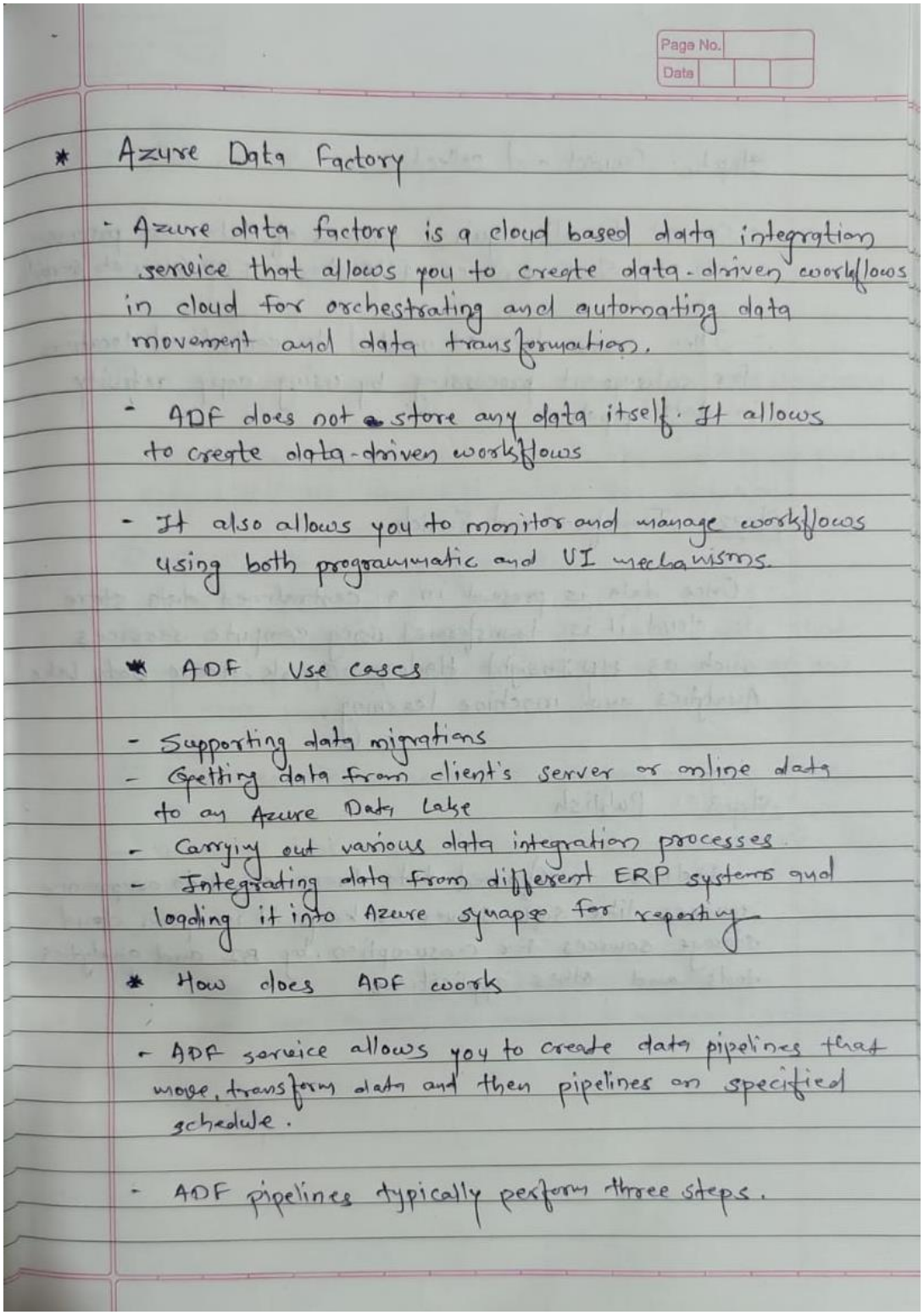
Data Engineering Batch 1 **Day 22:20/02/2024**

Name: Pradip Bochare

- **Notes**



Page No.
Date

* Azure Data Factory

- Azure data factory is a cloud based data integration service that allows you to create data-driven workflows in cloud for orchestrating and automating data movement and data transformation.

- ADF does not store any data itself. It allows to create data-driven workflows

- It also allows you to monitor and manage workflows using both programmatic and UI mechanisms.

* ADF Use cases

- Supporting data migrations
- Getting data from 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

* How does ADF work

- ADF service allows you to create data pipelines that move, transform data and then pipelines on specified schedule.

