Data Ingestion End-to-End Pipeline

Course-end Project 1

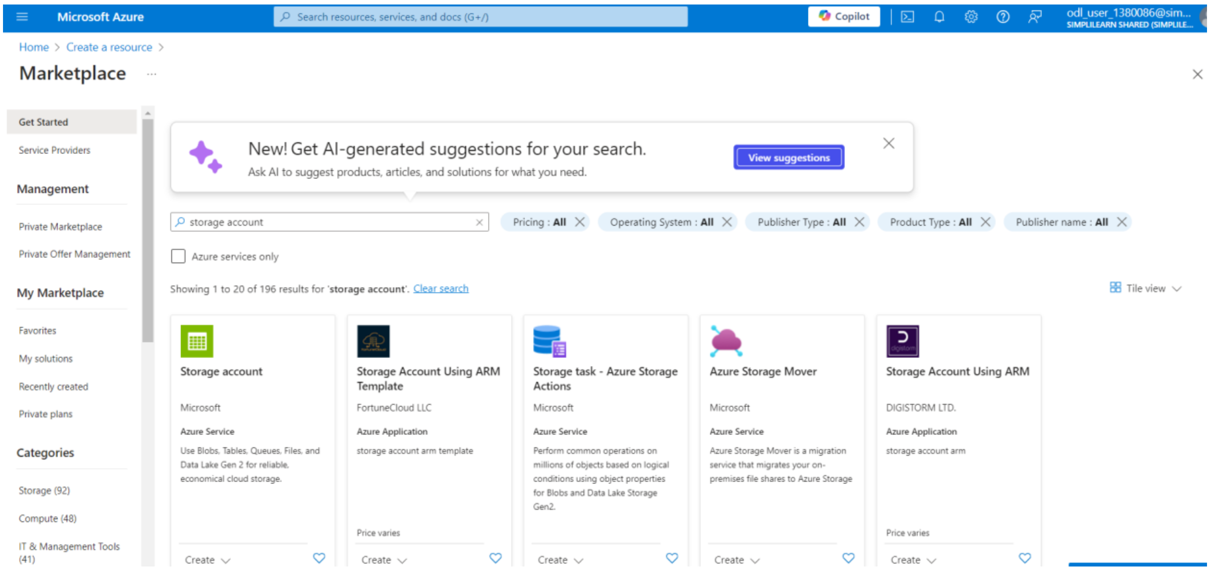
Description

Your company is looking for a data engineer and is inviting candidates to apply for this position by providing a portal where applicants can add their credentials.

As thousands of candidates have applied for this position, the company has a huge amount of data that it needs to upload to its website. This data is moved to Azure Data Lake Storage parallelly. The company wants to save the contents of all CSV files to Delta Lake of Azure Databricks so that these files can be retrieved and accessed from Azure Databricks when required.

Step 1: Create a Landing Storage Account in Azure

1. **Log in to the Azure Portal**.
2. **Create a Storage Account**:
   - Go to **"Create a resource"** and search for **"Storage account"**.
   - Click on **"Create"**.
   - Fill in the required details (Subscription, Resource group, Storage account name, Region, Performance, and Replication).
   - Click on **"Review + create"** and then **"Create"**.



New! Get AI-generated suggestions for your search.

Ask AI to suggest products, articles, and solutions for what you need.

storage account

□ Azure services only

Showing 1 to 20 of 196 results for 'storage account'. Clear search

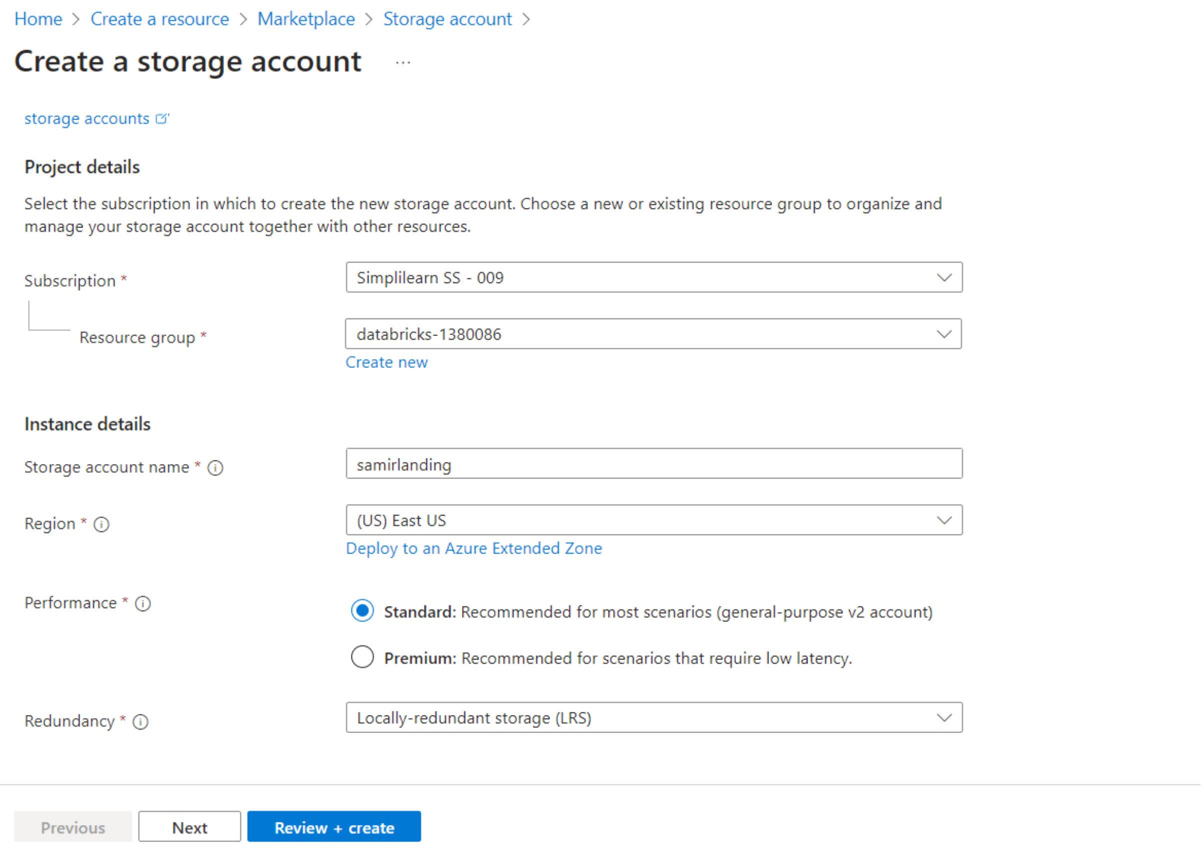
Storage account
Microsoft
Azure Service
Use Blobs, Tables, Queues, Files, and Data Lake Gen 2 for reliable, economical cloud storage.

Storage Account Using ARM Template
FortuneCloud LLC
Azure Application
storage account arm template
Price varies

Storage task - Azure Storage Actions
Microsoft
Azure Service
Perform common operations on millions of objects based on logical conditions using object properties for Blobs and Data Lake Storage Gen2.

Azure Storage Mover
Microsoft
Azure Service
Azure Storage Mover is a migration service that migrates your on-premises file shares to Azure Storage

Storage Account Using ARM
DIGISTORM LTD.
Azure Application
storage account arm
Price varies



Create a storage account ···

storage accounts

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * Simplilearn SS - 009

Resource group * databricks-1380086

Create new

Instance details

Storage account name * samirlanding

Region * (US) East US

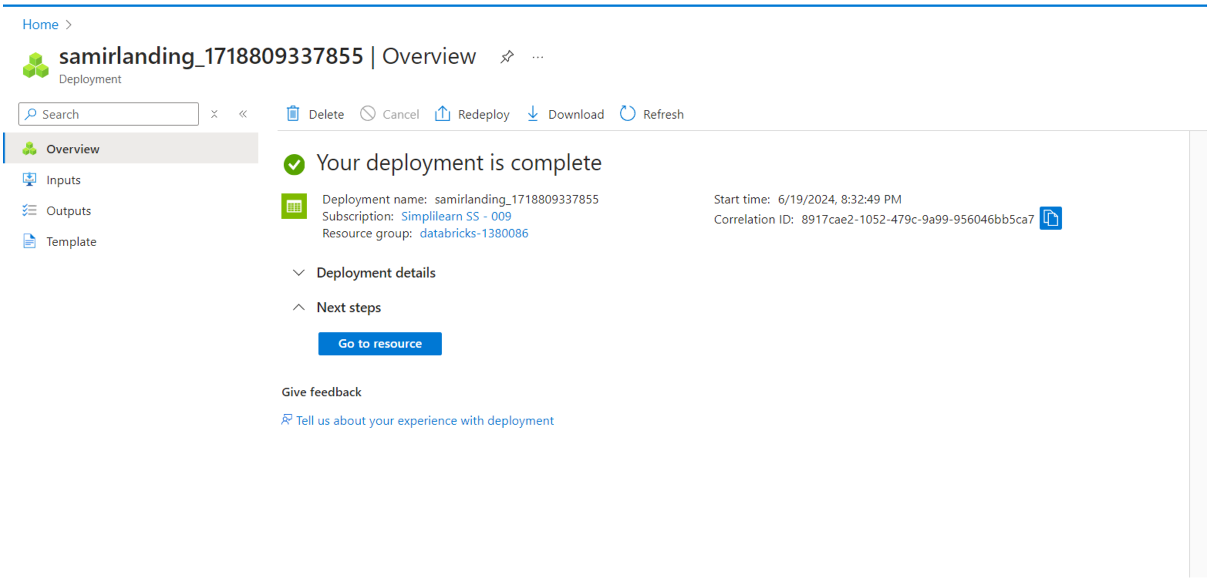
Deploy to an Azure Extended Zone

Performance *

● **Standard:** Recommended for most scenarios (general-purpose v2 account)

○ **Premium:** Recommended for scenarios that require low latency.

Redundancy * Locally-redundant storage (LRS)



samirlanding_1718809337855 | Overview

Deployment

Your deployment is complete

Deployment name: samirlanding_1718809337855
Subscription: Simplilearn SS - 009
Resource group: databricks-1380086

Start time: 6/19/2024, 8:32:49 PM
Correlation ID: 8917cae2-1052-479c-9a99-956046bb5ca7

Deployment details

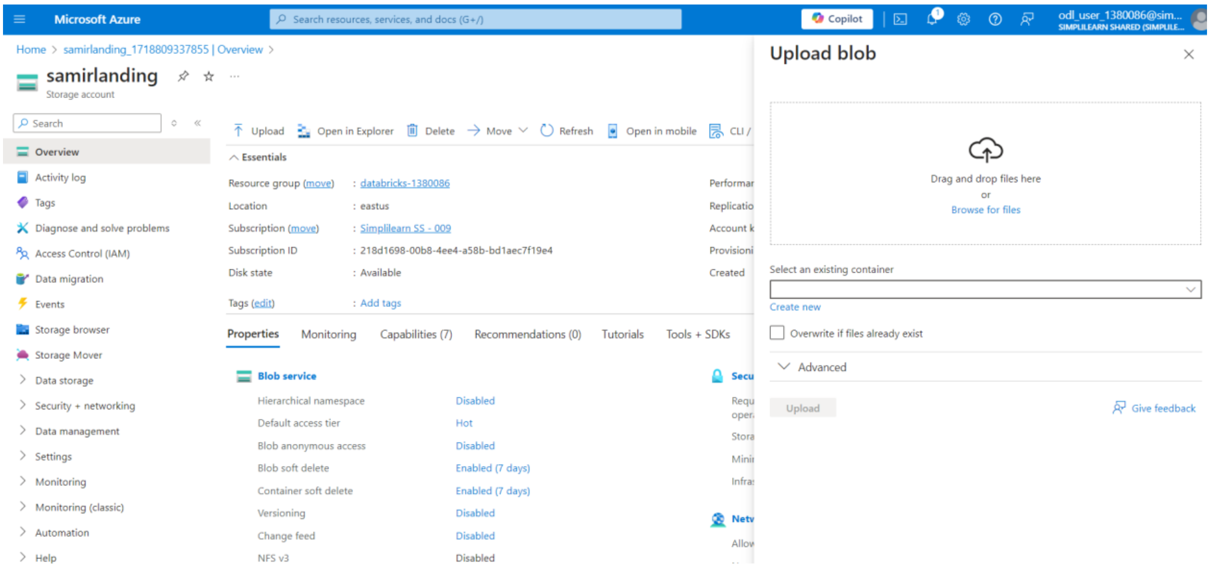
Next steps

Give feedback

Tell us about your experience with deployment

Step 2: Store the CSV Files in the Landing Storage Account

1. **Upload CSV Files**:
   - Navigate to the created storage account.
   - Go to **"Containers"** and create a new container.
   - Upload the CSV files to this container.



Microsoft Azure

Home > samirlanding_1718809337855 | Overview >

samirlanding
Storage account

Upload | Open in Explorer | Delete | Move | Refresh | Open in mobile | CLI /

Essentials

Resource group (move) : databricks-1380086
Location : eastus
Subscription (move) : Simplilearn SS - 009
Subscription ID : 218d1698-00b8-4ee4-a58b-bd1aec7f19e4
Disk state : Available

Tags (edit) : Add tags

Properties | Monitoring | Capabilities (7) | Recommendations (0) | Tutorials | Tools + SDKs

Blob service

Hierarchical namespace	Disabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)
Container soft delete	Enabled (7 days)
Versioning	Disabled
Change feed	Disabled
NFS v3	Disabled

Upload blob

Drag and drop files here
or
Browse for files

Select an existing container

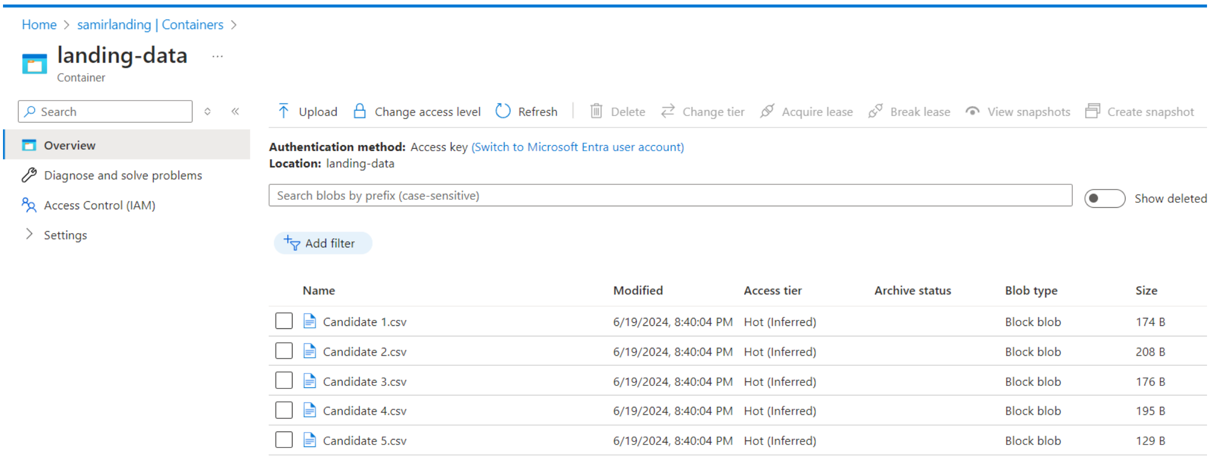
Create new

□ Overwrite if files already exist

Advanced

Upload

Give feedback



Home > samirlanding | Containers >

landing-data
Container

Upload | Change access level | Refresh | Delete | Change tier | Acquire lease | Break lease | View snapshots | Create snapshot

