Name
Dev	CookbookRG	BigDataAnalytics
UAT	UATCookbookRG	UATBigDataAnalytics
Production	ProdCookbookRG	ProdBigDataAnalytics

Chapter 10: Understanding Security and Monitoring in Azure Databricks




e1 | Access Control (IA
Download role assignments
ccess
Role assignments
Roles
ss
level of access to this resource.
my access
ccess
he level of access a user, group,
rincipal, or managed identity has to
urce. Learn more
roup, or service principal
by name or email address
Add role assignment
Role
Select a role
Select a role
Storage Account Backup Contributor Role
Storage Account Contributor
Storage Account Key Operator Service Role
Storage Blob Data Contributor
Storage Blob Data Owner
Storage Blob Data Reader
Storage Blob Delegator
Storage File Data SMB Share Contributor
Storage File Data SMB Share Elevated Contributor

Add role assignment

Role

Storage Blob Data Reader

Assign access to

User, group, or service principal

Select

sqluser

No users, groups, or service principals found.

Selected members:

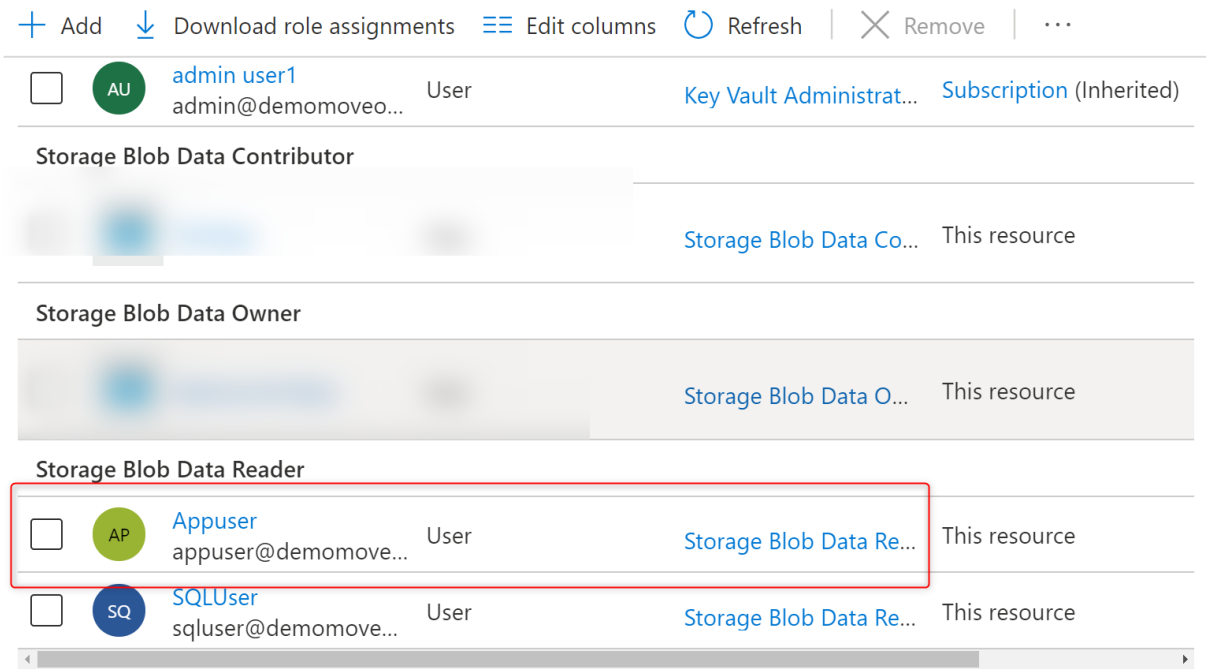
SQ

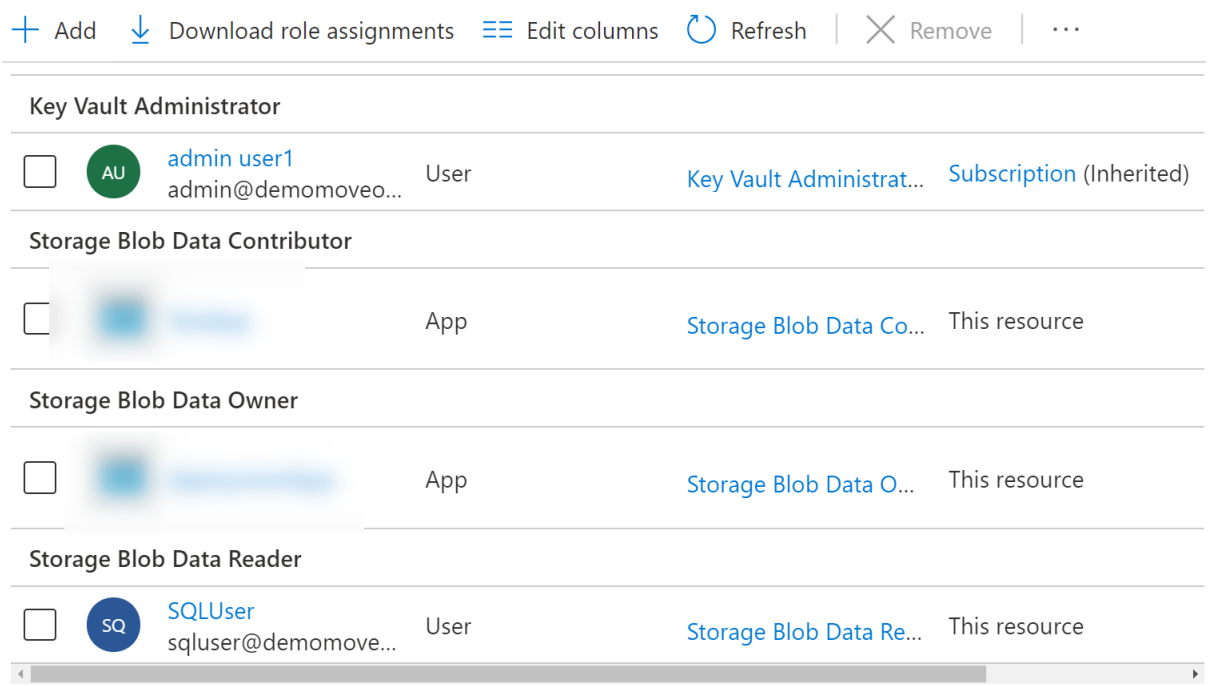
SQLUser
sqluser@demomoveorg.onmicrosoft.c...

Remove

Save

Discard


Add
Download role assignments
Edit columns
Refresh
Remove
admin user1
admin@demomoveo...
User
Key Vault Administrat...
Subscription (Inherited)
Storage Blob Data Contributor
Storage Blob Data Co...
This resource
Storage Blob Data Owner
Storage Blob Data O...
This resource
Storage Blob Data Reader
Appuser
appuser@demomove...
User
Storage Blob Data Re...
This resource
SQLUser
sqluser@demomove...
User
Storage Blob Data Re...
This resource


Add
Download role assignments
Edit columns
Refresh
Remove
Key Vault Administrator
admin user1
admin@demomoveo...
User
Key Vault Administrat...
Subscription (Inherited)
Storage Blob Data Contributor
App
Storage Blob Data Co...
This resource
Storage Blob Data Owner
App
Storage Blob Data O...
This resource
Storage Blob Data Reader
SQLUser
sqluser@demomove...
User
Storage Blob Data Re...
This resource

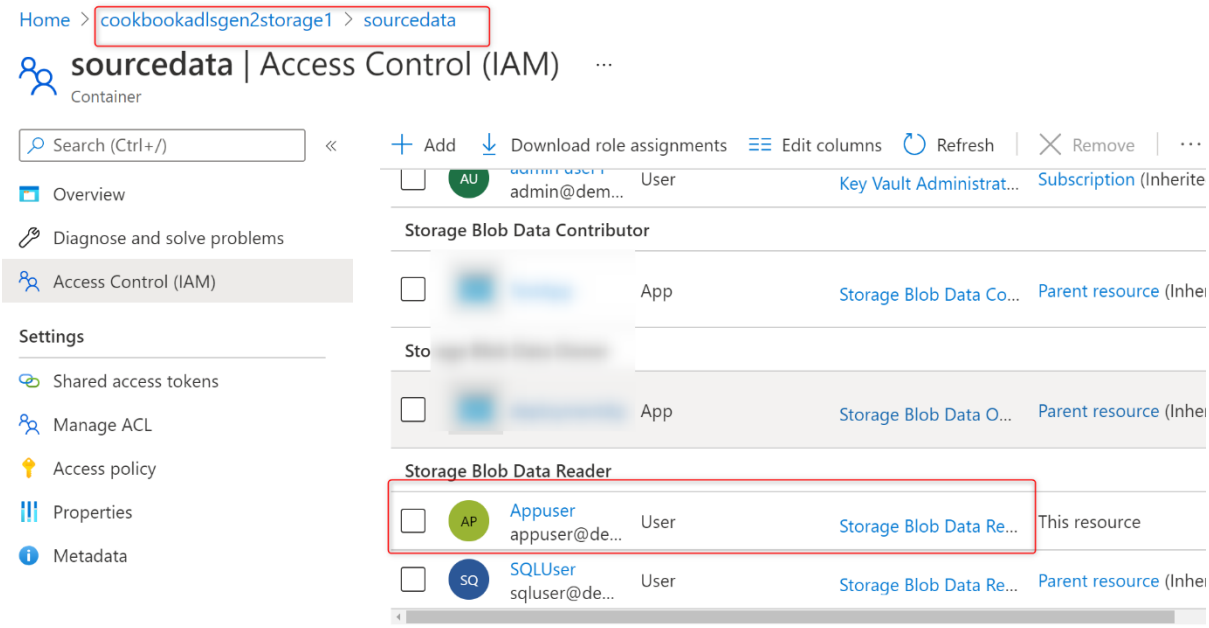
Storage Blob Data Owner

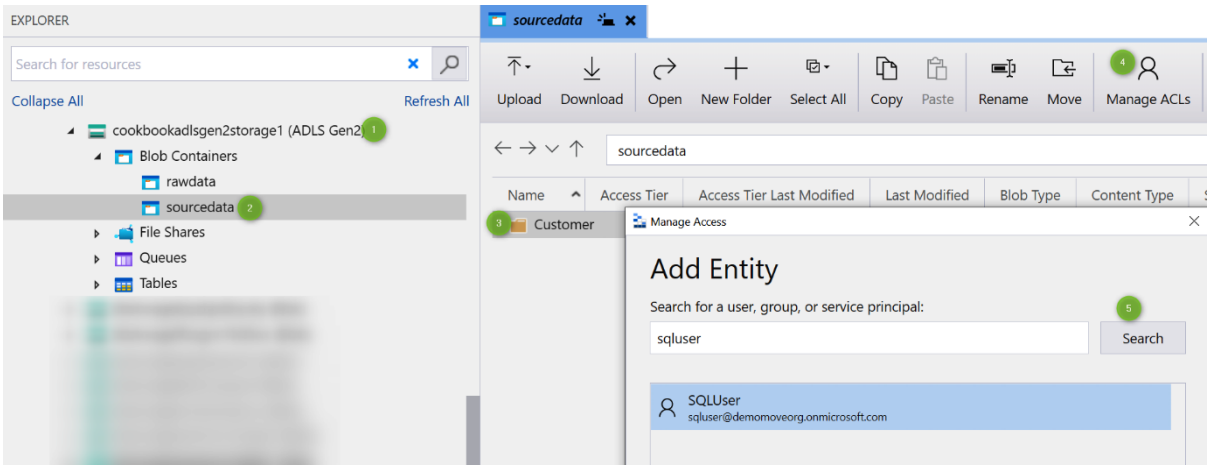
deploymentAp App Storage Blob Data O...

