

# Azure Databricks Coding Challenge

**Name : Mitushi Vishwakarma**

## **Question 3 :Execute & explain, Azure datafactory and its copy activity.**

Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.

- ADF does not store any data itself.
- It allows you to create data-driven workflows to orchestrate the movement of data between supported data stores and then process the data using compute services in other regions or in an on-premise environment.
- It also allows you to monitor and manage workflows using both programmatic and UI mechanisms.

**Copy Activity** in Azure Data Factory copies data from a source data store to a sink data store.

Azure supports various data stores such as source or sink data stores like

Azure Blob storage,

Azure Cosmos DB (DocumentDB API),

Azure Data Lake Store,

Oracle,

Cassandra, etc.

**Copy Activity** : Creating two storage accounts “storageoneadfcopy” and “storagesecondadfcopy” and performing copy activity in data factory.

Created first blob storage account and container that will be the source store :

The screenshot displays the Microsoft Azure portal interface. At the top, the navigation bar includes the Microsoft Azure logo, an 'Upgrade' button, a search bar, and user information for 'mitushi.vishjs@gmail.com'. The main content area shows the 'Overview' page for a deployment named 'storageoneadfcopy\_1708495292918'. The deployment status is 'Complete', indicated by a green checkmark. Key details include the deployment name, subscription 'Free Trial', resource group 'ADFCopyActivity', start time '21/2/2024, 11:31:41 am', and correlation ID '65be3afa-f97c-4167-8c00-60d0eaf0ca5b'. A 'Go to resource' button is visible. The left sidebar contains navigation links for 'Overview', 'Inputs', 'Outputs', and 'Template'. On the right, there are panels for 'Cost Management' and 'Microsoft Defender for Cloud'.

## Created container in storage account :

Home > storageaccount\_1708495292918 | Overview > storageaccount\_1708495292918 | Containers

Storage account

Search containers by prefix

Name	Last modified	Anonymous access level	Lease state
\$logs	21/2/2024, 11:32:15 am	Private	Available
sourcecontainer	21/2/2024, 11:34:05 am	Private	Available

## Uploaded file in container :

Home > storageaccount\_1708495292918 | Overview > storageaccount\_1708495292918 | Containers > sourcecontainer

Container

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
jobs_in_data - Copy.csv	21/2/2024, 11:47:45 ...	Hot (inferred)		Block blob	1.22 KiB	Available

## Created second blob storage account and container that will be the sink store :

Home > storagesecondadfcopy\_1708496462127 | Overview

Deployment

Search

Deployment name: storagesecondadfcopy\_1708496462127

Start time: 21/2/2024, 11:51:08 am

Subscription: Free Trial

Resource group: ADFCopyActivity

Correlation ID: e064eb27-e2bc-43e8-b686-28f0dd2d657a

Deployment details

Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

Cost Management

Get notified to stay within your budget and prevent unexpected charges on your bill.

Set up cost alerts >

Microsoft Defender for Cloud

Secure your apps and infrastructure

Go to Microsoft Defender for Cloud >

Free Microsoft tutorials

## Created container in storage account :

The screenshot shows the 'Containers' page in the Microsoft Azure portal. The breadcrumb navigation is 'Home > storageesecondadfcopy\_1708496462127 | Overview > storageesecondadfcopy'. The page title is 'storageesecondadfcopy | Containers'. On the left, there is a sidebar with navigation options: Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, and Fuente. The main content area has a search bar and a table of containers. The table has columns: Name, Last modified, Anonymous access level, and Lease state. There are two containers listed: '\$logs' and 'sinkcontainer', both created on 21/2/2024 at 11:51:38 am and 11:53:50 am respectively, with 'Private' anonymous access level and 'Available' lease state.

Name	Last modified	Anonymous access level	Lease state
\$logs	21/2/2024, 11:51:38 am	Private	Available
sinkcontainer	21/2/2024, 11:53:50 am	Private	Available

## Empty Container :

The screenshot shows the 'sinkcontainer' page in the Microsoft Azure portal. The breadcrumb navigation is 'Home > storageesecondadfcopy\_1708496462127 | Overview > storageesecondadfcopy | Containers > sinkcontainer'. The page title is 'sinkcontainer'. On the left, there is a sidebar with navigation options: Overview, Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens, Access policy, Properties, and Metadata. The main content area has a search bar and a table of blobs. The table has columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. There are no results shown in the table.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
No results						

## Created Data Factory named “1046DataFactory “ :

The screenshot shows the 'Overview' page for a Data Factory named 'Microsoft.DataFactory-20240221115758'. The breadcrumb navigation is 'Home > Microsoft.DataFactory-20240221115758 | Overview'. The page title is 'Microsoft.DataFactory-20240221115758 | Overview'. On the left, there is a sidebar with navigation options: Overview, Inputs, Outputs, and Template. The main content area has a search bar and a deployment status section. The deployment status is 'Your deployment is complete'. It shows the deployment name, subscription, resource group, start time, and correlation ID. There are also links for 'Deployment details', 'Next steps', and 'Go to resource'. On the right, there are sections for 'Cost management', 'Microsoft Defender for Cloud', and 'Free Microsoft tutorials'.

