

**NAME : AKULA SHARATH CHANDRA**

**BATCH: DATA ENGINEERING**

**DATE :20-02-2024**

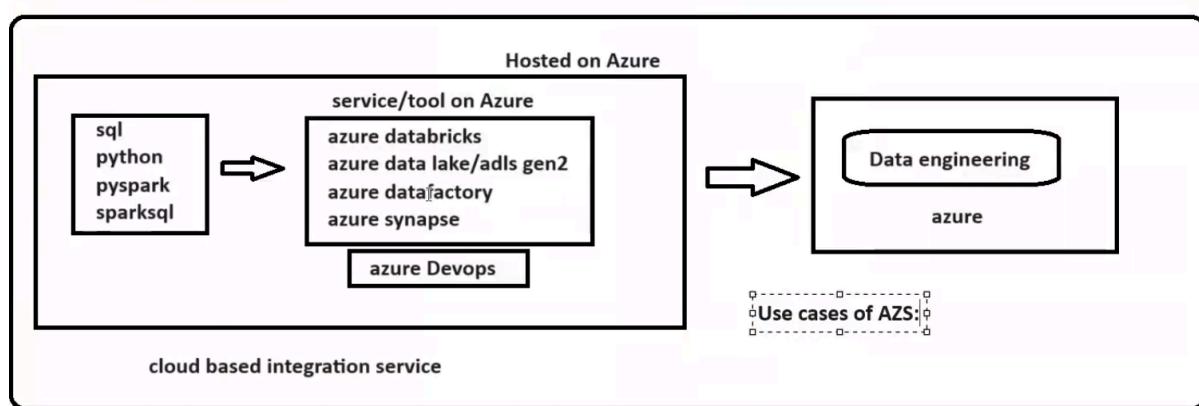
**TOPIC : AZURE DATA FACTORY**

**1) DATA FACTORY:** 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.

## **2) ADF USE CASES:**

ADF can be used for:

- Supporting data migrations
- Getting data from a client's server or online data to an Azure Data Lake.
- Carrying out various data integration processes.
- Integrating data from different ERP systems and loading it into Azure.
- Synapse for reporting.



## **3) HOW DOES ADF WORKS:**

The Data Factory service allows you to create data pipelines that move and transform data and then run the pipelines on a specified schedule (hourly, daily, weekly, etc.).

It works on 3 steps

## **Step 1: Connect and Collect**

Connect to all the required sources of data and processing such as SaaS services, file shares, FTP, and web services. Then, move the data as needed to a centralized location for subsequent processing by using the Copy Activity in a data pipeline to move data from both on-premise and cloud source data stores to a centralization data store in the cloud for further analysis.

## **Step 2: Transform and Enrich**

Once data is present in a centralized data store in the cloud, it is transformed using compute services such as HDInsight Hadoop, Spark, Azure Data Lake Analytics, and Machine Learning.

## **Step 3: Publish**

Deliver transformed data from the cloud to on-premise sources like SQL Server or keep it in your cloud storage sources for consumption by BI and analytics tools and other applications.

**4) DATA MIGRATION:**Data migration occurs between two cloud data stores and between an on-premise data store and a cloud data store.

### **COPY ACTIVITY:**

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.

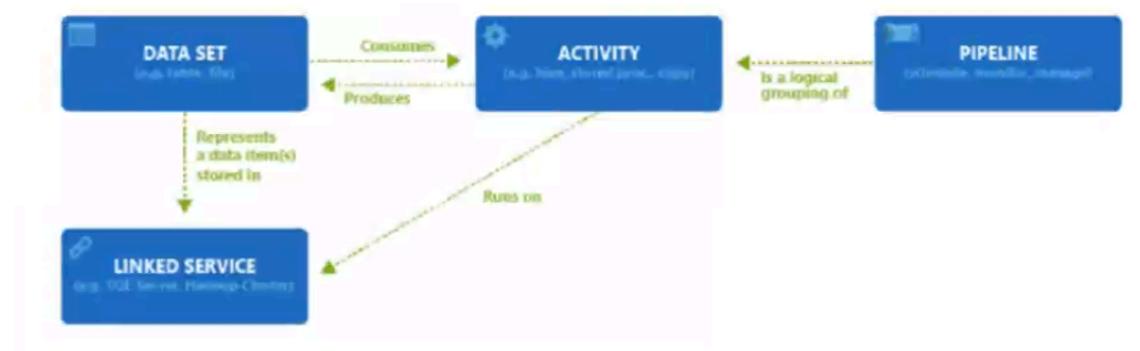
→ Azure Data Factory supports transformation activities such as Hive, MapReduce, Spark, etc that can be added to pipelines either individually or chained with other activities.

## 5) AZURE DATA FACTORY KEY COMPONENTS:

Azure Data Factory has four key components that work together to define input and output data, processing events, and the schedule and resources required to execute the desired data flow.