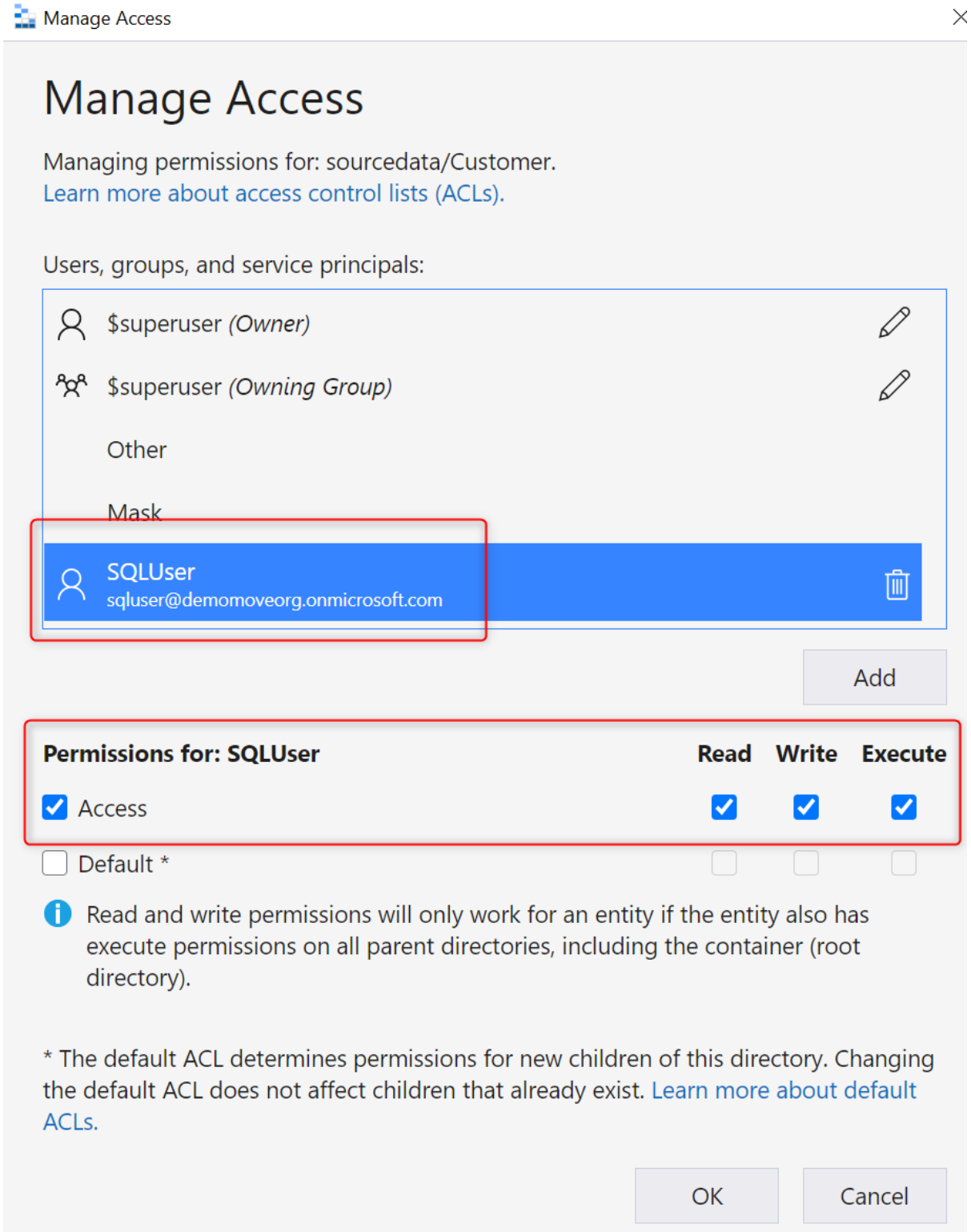
Storage Blob Data Reader

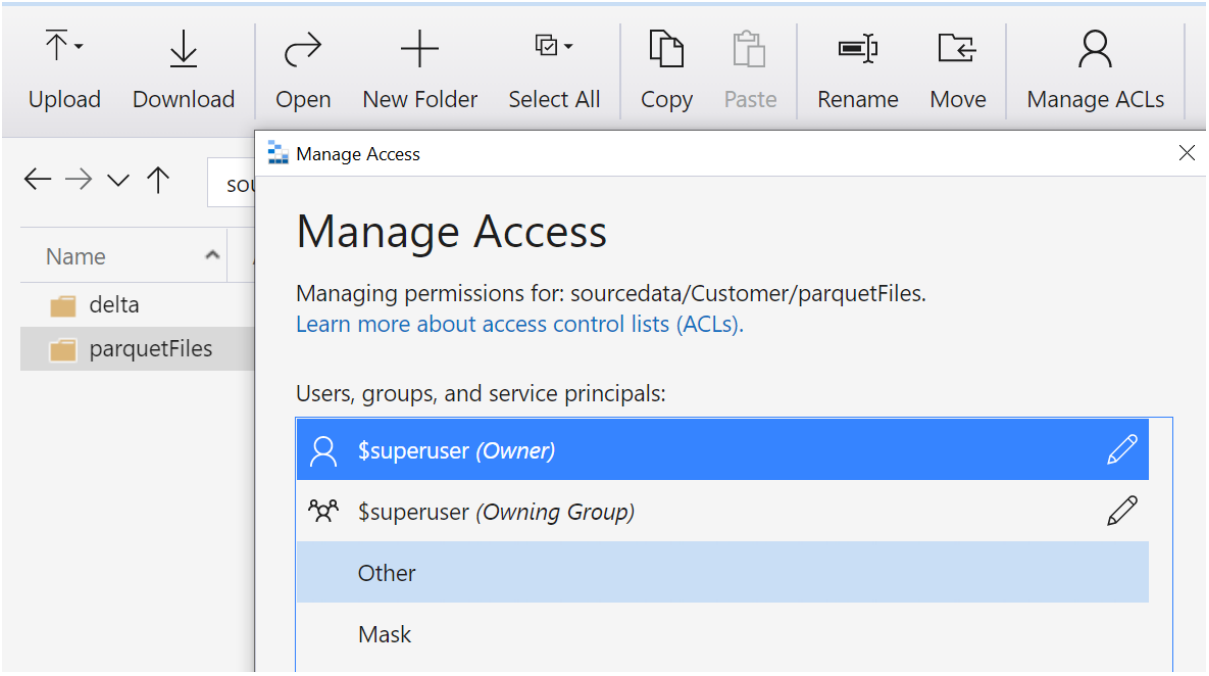
SQLUser sqluser@de... User Storage Blob Data Re...