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: landing-data

Search blobs by prefix (case-sensitive)

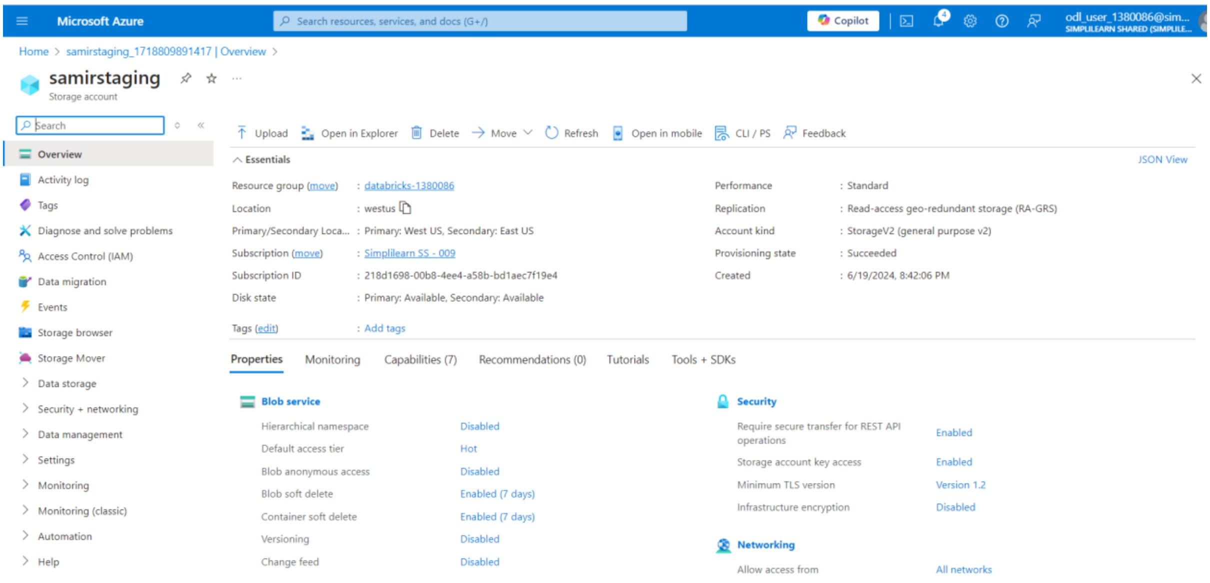
Show deleted

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size
Candidate 1.csv	6/19/2024, 8:40:04 PM	Hot (Inferred)		Block blob	174 B
Candidate 2.csv	6/19/2024, 8:40:04 PM	Hot (Inferred)		Block blob	208 B
Candidate 3.csv	6/19/2024, 8:40:04 PM	Hot (Inferred)		Block blob	176 B
Candidate 4.csv	6/19/2024, 8:40:04 PM	Hot (Inferred)		Block blob	195 B
Candidate 5.csv	6/19/2024, 8:40:04 PM	Hot (Inferred)		Block blob	129 B

Step 3: Create a Staging Storage Account in Azure

Repeat the steps from Step 1 to create another storage account that will be used for staging.



Essentials

Resource group (move) : databricks-1380086
Location : westus
Primary/Secondary Loca... : Primary: West US, Secondary: East US
Subscription (move) : Simplilearn SS - 009
Subscription ID : 218d1698-00b8-4ee4-a58b-bd1aec7f19e4
Disk state : Primary: Available, Secondary: Available
Tags (edit) : Add tags

Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 6/19/2024, 8:42:06 PM

Blob service

Hierarchical namespace	Disabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)
Container soft delete	Enabled (7 days)
Versioning	Disabled
Change feed	Disabled

Security

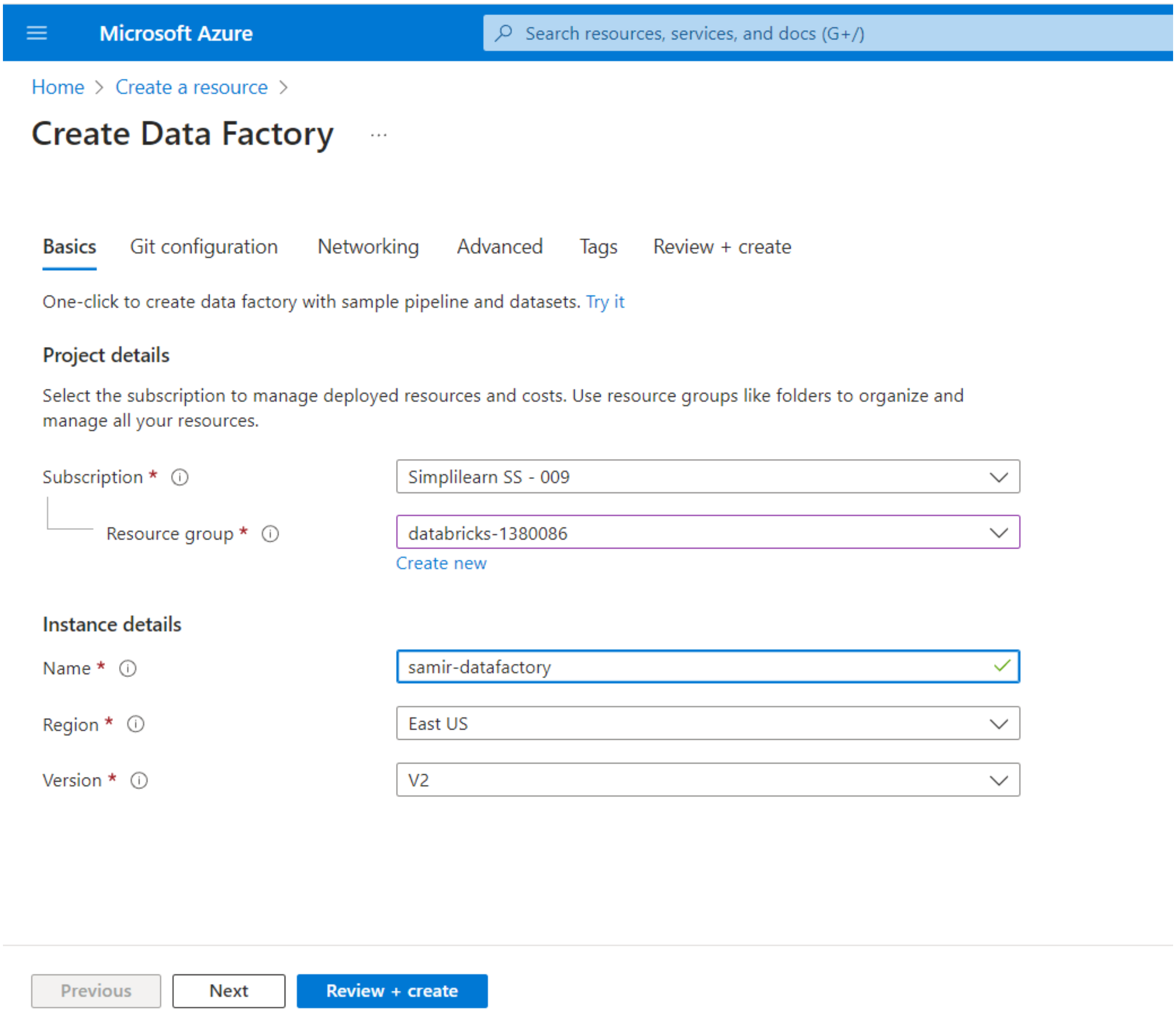
Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled

Networking

Allow access from	All networks

Step 4: Create an Azure Data Factory Resource

1. **Create Azure Data Factory**:
   - Go to **"Create a resource"** and search for **"Data Factory"**.
   - Click on **"Create"**.
   - Fill in the required details (Subscription, Resource group, Data Factory name, Version, and Region).
   - Click on **"Review + create"** and then **"Create"**.



Microsoft Azure

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

Home > Create a resource >

Create Data Factory ...

Basics | Git configuration | Networking | Advanced | Tags | Review + create

One-click to create data factory with sample pipeline and datasets. Try it

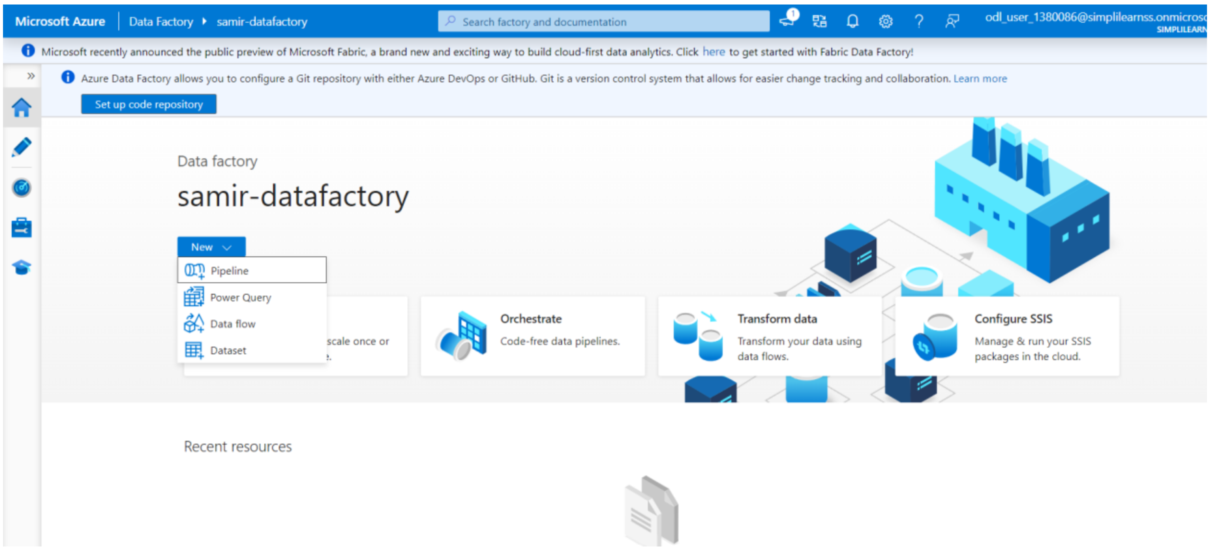
Project details

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

Subscription *	Simplilearn SS - 009
Resource group *	databricks-1380086
	Create new

Instance details

Name *	samir-datafactory
Region *	East US
Version *	V2



Microsoft Azure | Data Factory ▸ samir-datafactory

Search factory and documentation

odl_user_1380086@simplilearnss.onmicros
SIMPLILEARN

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration. Learn more

Set up code repository

Data factory

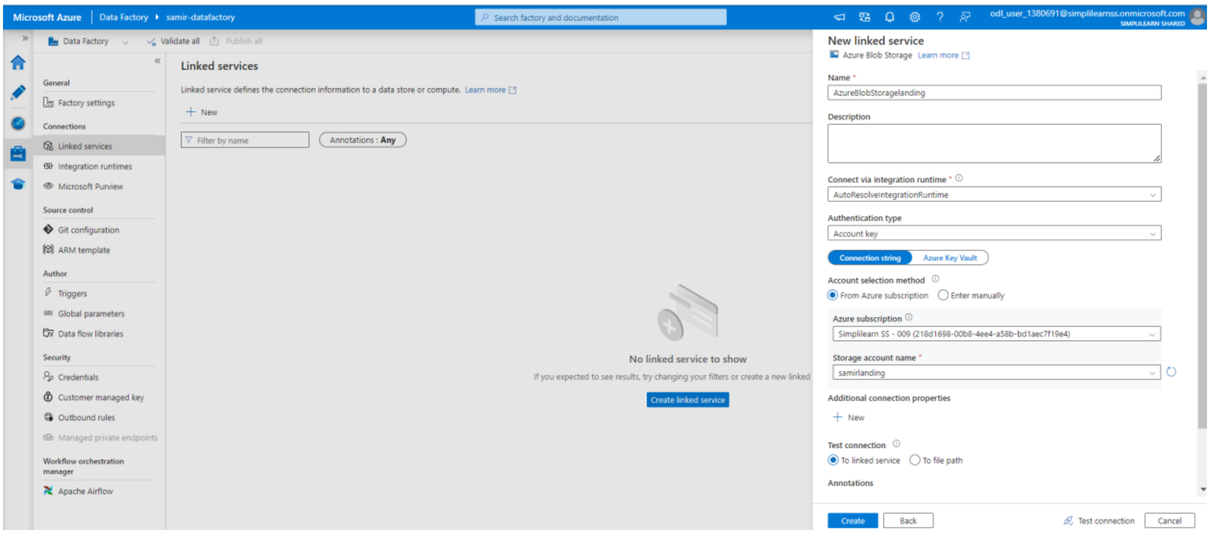
samir-datafactory

New
Pipeline
Power Query
Data flow
Dataset
scale once or
Orchestrate
Code-free data pipelines.
Transform data
Transform your data using data flows.
Configure SSIS
Manage & run your SSIS packages in the cloud.
Recent resources

Step 5: Create Linked Services for the Storage Accounts

1. **In Azure Data Factory**:
   - Go to **"Manage"** on the left panel.
   - Under **"Connections"**, click on **"New"**.