**Your deployment is complete**

Deployment name : Microsoft.DataFactory-20240221115758  
Subscription : Free Trial  
Resource group : ADFCopyActivity

Start time : 21/2/2024, 11:59:59 am  
Correlation ID : 42df5b04-f8b5-4b93-9f06-cf6f4f2d53f3

**Deployment details**

**Next steps**

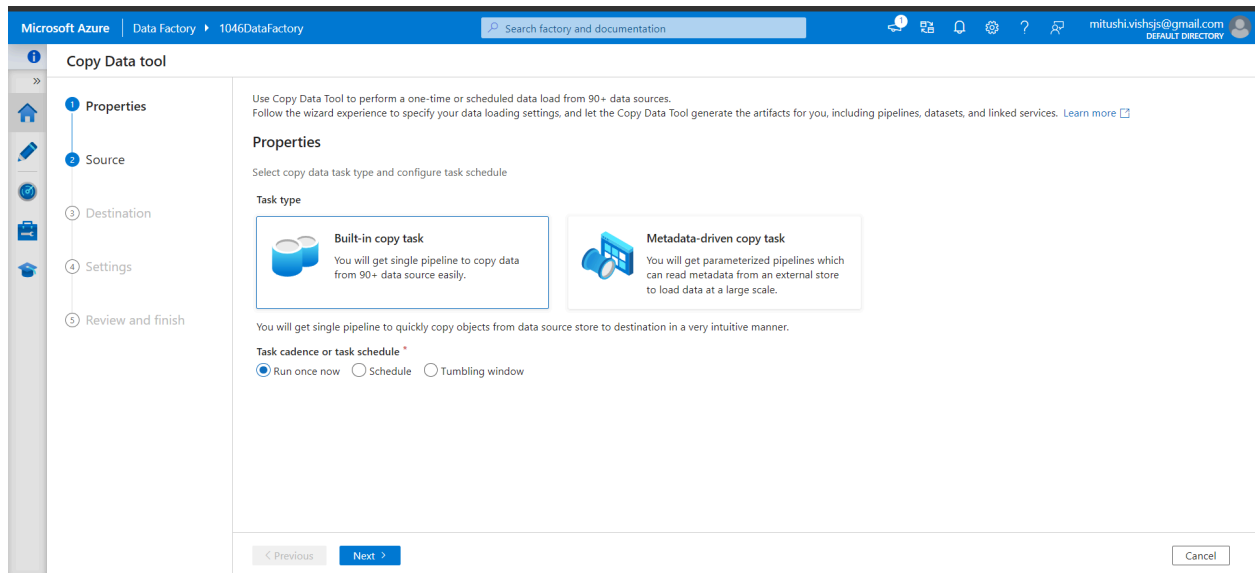
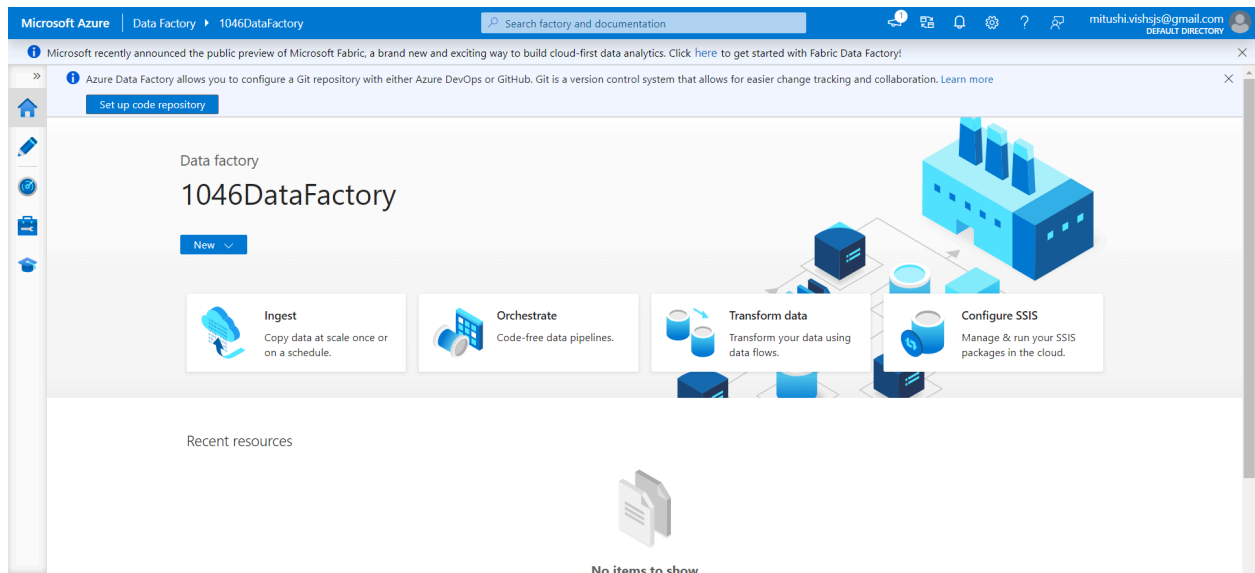
[Go to resource](#)