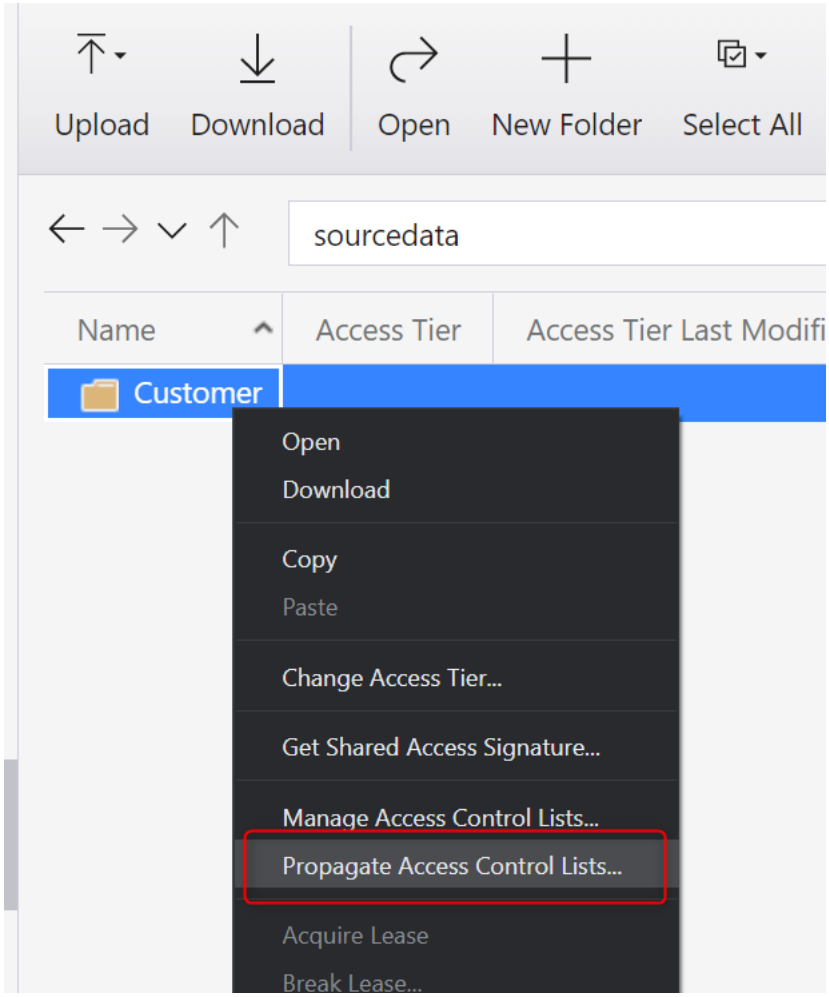
Storage Blob Data Reader

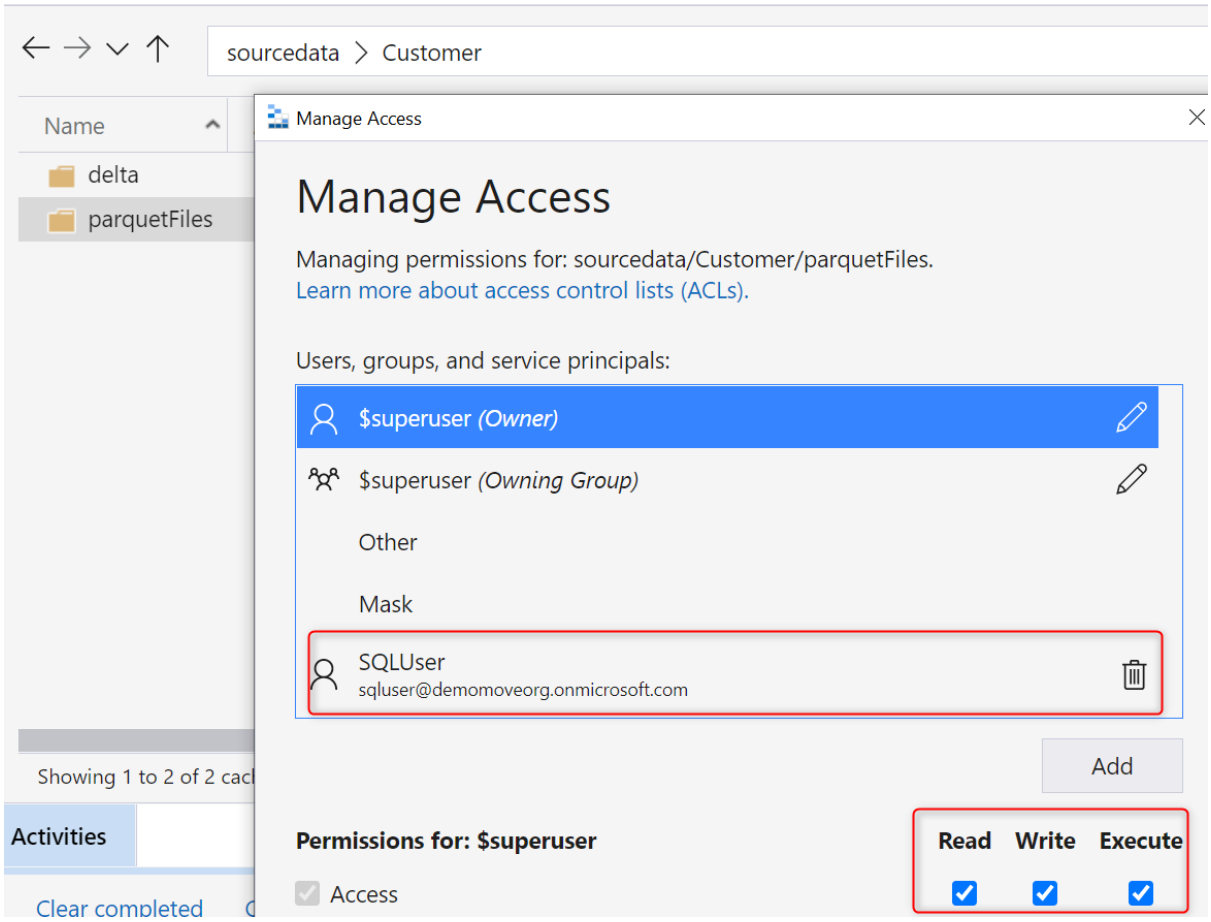


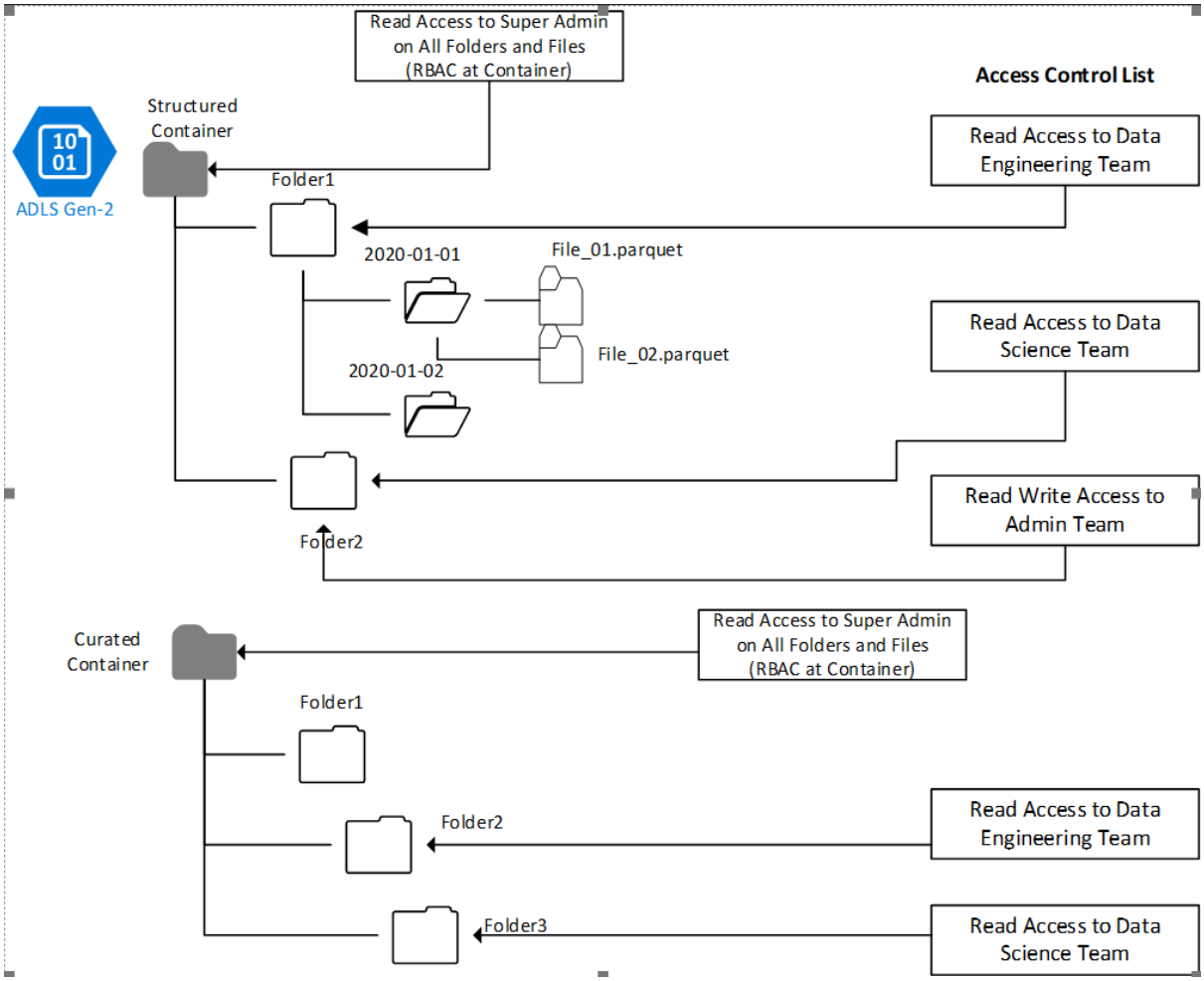


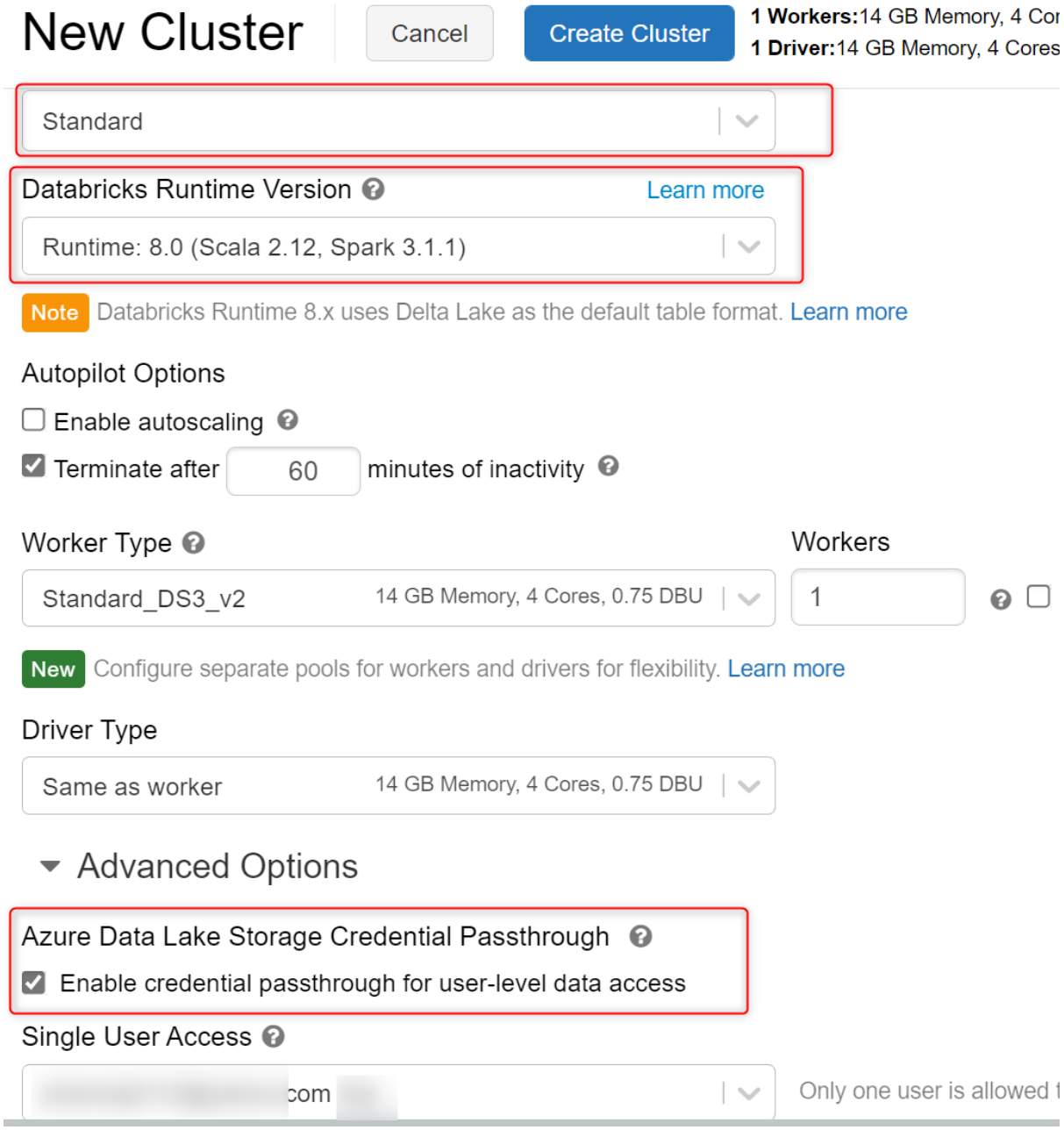

Manage Access
Manage Access
Managing permissions for: sourcedata/Customer.
Learn more about access control lists (ACLs).
Users, groups, and service principals:
$superuser (Owner)
$superuser (Owning Group)
Other
Mask
SQLUser
sqluser@demomoveorg.onmicrosoft.com
Add
Permissions for: SQLUser
Read
Write
Execute
Access
Default *
Read and write permissions will only work for an entity if the entity also has execute permissions on all parent directories, including the container (root directory).
* The default ACL determines permissions for new children of this directory. Changing the default ACL does not affect children that already exist. Learn more about default ACLs.
OK
Cancel

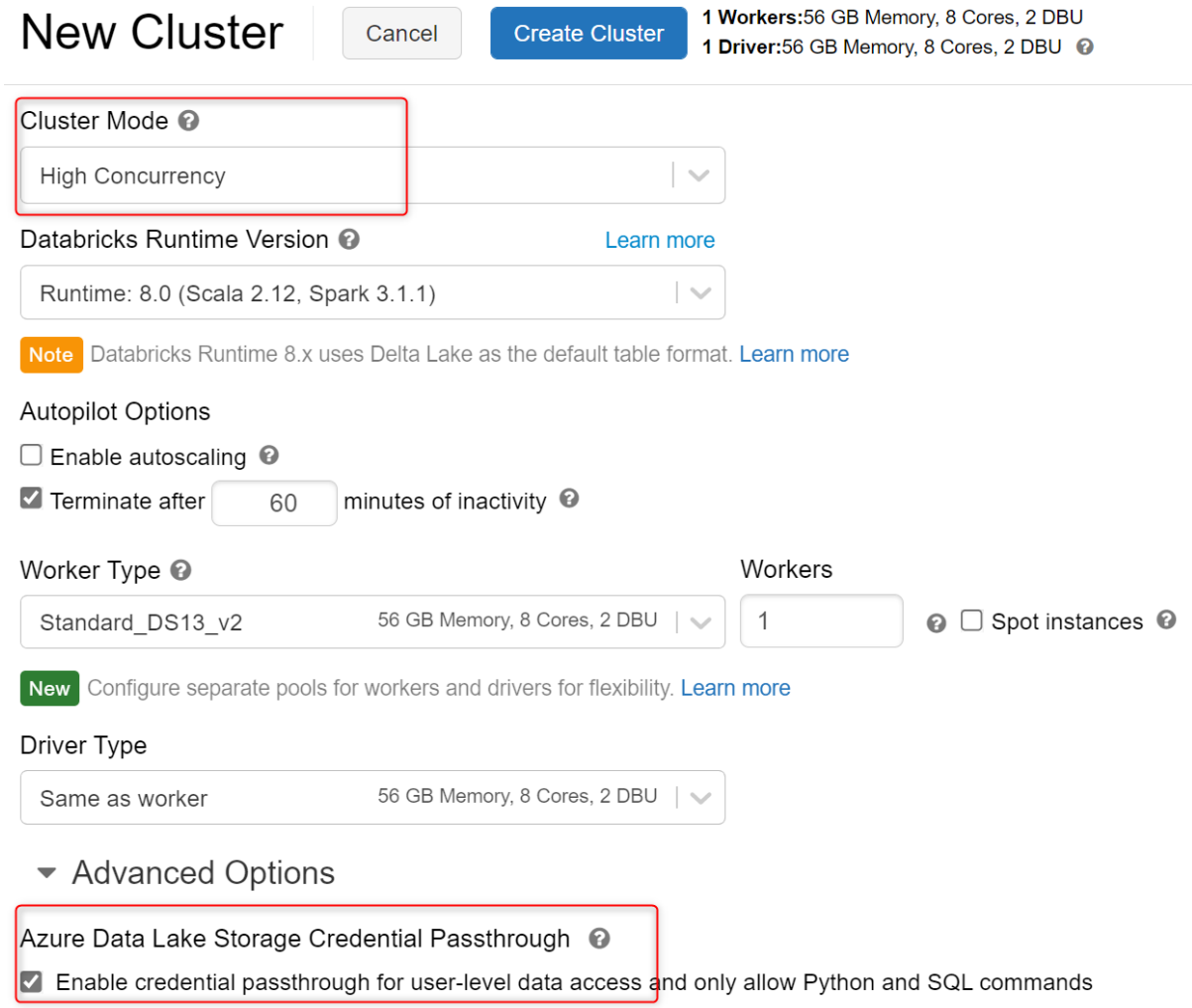

Upload
Download
Open
New Folder
Select All
Copy
Paste
Rename
Move
Manage ACLs
Manage Access
Manage Access
Managing permissions for: sourcedata/Customer/parquetFiles.
Learn more about access control lists (ACLs).
Users, groups, and service principals:
$superuser (Owner)
$superuser (Owning Group)
Other
Mask
Name
delta
parquetFiles