- Choose **"Azure Blob Storage"** and configure the linked service for both landing and staging storage accounts by providing the necessary details



Linked services

Linked service defines the connection information to a data store or compute. Learn more

No linked service to show

If you expected to see results, try changing your filters or create a new linked

New linked service

Azure Blob Storage Learn more

Name *: AzureBlobStoragelanding

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Simplilearn SS - 009 (218d1698-00b8-4ee4-a58b-bd1aec7f19e4)

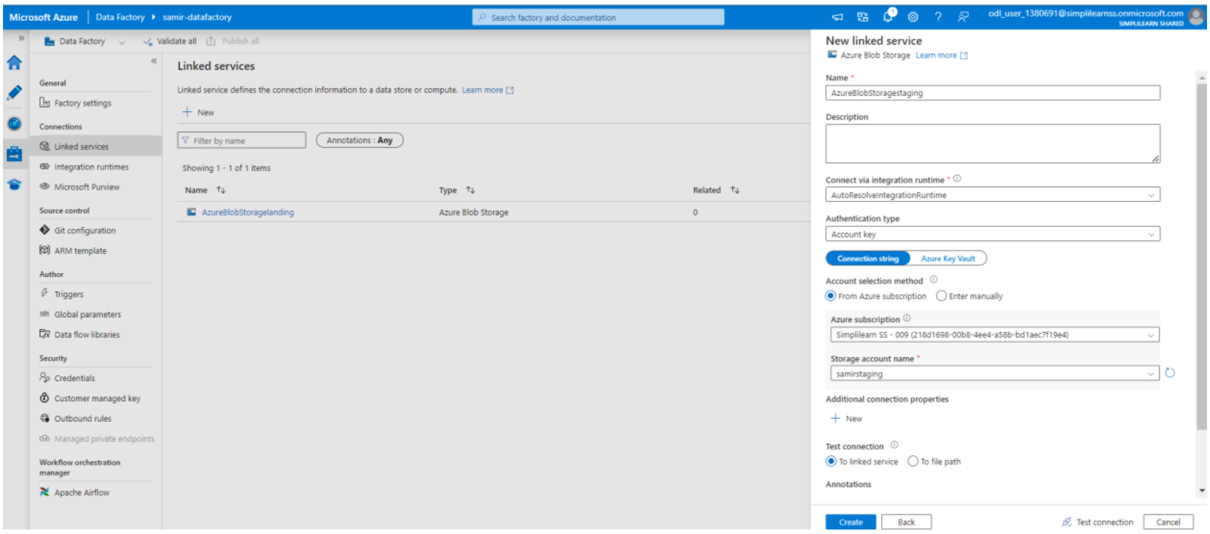
Storage account name *: samirlanding

Additional connection properties

Test connection: ● To linked service ○ To file path

Annotations

Create | Back | Test connection | Cancel



Linked services

Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 1 of 1 items

Name	Type	Related
AzureBlobStoragelanding	Azure Blob Storage	0

New linked service

Azure Blob Storage Learn more

Name *: AzureBlobStoragestaging

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Simplilearn SS - 009 (218d1698-00b8-4ee4-a58b-bd1aec7f19e4)

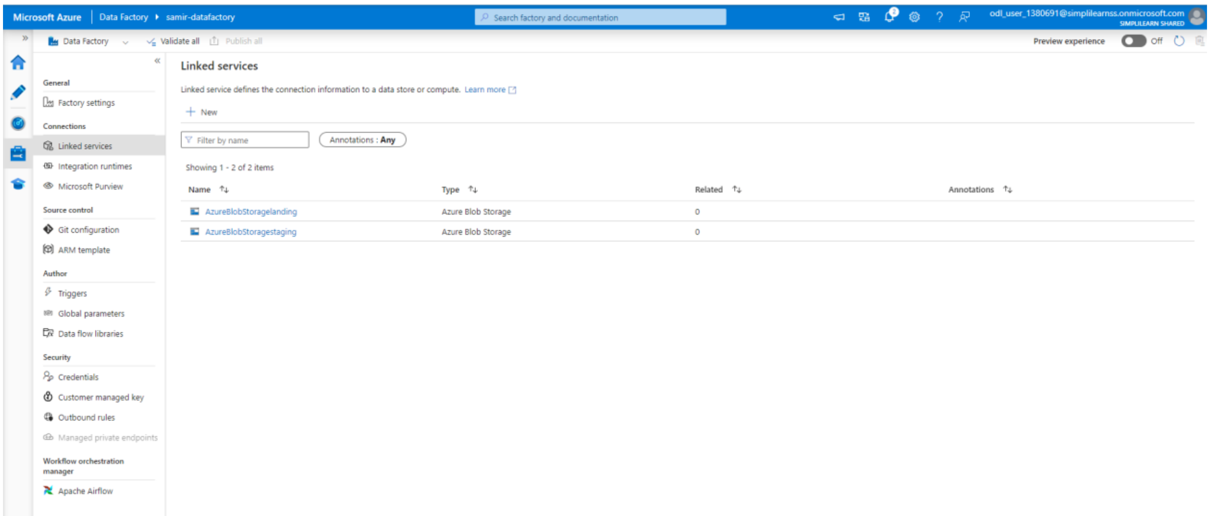
Storage account name *: samirstaging

Additional connection properties

Test connection: ● To linked service ○ To file path

Annotations

Create | Back | Test connection | Cancel



Linked services

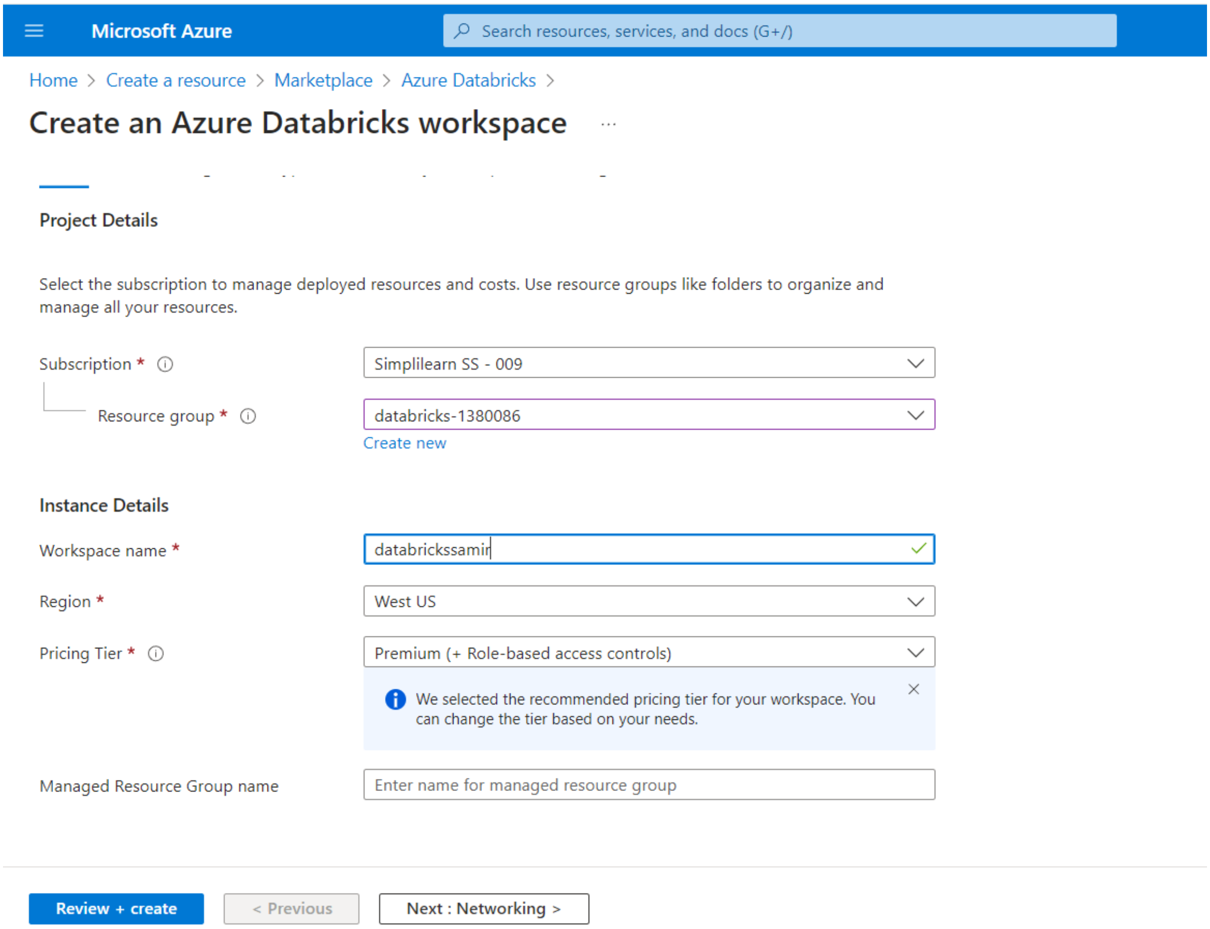
Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 2 of 2 items

Name	Type	Related	Annotations
AzureBlobStoragelanding	Azure Blob Storage	0	
AzureBlobStoragestaging	Azure Blob Storage	0	

Step 6: Use Azure Databricks as a Part of the ADF Pipeline

1. **Create a Databricks Workspace**:
   - Go to **"Create a resource"** and search for **"Azure Databricks"**.
   - Click on **"Create"**.
   - Fill in the required details and create the workspace.
2. **Create a Databricks Cluster** within the workspace.


Microsoft Azure
Search resources, services, and docs (G+/)
Home > Create a resource > Marketplace > Azure Databricks >

Create an Azure Databricks workspace

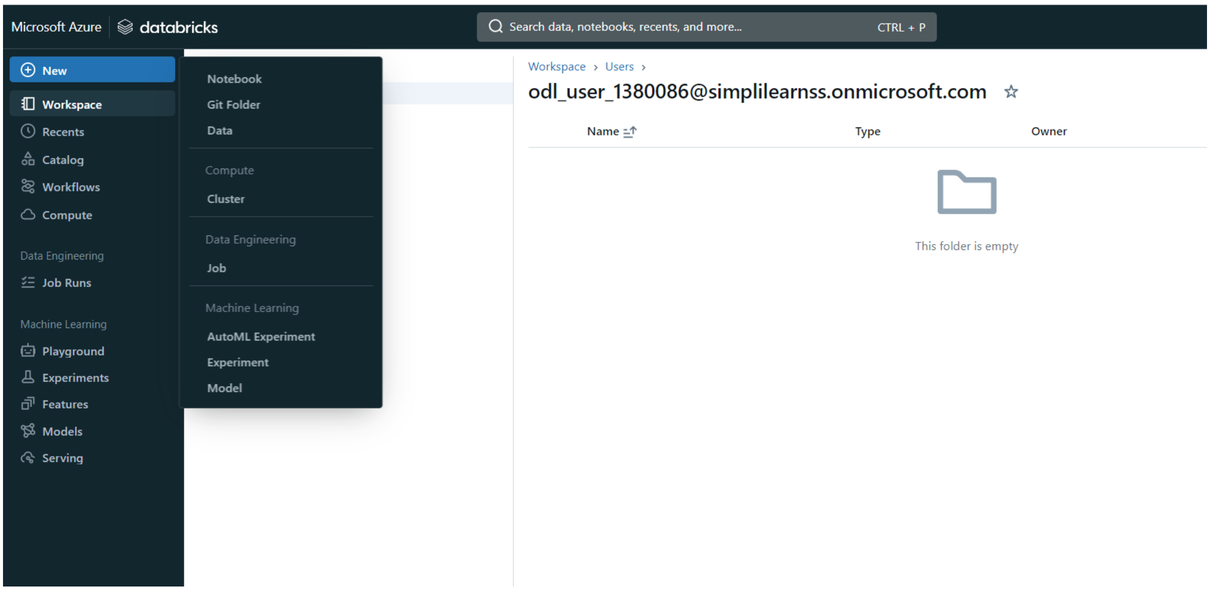
Project Details

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

Subscription *	Simplilearn SS - 009
Resource group *	databricks-1380086
	Create new

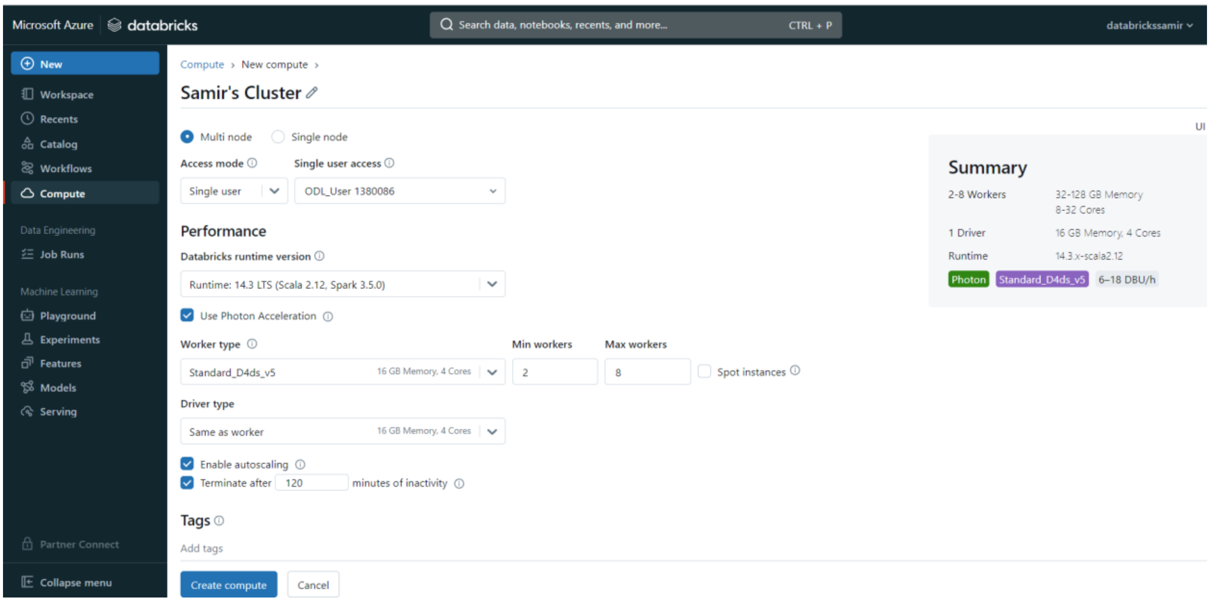
Instance Details

Workspace name *	databrickssamir
Region *	West US
Pricing Tier *	Premium (+ Role-based access controls)
	We selected the recommended pricing tier for your workspace. You can change the tier based on your needs.
Managed Resource Group name	Enter name for managed resource group



odl_user_1380086@simplilearnss.onmicrosoft.com

This folder is empty



Samir's Cluster

● Multi node ○ Single node

Access mode: Single user

Single user access: ODL_User 1380086

Performance

Databricks runtime version: Runtime: 14.3 LTS (Scala 2.12, Spark 3.5.0)

