

Azure Databricks Assignment - 7

Mitushi Vishwakarma

NOTES :

In strategy it is important to see distant things to get a close view

Azure Databricks

Day - 7

Azure Data factory

- cloud based data integration that allows to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.
- doesn't store data itself.
- allows to monitor and manage workflows using both programmatic and UI.

Use cases:-

- Supporting data migration
- getting data from a client's server or online data to an Azure Data Lake
- Integrating data from different systems and loading it into Azure Synapse for reporting

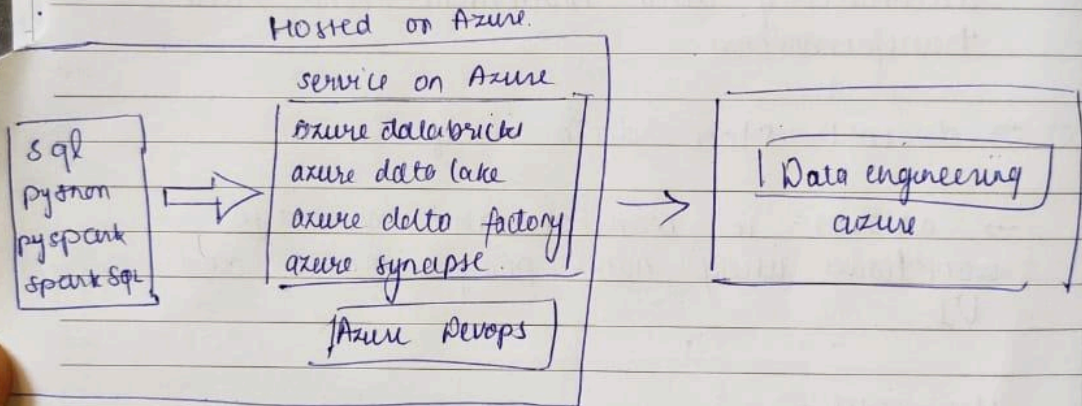
ERP (Enterprise Resource Planning)

14-11-2023
TUESDAY
WK 46 • DAY 318-047
14

44	45	46	47	48
W	T	F	S	S
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31				

How Azure Data Factory work?

- Allows you to create data pipeline that move and transform data and specified schedule.
- Data is produced by workflows in time slice data and we can specify the pipeline made as scheduled.



Step 1: Connect and Collect:

- Connect all the required sources of data and processing such as SaaS services, file shares, FTP and web services

Step 2: Transform and Enrich

- Once data is present in a

13-11-2023

13

NOVEMBER

WK 46 • DAY 317-048

MONDAY

2023

52	25	26	27	28	29	30	31
51	18	19	20	21	22	23	24
50	11	12	13	14	15	16	17
49	4	5	6	7	8	9	10
48	1	2	3				
WK	M	T	W	T	F	S	S
DECEMBER	2023						

26
18
12
5
S
2023

Centralized data store in the cloud. It is transformed using compute services such as HDInsight Hadoop, Spark, Azure data Lake Analytics.

12 SUNDAY

Step 3: Publish:-

→ Cloud → on premise sources like SQL Server.

Key Components

Four components that work together to define input and output data, processing events.

1. Datasets represent data structure within the data store.

Input dataset - Input for an activity in pipeline

Output dataset - output for the activity.

48	27	28	29	30
47	20	21	22	23
46	13	14	15	16
45	6	7	8	9
44	1	2	3	4
43				
42				
41				
40				
39				
38				
37				
36				
35				
34				
33				
32				
31				
30				
29				
28				
27				
26				
25				
24				
23				
22				
21				
20				
19				
18				
17				
16				
15				
14				
13				
12				
11				
10				
9				
8				
7				
6				
5				
4				
3				
2				
1				
S				
S				
F				
M				
W				
Th				
Fr				
Sa				
Su				

2023
SATURDAY
WK 45 • DAY 315-050
NOVEMBER

11-11-2023

2. A pipeline is a group of activities -

→ Group activities into a unit that together perform a task.

