

Assignment-22

Name- Ajay Chaudhary

Batch-Data engineering(Batch-1)

Handwritten notes-

What is Azure data factory ?

Azure data factory is a cloud based data integration service that allows you to create data driven workflow in the cloud for orchestrating & automating data movement & data transformation.

⇒ ADF does not store any data itself.

⇒ It allows you to monitor & manage workflow using both programmatic & UI mechanism.

ADF can be used for -

- Supporting data migration
- Getting data from a client server or online data to an ~~ADF~~ ADF.
- Carrying out various data integrated process.
- Integrating data from different ERP systems & loading it into Azure Synapse for reporting.

How does ADF work ?

The data factory service allow you to create pipeline that move & transform data & then run the pipelines on a specified schedule (hourly, weekly, monthly)

Azure Data Factory can perform in 3 steps -

- 1). Step 1 → Connect & collect
Connect to all the required source of data & processing such as SaaS service.
- 2). Step 2 → Transform & enrich
Once data present in centralized storage it is transformed using complete compute service such as HDInsight, MapReduce, Spark
- 3). Step 3 → Publish
Deliver transformed data from the cloud to on-premise.

Azure data factory key component ⇒

It has 4 key component that work together to define input & output data, processing events & the scheduled & resources required to execute the desired data flow.

- 1). Dataset represent data structure within the data store.
An input dataset represents the input for an activity in the pipeline.

An output dataset represent the output for the activity.

Example → An azure blob dataset operation the blob container & folder in the Azure blob storage from which the pipeline should read the data.

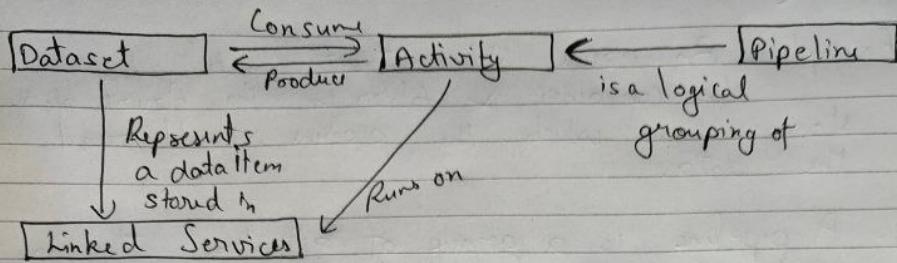
- 2). A pipeline is a group of activities. They are used to group activity into a unit that together perform a task.

A data factory may have one or more pipelines. Example, a pipeline could contain a group of activities that ingest data from an Azure blob & then runs a Hive query on an HDInsight cluster to partition the data.

- 3). Activities define the actions to perform on your data —
Currently, Azure Data factory supports two types of activities, data movements & data transform.

- 4). Linked services define the informed needed for Azure data factory to connect to external resources.

