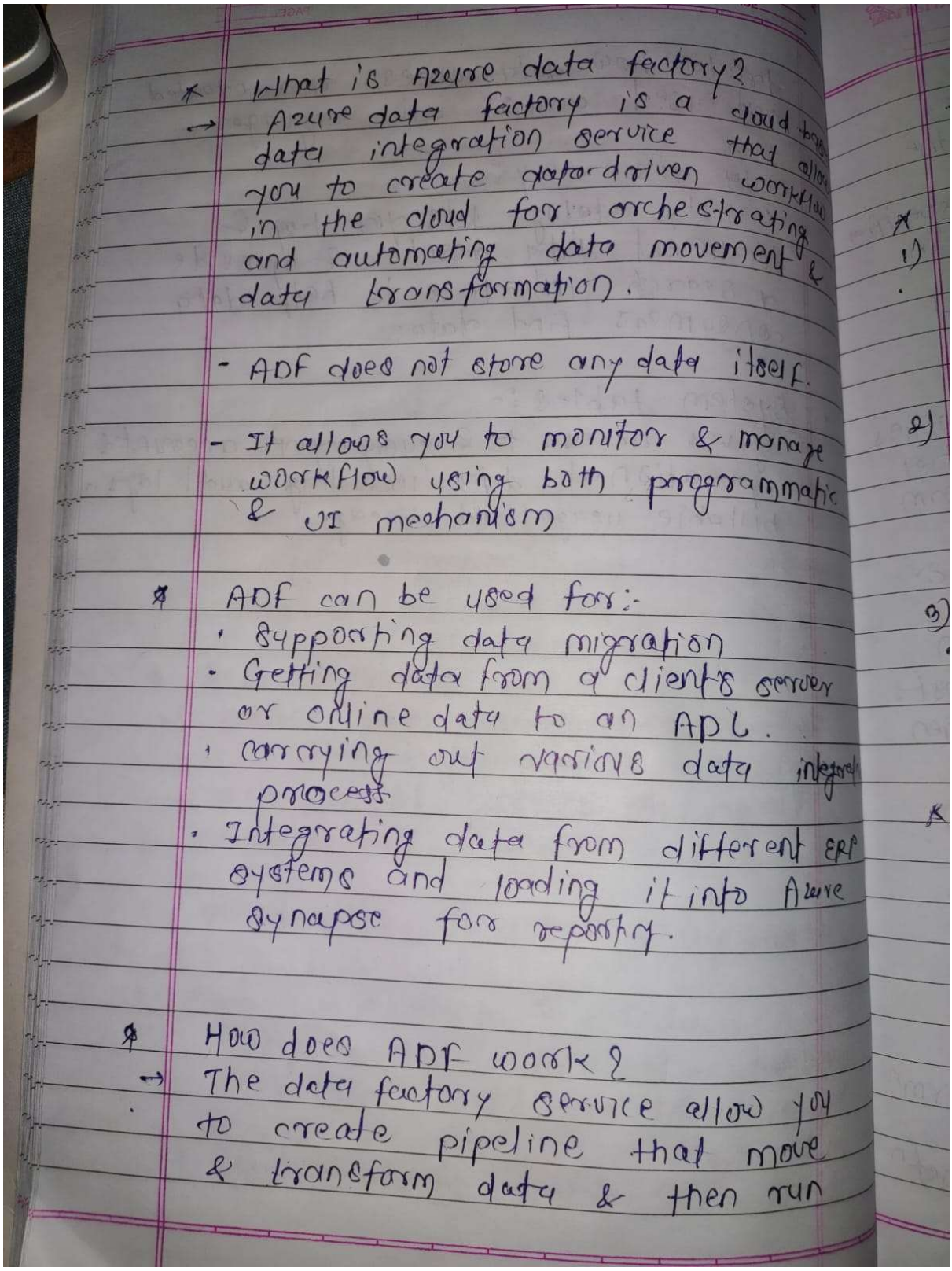
Name: Krushnakumar Patle

Email: krishnapatle128@gmail.com

Batch: Data Engineering Batch-1



* What is Azure 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 & 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's server or online data to an ADL.
  - carrying out various data integration process
  - Integrating data from different ERP systems and 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 etc).