→ may have one or more pipelines.

3. Activities defines the actions to perform on your data.

→ data movement

→ data transformation

4. Linked Services define the information, needed for Azure data factory to connect the external resources.

Copy Activity

Copy data from data source (source) to Sink data store.

following tools or APIs can be used to create data pipeline in ADF.

Azure portal

.NET API

Visual Studio

PowerShell

10-11-2023

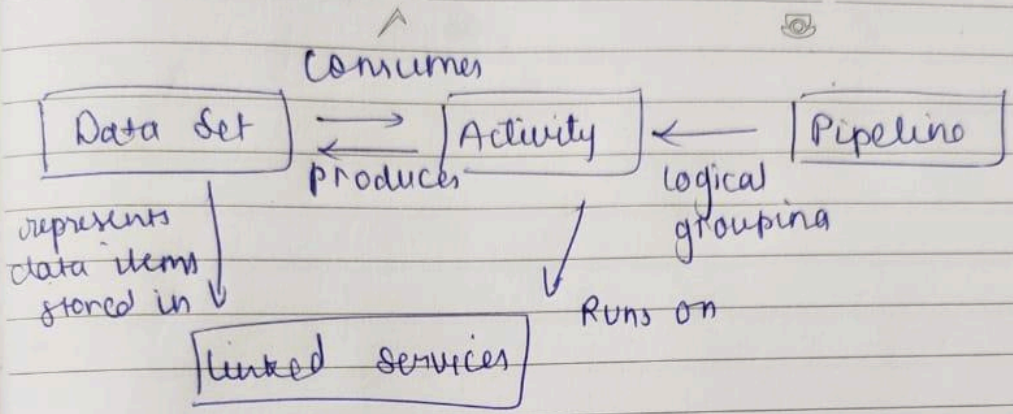
NOVEMBER

WK 45 • DAY 31/4/051

FRIDAY

48 49 50 51 52
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

The system of domination is founded on depriving nations of their true identity



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 shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, an Upgrade button, a search bar, and user information for 'mitushi.vishajp@gmail.com'. The main content area displays the 'Overview' page for a deployment named 'storageoneadfcopy_1708495292918'. The deployment is marked as 'complete' with a green checkmark. Key details include: Deployment name: storageoneadfcopy_1708495292918, 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. On the right, there are sections for 'Cost Management' and 'Microsoft Defender for Cloud'. The left sidebar shows navigation options like Overview, Inputs, Outputs, and Template.

Created container in storage account :

The screenshot shows the 'Containers' page for the 'storageoneadfcopy' storage account. A notification at the top right states 'Successfully created storage container' for 'sourcecontainer'. The main area lists containers with columns: Name, Last modified, Anonymous access level, and Lease state. Two containers are listed: '\$logs' and 'sourcecontainer', both created on 21/2/2024 at 11:32:15 am and 11:34:05 am respectively, with 'Private' access and 'Available' lease state. The left sidebar shows navigation options like Overview, Activity log, Tags, and Diagnose and solve problems.

Uploaded file in container :

The screenshot shows the 'sourcecontainer' page. A notification at the top right states 'Successfully uploaded blob(s)' for 'sourcecontainer'. The main area lists blobs with columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. One blob is listed: 'jobs_in_data - Copy.csv', uploaded on 21/2/2024 at 11:47:45, with 'Hot (Inferred)' access tier, 'Block blob' type, and '1.22 KiB' size. The left sidebar shows navigation options like Overview, Diagnose and solve problems, Access Control (IAM), and Settings.

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

The screenshot shows the 'Overview' page for a deployment named 'storageesecondadfcopy_1708496462127'. The deployment is complete, with a green checkmark icon. The deployment details show the name, subscription (Free Trial), resource group (ADFCopyActivity), start time (21/2/2024, 11:51:08 am), and correlation ID (e064eb27-e2bc-43e8-b686-28f0dd2d657a). The 'Next steps' section includes a 'Go to resource' button. The left sidebar shows the navigation menu with 'Overview', 'Inputs', 'Outputs', and 'Template'. The right sidebar contains links for 'Cost Management', 'Microsoft Defender for Cloud', and 'Free Microsoft tutorials'.

