

Asmita Porwal
Data Engineering
Batch-1
21/02/2024

Coding Challenge -4 Question-3

Execute & explain, Azure datafactory and its copy activity.

Azure Data Factory (ADF):

Azure Data Factory is a cloud-based data integration service provided by Microsoft Azure.

It allows you to create, schedule, and manage data pipelines that can move data between supported on-premises and cloud-based data stores.

ADF enables you to compose data storage, processing, and movement services into streamlined, scalable, and reliable data production pipelines.

Key components of Azure Data Factory include:

Pipelines: A collection of data processing activities, including data movement and transformation, that are executed sequentially.

Datasets: Represent the data structures within data stores, such as tables, files, or folders. Datasets define the schema and location of the data.

Linked Services: Define the connection information to connect to external data stores or compute services.

Activities: The processing steps in a pipeline, such as copying data or transforming data.

Copy Activity:

Copy Activity is one of the key activities within Azure Data Factory.

It allows you to copy data from one data store to another.

This could involve moving data between on-premises and cloud storage, copying data between cloud data stores, or performing transformations during the copy process.

Key features of Copy Activity:

Source and Destination: You define the source and destination linked services for the data copy operation. These linked services contain the connection information for the data stores.

Mappings: You can specify mappings to transform data during the copy process. This is useful when the schema of the source and destination data is different.

Data Movement Modes: Copy Activity supports different data movement modes, including bulk insert, table-to-table insert, and stored procedure execution.

Data Integration Runtime: The copy operation is executed by a Data Integration Runtime (DIR), which is an infrastructure responsible for moving data securely and efficiently across network boundaries.

Monitoring and Logging: Azure Data Factory provides monitoring and logging capabilities, allowing you to track the progress and status of your data copy operations.

Created the container inside the storage

The screenshot shows the Azure Storage portal interface for a storage account named 'storage1060'. The left sidebar contains a navigation menu with options like Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Data storage (highlighted), Containers (selected), File shares, Queues, Tables, Security + networking, and Networking. The main content area is titled 'storage1060 | Containers' and includes a search bar and a table of containers. The table has columns for Name, Last modified, Anonymous access level, and Lease state. Two containers are listed: '\$logs' and 'containerofstorage1060', both with a Private access level and Available lease state.

Name	Last modified	Anonymous access level	Lease state
\$logs	16/02/2024, 14:09:54	Private	Available
containerofstorage1060	21/02/2024, 09:57:08	Private	Available

Uploaded the files Marks_data.csv

Home > storage1060 | Containers >

containerofstorage1060 Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: containerofstorage1060

Search blobs by prefix (case-sensitive) ☐ Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> output					
<input type="checkbox"/> Marks_data.csv	21/02/2024, 09:57:35	Hot (Inferred)		Block blob	88 B

Step 1: Start the copy data Tool

1. On the home page of Azure Data Factory, select the **Ingest** tile to start the Copy Data tool.

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!

Data factory

datafactory1060

New

Ingest
Copy data at scale once or on a schedule.

Orchestrate
Code-free data pipelines.

Transform data
Transform your data using data flows.