* ADF can perform in 3 steps:-

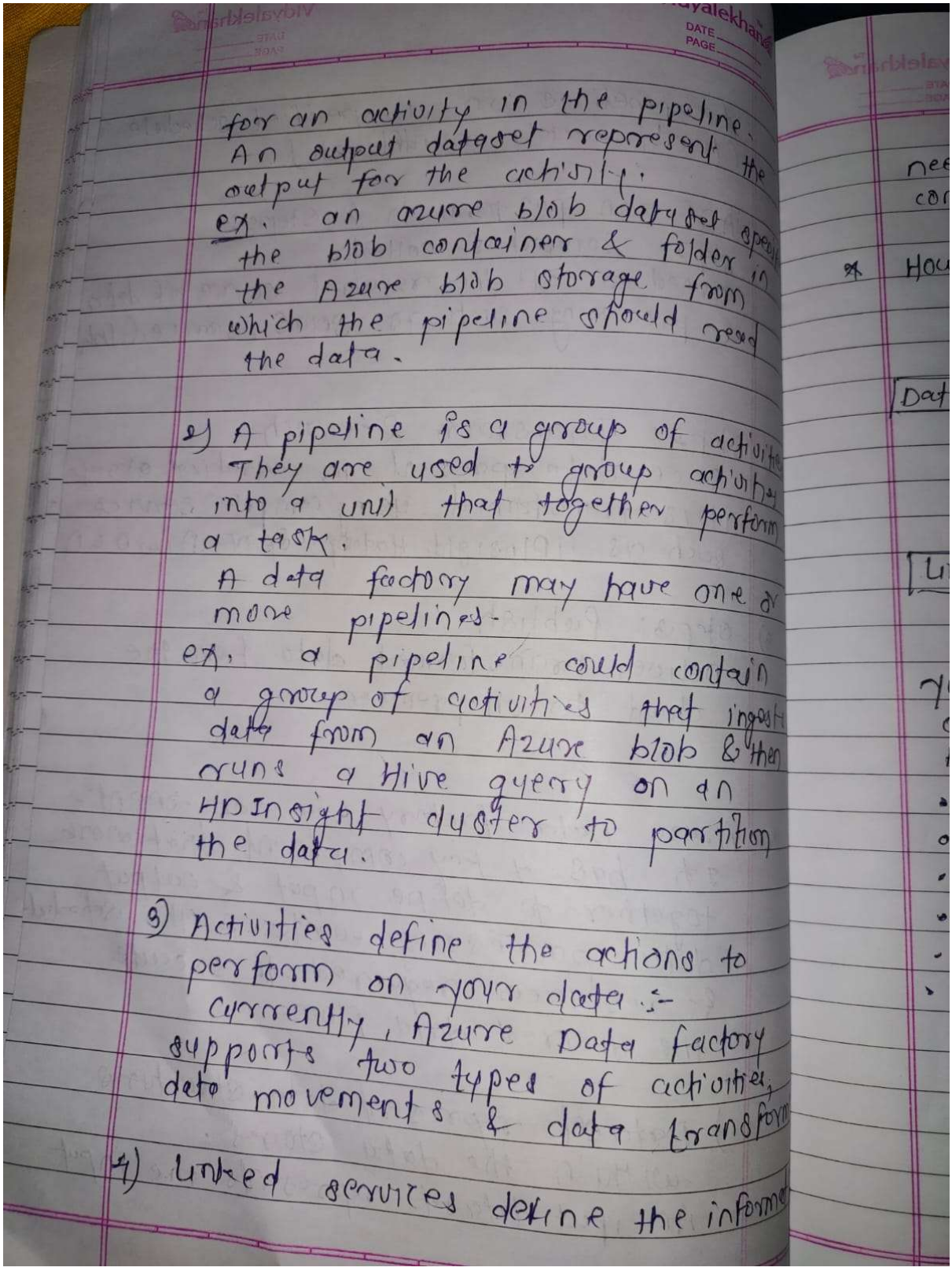
1) Step 1: Connect & collect
Connect to all the required source of data & processing such as SaaS, Source, link

2) Step 2: Transform & enrich
Once data present in centralized storage it is transformed using compute service such as HDInsight, Hadoop, Spark, Azure DL

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 schedule & resources required to execute the desired data flow.

1) Dataset represent data structures within the data stores.
An input dataset represents the input



for an activity in the pipeline. An output dataset represent the output for the activity.
ex. an azure blob dataset specifies the blob container & folder in the Azure blob storage from which the pipeline should read the data.

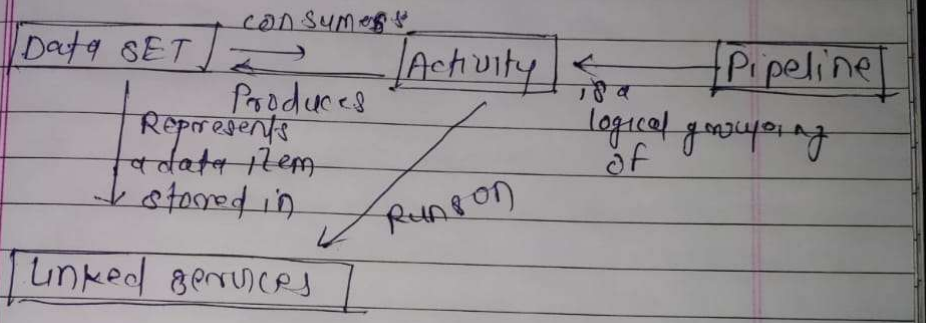
2) A pipeline is a group of activities. They are used to group activities into a unit that together perform a task.
A data factory may have one or more pipelines.
ex. a pipeline could contain a group of activities that ingests data from an Azure blob & then runs a Hive query on an HDInsight cluster to 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 transformations