How ADF component work together -

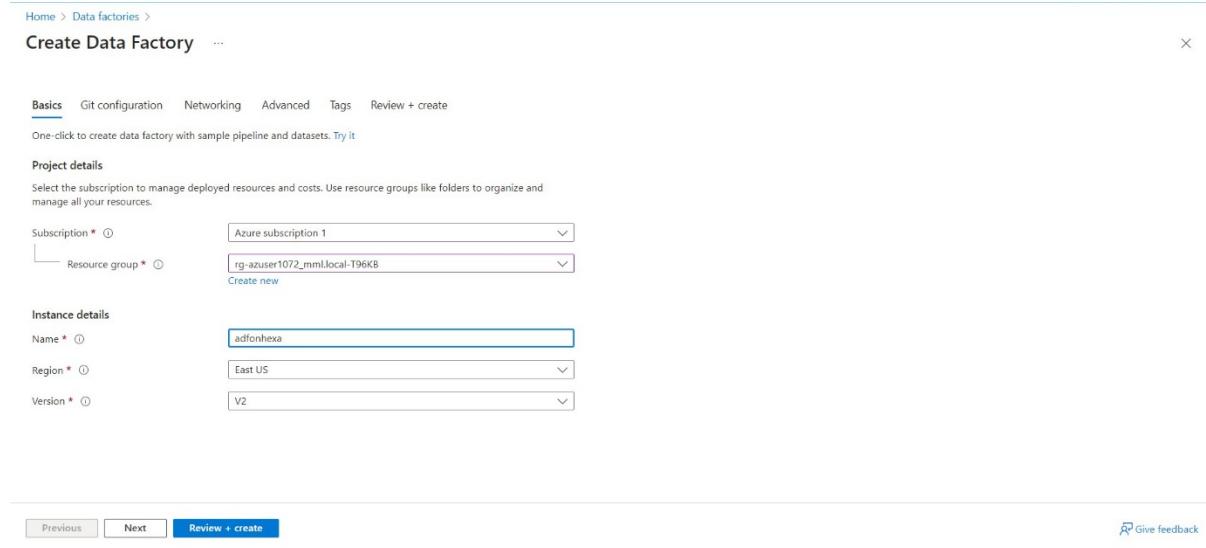
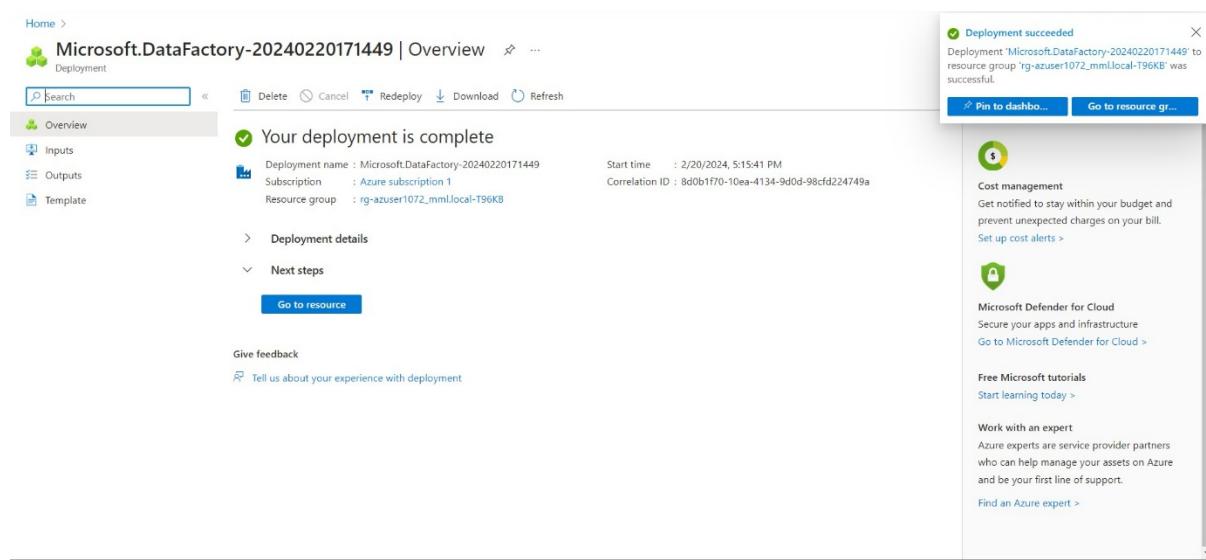


You can use one of the following tools or APIs to create data pipeline in Azure Data factory

- Azure portal
- Visual Studio
- Powershell
- .NET API
- REST API
- Azure resources management template.

Copy Activity:

1. Create a data factory

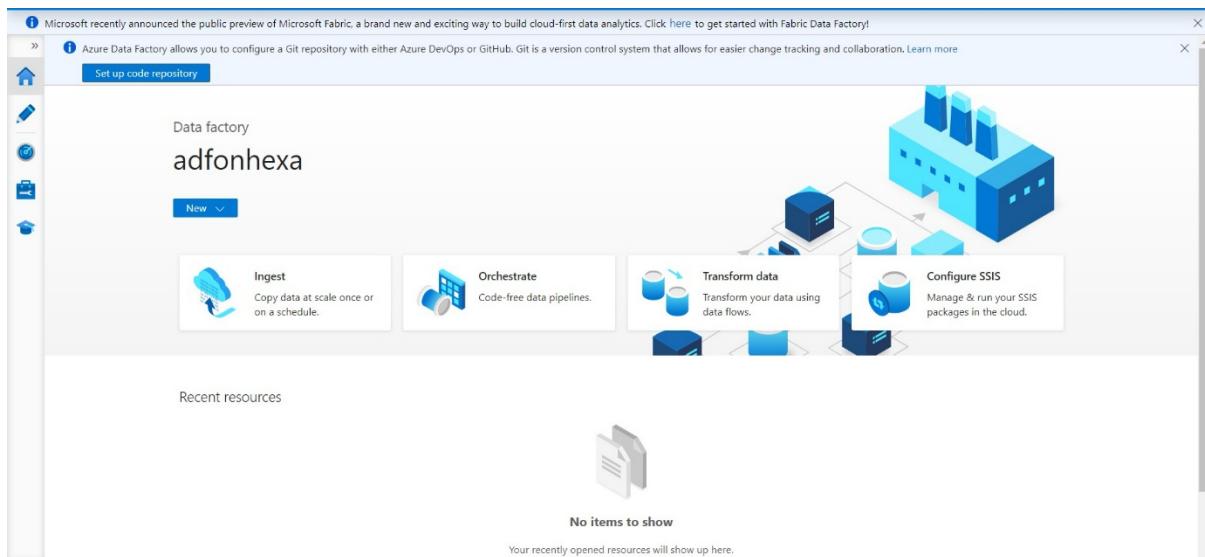



Use the copy data tool to copy data

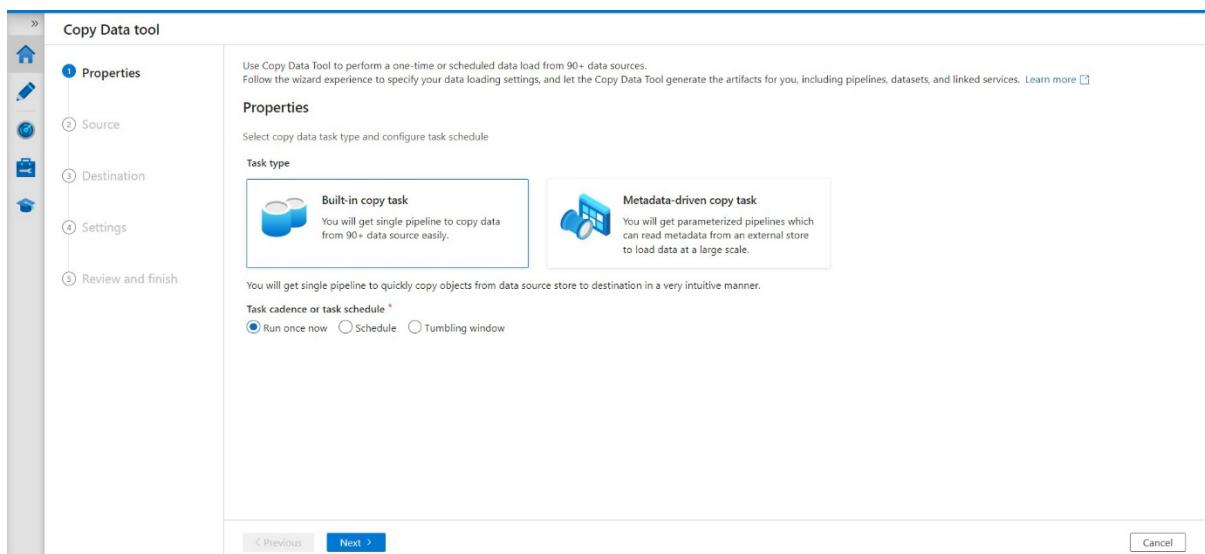
The steps below will walk you through how to easily copy data with the copy data tool in Azure Data Factory.

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.



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**.

Home > Storage accounts >

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review

Resource group * rg-azuser1072_mml.local-T96KB [Create new](#)

Instance details

Storage account name * 1ststorageacc

Region * (Asia Pacific) Central India [Deploy to an edge zone](#)

Performance * Standard: Recommended for most scenarios (general-purpose v2 account) Premium: Recommended for scenarios that require low latency.