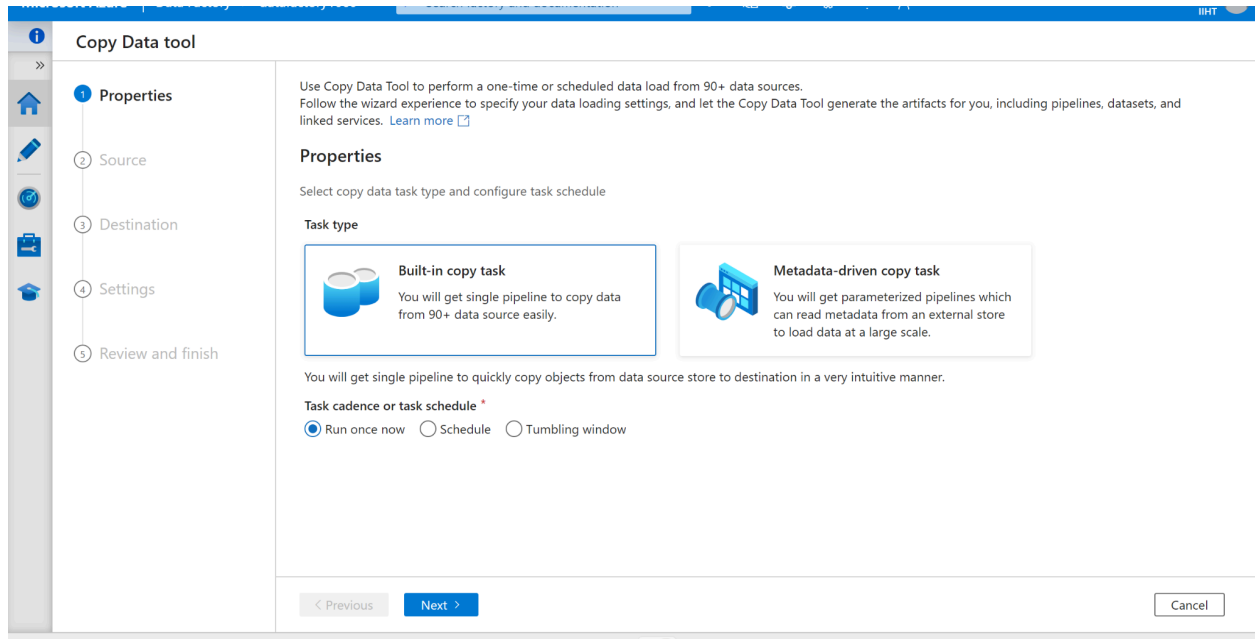
Configure SSIS
Manage & run your SSIS packages in the cloud.

Recent resources

No items to show

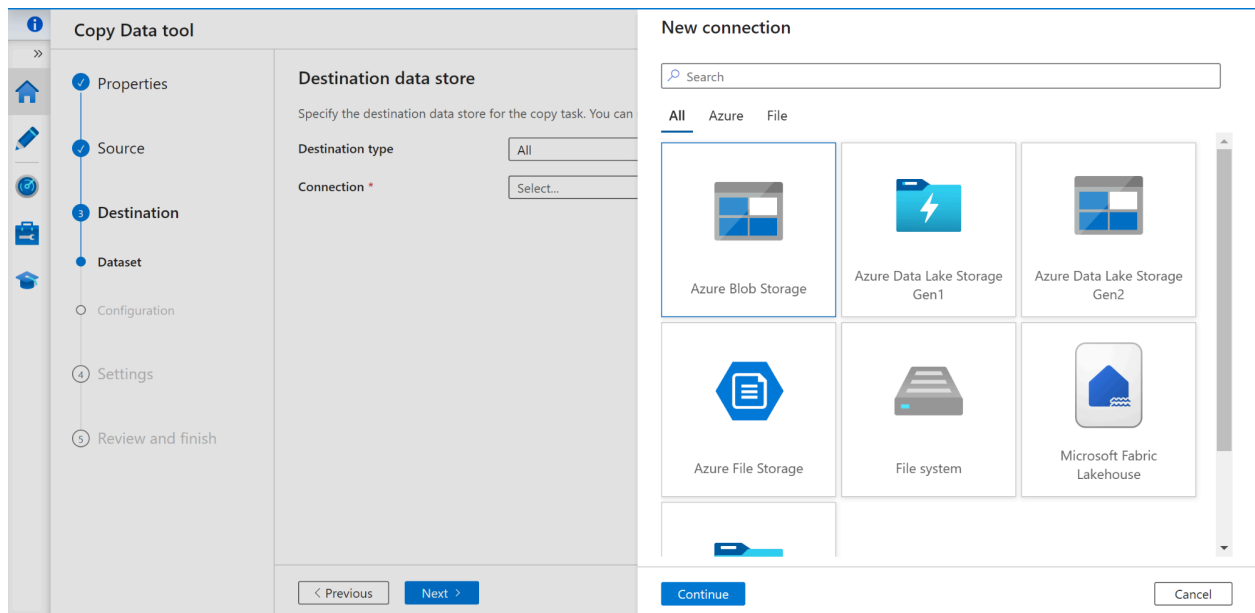
Your recently opened resources will show up here.