☑ Use Photon Acceleration

Worker type: Standard_D4ds_v5 16 GB Memory, 4 Cores

Min workers: 2

Max workers: 8

□ Spot instances

Driver type: Same as worker 16 GB Memory, 4 Cores

☑ Enable autoscaling

☑ Terminate after 120 minutes of inactivity

Tags

Add tags

Create compute Cancel

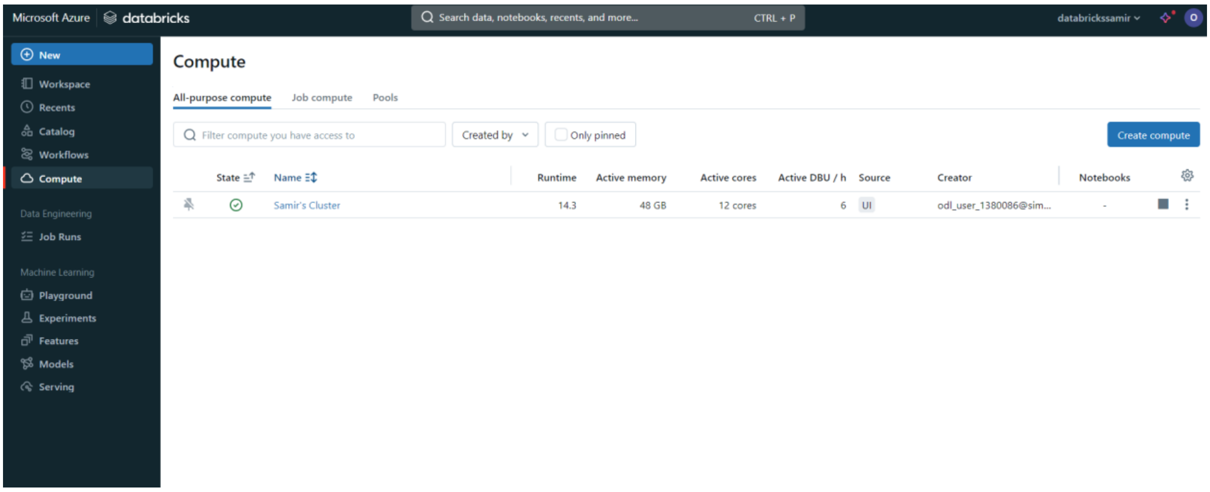
Summary

2-8 Workers 32-128 GB Memory 8-32 Cores

1 Driver 16 GB Memory, 4 Cores

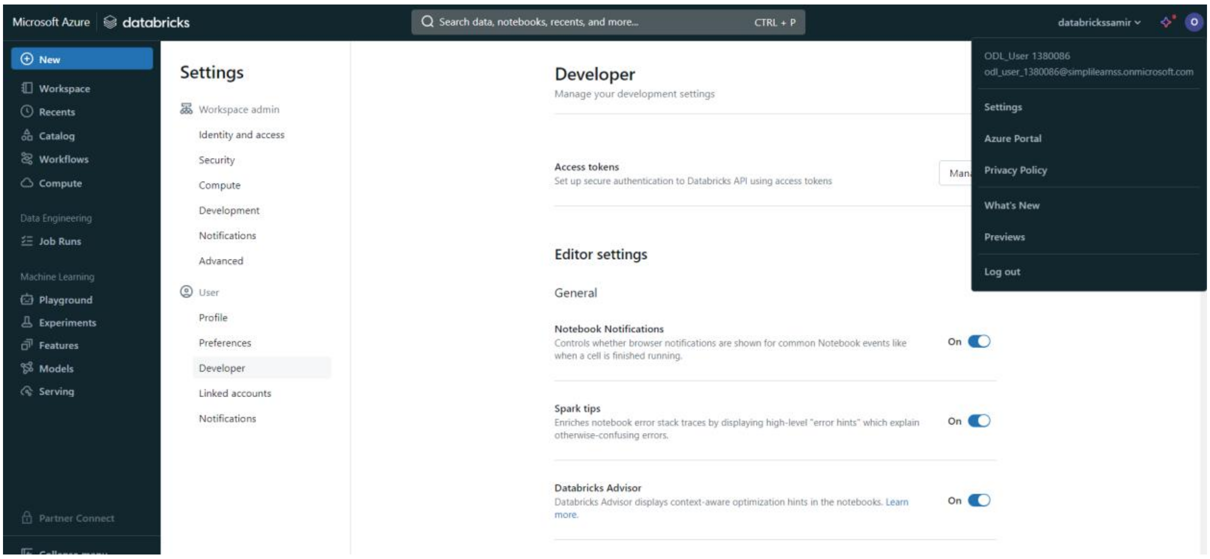
Runtime 14.3.x-scala2.12

Photon Standard_D4ds_v5 6–18 DBU/h



Compute

State	Name	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	Notebooks
✓	Samir's Cluster	14.3	48 GB	12 cores	6	UI	odl_user_1380086@sim...	-



Developer

Manage your development settings

Access tokens

Set up secure authentication to Databricks API using access tokens

Editor settings

General

Notebook Notifications

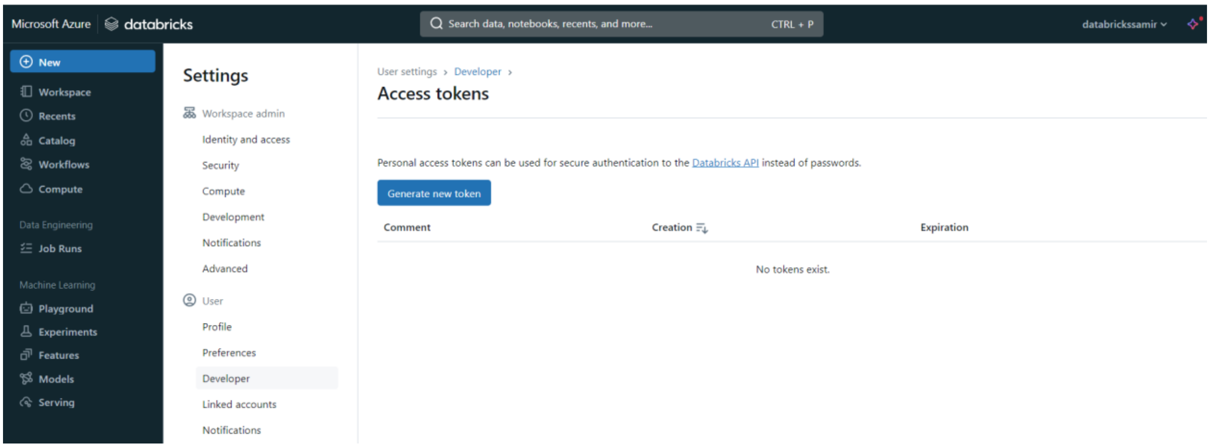
Controls whether browser notifications are shown for common Notebook events like when a cell is finished running. On

Spark tips

Enriches notebook error stack traces by displaying high-level "error hints" which explain otherwise-confusing errors. On

Databricks Advisor

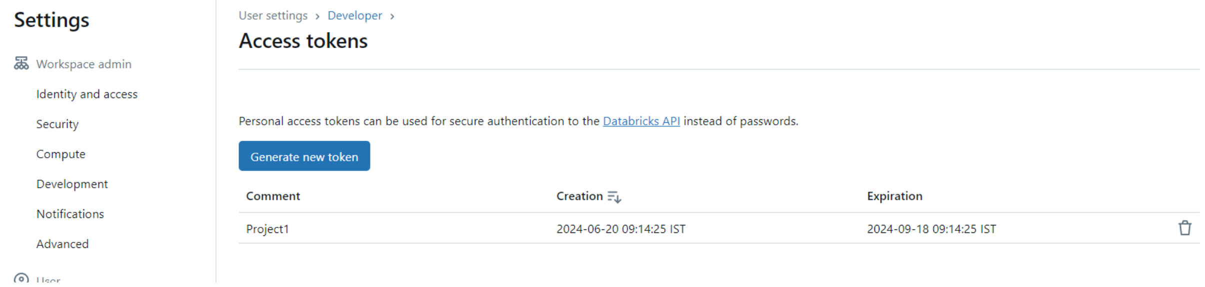
Databricks Advisor displays context-aware optimization hints in the notebooks. Learn more. On



Access tokens

Personal access tokens can be used for secure authentication to the Databricks API instead of passwords.

Comment	Creation	Expiration
No tokens exist.		



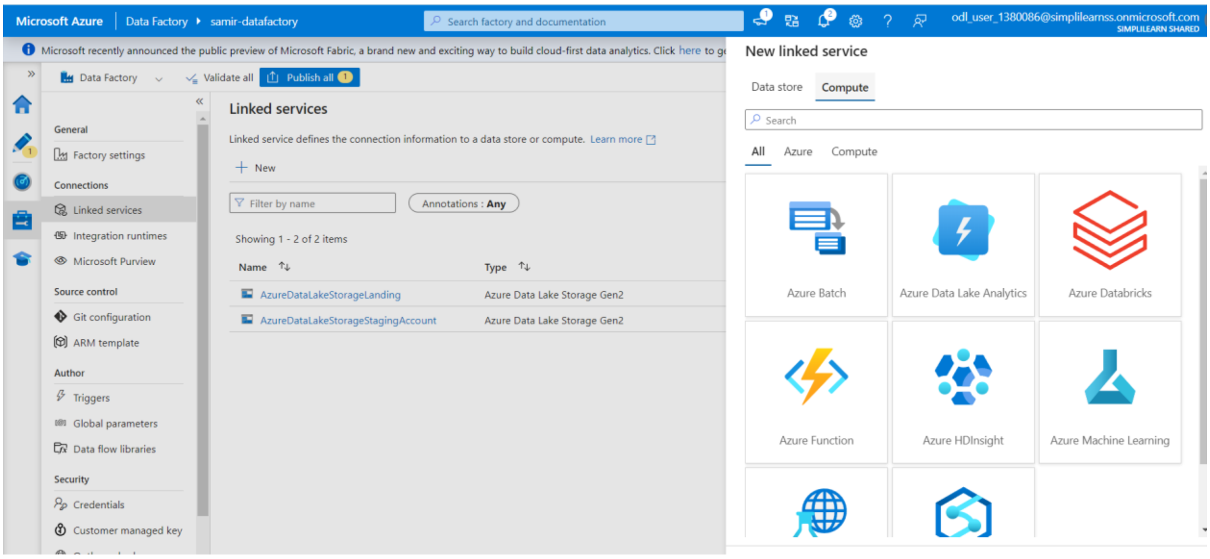
Access tokens

Personal access tokens can be used for secure authentication to the Databricks API instead of passwords.

Comment	Creation	Expiration
Project1	2024-06-20 09:14:25 IST	2024-09-18 09:14:25 IST

Step 7: Create a Linked Service in ADF for Databricks

1. **In Azure Data Factory**:
   - Go to **"Manage"** on the left panel.
   - Under **"Connections"**, click on **"New"**.
   - Choose **"Azure Databricks"** and configure the linked service by providing the necessary details (e.g., Databricks workspace URL, token).



Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to g

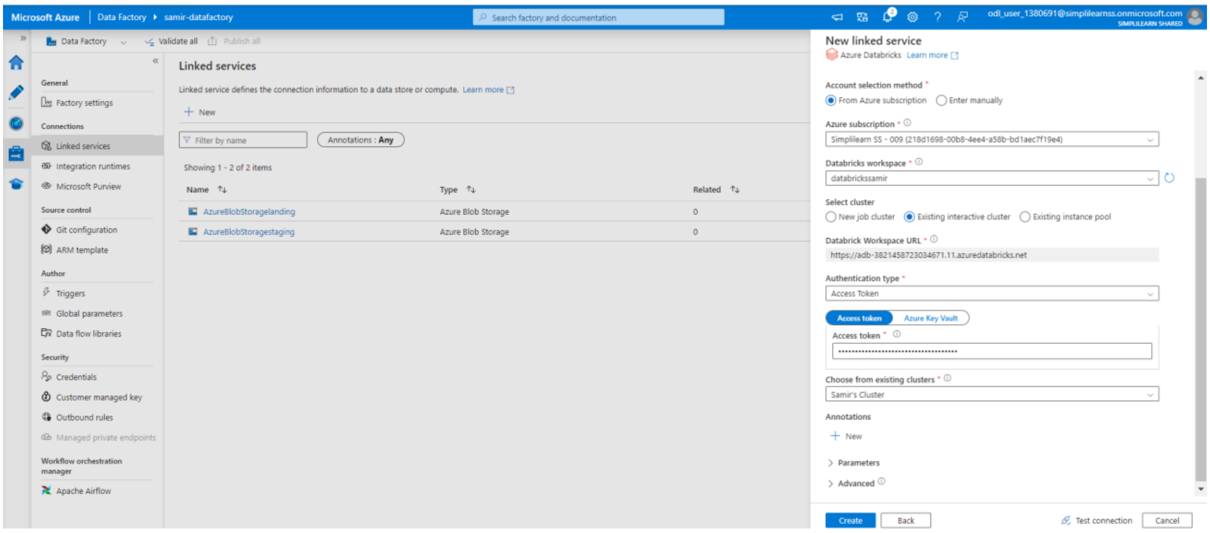
Linked services

Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 2 of 2 items

Name	Type
AzureDataLakeStorageLanding	Azure Data Lake Storage Gen2
AzureDataLakeStorageStagingAccount	Azure Data Lake Storage Gen2

New linked service



Linked services

Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 2 of 2 items

Name	Type	Related
AzureBlobStoragelanding	Azure Blob Storage	0
AzureBlobStoragestaging	Azure Blob Storage	0

New linked service

Azure Databricks Learn more

Account selection method *
● From Azure subscription ○ Enter manually

Azure subscription *
Simplilearn SS - 009 (218d1698-00b8-4ee4-a58b-bd1aec7f19e4)

Databricks workspace *
databrickssamir

Select cluster
○ New job cluster ● Existing interactive cluster ○ Existing instance pool

Databrick Workspace URL *
https://adb-3821458723034671.11.azuredatabricks.net

Authentication type *
Access Token

Access token | Azure Key Vault

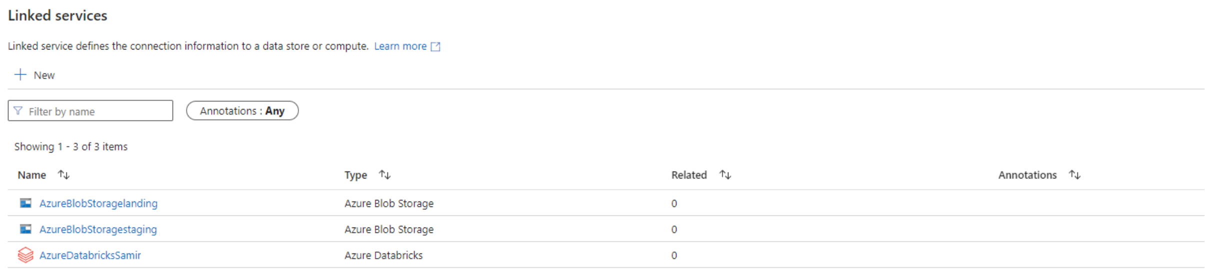
Access token *
••••••••••••••••••••••••••••••••••••

Choose from existing clusters *
Samir's Cluster

Annotations

Parameters

Advanced



Linked services

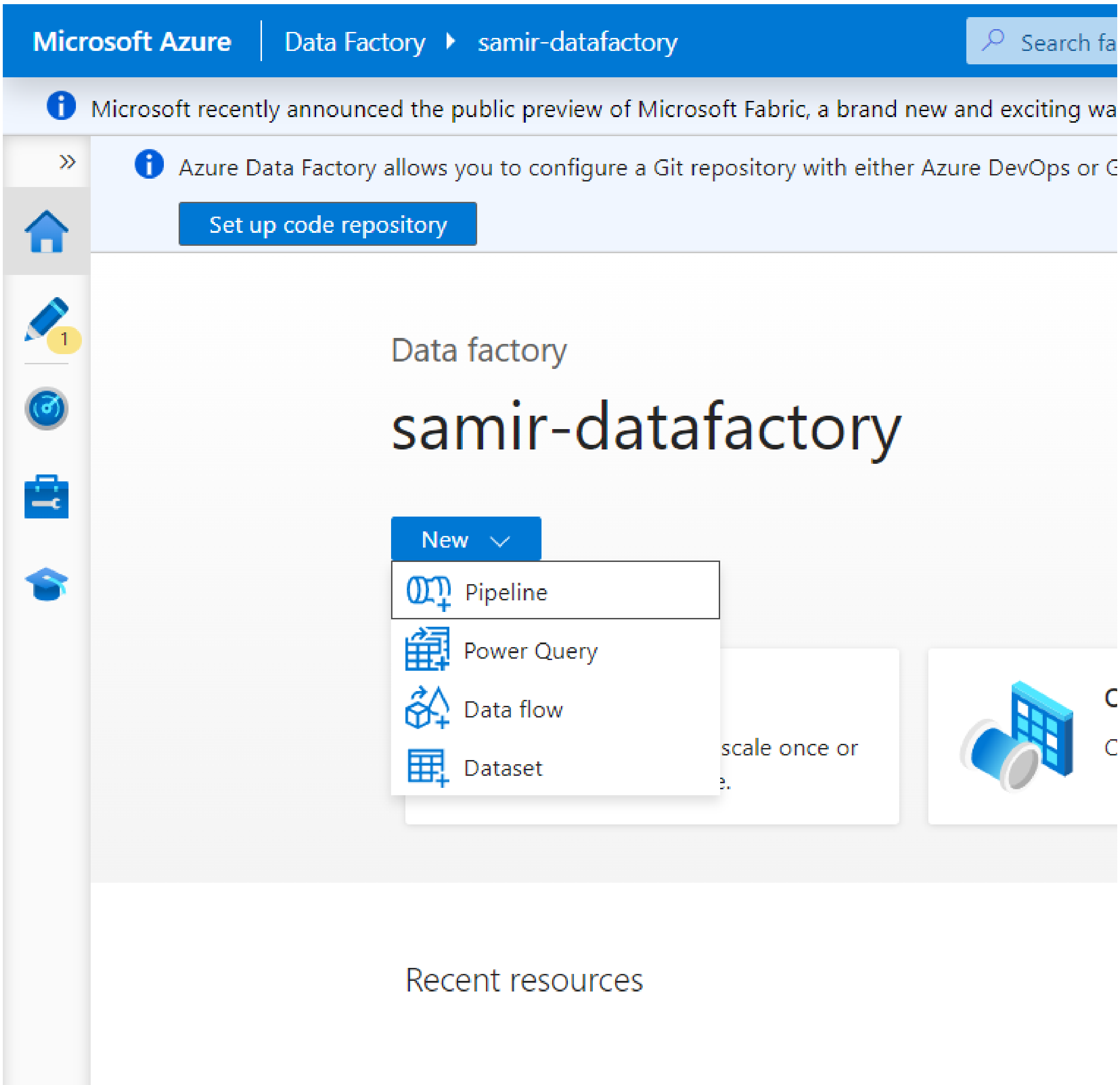
Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 3 of 3 items

Name	Type	Related	Annotations
AzureBlobStoragelanding	Azure Blob Storage	0	
AzureBlobStoragestaging	Azure Blob Storage	0	
AzureDatabricksSamir	Azure Databricks	0	

Step 8: Create an Azure Data Factory Pipeline

1. **Create a Pipeline**:
   - Go to **"Author"** on the left panel.
   - Click on the **"Pipeline"** icon and then **"New pipeline"**.
   - Add a **"Data Flow"** activity from the activities pane.