Redundancy * Geo-redundant storage (GRS) Make read access to data available in the event of regional unavailability.

[Review](#) [< Previous](#) [Next : Advanced >](#) [Give feedback](#)

Home >

1ststorageacc_1708429817235 | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Deployment succeeded
Deployment '1ststorageacc_1708429817235' to resource group 'rg-azuser1072_mml.local-T96KB' was successful.

Go to resource Pin to dashboard

Overview

Your deployment is complete

Deployment name: 1ststorageacc_1708429817235
Subscription: Azure subscription 1
Resource group: rg-azuser1072_mml.local-T96KB

Start time: 2/20/2024, 5:20:21 PM
Correlation ID: 8be48lba-3021-4b10-bf20-1e11c6a25eb4

Deployment details

Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

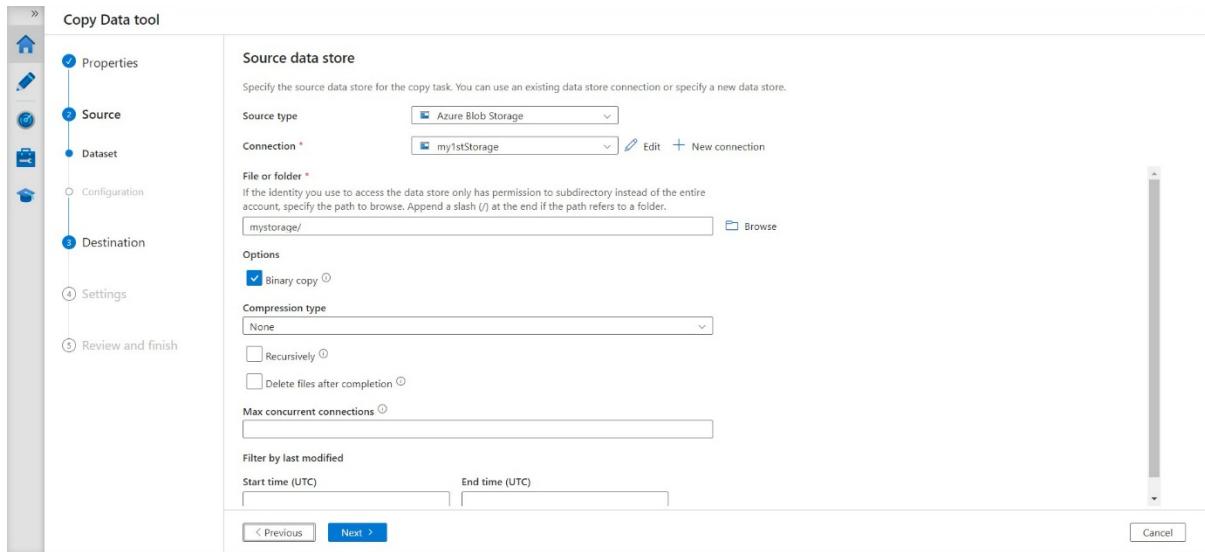
Cost Management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >

Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

Free Microsoft tutorials
Start learning today >

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >

1. 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**.
 1. Select the newly created connection in the **Connection** block.
 2. In the **File or folder** section, select **Browse** to navigate to the **mystorage/input** folder, select the file, and then click **OK**.
 3. Select the **Binary copy** checkbox to copy file as-is, and then select **Next**.



Deployment succeeded

Deployment '2ndstoageacc_1708430325923' to resource group 'rg-azuser1072_mmilocal-T96KB' was successful.

[Go to resource](#) [Pin to dashboard](#)

Your deployment is complete

Deployment name: 2ndstoageacc_1708430325923
Subscription: Azure subscription 1
Resource group: rg-azuser1072_mmilocal-T96KB

Start time: 2/20/2024, 5:28:50 PM
Correlation ID: 84b255ba-98a3-4132-81ac-df18466da5ce

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 >](#)