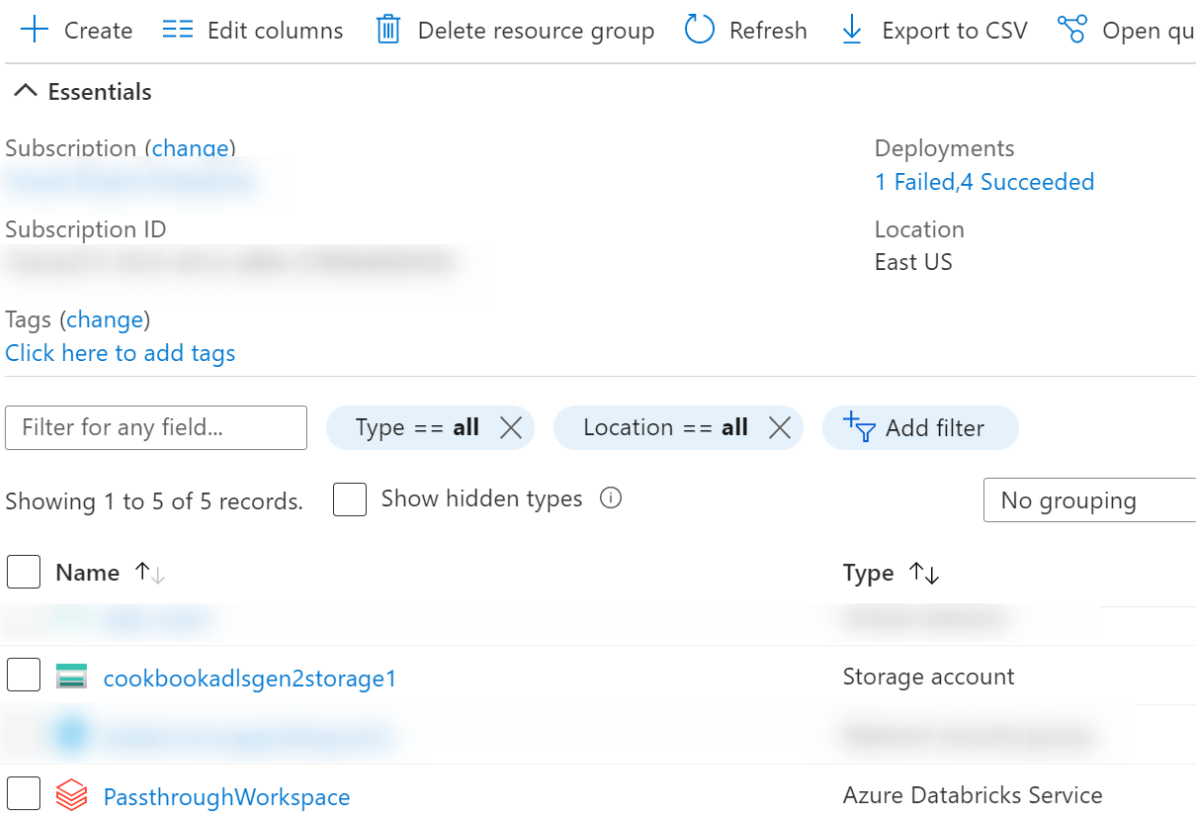

Upload
Download
Open
New Folder
Select All
sourcedata
Name
Access Tier
Access Tier Last Modifi
Customer
Open
Download
Copy
Paste
Change Access Tier...
Get Shared Access Signature...
Manage Access Control Lists...
Propagate Access Control Lists...
Acquire Lease
Break Lease...

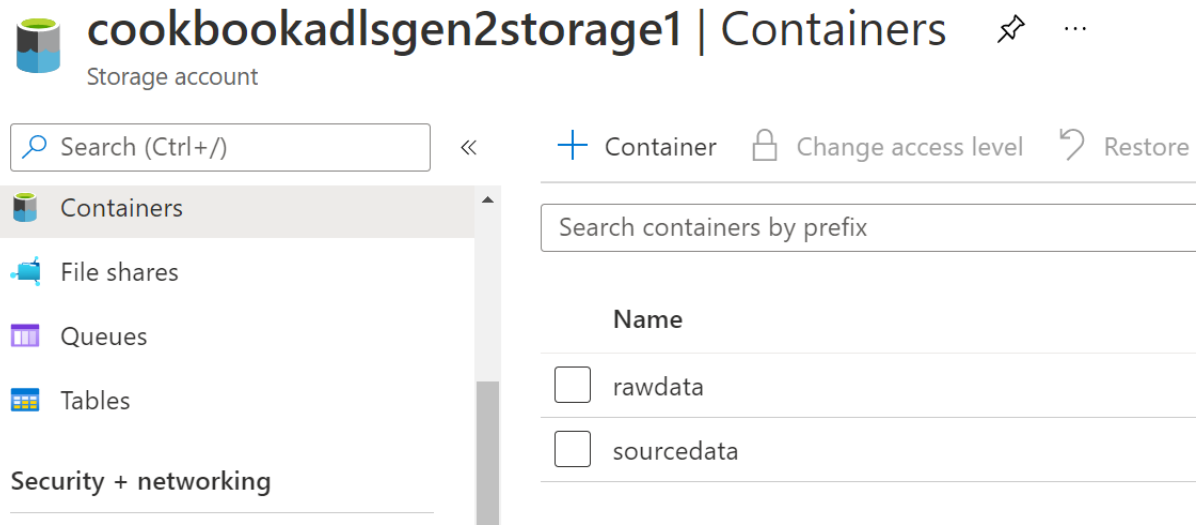

sourcedata > Customer
Name
delta
parquetFiles
Showing 1 to 2 of 2 cach
Activities
Clear completed
Manage Access
Manage Access
Managing permissions for: sourcedata/Customer/parquetFiles.
Learn more about access control lists (ACLs).
Users, groups, and service principals:
$superuser (Owner)
$superuser (Owning Group)
Other
Mask
SQLUser
sqluser@demomoveorg.onmicrosoft.com
Add
Permissions for: $superuser
Read
Write
Execute
Access

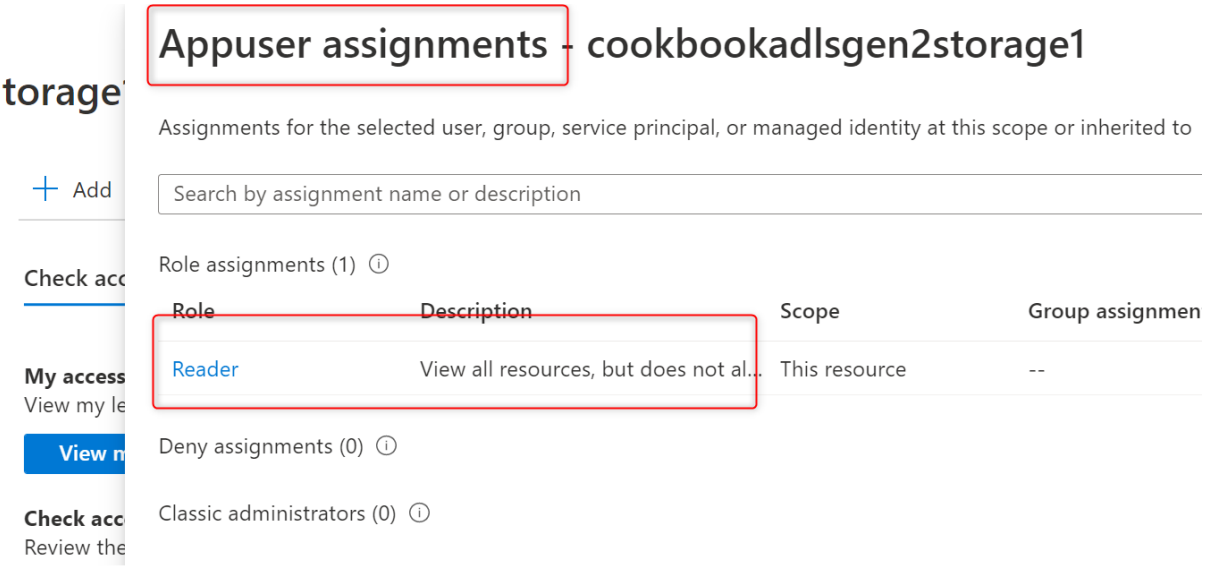

Read Access to Super Admin on All Folders and Files (RBAC at Container)
Access Control List
Structured Container
ADLS Gen-2
Read Access to Data Engineering Team
Folder1
2020-01-01
File_01.parquet
Read Access to Data Science Team
2020-01-02
File_02.parquet
Read Write Access to Admin Team
Folder2
Curated Container
Read Access to Super Admin on All Folders and Files (RBAC at Container)
Folder1
Folder2
Read Access to Data Engineering Team
Folder3
Read Access to Data Science Team

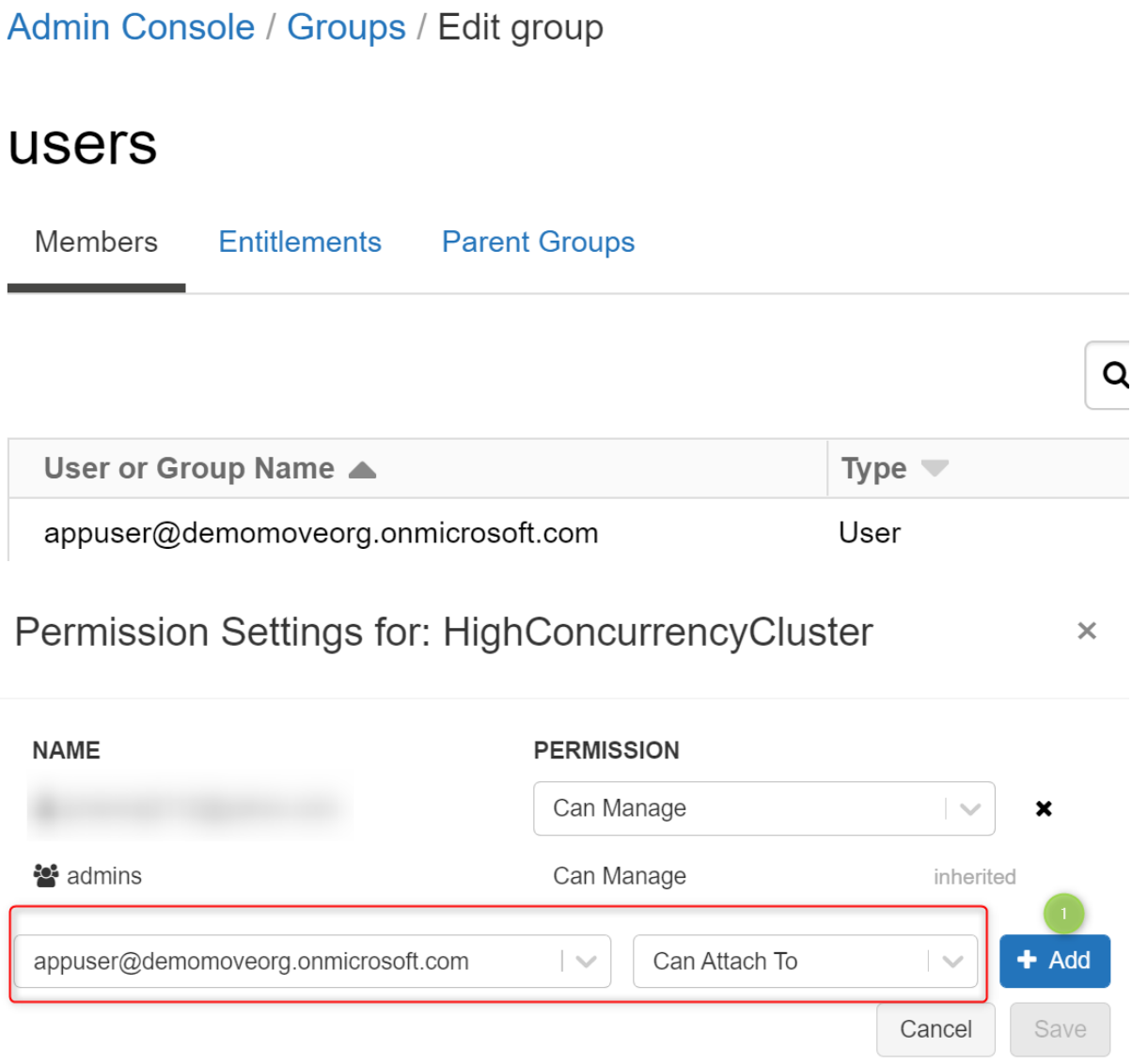

New Cluster
Cancel
Create Cluster
1 Workers:14 GB Memory, 4 Cor
1 Driver:14 GB Memory, 4 Cores
Standard
Databricks Runtime Version
Learn more
Runtime: 8.0 (Scala 2.12, Spark 3.1.1)
Note
Databricks Runtime 8.x uses Delta Lake as the default table format. Learn more
Autopilot Options
Enable autoscaling
Terminate after
60
minutes of inactivity
Worker Type
Workers
Standard_DS3_v2
14 GB Memory, 4 Cores, 0.75 DBU
1
New
Configure separate pools for workers and drivers for flexibility. Learn more
Driver Type
Same as worker
14 GB Memory, 4 Cores, 0.75 DBU
Advanced Options
Azure Data Lake Storage Credential Passthrough
Enable credential passthrough for user-level data access
Single User Access
com
Only one user is allowed t