Microsoft Azure | Data Factory ▸ samir-datafactory

Search fa

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting wa

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or G

Set up code repository

Data factory

samir-datafactory

New ⌵

- Pipeline
- Power Query
- Data flow
- Dataset

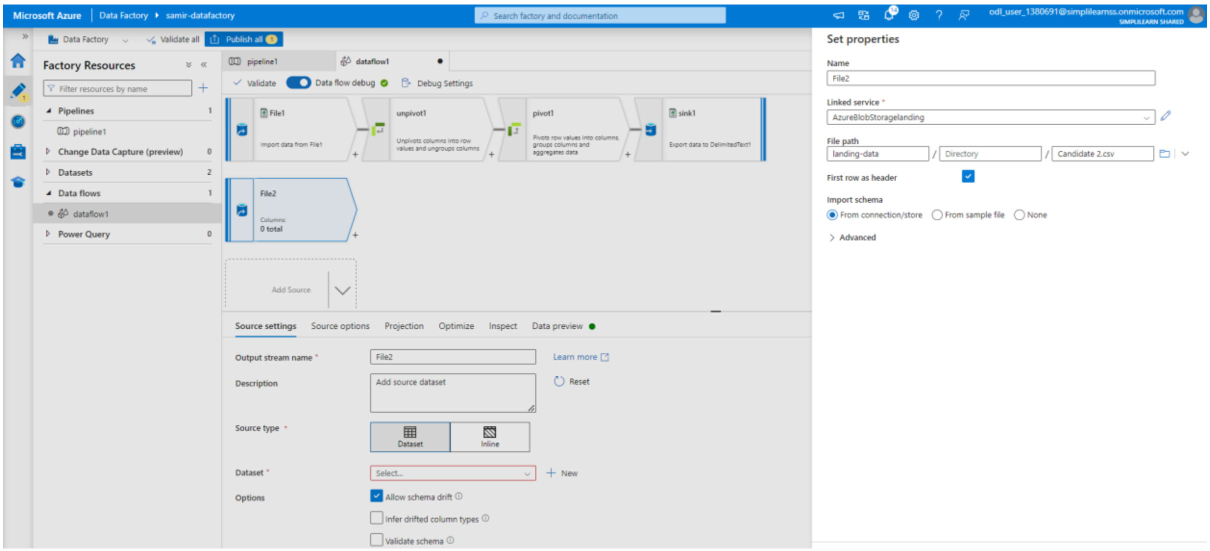
scale once or

e.

C

C

Recent resources



Microsoft Azure | Data Factory ▸ samir-datafactory

Search factory and documentation

odl_user_1380691@simplilearnss.onmicrosoft.com
SIMPLILEARN SHARED

Data Factory ⌵ | Validate all | Publish all 1

Factory Resources

Filter resources by name

Pipelines	1
pipeline1	
Change Data Capture (preview)	0
Datasets	2
Data flows	1
dataflow1	
Power Query	0

pipeline1 | dataflow1 ●

✓ Validate | Data flow debug | Debug Settings

File1
Import data from File1

unpivot1
Unpivots columns into row values and ungroups columns

pivot1
Pivots row values into columns, groups columns and aggregates data

sink1
Export data to DelimitedText1

File2
Columns:
0 total

Add Source

Source settings | Source options | Projection | Optimize | Inspect | Data preview ●

Output stream name * File2 | Learn more

Description: Add source dataset | Reset

Source type * [✓ Dataset] Inline

Dataset * Select... | + New

Options
☑ Allow schema drift
☐ Infer drifted column types
☐ Validate schema

Set properties

Name: File2

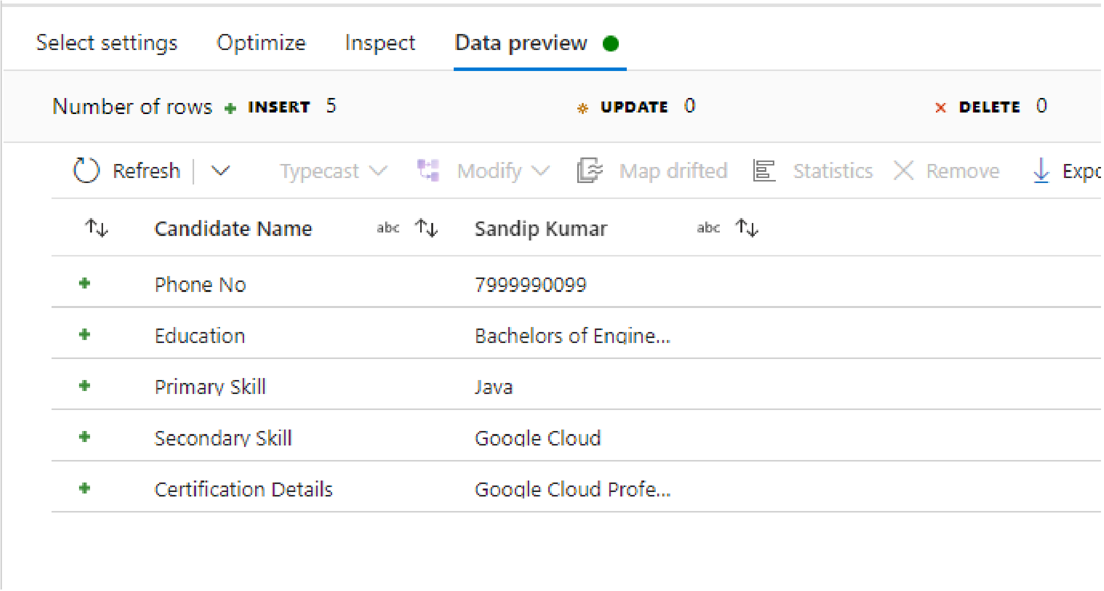
Linked service *: AzureBlobStoragelanding

File path: landing-data / Directory / Candidate 2.csv

First row as header ☑

Import schema
● From connection/store ○ From sample file ○ None

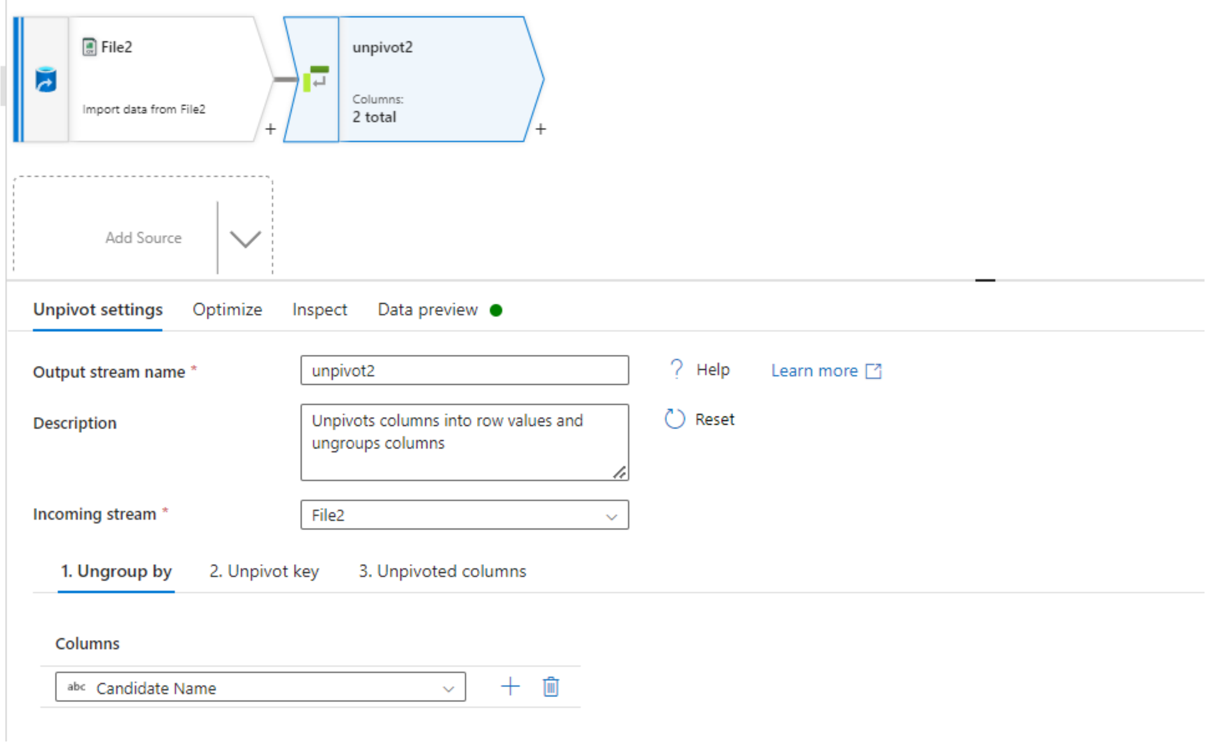
› Advanced



Number of rows ✚ INSERT 5 ✳ UPDATE 0 × DELETE 0

	Candidate Name	Sandip Kumar
✚	Phone No	7999990099
✚	Education	Bachelors of Engine...
✚	Primary Skill	Java
✚	Secondary Skill	Google Cloud
✚	Certification Details	Google Cloud Profe...

The data has to be transposed. Since there’s no way to directly transpose a table in ADF, we apply a series of Unpivot and Pivot to Transpose

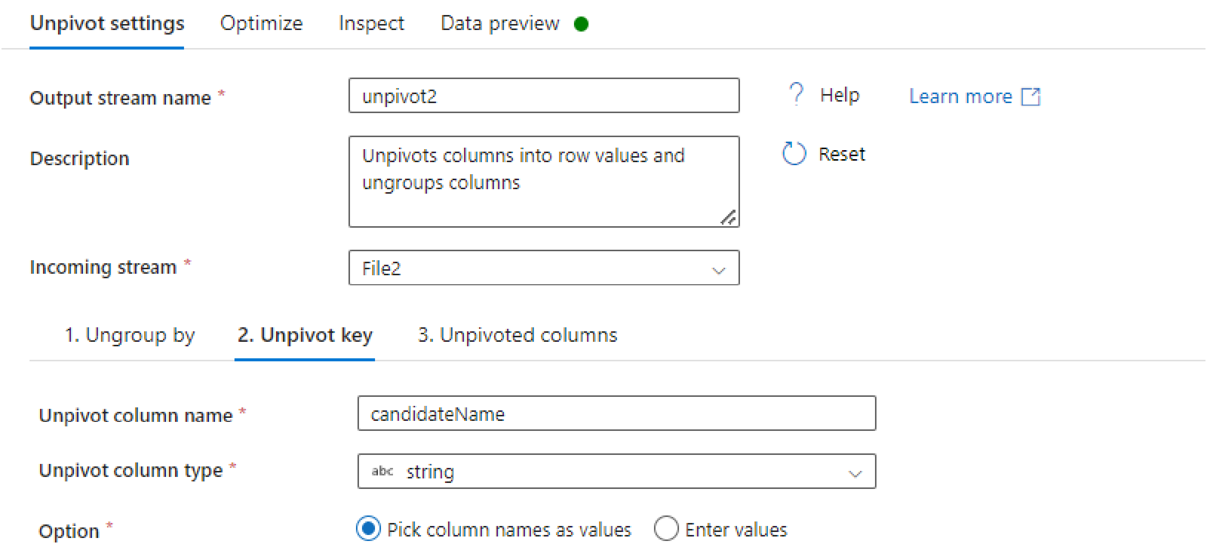


Output stream name * unpivot2

Description: Unpivots columns into row values and ungroups columns

Incoming stream * File2

Columns: Candidate Name



Output stream name * unpivot2

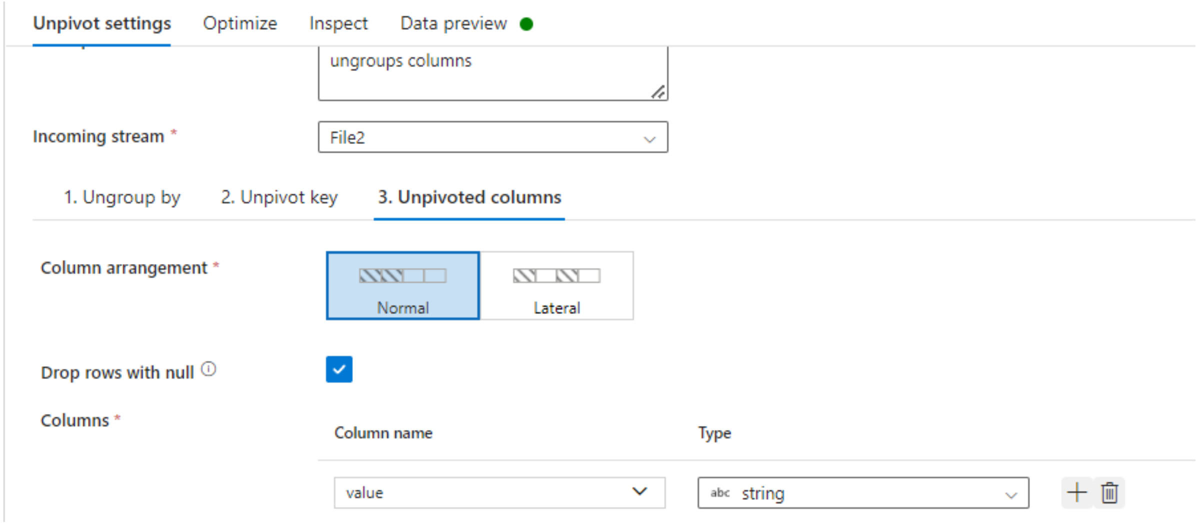
Description: Unpivots columns into row values and ungroups columns