Created container in storage account :

The screenshot shows the 'Containers' page for the storage account 'storageesecondadfcopy'. The page displays a table of containers. The table has columns for 'Name', 'Last modified', 'Anonymous access level', and 'Lease state'. There are two containers listed: '\$logs' and 'sinkcontainer'. The 'sinkcontainer' is the one created. The left sidebar shows the navigation menu with 'Overview', 'Activity log', 'Tags', 'Diagnose and solve problems', 'Access Control (IAM)', 'Data migration', and 'Events'. The right sidebar contains links for 'Cost Management', 'Microsoft Defender for Cloud', and 'Free Microsoft tutorials'.

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' container details page. The page displays the 'Overview' tab, which includes the 'Authentication method' (Access key) and 'Location' (sinkcontainer). The 'Search blobs by prefix' field is empty. The 'Settings' section on the left includes 'Shared access tokens', 'Access policy', 'Properties', and 'Metadata'. The right sidebar contains links for 'Cost Management', 'Microsoft Defender for Cloud', and 'Free Microsoft tutorials'.

Created Data Factory named “1046DataFactory” :

The screenshot shows the 'Overview' page for a newly created Azure Data Factory. The deployment is complete. Key details include:

- Deployment name:** Microsoft.DataFactory-20240221115758
- Subscription:** Free Trial
- Resource group:** ADFCopyActivity
- Start time:** 21/2/2024, 11:59:59 am
- Correlation ID:** 42df5b04-fbb5-4b93-9f06-cf64f2d53f3

On the right, there are promotional tiles for 'Cost management', 'Microsoft Defender for Cloud', and 'Free Microsoft tutorials'.

Selecting Ingest to copy data :

The screenshot shows the '1046DataFactory' overview page. The 'Ingest' option is highlighted, which is described as 'Copy data at scale once or on a schedule.' Other options visible include 'Orchestrate', 'Transform data', and 'Configure SSIS'.

The screenshot shows the 'Copy Data tool' configuration wizard. The 'Properties' step is selected in the left-hand navigation pane. The main content area shows the 'Task type' selection, with 'Built-in copy task' chosen. Below this, the 'Task cadence or task schedule' is set to 'Run once now'.

Task type options:

- Built-in copy task:** You will get single pipeline to copy data from 90+ data source easily.
- Metadata-driven copy task:** You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

Task cadence or task schedule:

- ☒ Run once now
- ☐ Schedule
- ☐ Tumbling window

Navigation buttons at the bottom include '< Previous', 'Next >', and 'Cancel'.

Selecting Source data store :

Microsoft Azure | Data Factory | 1046DataFactory

Search factory and documentation

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

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

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)

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: [masked]

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:** Two text input fields for 'Start time (UTC)' and 'End time (UTC)'.

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

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

- 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)
- My (PostgreSQL icon)
- My (Amazon Redshift 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

mitushi.vishjs@gmail.com
DEFAULT DIRECTORY

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

mitushi.vishjs@gmail.com
DEFAULT DIRECTORY

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

1

Block size (MB)

1

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

Summary

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

Azure Blob Storage → Azure Blob Storage

Properties

Task name: CopyActivity

Task description

Source

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 the title 'Deployment complete'. Below this is a table showing the deployment steps and their status. The table has two columns: 'Deployment step' and 'Status'. The steps listed are 'Validating copy runtime environment', 'Creating datasets', 'Creating pipelines', and 'Running pipelines', all of which are marked as '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 three buttons: 'Finish', 'Edit pipeline', and 'Monitor'.

Microsoft Azure | Data Factory | 1046DataFactory | Search factory and documentation | mitushi.vishjs@gmail.com | DEFAULT DIRECTORY

Copy Data tool

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/#