- >**Datasets represent data structures within the data stores:** Input datasets provide activity input in a pipeline, while output datasets represent activity output. For instance, an Azure Blob dataset specifies the source in Blob Storage for reading, and an Azure SQL Table dataset designates the table for writing activity output.
- >**A pipeline is a group of activities:** A pipeline in Azure Data Factory is a collection of activities that collaboratively execute a task. Multiple pipelines can exist in a data factory.
- >**Activities define the actions to perform on your data:** Activities, specifying data actions, come in two types: data movement and data transformation.
- >**Linked services define the information needed for Azure Data Factory to connect to external resources:** Linked services provide the necessary information, like connection strings, for Data Factory to connect to external resources, such as Azure Storage.

## 6) HOW DOES DATA FACTORY COMPONENTS WORKS



We Can use one of the following tools or APIs to create data pipelines in Azure Data Factory:

→LINKED SERVICE

→Azure portal

→Visual Studio

→PowerShell

→NET API

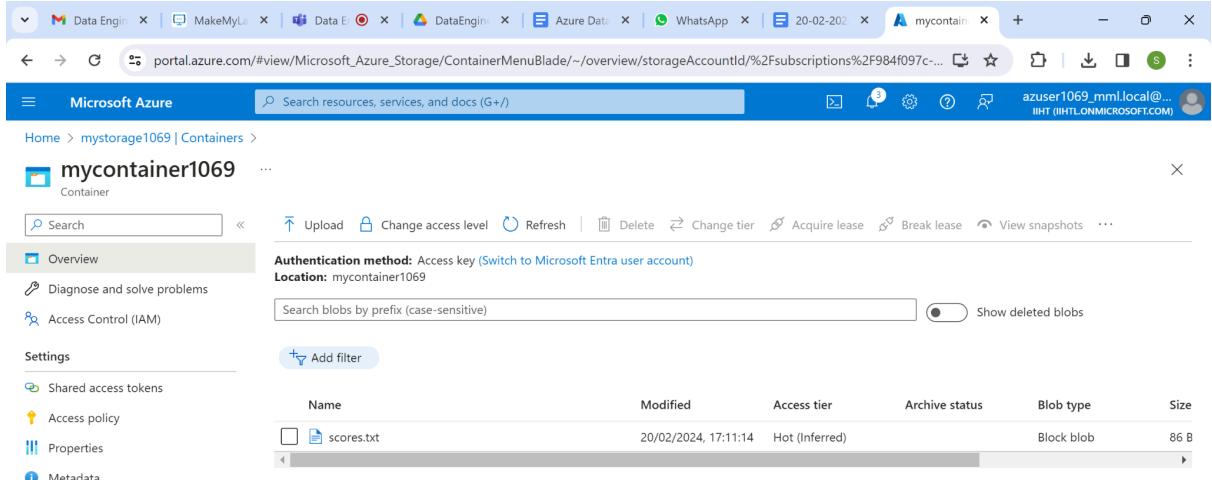
→Azure Resource Manager template

## 7)CREATING A DATA FACTORY ACCOUNT:

→ to create a data factory account firstly we need to create two storage accounts one is for source group and one is for destination group.

The screenshot shows the Microsoft Azure portal with a deployment overview page. The title is "mystorage1069\_1708429139385 | Overview". The main message is "Your deployment is complete". Deployment details include: Deployment name: mystorage1069\_170842..., Start time: 20/02/2024, 17:09:02, Subscription: Azure subscription 1, Correlation ID: e6243ef6-468d-4182-9b55-0a9ed602e2l, and Resource group: rg-azuser1069\_mml.local... . There are links for "Deployment details" and "Next steps", and a "Go to resource" button. Below the main message, there are links for "Give feedback" and "Tell us about your experience with deployment". To the right, there are promotional cards for "Cost Management", "Microsoft Defender for Cloud", "Free Microsoft tutorials", and "Work with an expert". The bottom of the screen shows the Windows taskbar with various pinned icons and system status information.

→ Then i have created a container and uploaded a file in it .



The screenshot shows the Microsoft Azure Storage Container blade for 'mycontainer1069'. The left sidebar has sections for Overview, Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Access policy, Properties, Metadata), and a blob list. The main area displays a table with one row for 'scores.txt'. The table columns are Name, Modified, Access tier, Archive status, Blob type, and Size. The file 'scores.txt' was modified on 20/02/2024, 17:11:14, is in the Hot (Inferred) tier, has an Archive status of Not yet archived, is a Block blob, and is 86 B in size. A search bar at the top right allows searching by blob prefix (case-sensitive). A 'Show deleted blobs' toggle switch is also present.

https://portal.azure.com/#

34°C Partly sunny

Search

5:12 PM 2024-02-20

→ Then again, I have created another storage account for the destination type .