Incoming stream * File2

Unpivot column name * candidateName

Unpivot column type * abc string

Option * ● Pick column names as values ○ Enter values



ungroups columns

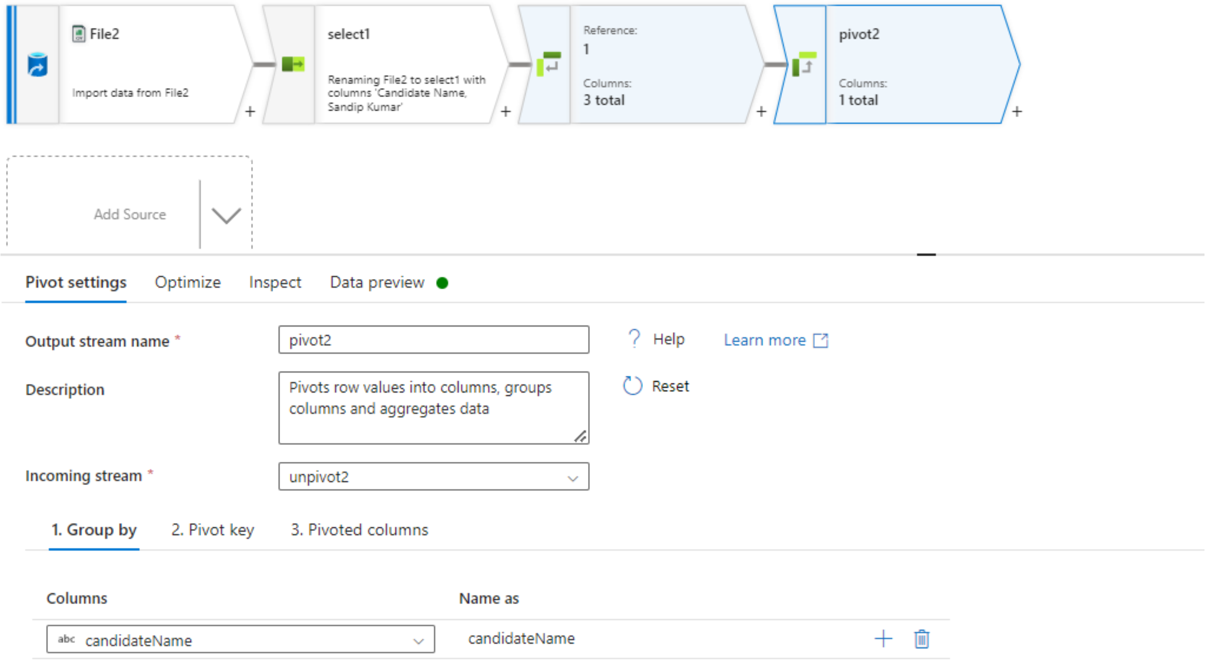
Incoming stream * File2

Column arrangement * [✓ Normal] Lateral

Drop rows with null ⓘ ☑

Columns *

Column name	Type
value	abc string



File2

Import data from File2

select1

Renaming File2 to select1 with columns 'Candidate Name, Sandip Kumar'

Reference:
1
Columns:
3 total

pivot2
Columns:
1 total

Add Source

Pivot settings | Optimize | Inspect | Data preview ●

Output stream name * pivot2

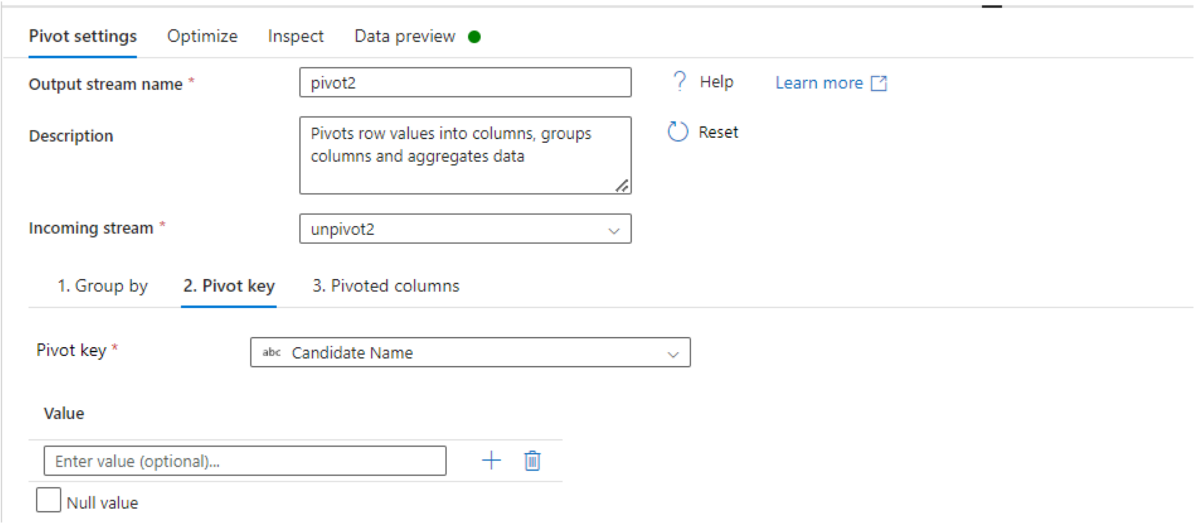
Description: Pivots row values into columns, groups columns and aggregates data

Incoming stream * unpivot2

? Help | Learn more | Reset

1. Group by | 2. Pivot key | 3. Pivoted columns

Columns	Name as
abc candidateName	candidateName



Pivot settings | Optimize | Inspect | Data preview ●

Output stream name * pivot2

Description: Pivots row values into columns, groups columns and aggregates data

Incoming stream * unpivot2

? Help | Learn more | Reset

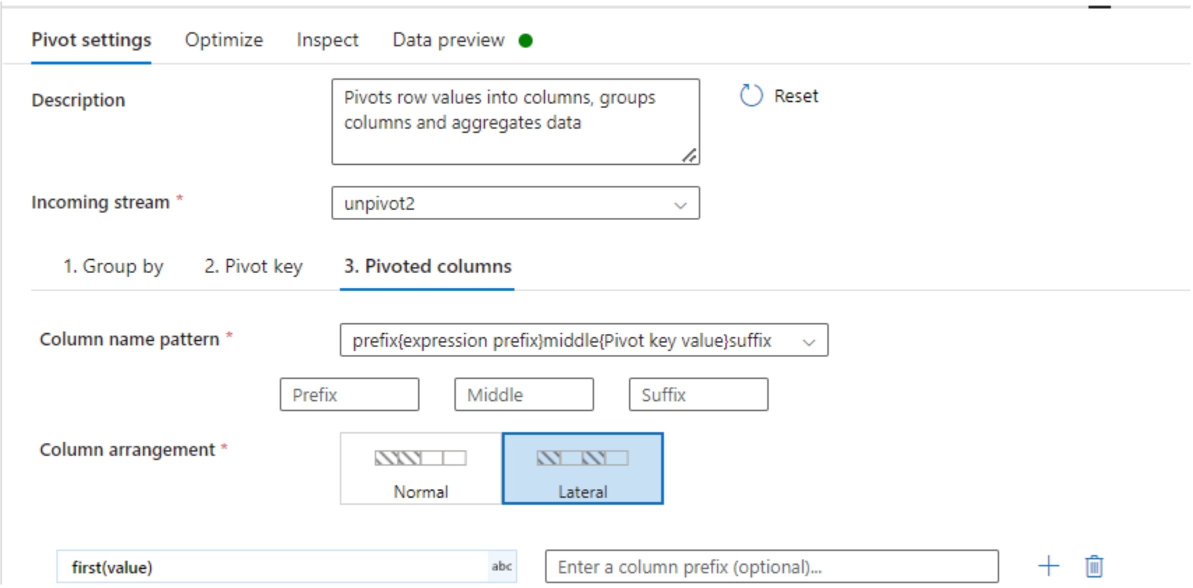
1. Group by | 2. Pivot key | 3. Pivoted columns

Pivot key * abc Candidate Name

Value

Enter value (optional)...

□ Null value



Description: Pivots row values into columns, groups columns and aggregates data

Incoming stream *: unpivot2

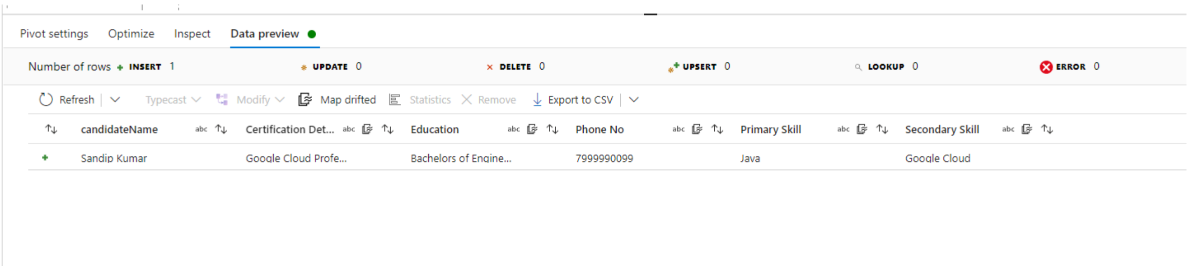
1. Group by | 2. Pivot key | 3. Pivoted columns

Column name pattern *: prefix{expression prefix}middle{Pivot key value}suffix

Prefix | Middle | Suffix

Column arrangement *: Normal | Lateral

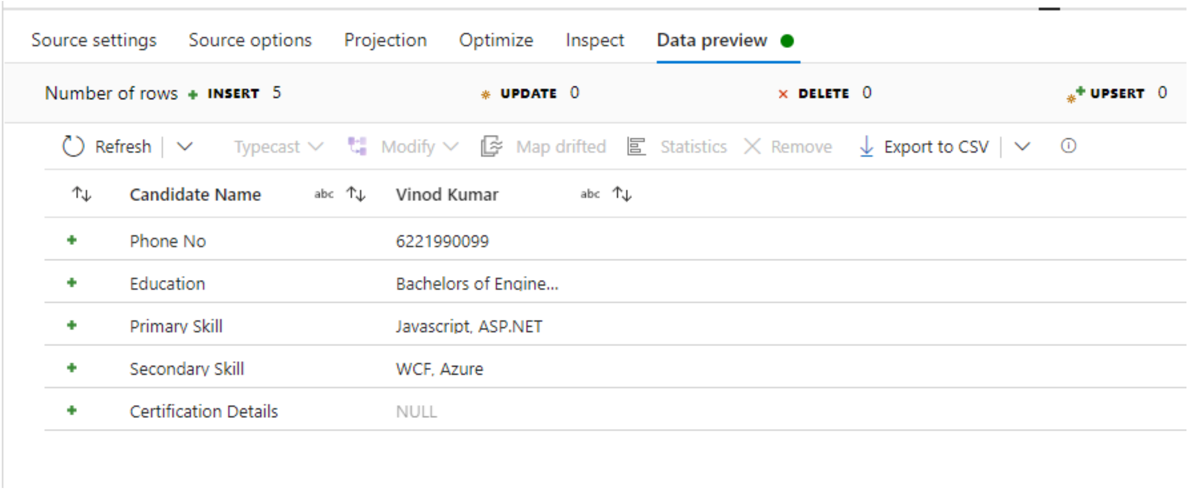
first(value) abc | Enter a column prefix (optional)...



Number of rows INSERT 1 | UPDATE 0 | DELETE 0 | UPSERT 0 | LOOKUP 0 | ERROR 0

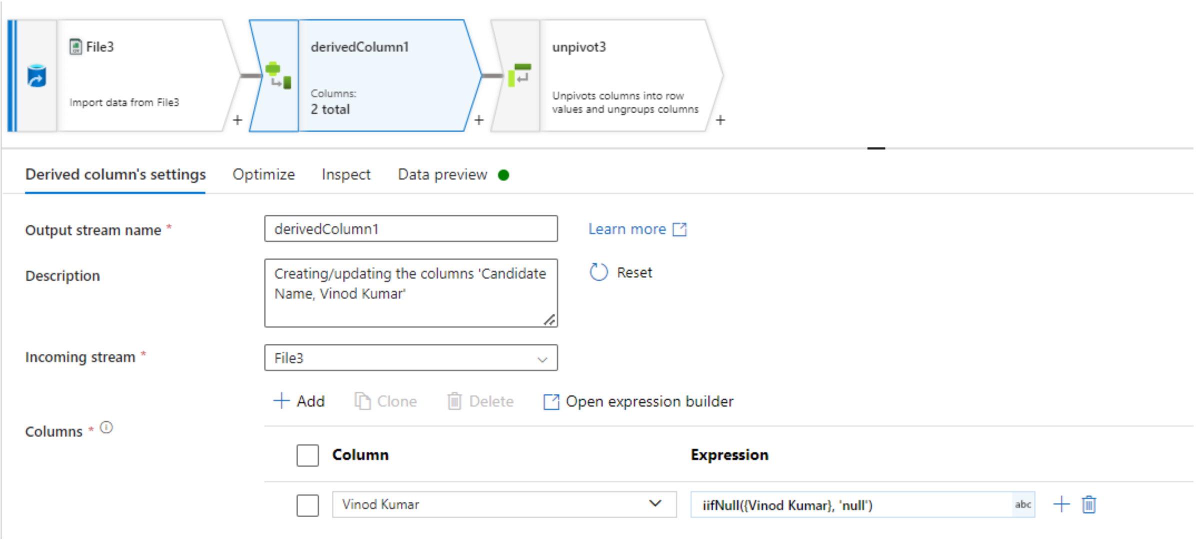
candidateName	Certification Det...	Education	Phone No	Primary Skill	Secondary Skill
Sandip Kumar	Google Cloud Profe...	Bachelors of Engine...	7999990099	Java	Google Cloud

One of files has Null Values in one of the columns. Use a Derived Column to replace null values with a string



Number of rows INSERT 5 | UPDATE 0 | DELETE 0 | UPSERT 0

Candidate Name	Vinod Kumar
Phone No	6221990099
Education	Bachelors of Engine...
Primary Skill	Javascript, ASP.NET
Secondary Skill	WCF, Azure
Certification Details	NULL



Output stream name * derivedColumn1

Learn more

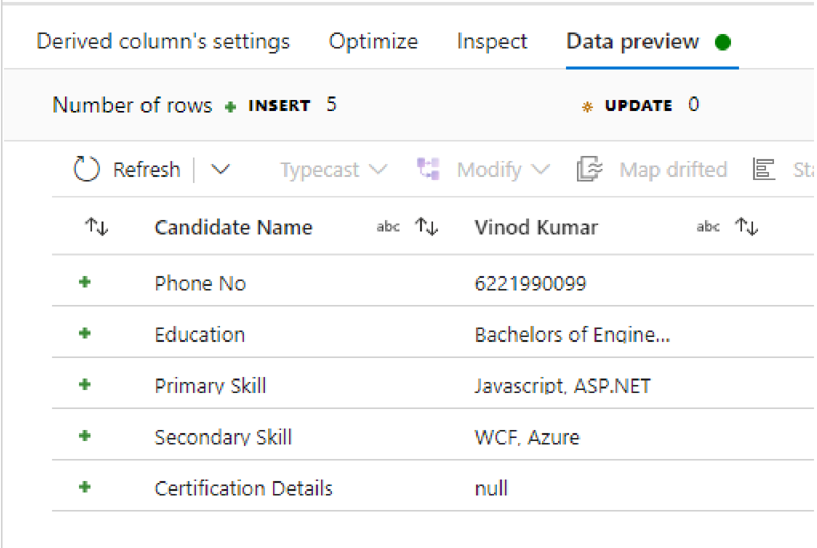
Description: Creating/updating the columns 'Candidate Name, Vinod Kumar'