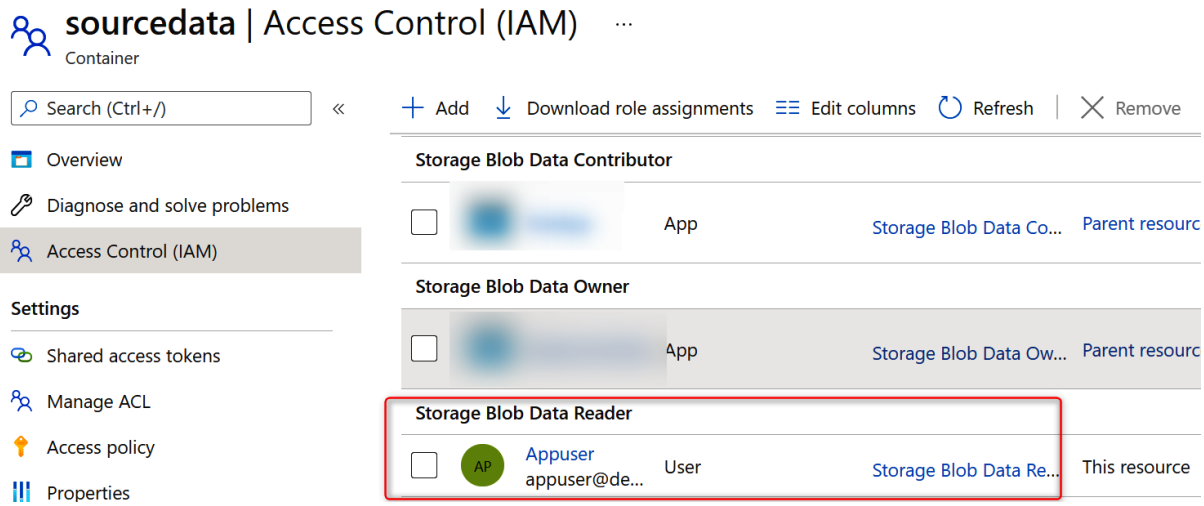

New Cluster
Cancel
Create Cluster
1 Workers:56 GB Memory, 8 Cores, 2 DBU
1 Driver:56 GB Memory, 8 Cores, 2 DBU
Cluster Mode
High Concurrency
Databricks Runtime Version
Learn more
Runtime: 8.0 (Scala 2.12, Spark 3.1.1)
Note
Databricks Runtime 8.x uses Delta Lake as the default table format. Learn more
Autopilot Options
Enable autoscaling
Terminate after
60
minutes of inactivity
Worker Type
Standard_DS13_v2
56 GB Memory, 8 Cores, 2 DBU
Workers
1
Spot instances
New
Configure separate pools for workers and drivers for flexibility. Learn more
Driver Type
Same as worker
56 GB Memory, 8 Cores, 2 DBU
Advanced Options
Azure Data Lake Storage Credential Passthrough
Enable credential passthrough for user-level data access and only allow Python and SQL commands


Create
Edit columns
Delete resource group
Refresh
Export to CSV
Open qu
Essentials
Subscription (change)
Deployments
1 Failed,4 Succeeded
Subscription ID
Location
East US
Tags (change)
Click here to add tags
Filter for any field...
Type == all
Location == all
Add filter
Showing 1 to 5 of 5 records.
Show hidden types
No grouping
Name
Type
cookbookadlsgen2storage1
Storage account
PassthroughWorkspace
Azure Databricks Service


cookbookadlsgen2storage1 | Containers
Storage account
Search (Ctrl+/)
Containers
File shares
Queues
Tables
Security + networking
Container
Change access level
Restore
Search containers by prefix
Name
rawdata
sourcedata


Appuser assignments - cookbookadlsgen2storage1
Assignments for the selected user, group, service principal, or managed identity at this scope or inherited to
Add
Search by assignment name or description
Role assignments (1)
Role
Description
Scope
Group assignmen
Reader
View all resources, but does not al...
This resource
--
Deny assignments (0)
Classic administrators (0)


Admin Console / Groups / Edit group
users
Members
Entitlements
Parent Groups
User or Group Name
Type
appuser@demomoveorg.onmicrosoft.com
User
Permission Settings for: HighConcurrencyCluster
NAME
PERMISSION
Can Manage
admins
Can Manage
inherited
appuser@demomoveorg.onmicrosoft.com
Can Attach To
1
Add
Cancel
Save



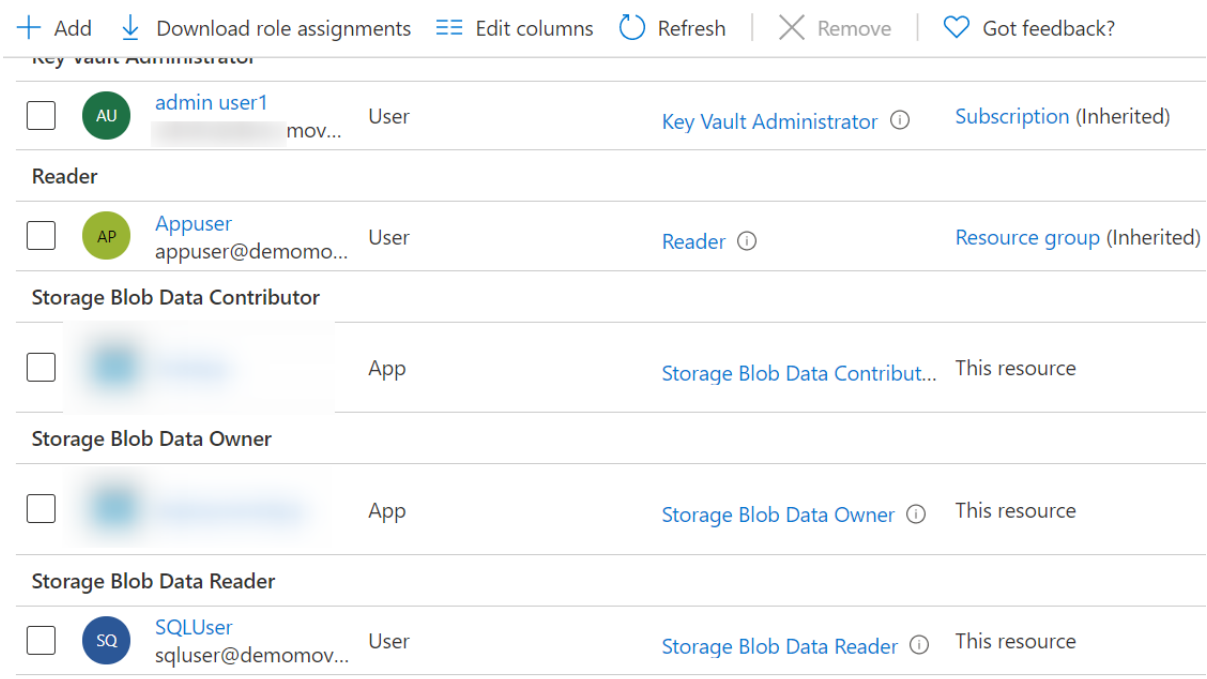

```
1  #Reading the files from ADLS Gen-2 using AAD authentication of the user.
2  display(dbutils.fs.ls("abfss://sourcedata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles"))
```