- ADF pipelines typically perform three steps.

Page No.
Date

Step 1: Connect and collect

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

- Then move data as needed to centralized location for subsequent processing by using copy activity in data pipeline

Step 2: Transform and Enrich

- Once data is present in a centralized data store in cloud it is transformed using compute services such as HD Insight Hadoop, Sparks, Azure Data Lake Analytics and machine learning.

Step 3: Publish

- Delivers transformed data from 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.

Page No.
Date

* Azure Data Factory key components.

- ADF has four key components that work together to define & output data, processing events & schedule & resources required to execute desired data flow.

(1) Datasets represent data structure within data stores.

- input dataset represents input for activity in pipeline output dataset represents output for activity.

(2) Pipeline is a group of activities.

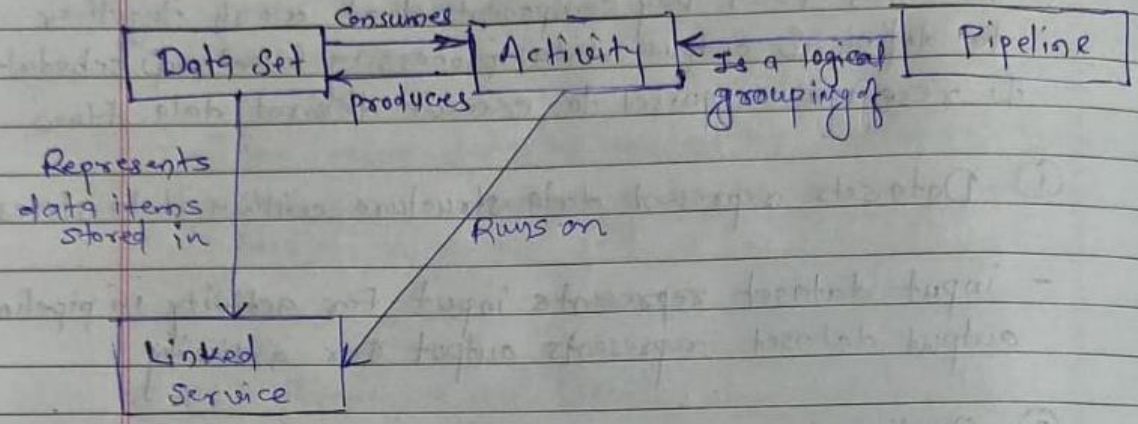
- They are used to group activities into a unit that together performs task. A data factory may have one or more pipelines.

(3) Activities define actions to perform on your data.

- Currently ADF supports two types of activities data movement and data transformation.

(4) Linked Services define information needed for Azure data factory to connect to external resources