**Cost management**  
Get notified to stay within your budget and prevent unexpected charges on your bill.  
[Set up cost alerts >](#)

**Microsoft Defender for Cloud**  
Secure your apps and infrastructure  
[Go to Microsoft Defender for Cloud >](#)

**Free Microsoft tutorials**  
[Start learning today >](#)

## Selecting Ingest to copy data :



## Selecting Source data store :

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

mitushi.vishjs@gmail.com  
DEFAULT DIRECTORY

### Copy Data tool

Properties  
Source  
Dataset  
Configuration  
Destination  
Settings  
Review and finish

#### Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new one.

Source type: All

Connection: Select... [+ New connection](#)

< Previous Next >

#### New connection

Search

All Azure Database File Generic protocol NoSQL Services and apps

Amazon Redshift Amazon S3 Amazon S3 Compatible

Apache Impala **Azure Blob Storage** Azure Cosmos DB for MongoDB

Azure Cosmos DB for NoSQL Azure Data Explorer (Kusto) Azure Data Lake Storage Gen1

Continue Cancel

## Linked first storage account :

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

mitushi.vishjs@gmail.com  
DEFAULT DIRECTORY

### Copy Data tool

Properties  
Source  
Dataset  
Configuration  
Destination  
Settings  
Review and finish

#### Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new one.

Source type: All

Connection: AzureBlobStorage1 [Edit](#) [+ New](#)

**File or folder \***  
If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse. Append a slash (/) at the end if the path refers to a folder.

Options  
☐ Binary copy  
☒ Recursively  
☐ Enable partitions discovery

Max concurrent connections

Filter by last modified  
Start time (UTC) End time (UTC)

< Previous Next >

#### Edit linked service

Azure Blob Storage [Learn more](#)

Name: AzureBlobStorage1

Description

Connect via integration runtime: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string Azure Key Account key

Account selection method: Enter manually

Storage account name: storageoneadfcopy

Storage account key:   
Storage account key:

Storage account key:   
Storage account key:

Apply Cancel Test connection

Provided the path for source data store files :

The screenshot shows the 'Copy Data tool' configuration page in Microsoft Azure Data Factory. The left sidebar indicates the 'Source' step is selected. The main panel is titled 'Source data store' and contains the following fields:

- Source type:** A dropdown menu set to 'All'.
- Connection \*:** A dropdown menu set to 'AzureBlobStorage1', with 'Edit' and '+ New connection' links.
- File or folder \*:** A text input field containing 'sourcecontainer/jobs\_in\_data - Copy.csv', with a 'Browse' button.
- Options:** A group of checkboxes including 'Binary copy' (unchecked), 'Recursively' (checked), and 'Enable partitions discovery' (unchecked).
- Max concurrent connections:** A text input field.
- Filter by last modified:** A section with 'Start time (UTC)' and 'End time (UTC)' input fields.

At the bottom of the configuration panel are '< Previous' and 'Next >' buttons. A 'Cancel' button is located at the bottom right of the main panel.

Creating Destination data store connection :

The screenshot shows the 'Copy Data tool' configuration page in Microsoft Azure Data Factory, with the 'Destination' step selected. The main panel is titled 'Destination data store' and contains the following fields:

- Destination type:** A dropdown menu set to 'All'.
- Connection \*:** A dropdown menu set to 'Select...', with a '+ New connection' link.