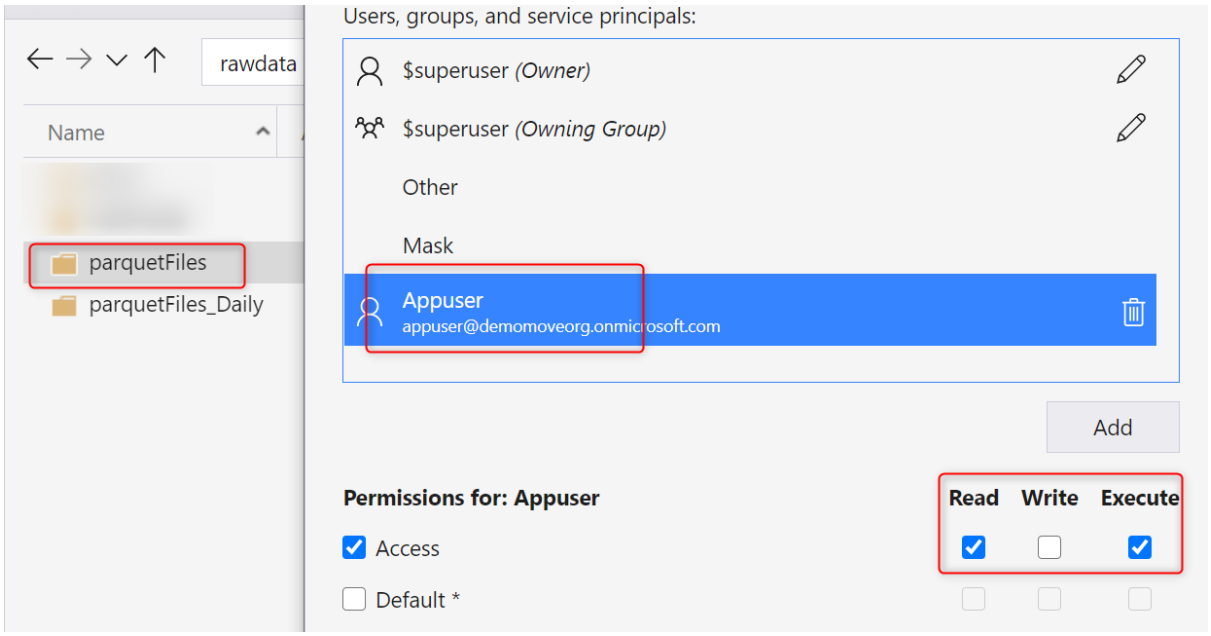
▸ (3) Spark Jobs

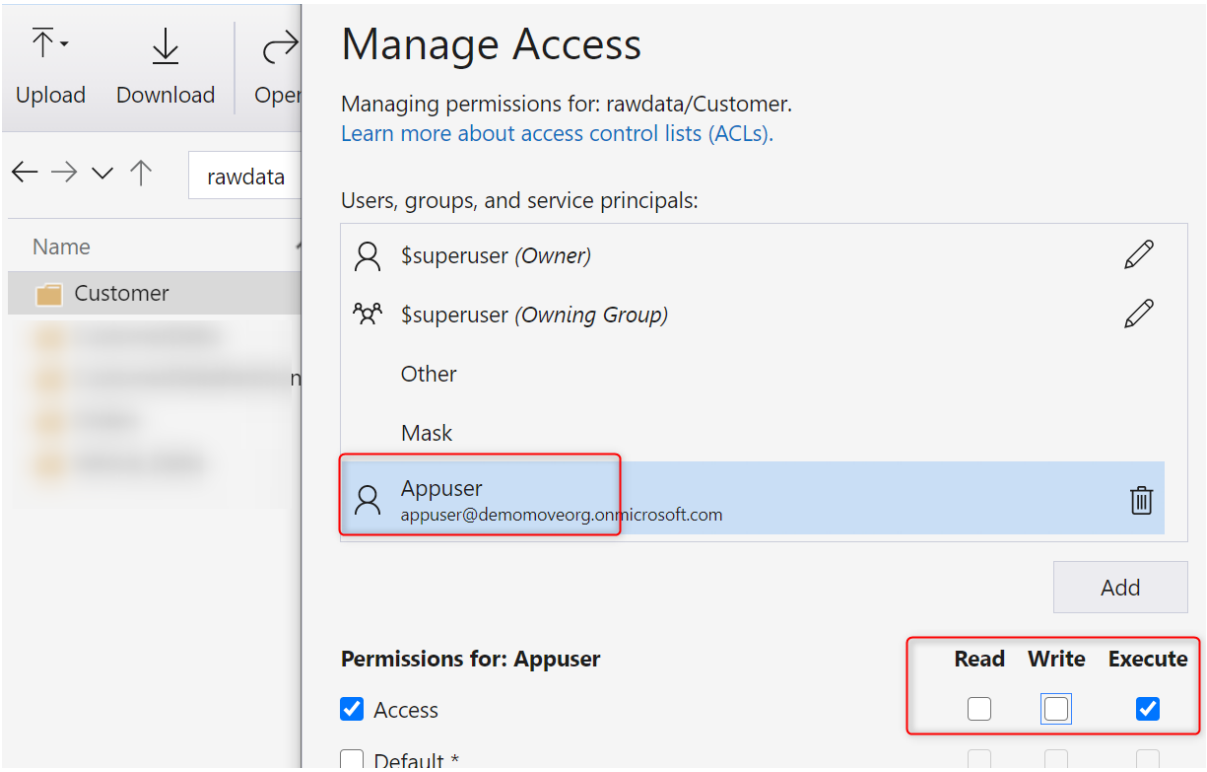
	path	name
1	abfss://sourcedata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00000-tid-7877307066025976411-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-56-1-c000.snappy.parquet	part-00000-tid-7877
2	abfss://sourcedata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00001-tid-7877307066025976411-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-57-1-c000.snappy.parquet	part-00001-tid-7877
3	abfss://sourcedata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00002-tid-7877307066025976411-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-58-1-c000.snappy.parquet	part-00002-tid-7877
4	abfss://sourcedata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00003-tid-7877307066025976411-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-59-1-c000.snappy.parquet	part-00003-tid-7877

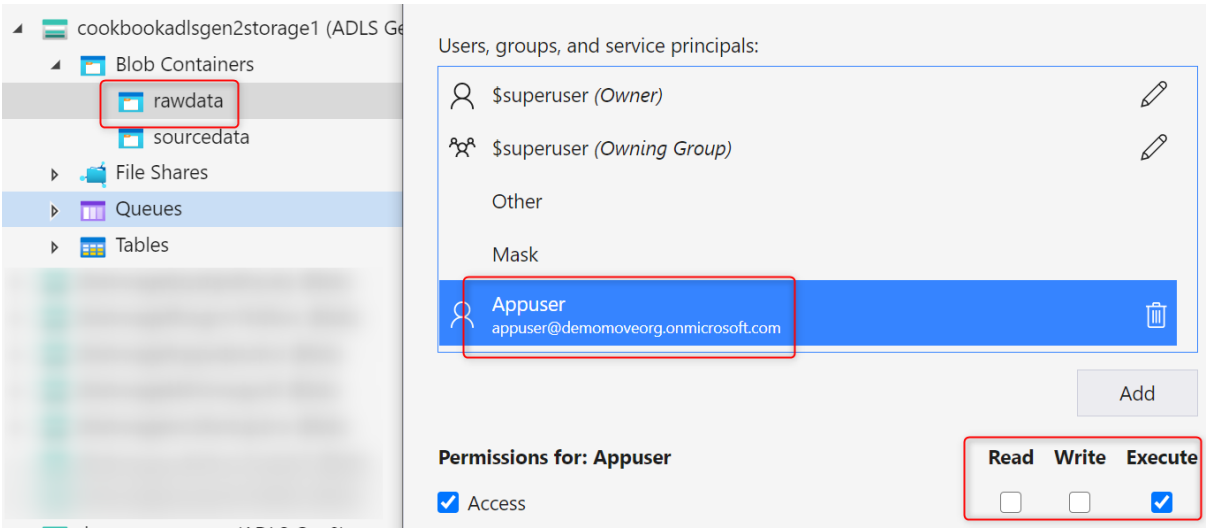
```
Operation failed: &#34;This request is not authorized to perform this operation using this permission.&#34;, 403, GET, https://cookbookadlsgen2storage1.dfs.core.windows.net/rawdata?upn=false&amp;resource=filesystem&amp;maxResults=500&amp;timeout=90&amp;recursive=false, AuthorizationPermissionMismatch, &#34;This request is not authorized to perform this operation using this permission. RequestId:3bc110f1-001f-008c-32a1-6c2309000000 Time:2021-06-29T04:45:08.9110582Z&#34;
Command took 0.68 seconds -- by appuser@demomoveorg.onmicrosoft.com at 6/29/2021,
```

```
Caused by: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 15.0 failed 4 times, most recent failure
Lost task 0.3 in stage 15.0 (TID 26) (10.139.64.5 executor 0): Operation failed: "This request is not authorized to perform this operation using this permission.", 403, PUT, https://cookbookadlsgen2storage1.dfs.core.windows.net/sourcedata/Customer/delta/part-00000-54a93d8-1f28-48b2-8a33-a5b9282b3863-c000.snappy.parquet?resource=file&timeout=90, AuthorizationPermissionMismatch, "This request is not authorized to perform this operation using this permission. RequestId:23f003f7-601f-0033-462b-8914ac000000 Time:2021-08-04T12:22:50.6574974Z"
```


Add
Download role assignments
Edit columns
Refresh
Remove
Got feedback?
admin user1
User
Key Vault Administrator
Subscription (Inherited)
Reader
Appuser
appuser@demomo...
User
Reader
Resource group (Inherited)
Storage Blob Data Contributor
App
Storage Blob Data Contribut...
This resource
Storage Blob Data Owner
App
Storage Blob Data Owner
This resource
Storage Blob Data Reader
SQLUser
sqluser@demomov...
User
Storage Blob Data Reader
This resource


rawdata
Name
parquetFiles
parquetFiles_Daily
Users, groups, and service principals:
$superuser (Owner)
$superuser (Owning Group)
Other
Mask
Appuser
appuser@demomoveorg.onmicrosoft.com
Add
Permissions for: Appuser
Read
Write
Execute
Access
Default *



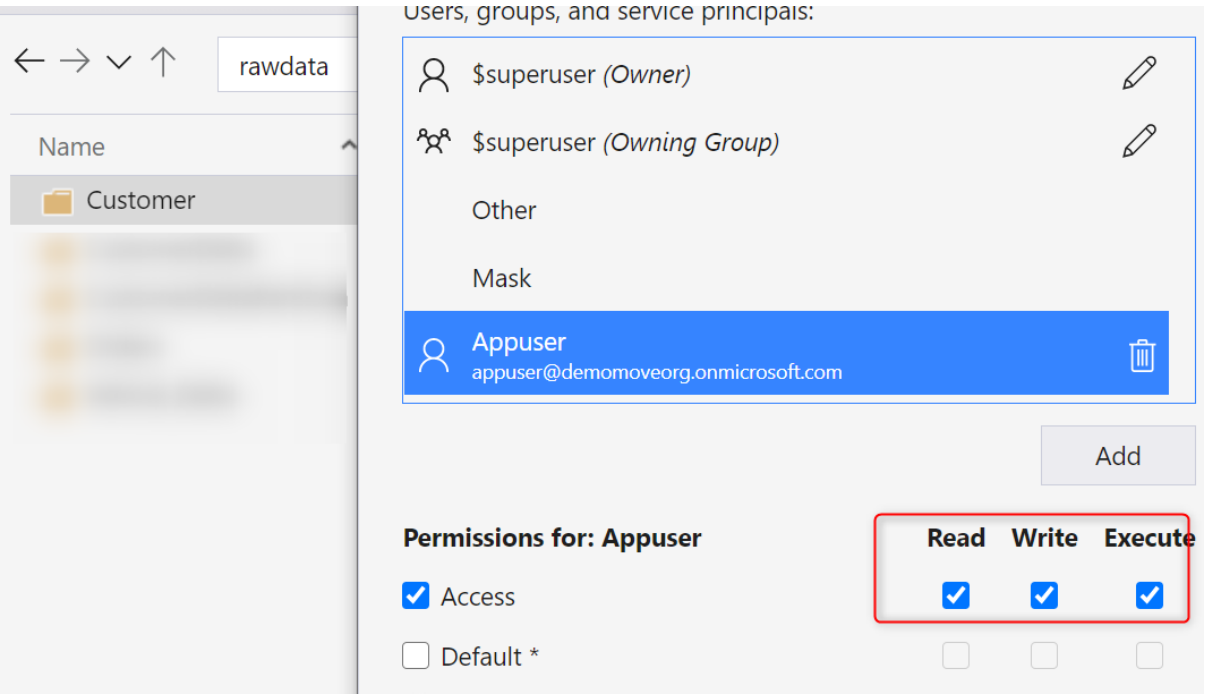



```
#Reading the files from ADLS Gen-2 using AAD authentication of the user with ACLs granted to folders
display(dbutils.fs.ls("abfss://rawdata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles"))
```

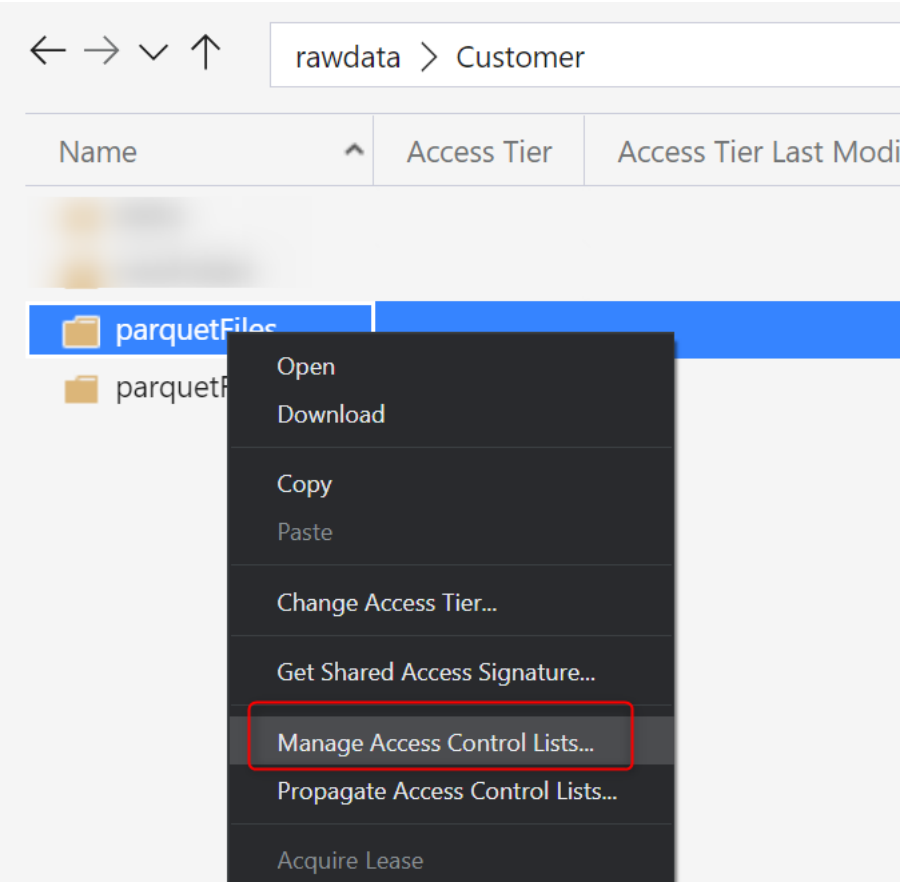
▸ (2) Spark Jobs

	path	name
1	abfss://rawdata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00000-tid-787730706602597641l-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-56-1-c000.snappy.parquet	part-00000-tid-7877307(
2	abfss://rawdata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00001-tid-787730706602597641l-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-57-1-c000.snappy.parquet	part-00001-tid-7877307(
3	abfss://rawdata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00002-tid-787730706602597641l-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-58-1-c000.snappy.parquet	part-00002-tid-7877307(
4	abfss://rawdata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00003-tid-787730706602597641l-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-59-1-c000.snappy.parquet	part-00003-tid-7877307(
5	abfss://rawdata@cookbookadlsgen2storage1.dfs.core.windows.net/Customer/parquetFiles/part-00004-tid-787730706602597641l-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-60-1-c000.snappy.parquet	part-00004-tid-7877307(