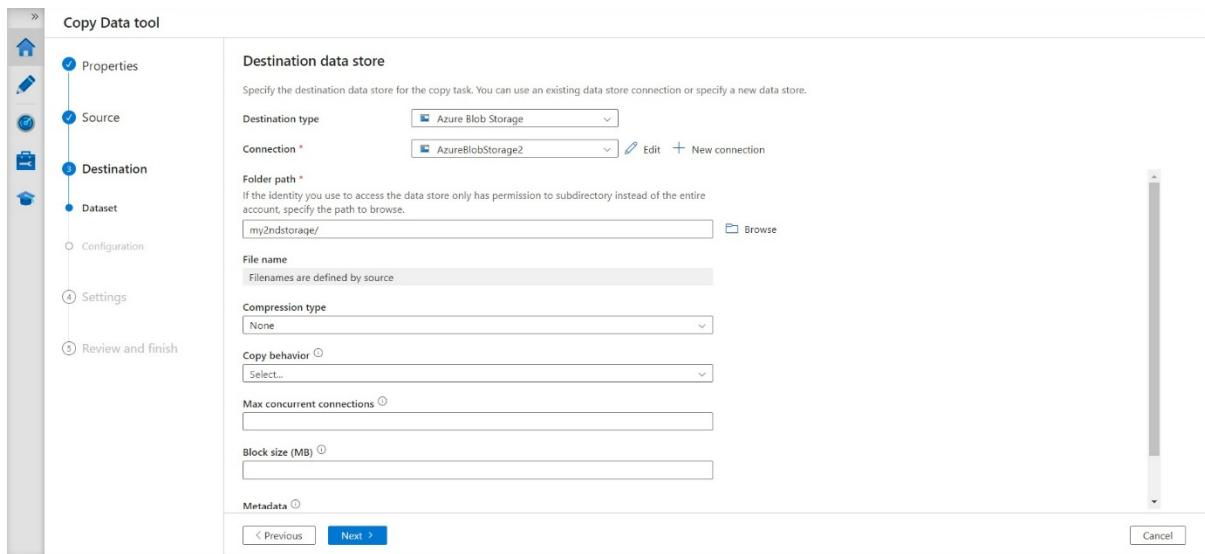
Microsoft Defender for Cloud
Secure your apps and infrastructure.
[Go to Microsoft Defender for Cloud >](#)

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

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
[Find an Azure expert >](#)

Step 3: Complete destination configuration

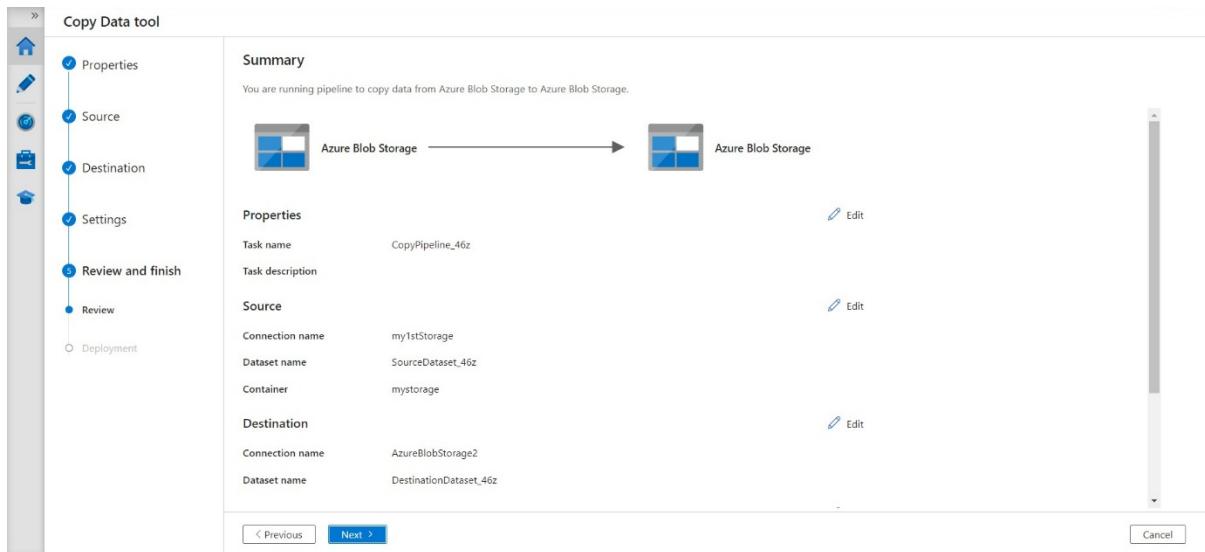
1. Select the **AzureBlobStorage** connection that you created in the **Connection** block.
2. In the **Folder path** section, enter **my2ndstorage/output** for the folder path.