Reset

Incoming stream * File3

Columns *

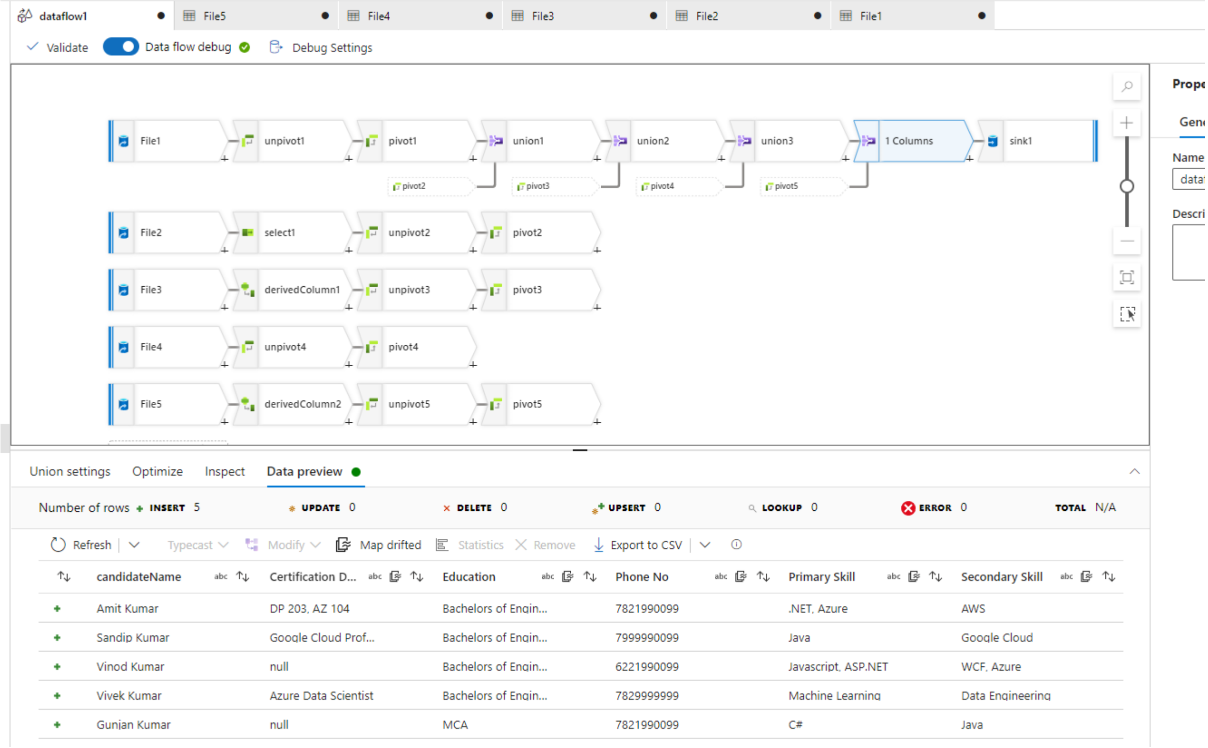
☐	Column	Expression
☐	Vinod Kumar	iifNull((Vinod Kumar), 'null')



Number of rows + INSERT 5 ✱ UPDATE 0

	Candidate Name	Vinod Kumar
+	Phone No	6221990099
+	Education	Bachelors of Engine...
+	Primary Skill	Javascript, ASP.NET
+	Secondary Skill	WCF, Azure
+	Certification Details	null

The final Data Flow looks like this:-

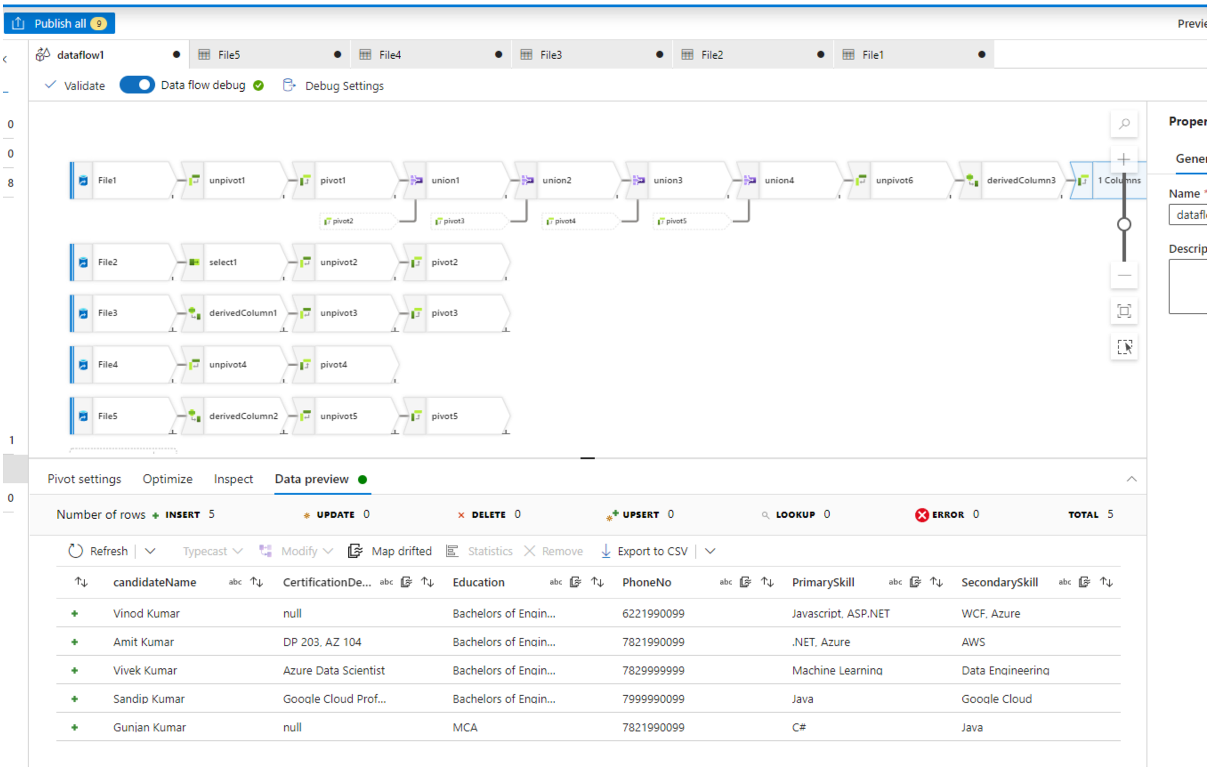


Number of rows INSERT 5 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 ERROR 0 TOTAL N/A

	candidateName	Certification D...	Education	Phone No	Primary Skill	Secondary Skill
+	Amit Kumar	DP 203, AZ 104	Bachelors of Engin...	7821990099	.NET, Azure	AWS
+	Sandip Kumar	Google Cloud Prof...	Bachelors of Engin...	7999990099	Java	Google Cloud
+	Vinod Kumar	null	Bachelors of Engin...	6221990099	Javascript, ASP.NET	WCF, Azure
+	Vivek Kumar	Azure Data Scientist	Bachelors of Engin...	7829999999	Machine Learning	Data Engineering
+	Gunjan Kumar	null	MCA	7821990099	C#	Java

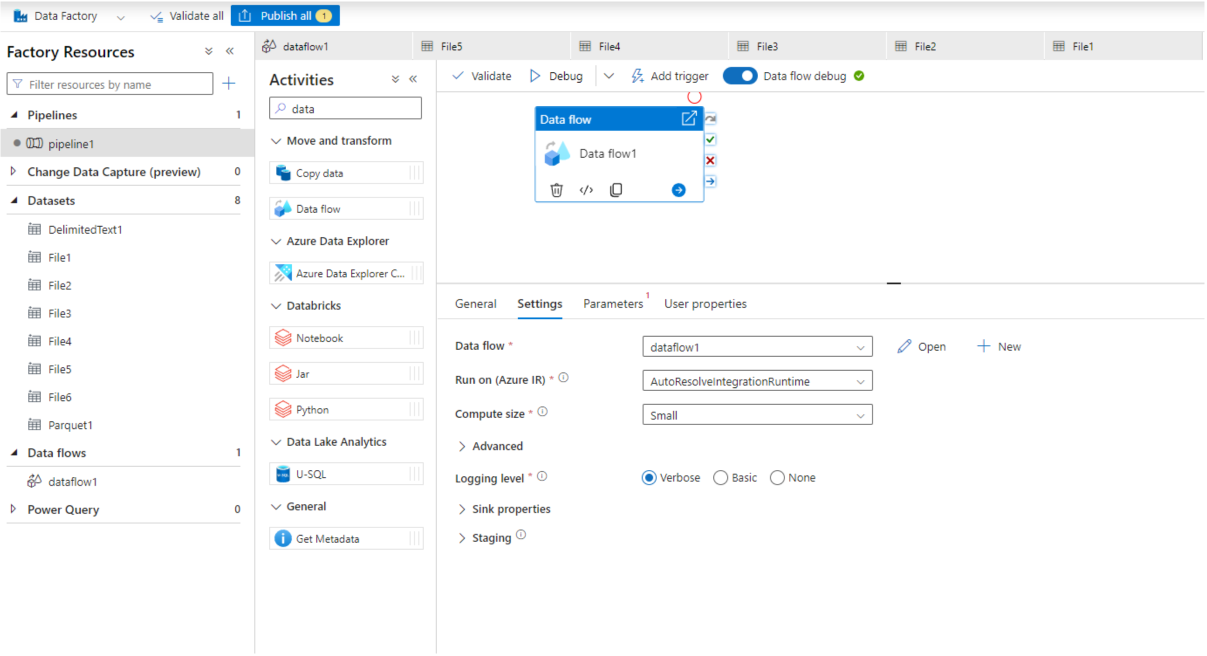
Step 9: Convert CSV Files to Parquet Files in Staging Storage

- The Column Names have spaces in between, and therefore can’t be directly converted to parquet. Use “replace” function is Select to convert the columns
- Publish the pipeline and put a manual trigger to run the pipeline



Number of rows INSERT 5 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 ERROR 0 TOTAL 5

	candidateName	CertificationDe...	Education	PhoneNo	PrimarySkill	SecondarySkill
+	Vinod Kumar	null	Bachelors of Engin...	6221990099	Javascript, ASP.NET	WCF, Azure
+	Amit Kumar	DP 203, AZ 104	Bachelors of Engin...	7821990099	.NET, Azure	AWS
+	Vivek Kumar	Azure Data Scientist	Bachelors of Engin...	7829999999	Machine Learning	Data Engineering
+	Sandip Kumar	Google Cloud Prof...	Bachelors of Engin...	7999990099	Java	Google Cloud
+	Gunjan Kumar	null	MCA	7821990099	C#	Java

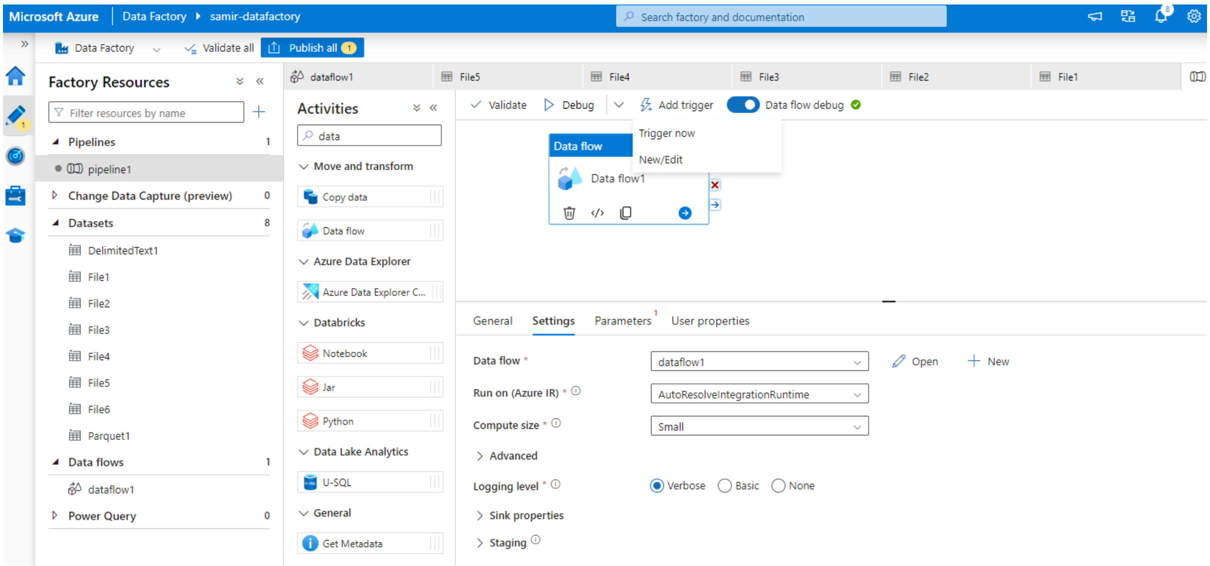


Data flow * dataflow1

Run on (Azure IR) * AutoResolveIntegrationRuntime

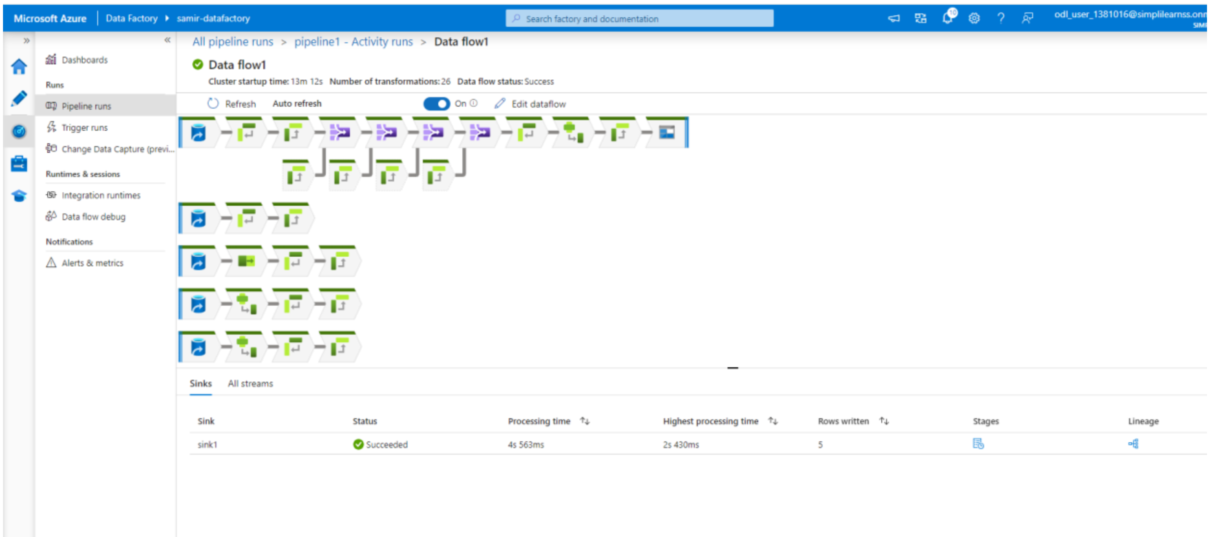
Compute size * Small

Logging level * ● Verbose ○ Basic ○ None



General | **Settings** | Parameters | User properties

- Data flow *: dataflow1
- Run on (Azure IR) *: AutoResolveIntegrationRuntime
- Compute size *: Small
- Advanced
- Logging level *: ● Verbose ○ Basic ○ None
- Sink properties
- Staging