Operation failed: "This request is not authorized to perform this operation using this permission.", 403, GET, https://cookbookadlsgen2storage1.dfs.core.windows.net/rawdata?upn=false&resource=filesystem&maxResults=500&directory=Customer/parquetFiles_Daily&timeout=90&recursive=false, AuthorizationPermissionMismatch, "This request is not authorized to perform this operation using this permission. RequestId:6afba26f-801f-0066-29a8-6c0427000000 Time:2021-06-29T05:35:43.8238029Z"



Caused by: Operation failed: "This request is not authorized to perform this operation using this permission.", 403, GET, https://cookbookadlsgen2storage1.dfs.core.windows.net/rawdata/Customer/parquetFiles/part-00000-tid-7877307066025976411-3a62bd06-bc9c-467f-ae7b-d5db8ca40833-56-1-c000.snappy.parquet?timeout=90, AuthorizationPermissionMismatch, "This request is not authorized to perform this operation using this permission. RequestId:d5024788-701f-0096-55aa-6c42d6000000 Time:2021-06-29T05:51:03.3595431Z"



+ Subnet + Gateway subnet Refresh | Manage users

Search subnets

Name	IPv4	IPv6
default	10.0.0.0/24	-
private-subnet	10.0.2.0/24	-
public-subnet	10.0.1.0/24	-

Create an Azure Databricks workspace ...

manage all your resources.

Subscription * [redacted]

Resource group * CoookbookStorageRG

Create new

Instance Details

Workspace name * vnet-adbworkspace

Region * East US

Pricing Tier * Standard (Apache Spark, Secure with Azure AD)

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) ○ Yes ◉ No

Deploy Azure Databricks workspace in your own Virtual Network (VNet) ◉ Yes ○ No

Virtual Network * adb-vnet1

Two new subnets will be created in your Virtual Network

Implicit delegation of both subnets will be done to Azure Databricks on your behalf

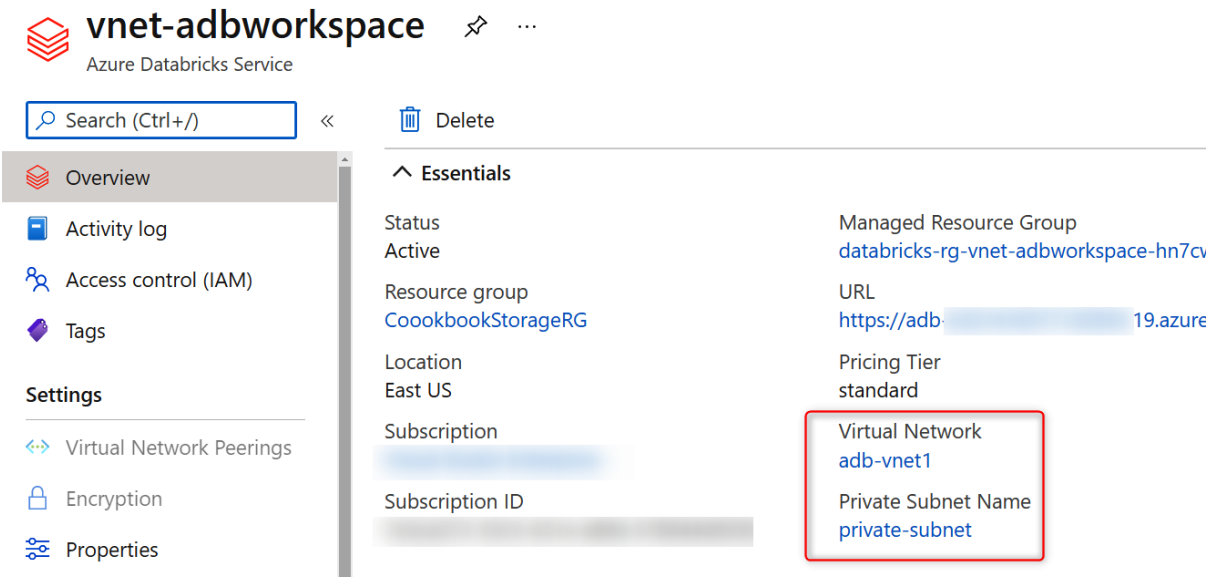
Public Subnet Name * public-subnet

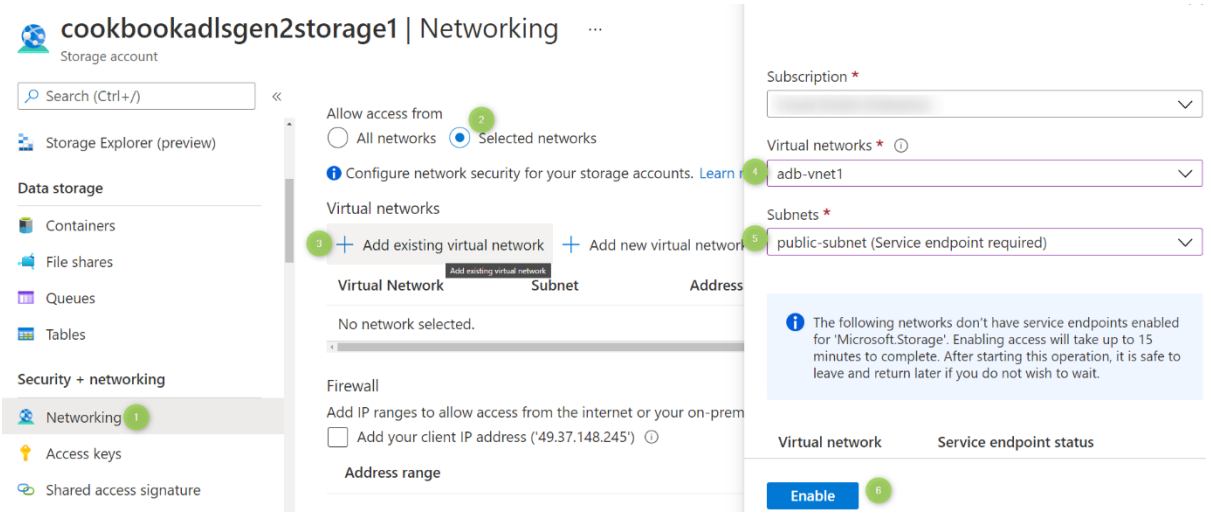
Public Subnet CIDR Range * 10.0.1.0/24

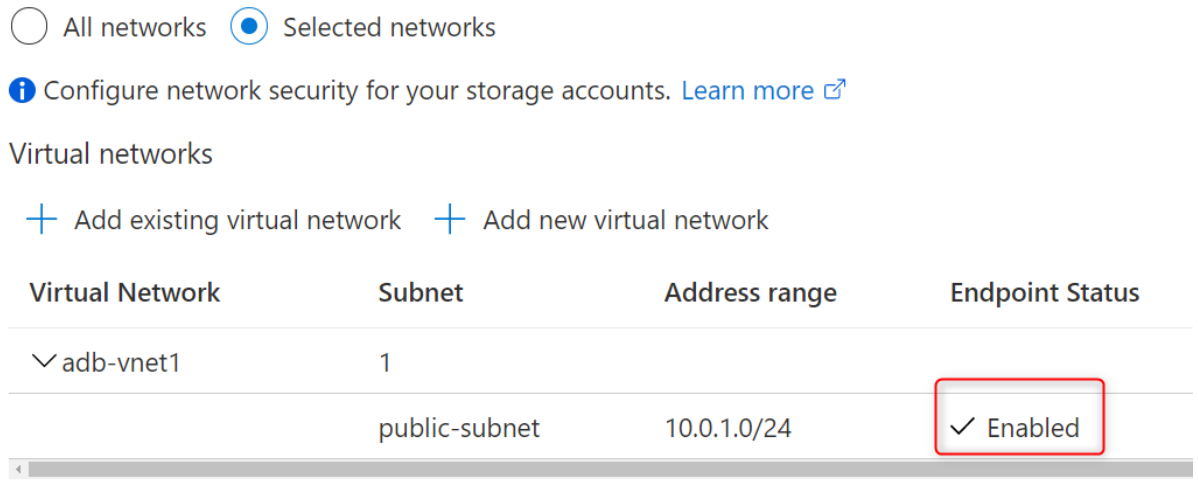
Private Subnet Name * private-subnet

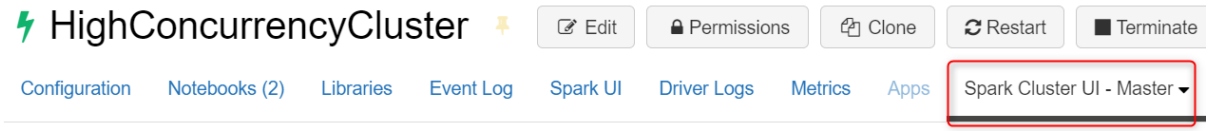
Private Subnet CIDR Range * 10.0.2.0/24

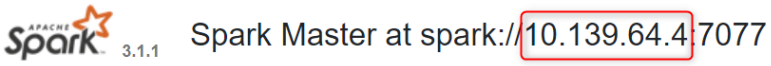
Review + create | < Previous | Next : Advanced >


vnet-adbworkspace
Azure Databricks Service
Search (Ctrl+/)
Delete
Overview
Activity log
Access control (IAM)
Tags
Settings
Virtual Network Peerings
Encryption
Properties
Essentials
Status
Active
Managed Resource Group
databricks-rg-vnet-adbworkspace-hn7cv
Resource group
CoookbookStorageRG
URL
https://adb
19.azure
Location
East US
Pricing Tier
standard
Subscription
Virtual Network
adb-vnet1
Subscription ID
Private Subnet Name
private-subnet


cookbookadlsgen2storage1 | Networking
Storage account
Search (Ctrl+/)
Storage Explorer (preview)
Data storage
Containers
File shares
Queues
Tables
Security + networking
Networking
Access keys
Shared access signature
Allow access from
All networks
Selected networks
Configure network security for your storage accounts. Learn
Virtual networks
Add existing virtual network
Add new virtual network
Virtual Network
Subnet
Address
No network selected.
Firewall
Add IP ranges to allow access from the internet or your on-prem
Add your client IP address ('49.37.148.245')
Address range
Subscription *
Virtual networks *
adb-vnet1
Subnets *
public-subnet (Service endpoint required)
The following networks don't have service endpoints enabled for 'Microsoft.Storage'. Enabling access will take up to 15 minutes to complete. After starting this operation, it is safe to leave and return later if you do not wish to wait.
Virtual network
Service endpoint status
Enable


All networks
Selected networks
Configure network security for your storage accounts. Learn more
Virtual networks
Add existing virtual network
Add new virtual network
Virtual Network
Subnet
Address range
Endpoint Status
adb-vnet1
1
public-subnet
10.0.1.0/24
Enabled

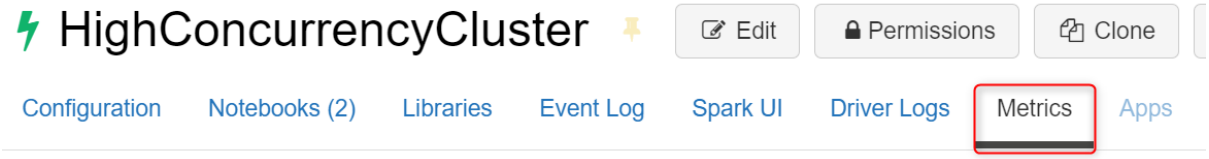




URL: spark://10.139.64.4:7077
Alive Workers: 2
Cores in use: 8 Total, 8 Used
Memory in use: 17.8 GiB Total, 14.2 GiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