→ After that I have created an empty container for my destination storage account .

Screenshot of Microsoft Azure portal showing the deployment details for 'adfdeststorage1069\_1708429473024'. The deployment was successful.

**Deployment Details:**

- Deployment name: adfdeststorage1069\_1708429473024
- Subscription: Azure subscription 1
- Resource group: rg-azuser1069\_mml.local-6B597
- Start time: 20/02/2024, 17:14:36
- Correlation ID: f14ac586-db61-474d-9cc6-e94e9c522

**Next steps:**

- Go to resource
- Give feedback
- Tell us about your experience with deployment

**Right-hand sidebar:**

- 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.

Screenshot of Microsoft Azure portal showing the containers for the storage account 'adfdeststorage1069'.

**Containers:**

Name	Last modified	Anonymous access level	Lease state
\$logs	20/02/2024, 17:15:06	Private	Available
adfdeststoragecontainer	20/02/2024, 17:15:55	Private	Available

**Left sidebar:**

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser
- Storage Mover
- Containers**
- File shares
- Queues
- Tables

**Right-hand sidebar:**

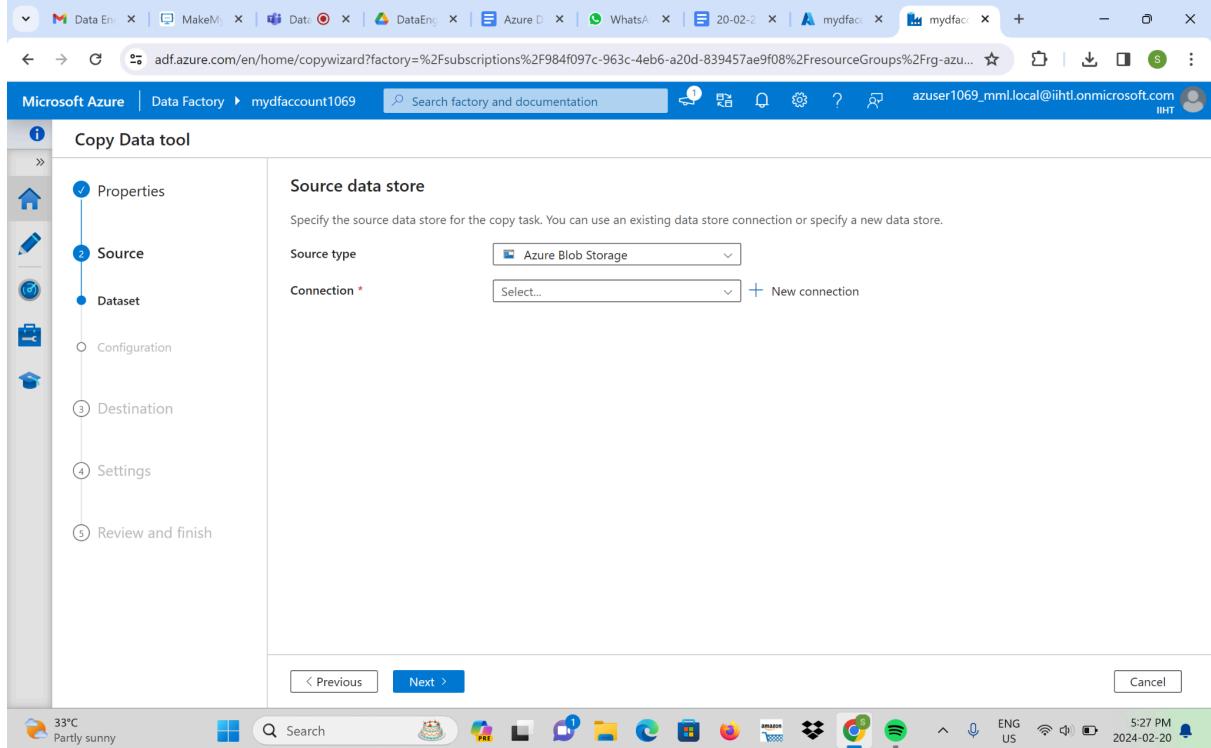
→ Then I created a data factory account and then I clicked on launch studio.