At the bottom of the configuration panel are '< Previous' and 'Next >' buttons. A 'New connection' dialog is open on the right side of the screen, displaying a grid of available data store connections:

- Azure Blob Storage**
- Azure Cosmos DB for MongoDB**
- Azure Cosmos DB for NoSQL**
- Azure Data Explorer (Kusto)**
- Azure Data Lake Storage Gen1**
- Azure Data Lake Storage Gen2**
- My** (MySQL icon)
- Elephant** (Hadoop icon)
- Triangle** (Amazon S3 icon)

The dialog has a 'Continue' button at the bottom left and a 'Cancel' button at the bottom right.

Linked second storage account in the destination connection :

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

Copy Data tool

Properties

Source

Destination

Dataset

Configuration

Settings

Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: All

Connection: Select... + New connection

New connection

Azure Blob Storage Learn more

Account key

Connection string Azure Key Vault

Account selection method

From Azure subscription Enter manually

Azure subscription

Select all

Storage account name \*

storagesecondadcopy

Additional connection properties

+ New

Test connection

To linked service To file path

Annotations

+ New

Parameters

Advanced

Create Back Test connection Cancel

Provided the path where the files will be copied :

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

Copy Data tool

Properties

Source

Destination

Dataset

Configuration

Settings

Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: All

Connection: AzureBlobStorage2 Edit + New connection

Folder path \*

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

sinkcontainer/ Browse

File name

Copy behavior

Select...

Max concurrent connections

Block size (MB)

Metadata

< Previous Next > Cancel

Review and submit :

The screenshot shows the 'Copy Data tool' configuration in the 'Review and finish' stage. The left sidebar lists the steps: Properties, Source, Destination, Settings, Review and finish (selected), Review, and Deployment. The main area displays a summary of the pipeline: 'You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage.' Below this is a diagram showing data flow from 'Azure Blob Storage' to 'Azure Blob Storage'. The 'Properties' section shows 'Task name' as 'CopyActivity'. The 'Source' section lists 'Connection name' as 'AzureBlobStorage1', 'Dataset name' as 'SourceDataset\_skn', 'Column delimiter' as ',', 'Escape character' as '\\', 'Quote char' as '-', and 'First row as header' as 'true'. Navigation buttons at the bottom include '< Previous', 'Next >', and 'Cancel'.

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

mitushi.vishjs@gmail.com  
DEFAULT DIRECTORY

**Copy Data tool**

Properties

Source

Destination

Settings

**Review and finish**

Review

Deployment

**Summary**

You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage.

Azure Blob Storage → Azure Blob Storage

**Properties** [Edit](#)

Task name: CopyActivity

Task description

**Source** [Edit](#)

Connection name: AzureBlobStorage1

Dataset name: SourceDataset\_skn

Column delimiter: ,

Escape character: \

Quote char: -

First row as header: true

< Previous Next > Cancel

Data Copied successfully :

The screenshot shows the 'Deployment complete' screen of the 'Copy Data tool'. The left sidebar lists the steps: Properties, Source, Destination, Settings, Review and finish (selected), Review, and Deployment. The main area displays 'Deployment complete' with a table of deployment steps and their status. The table shows four steps: 'Validating copy runtime environment', 'Creating datasets', 'Creating pipelines', and 'Running pipelines', all with a status of 'Succeeded'. Below the table, a message states: 'Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.' At the bottom, there are buttons for 'Finish', 'Edit pipeline', and 'Monitor'.

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

mitushi.vishjs@gmail.com  
DEFAULT DIRECTORY

**Copy Data tool**

Properties

Source

Destination

Settings

**Review and finish**

Review

Deployment

**Deployment complete**

Deployment step	Status
Validating copy runtime environment	✓ Succeeded
> Creating datasets	✓ Succeeded
> Creating pipelines	✓ Succeeded
> Running pipelines	✓ Succeeded

Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

Finish Edit pipeline Monitor



Checking Copied data file :

Microsoft Azure

Upgrade

Search resources, services, and docs (G+)

mitushi.vishjs@gmail.c...

DEFAULT DIRECTORY

Home > storagesecondadcopy > Containers >

sinkcontainer

Container

Search

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Create snapshot

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

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

Location: sinkcontainer

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> jobs_in_data - Copy.csv	21/2/2024, 12:30:01 ...	Hot (Inferred)		Block blob	1.22 KiB	Available	...

https://portal.azure.com/#