4) Linked services define the information

needed for Azure data factory to connect to external resources. :-

* How ADF ~~work~~ component coork together :-

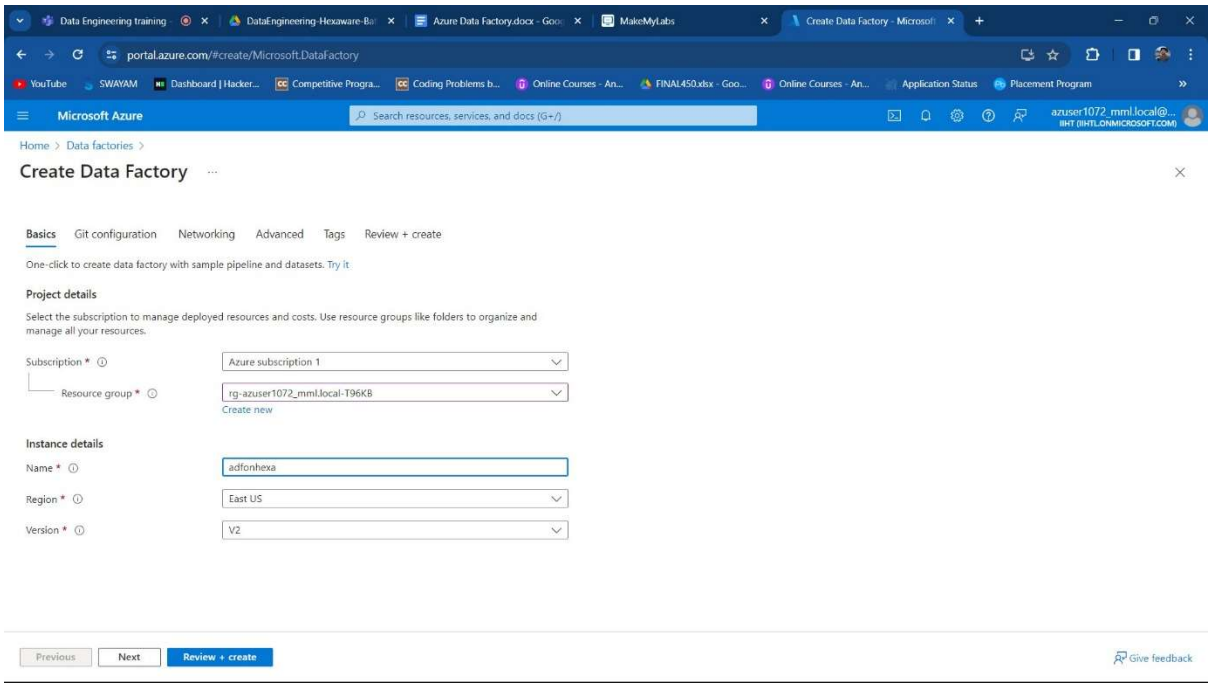


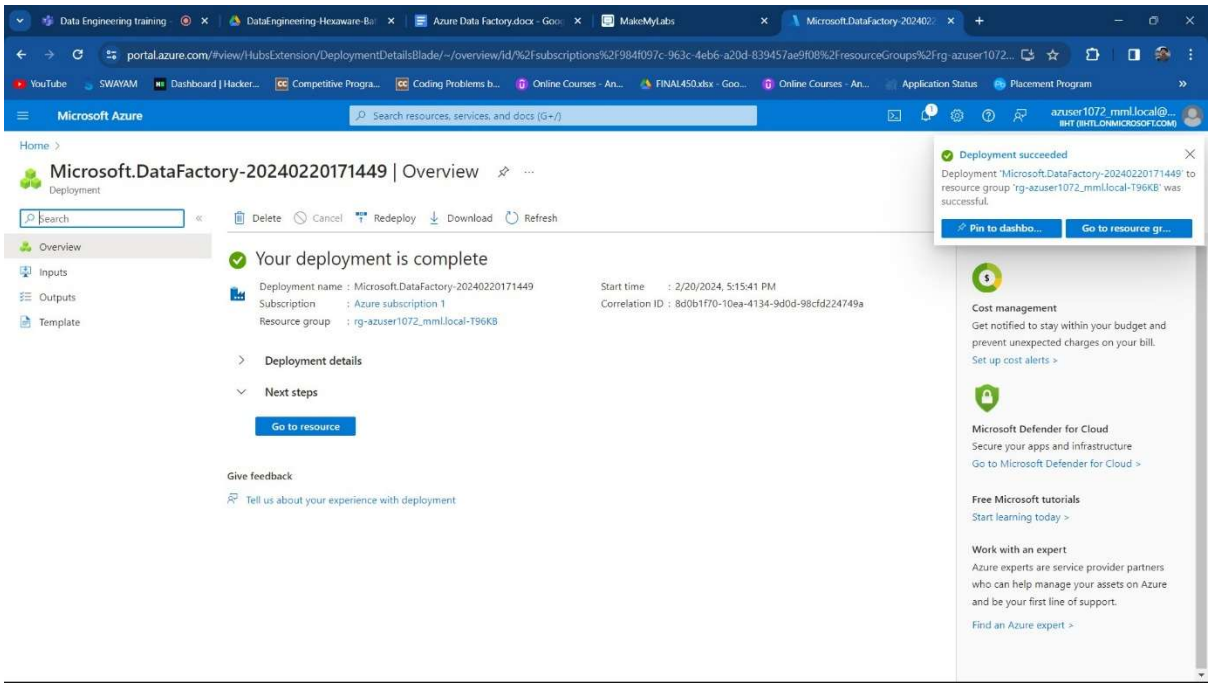
You can use one of the following tools or APIs to create data pipeline in Azure Data factory.