2. On the Properties page of the Copy Data tool, choose **Built-in** copy task under Task type, then select Next.



Step 2: Complete source configuration

1. Click + Create new connection to add a connection.
2. Select the linked service type that you want to create for the source connection. In this tutorial, we use Azure Blob Storage. Select it from the gallery, and then select Continue.



3. On the New connection (Azure Blob Storage) page, specify a name for your connection. Select your Azure subscription from the Azure subscription list and your storage account from the Storage account name list, test connection, and then select Create.

4. Select the newly created connection in the Connection block.

5. In the File or folder section, select Browse to navigate to the **containerofstorage1060/Marks_data.csv** folder, select the **Marks_data.csv** file, and then click OK.

6. Select the Binary copy checkbox to copy file as-is, and then select Next.

The screenshot shows the 'Copy Data tool' configuration window. On the left is a vertical sidebar with a list of steps: Properties, Source (highlighted with a blue circle and dot), Dataset, Configuration, Destination, Settings, and Review and finish. The main area is titled 'Source data store' and contains the following fields and options:

- Source type:** A dropdown menu with 'All' selected.
- Connection *:** A dropdown menu with 'AzureBlobStorage1' selected, followed by 'Edit' and '+ New connection' links.
- File or folder *:** A text input field containing 'containerofstorage1060/Marks_data.csv' and a 'Browse' button.
- Options:**
 - ☒ Binary copy ⓘ
 - Compression type:** A dropdown menu with 'None' selected.
 - ☒ Recursively ⓘ
 - ☐ Delete files after completion ⓘ
 - Max concurrent connections ⓘ:** An empty text input field.

At the bottom of the window are three buttons: '< Previous', 'Next >' (highlighted in blue), and 'Cancel'.

Step 3: Complete destination configuration

1. Select the AzureBlobStorage connection that you created in the Connection block.
2. In the Folder path section, enter **containerofstorage1060/output** for the folder path.

Copy Data tool

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: AzureBlobStorage1

Folder path: containerofstorage1060/output

File name:

Compression type: None

Copy behavior: Select...

Max concurrent connections:

< Previous Next > Cancel

3. Leave other settings as default and then select Next.

Step 4: Review all settings and deployment

1. On the Settings page, specify a name for the pipeline and its description, then select Next to use other default configurations.

Copy Data tool

Settings

Enter name and description for the copy data task, more options for data movement

Task name: ADFHelloDemov

Task description:

Data consistency verification: ☐

Enable logging: ☐

Enable staging: ☐

> Advanced

< Previous Next > Cancel

2. On the Summary page, review all settings, and select Next.

3. On the Deployment complete page, select Monitor to monitor the pipeline that you created.

Copy Data tool

»

Properties

Source

Destination

Settings

Review and finish

Review

Deployment

Azure Blob Storage

→

Azure Blob Storage

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

Step 5: Monitor the running results

Output folder created and that marks_data.csv file is copied

Home > storage1060 | Containers >

containerofstorage1060

Container

Search

«

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: containerofstorage1060

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> output					
<input type="checkbox"/> Marks_data.csv	21/02/2024, 09:57:35	Hot (Inferred)		Block blob	88 B

Home > storage1060 | Containers >

containerofstorage1060

Container

Search

«

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: containerofstorage1060 / output

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> [.]					
<input type="checkbox"/> Marks_data.csv	21/02/2024, 10:01:44	Hot (Inferred)		Block blob	88 B