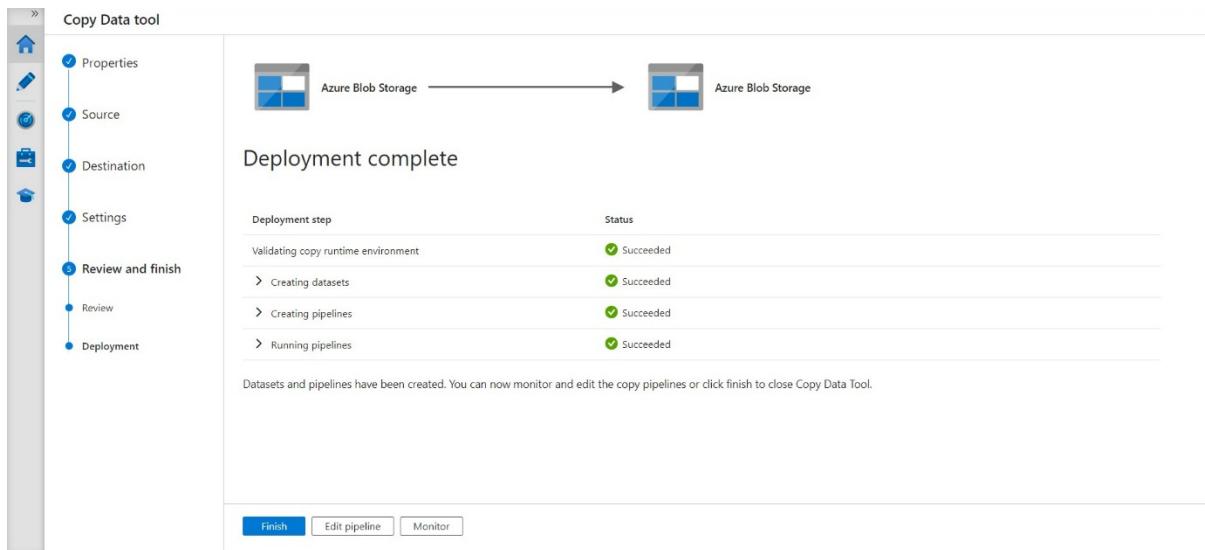


Step 4: Review all settings and deployment

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

1. On the **Summary** page, review all settings, and select **Next**.
2. On the **Deployment complete** page, select **Monitor** to monitor the pipeline that you created.





Step 5: Monitor the running results

1. The application switches to the **Monitor** tab. You see the status of the pipeline on this tab. Select **Refresh** to refresh the list. Click the link under **Pipeline name** to view activity run details or rerun the pipeline.

The screenshot shows the 'Monitor' tab of the application. The left sidebar has a 'Runs' section with 'Pipeline runs' selected, showing a list of runs. The main area is titled 'All pipeline runs > CopyPipeline_46z - Activity runs'. It includes buttons for 'Rerun', 'Cancel', 'Refresh', 'Update pipeline', and tabs for 'List' (selected) and 'Gantt'. A summary box for 'Copy data' shows 'Copy_46z' with a green checkmark and a 'refresh' icon. Below this is a table titled 'Activity runs' with one item: 'Copy_46z' (Status: Succeeded, Type: Copy data, Run start: 2/20/2024, Duration: 15s, Integration runtime: AutoResolveIntegration, User properties: my2ndstorage//). There are also links for 'Monitor in Azure Metrics' and 'Export to CSV'.

my2ndstorage 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)

Add filter

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: my2ndstorage

Search blobs by prefix (case-sensitive)

Show deleted blobs

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
student-dataset.csv	2/20/2024, 5:40:08 PM	Hot (Inferred)		Block blob	26.79 KiB	Available

Shared access tokens Access policy Properties Metadata

<https://portal.azure.com/#>