- Azure portal
- Visual studio
- Powershell
- .NET API
- REST API
- Azure resource management template

Copy Activity:

1. Create a data factory



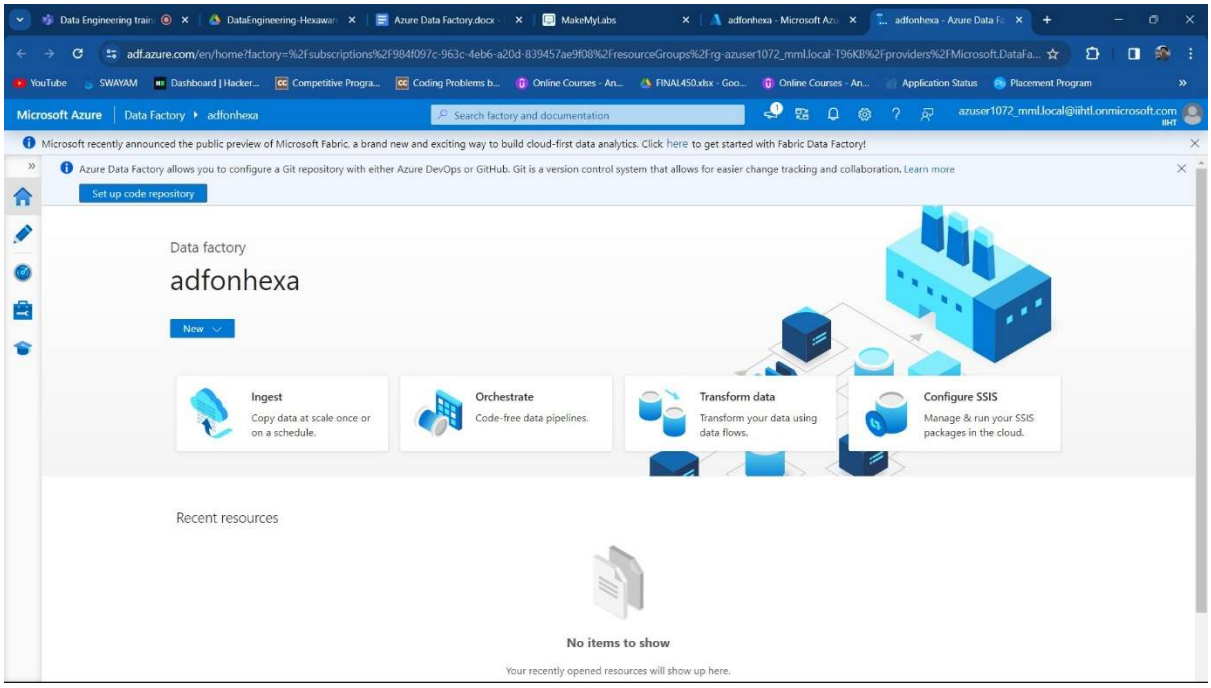


Use the copy data tool to copy data

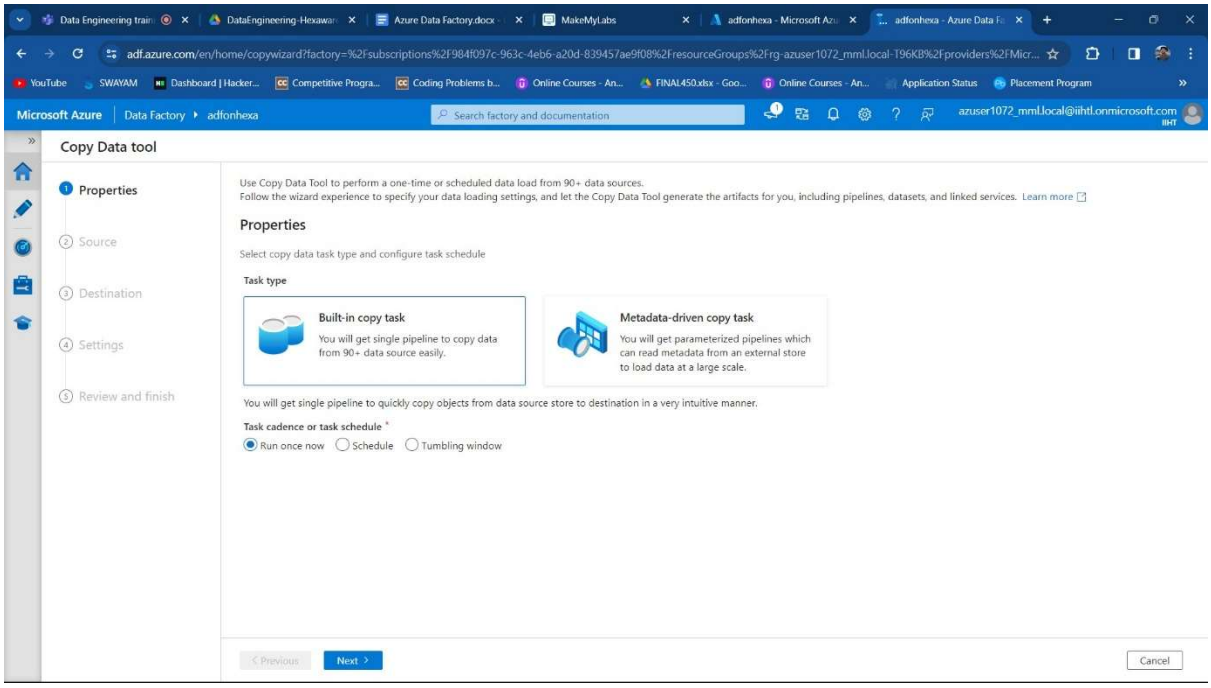
The steps below will walk you through how to easily copy data with the copy data tool in Azure Data Factory.