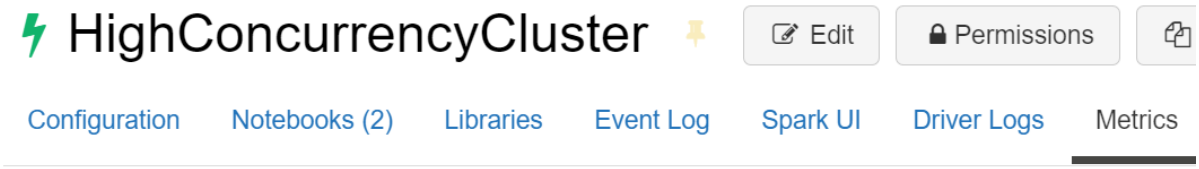
▾Workers (2)

Worker Id	Address	State	Cores
worker-20210805010536-10.139.64.6-35581	10.139.64.6:35581	ALIVE	4 (4 Used)
worker-20210805010540-10.139.64.5-41393	10.139.64.5:41393	ALIVE	4 (4 Used)



Live Metrics

Ganglia UI



Live Metrics

Ganglia UI

HighConcurrencyCluster

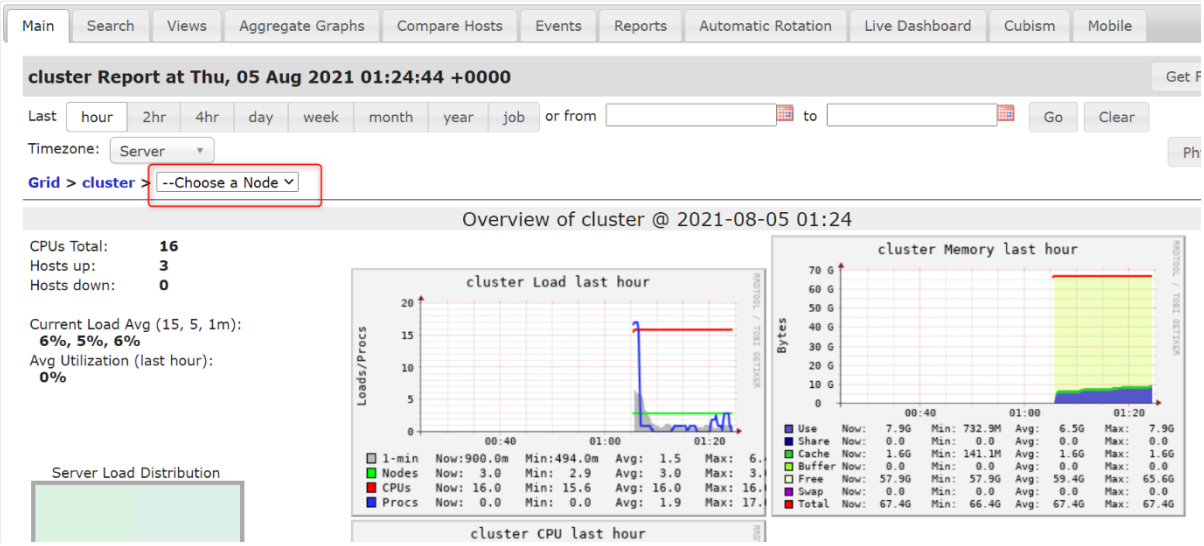
Edit | Permissions | Clone

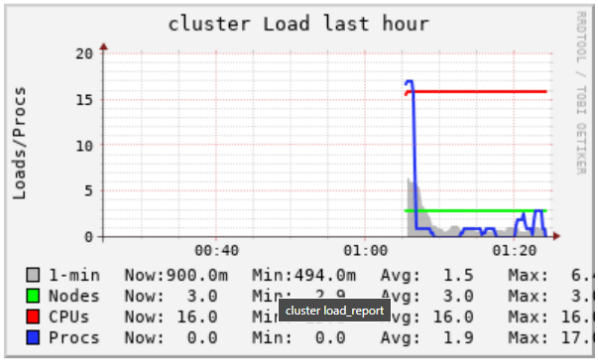
Configuration | Notebooks (2) | Libraries | Event Log | Spark UI | Driver Logs | Metrics | App

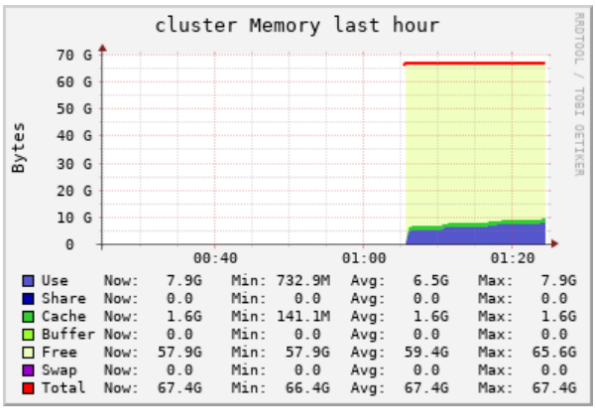
Ganglia UI

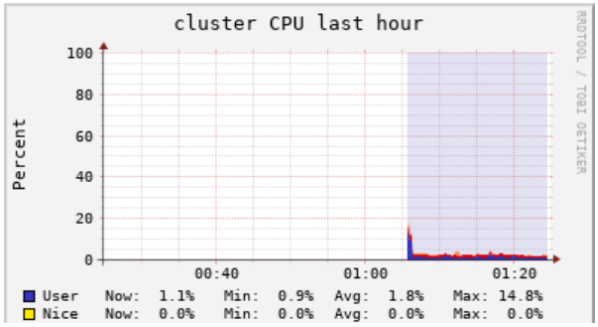
Historical Metrics Snapshots (5 files)

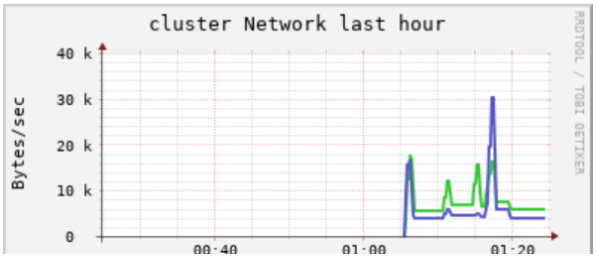
Name
2021-08-04 18:30:01 IST
2021-08-04 18:15:01 IST





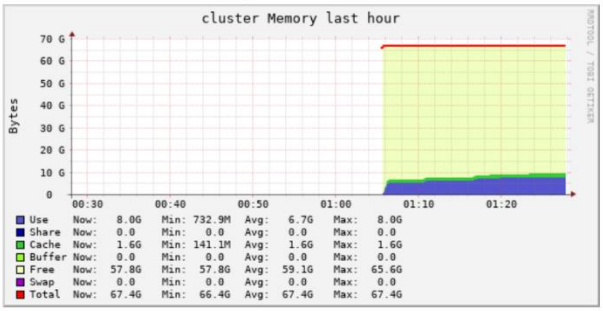




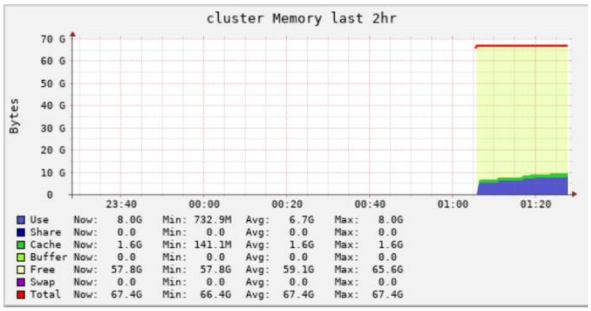


Cluster: cluster **Graph:** mem_report

hour + CSV JSON Inspect Hide/Show Events

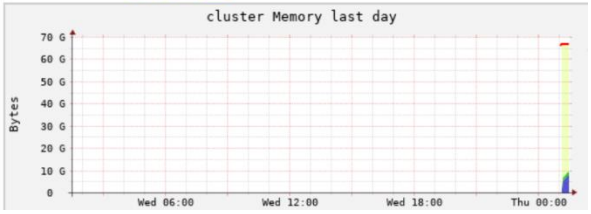


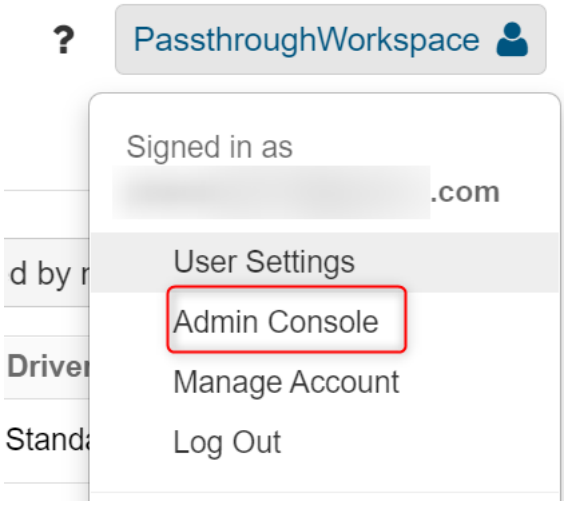
2hr + CSV JSON Inspect Hide/Show Events

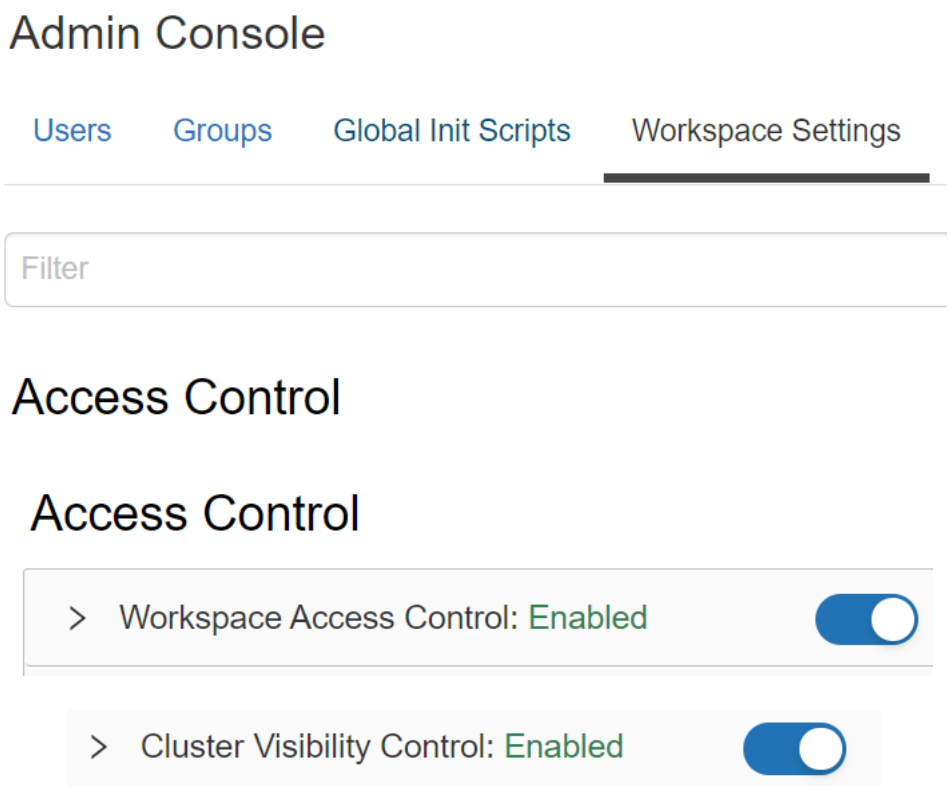


4hr + CSV JSON Inspect Hide/Show Events

day + CSV JSON Inspect Hide/Show Events




?
PassthroughWorkspace
Signed in as
.com
User Settings
Admin Console
Manage Account
Log Out


Admin Console
Users
Groups
Global Init Scripts
Workspace Settings
Filter
Access Control
Access Control
Workspace Access Control: Enabled
Cluster Visibility Control: Enabled
Who has access:
admins
Can Manage
.com)
Can Manage
No Permissions
Can Attach To
Can Restart
Can Manage
Add Users, Groups, and Service Principals:
Select User, Group or Service Principal...