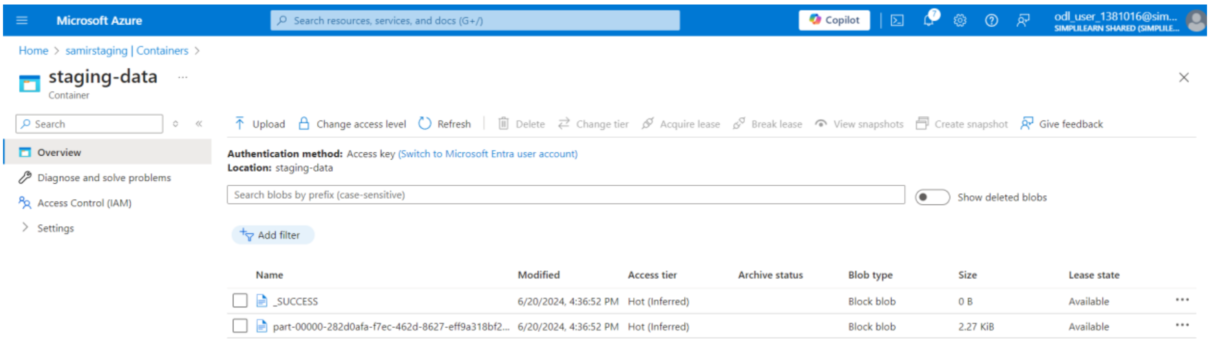
All pipeline runs > pipeline1 - Activity runs > Data flow1

Data flow1

Cluster startup time: 13m 12s Number of transformations: 26 Data flow status: Success

Sinks | All streams

Sink	Status	Processing time	Highest processing time	Rows written	Stages	Lineage
sink1	✓ Succeeded	4s 563ms	2s 430ms	5		



Home > samirstaging | Containers >

staging-data

Container

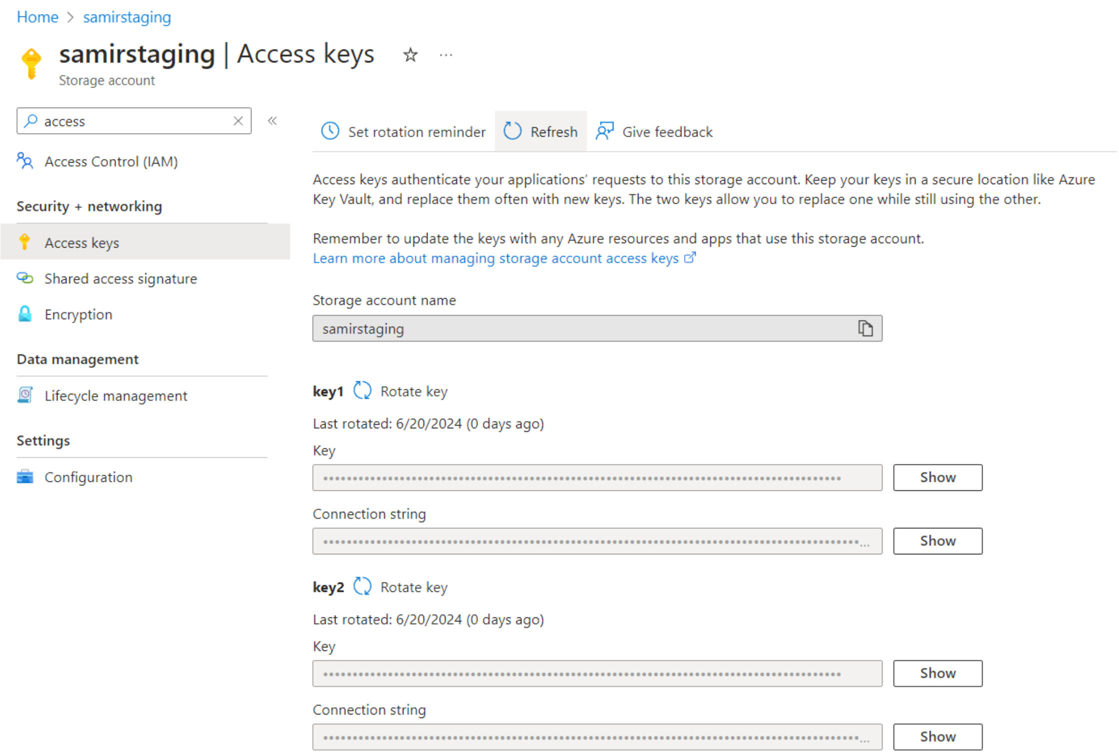
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: staging-data

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
_SUCCESS	6/20/2024, 4:36:52 PM	Hot (Inferred)		Block blob	0 B	Available
part-00000-282d0afa-f7ec-462d-8627-eff9a318bf2...	6/20/2024, 4:36:52 PM	Hot (Inferred)		Block blob	2.27 KiB	Available

The parquet file is stored in the staging area.

Step 10: Access Parquet Files from the Staging Account in Azure Databricks

1. **In Azure Databricks**:
   - Mount the staging storage account if needed.
   - Read the Parquet files into Databricks.



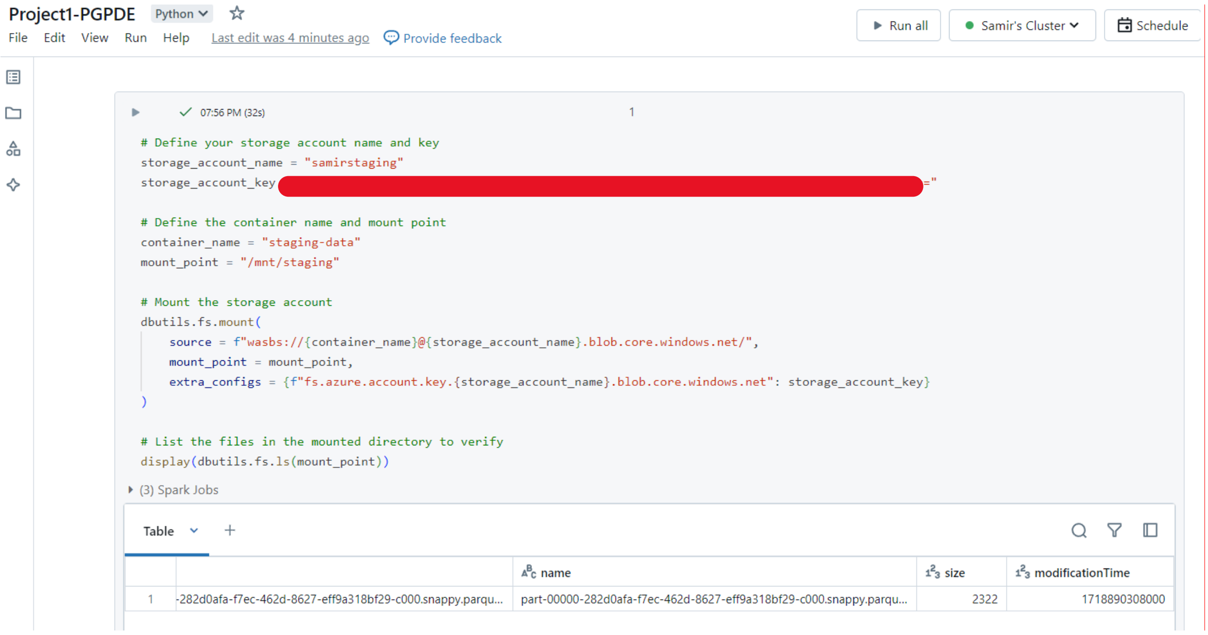
Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.

Remember to update the keys with any Azure resources and apps that use this storage account.
Learn more about managing storage account access keys

Storage account name: samirstaging

key1
Last rotated: 6/20/2024 (0 days ago)
Key: ••
Connection string: ••...

key2
Last rotated: 6/20/2024 (0 days ago)
Key: ••
Connection string: ••...



```
# Define your storage account name and key
storage_account_name = "samirstaging"
storage_account_key [redacted]"

# Define the container name and mount point
container_name = "staging-data"
mount_point = "/mnt/staging"

# Mount the storage account
dbutils.fs.mount(
    source = f"wasbs://{container_name}@{storage_account_name}.blob.core.windows.net/",
    mount_point = mount_point,
    extra_configs = {f"fs.azure.account.key.{storage_account_name}.blob.core.windows.net": storage_account_key}
)

# List the files in the mounted directory to verify
display(dbutils.fs.ls(mount_point))
```

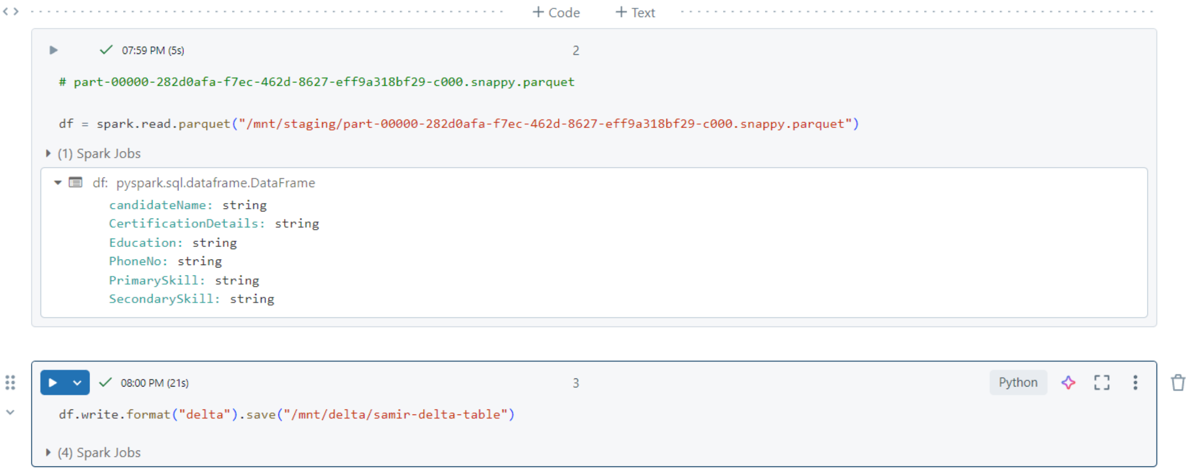
(3) Spark Jobs

Table

		name	size	modificationTime
1	-282d0afa-f7ec-462d-8627-eff9a318bf29-c000.snappy.parqu...	part-00000-282d0afa-f7ec-462d-8627-eff9a318bf29-c000.snappy.parqu...	2322	1718890308000

Step 11: Convert the Parquet Files to Azure Databricks Delta Tables

1. **In Azure Databricks**:
   - Convert Parquet to Delta.



```
# part-00000-282d0afa-f7ec-462d-8627-eff9a318bf29-c000.snappy.parquet

df = spark.read.parquet("/mnt/staging/part-00000-282d0afa-f7ec-462d-8627-eff9a318bf29-c000.snappy.parquet")
```

```
(1) Spark Jobs
df:  pyspark.sql.dataframe.DataFrame
    candidateName: string
    CertificationDetails: string
    Education: string
    PhoneNo: string
    PrimarySkill: string
    SecondarySkill: string
```

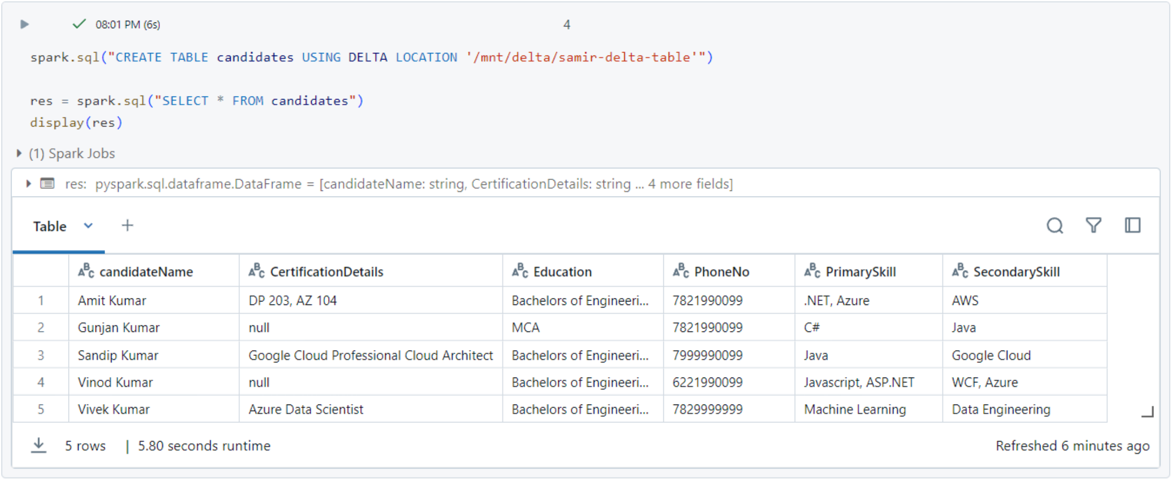
```
df.write.format("delta").save("/mnt/delta/samir-delta-table")
```

```
(4) Spark Jobs
```

Step 12: Store and Visualize the Data from Azure Databricks Delta Tables

1. **Create Delta Tables**:
   - Register the Delta table in the Databricks metastore.

2. **Visualize Data**:
   - Use Databricks notebooks to visualize the data



```
spark.sql("CREATE TABLE candidates USING DELTA LOCATION '/mnt/delta/samir-delta-table'")

res = spark.sql("SELECT * FROM candidates")
display(res)
```

```
(1) Spark Jobs
res:  pyspark.sql.dataframe.DataFrame = [candidateName: string, CertificationDetails: string ... 4 more fields]
```

	candidateName	CertificationDetails	Education	PhoneNo	PrimarySkill	SecondarySkill
1	Amit Kumar	DP 203, AZ 104	Bachelors of Engineeri...	7821990099	.NET, Azure	AWS
2	Gunjan Kumar	null	MCA	7821990099	C#	Java
3	Sandip Kumar	Google Cloud Professional Cloud Architect	Bachelors of Engineeri...	7999990099	Java	Google Cloud
4	Vinod Kumar	null	Bachelors of Engineeri...	6221990099	Javascript, ASP.NET	WCF, Azure
5	Vivek Kumar	Azure Data Scientist	Bachelors of Engineeri...	7829999999	Machine Learning	Data Engineering

5 rows | 5.80 seconds runtime

Refreshed 6 minutes ago