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.

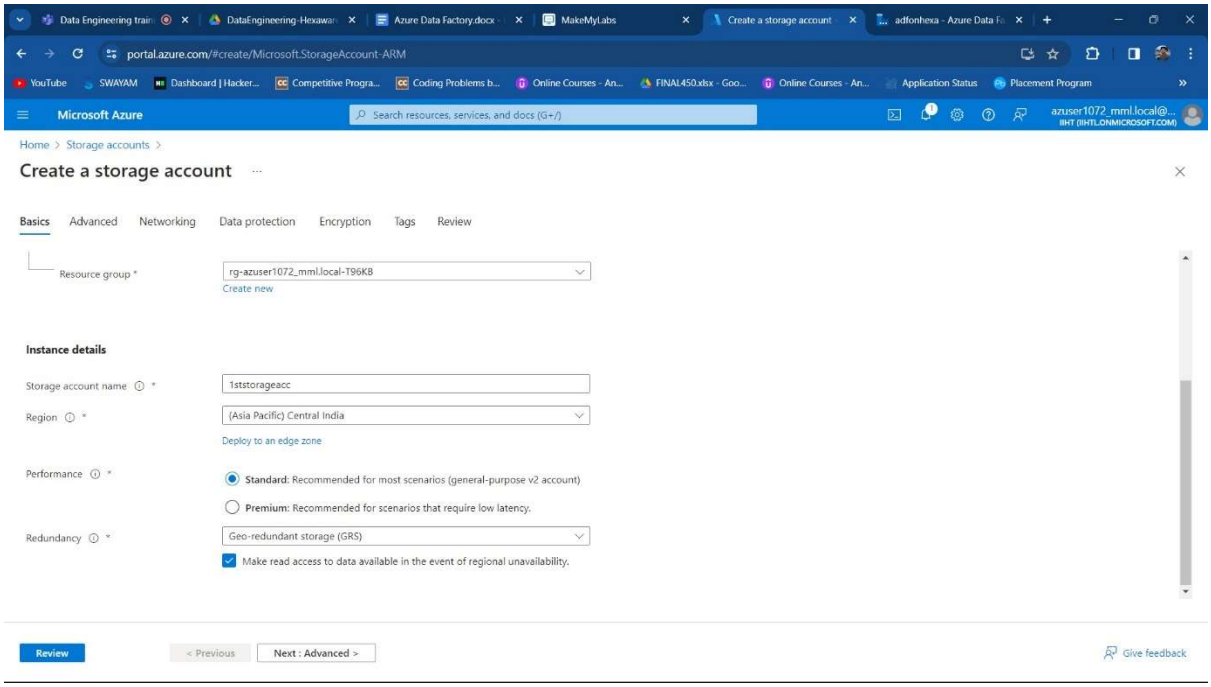


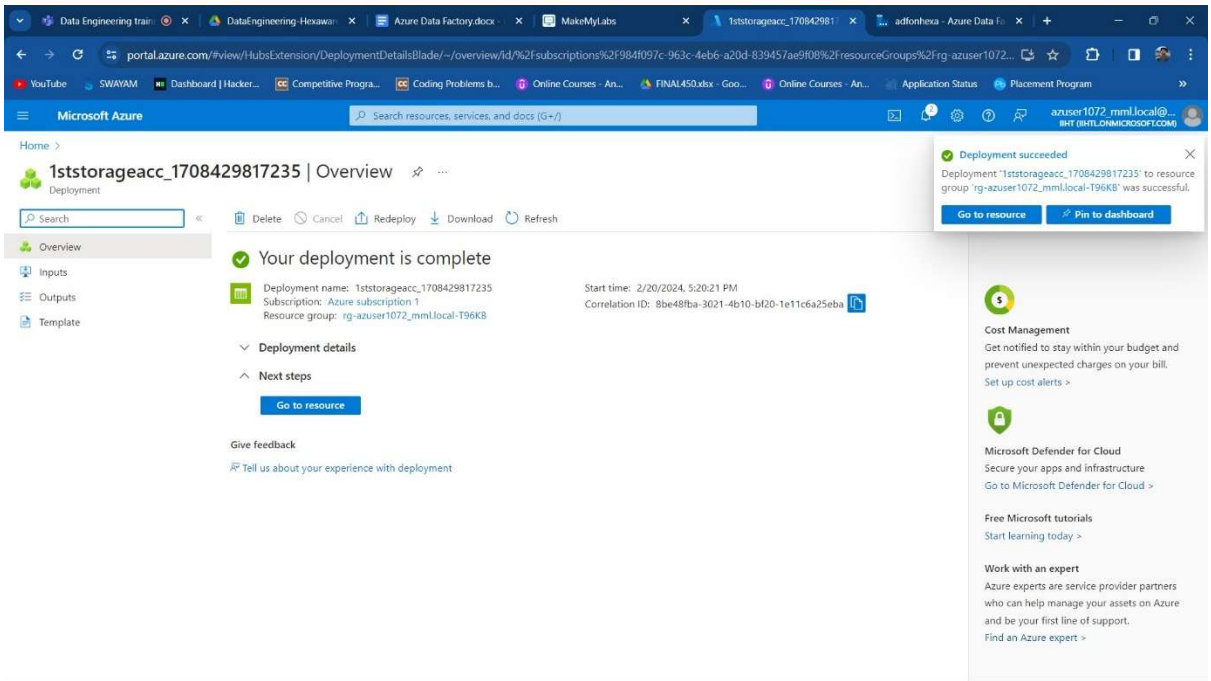
2. On the **Properties** page of the Copy Data tool, choose **Built-in copy task** under **Task type**, then select **Next**.