* How ADF components works together

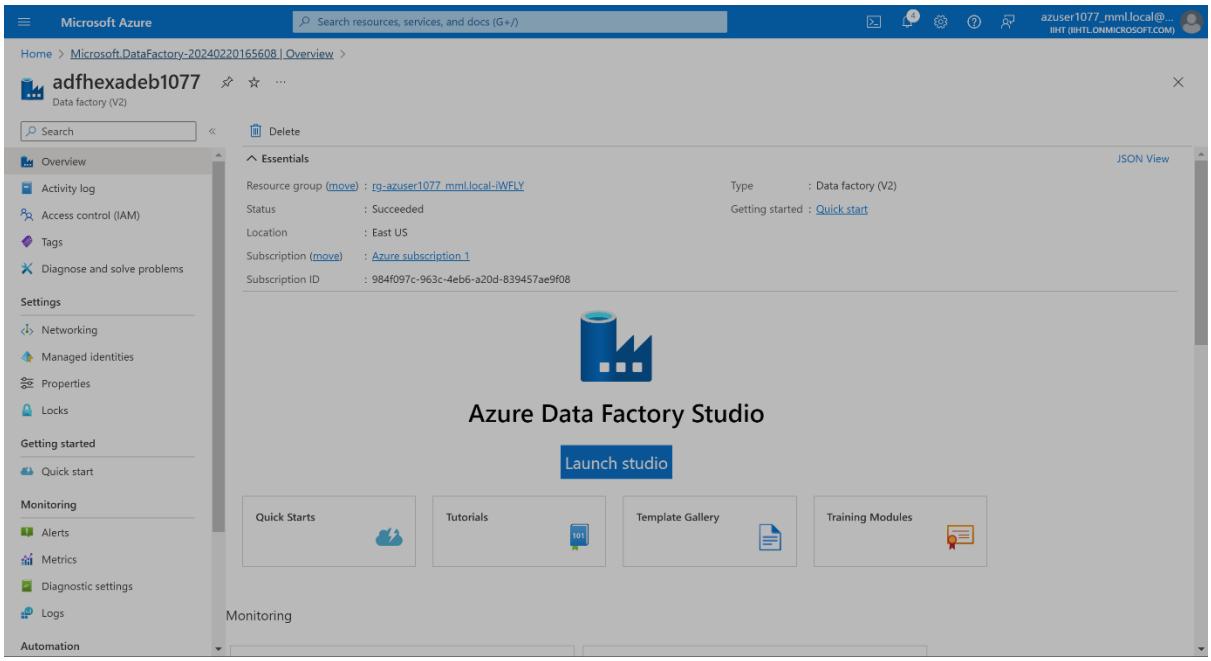


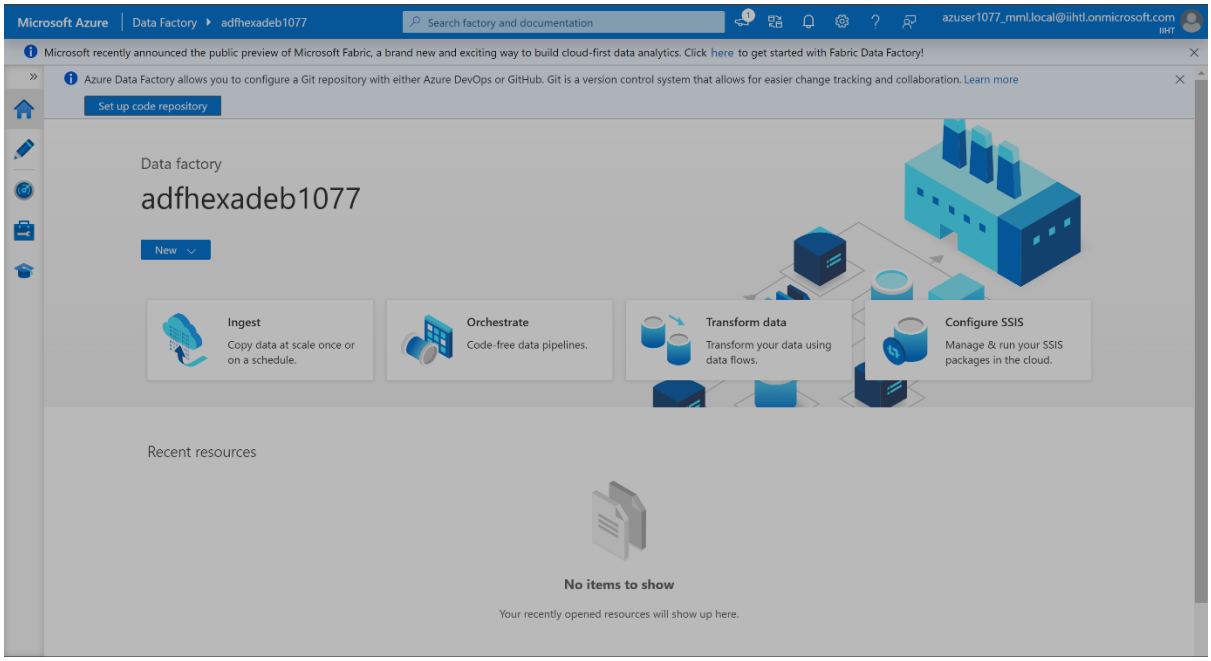
* You can also use one of following tools or APIs to create data pipelines in ADF.