The screenshot shows the Microsoft Azure portal with the URL [https://portal.azure.com/#@iiht.onmicrosoft.com/resource/subscriptions/984f097c-963c-4eb6-a20d-839457ae9f08/resourceGroups/rg-azuser1069\\_mmllocal-6BS97](https://portal.azure.com/#@iiht.onmicrosoft.com/resource/subscriptions/984f097c-963c-4eb6-a20d-839457ae9f08/resourceGroups/rg-azuser1069_mmllocal-6BS97). The page displays the 'mydfaccount1069' Data Factory (V2) settings. On the left, there's a navigation menu with sections like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Networking, Managed identities, Properties, Locks), Getting started (Quick start), Monitoring (Alerts, Metrics), and a weather widget (33°C, Partly sunny). The main content area shows the 'Essentials' section with details such as Resource group (move) to 'rg-azuser1069\_mmllocal-6BS97', Status 'Succeeded', Location 'East US', Subscription (move) to 'Azure subscription 1', and Subscription ID '984f097c-963c-4eb6-a20d-839457ae9f08'. A large blue icon of a factory building is centered. Below it, the text 'Azure Data Factory Studio' is displayed, followed by a prominent blue 'Launch studio' button. The top right corner shows the user's name 'azuser1069\_mml.local@iiht.onmicrosoft.com' and the date '2024-02-20'.

→ Then I have selected an option called ingestion to set the connection .

The screenshot shows the Microsoft Azure Data Factory studio interface at the URL [https://adf.azure.com/en/home?factory=%2Fsubscriptions%2F984f097c-963c-4eb6-a20d-839457ae9f08%2FresourceGroups%2Frg-azuser1069\\_mmllocal-6BS97](https://adf.azure.com/en/home?factory=%2Fsubscriptions%2F984f097c-963c-4eb6-a20d-839457ae9f08%2FresourceGroups%2Frg-azuser1069_mmllocal-6BS97). The page title is 'Data factory mydfaccount1069'. The left sidebar has icons for Home, Data flow, Pipeline, and New. The main area features a 3D factory building icon with four cards below it: 'Ingest' (Copy data at scale once or on a schedule.), 'Orchestrate' (Code-free data pipelines.), 'Transform data' (Transform your data using data flows.), and 'Configure SSIS' (Manage & run your SSIS packages in the cloud.). The bottom left shows a 'Recent resources' section with a document icon. The bottom right shows the date '2024-02-20' and time '5:22 PM'.

→ Then I have created a source data source and then I have selected azure blob storage as source type and created a connection by giving a source storage name and source container in it.



→ Later I have created a destination data source and then I have selected azure blob storage as destination type and created a connection by giving a destination storage name and destination container in it.

→ In the destination container there is no file present because we are copying source to destination type .So our motive is to copy the data from source and review at destination group . For that we need to give an empty destination container .

The screenshot shows the Microsoft Azure Data Factory Copy Data tool wizard. The left sidebar lists steps 1 through 5: Properties, Source, Destination, Dataset, Configuration, Settings, and Review and finish. Step 1 is selected. The main area is titled "Destination data store" and asks to specify the destination data store for the copy task. It shows "Destination type" as "Azure Blob Storage" and "Connection" as "Select...". On the right, there's a "New connection" section for "Azure Blob Storage" with fields for "Authentication type" (set to "Account key"), "Account selection method" (set to "From Azure subscription"), and "Azure subscription" (set to "Azure subscription 1 (984f097c-963c-4eb6-a20d-839457ae9f08)"). Below this are "Additional connection properties" and "Test connection" options. A "Create" button is at the bottom right.

→ Then we need to give the task name as “copyActivity” for the above process and click on the next .

The screenshot shows the Microsoft Azure Data Factory Copy Data tool wizard on the "Settings" step. The left sidebar shows steps 1 through 5. Step 4 is selected. The main area has a "Settings" title and a note: "Enter name and description for the copy data task, more options for data movement". It includes fields for "Task name" (set to "CopyPipeline\_j1c") and "Task description". Below these are checkboxes for "Data consistency verification", "Enable logging", and "Enable staging". A "Cancel" button is at the bottom right.

→ After giving the task name we need to review all the settings and click on finish .

The screenshot shows the Microsoft Azure Copy Data tool interface. On the left, a sidebar lists steps: Properties, Source, Destination, Settings, Review and finish, Review, and Deployment. The 'Review and finish' step is currently selected. In the main area, there is a diagram showing 'Azure Blob Storage' on the left connected by an arrow to 'Azure Blob Storage' on the right. Below the diagram, the text 'Deployment complete' is displayed. A table titled 'Deployment step' shows four rows with status 'Succeeded': 'Validating copy runtime environment', 'Creating datasets', 'Creating pipelines', and 'Running pipelines'. At the bottom, 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.' Three buttons are at the bottom: 'Finish' (highlighted in blue), 'Edit pipeline', and 'Monitor'.

→ Then we can see that datasets and pipelines are successfully created and running successfully.

→ Thereby we can see that in the destination container the file has been successfully copied from the source account .

The screenshot shows the Microsoft Azure Storage Container blade for the container 'adfdeststoragecontainer'. The left sidebar has 'Overview' selected. The main area displays a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size
scores.txt	20/02/2024, 17:35:58	Hot (Inferred)		Block blob	86 B