Step 2: Complete source configuration

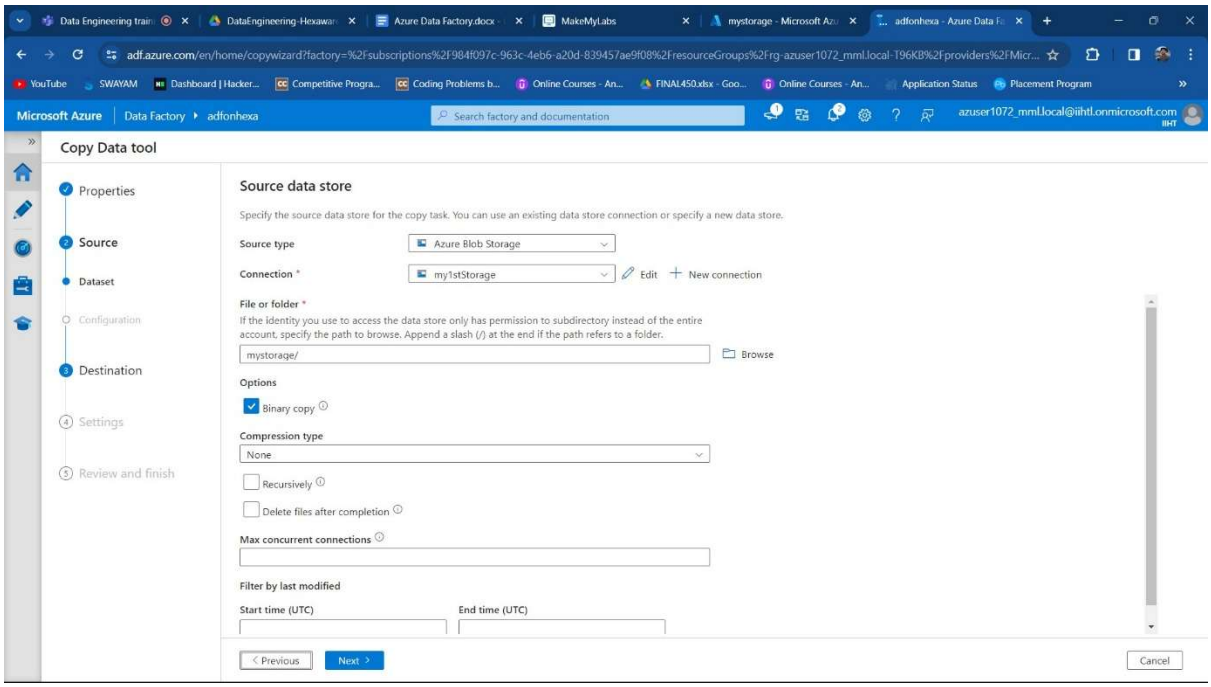
1. Click **+ Create new connection** to add a connection.
2. Select the linked service type that you want to create for the source connection. In this tutorial, we use **Azure Blob Storage**. Select it from the gallery, and then select **Continue**.

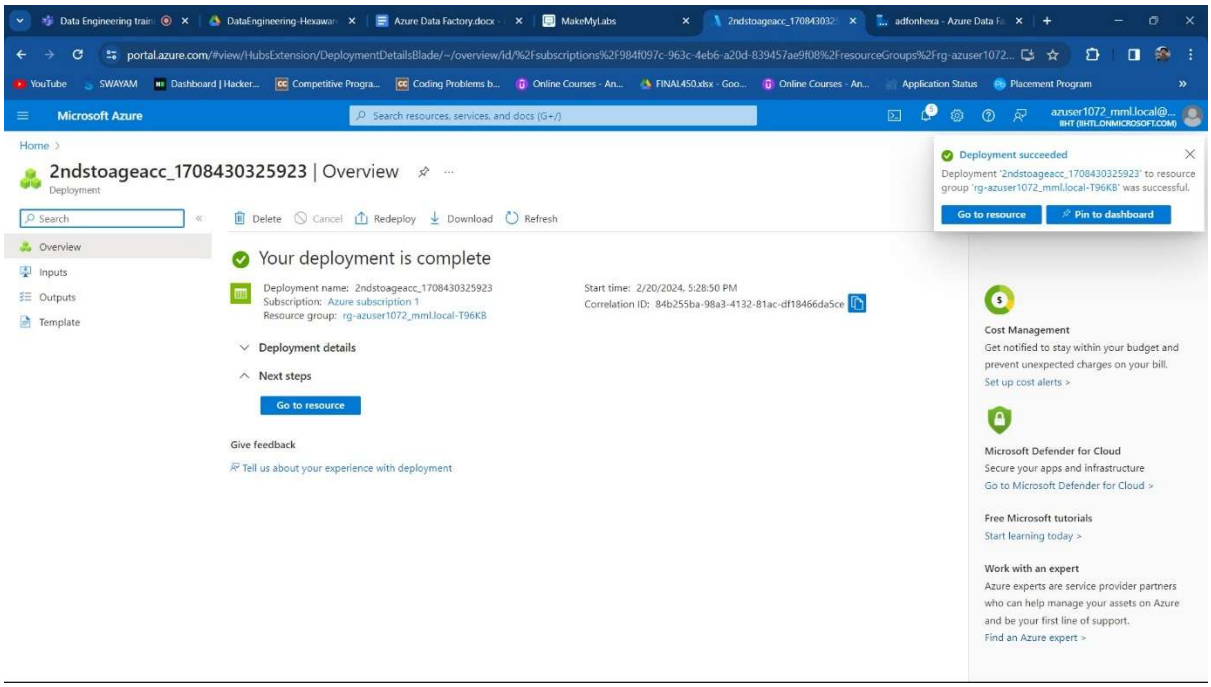




1. On the **New connection (Azure Blob Storage)** page, specify a name for your connection. Select your Azure subscription from the **Azure subscription** list and your storage account from the **Storage account name** list, test connection, and then select **Create**.

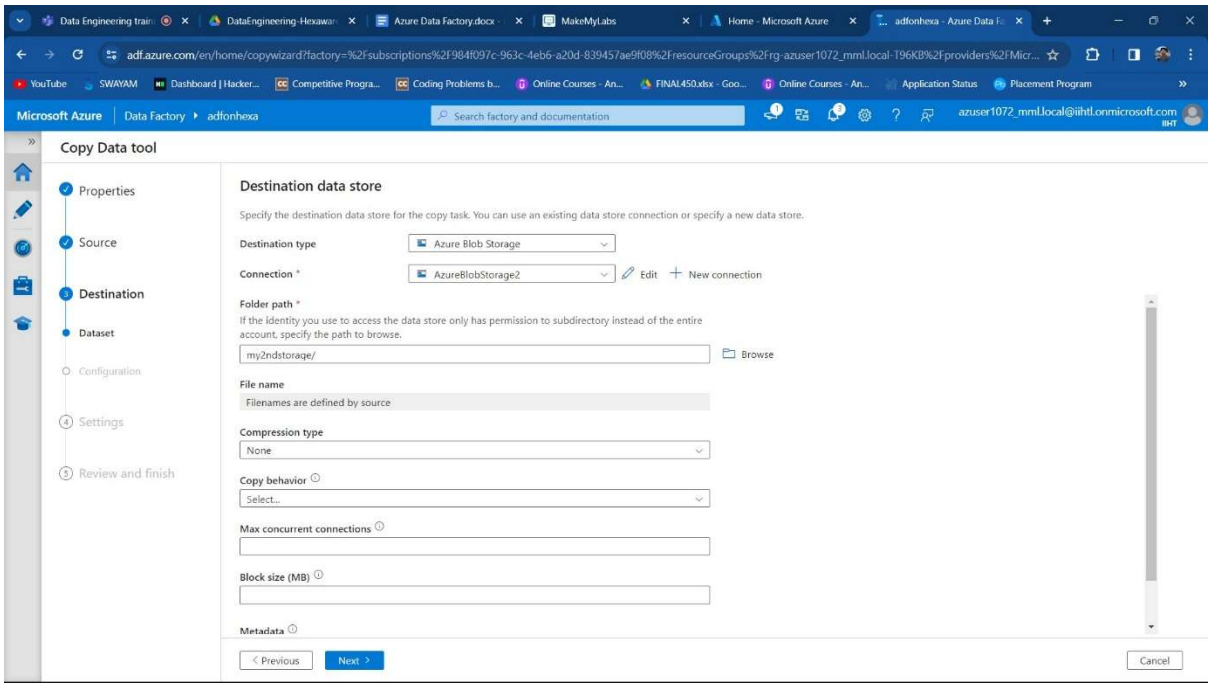
1. Select the newly created connection in the **Connection** block.
    2. In the **File or folder** section, select **Browse** to navigate to the **mystorage/input** folder, select the file, and then click **OK**.
    3. Select the **Binary copy** checkbox to copy file as-is, and then select **Next**.





Step 3: Complete destination configuration

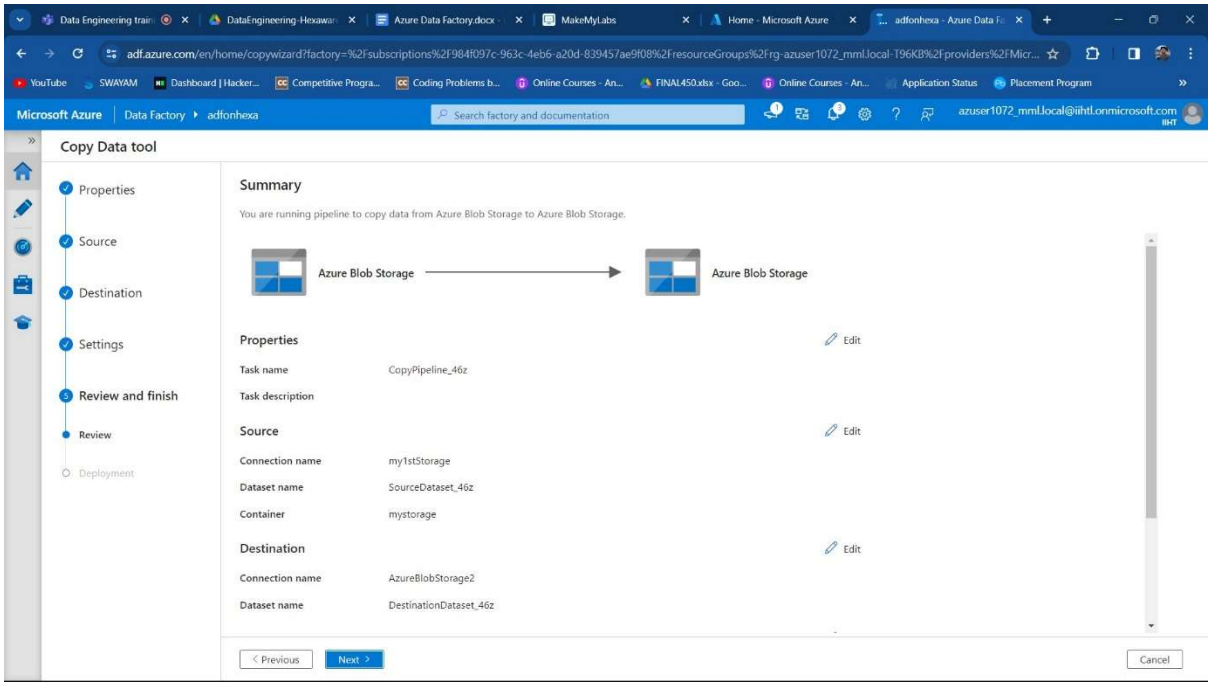
1. Select the **AzureBlobStorage** connection that you created in the **Connection** block.
2. In the **Folder path** section, enter **my2ndstorage/output** for the folder path.

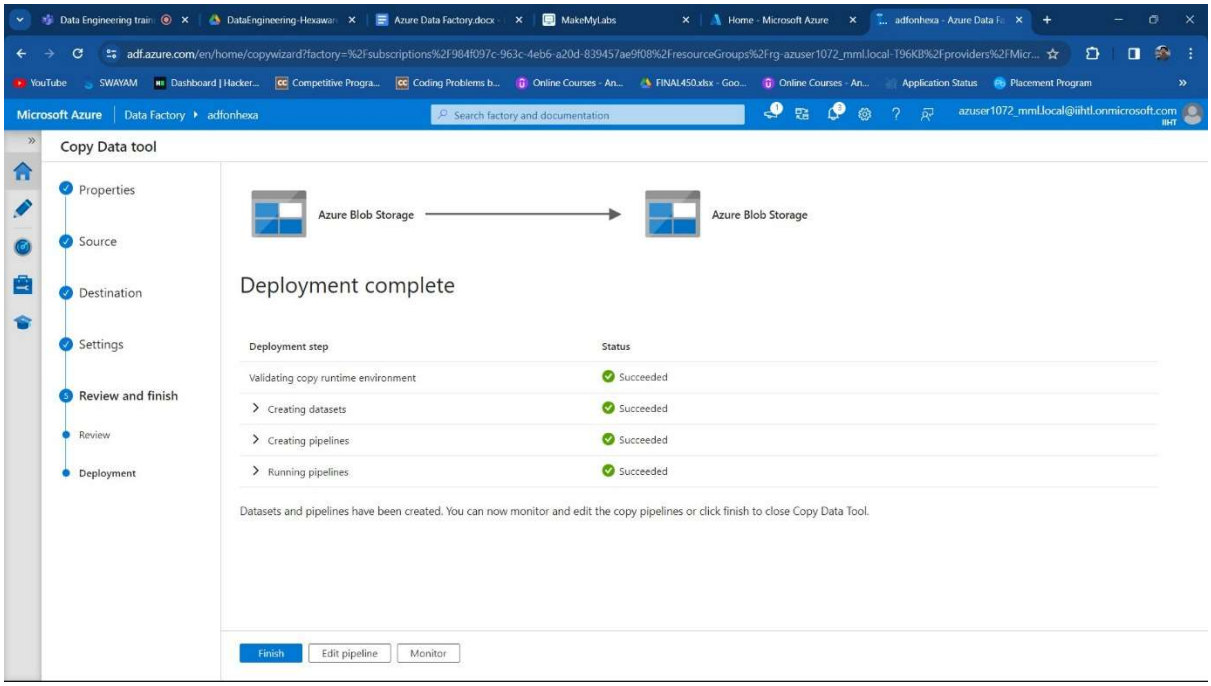


Step 4: Review all settings and deployment

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.

Day 22 Assessment