- Azure portal
- Visual Studio
- Powershell
- .NET API
- REST API

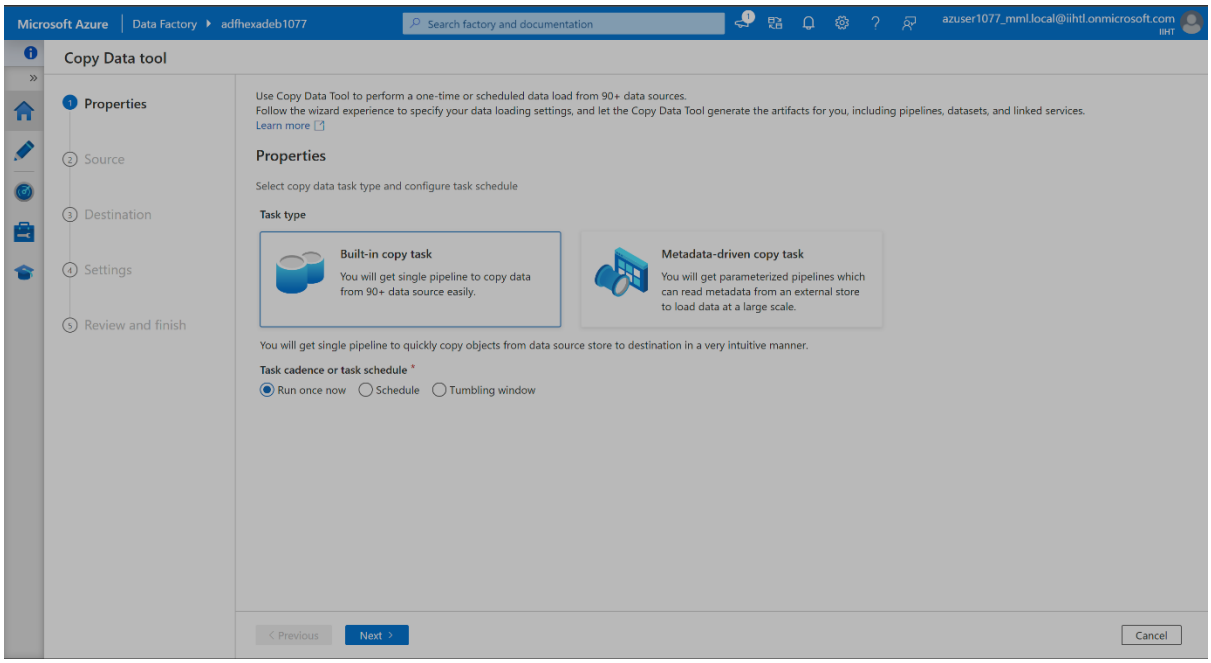
Azure Data Factory

- Making Azure Data Factory

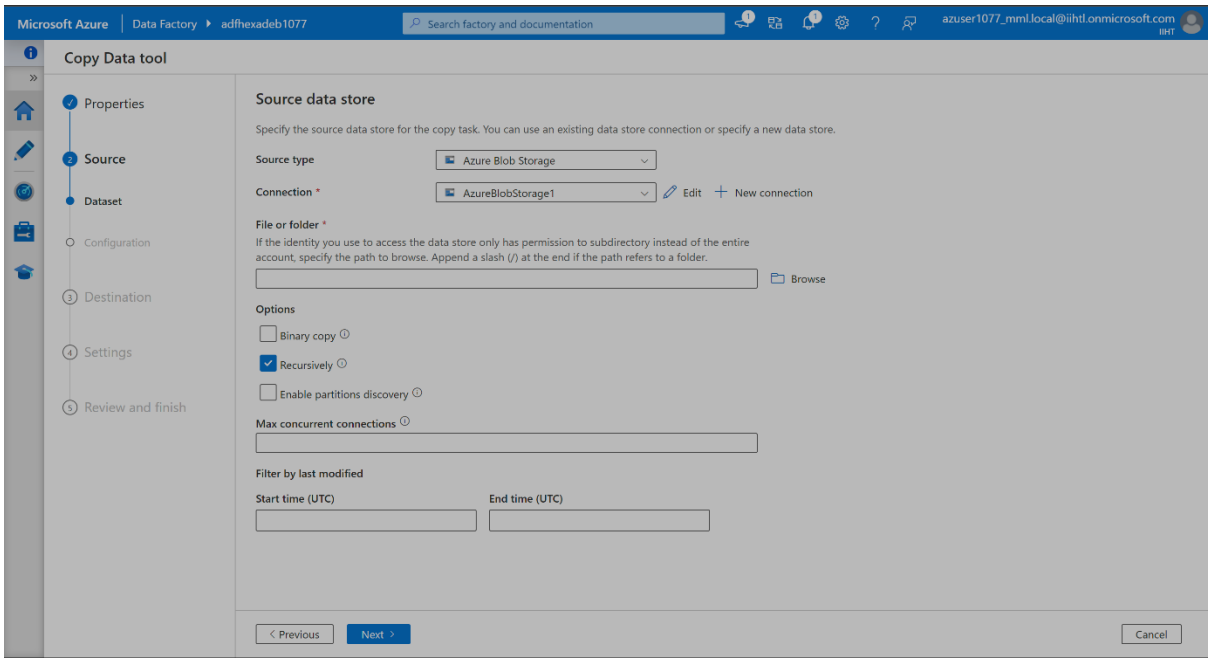


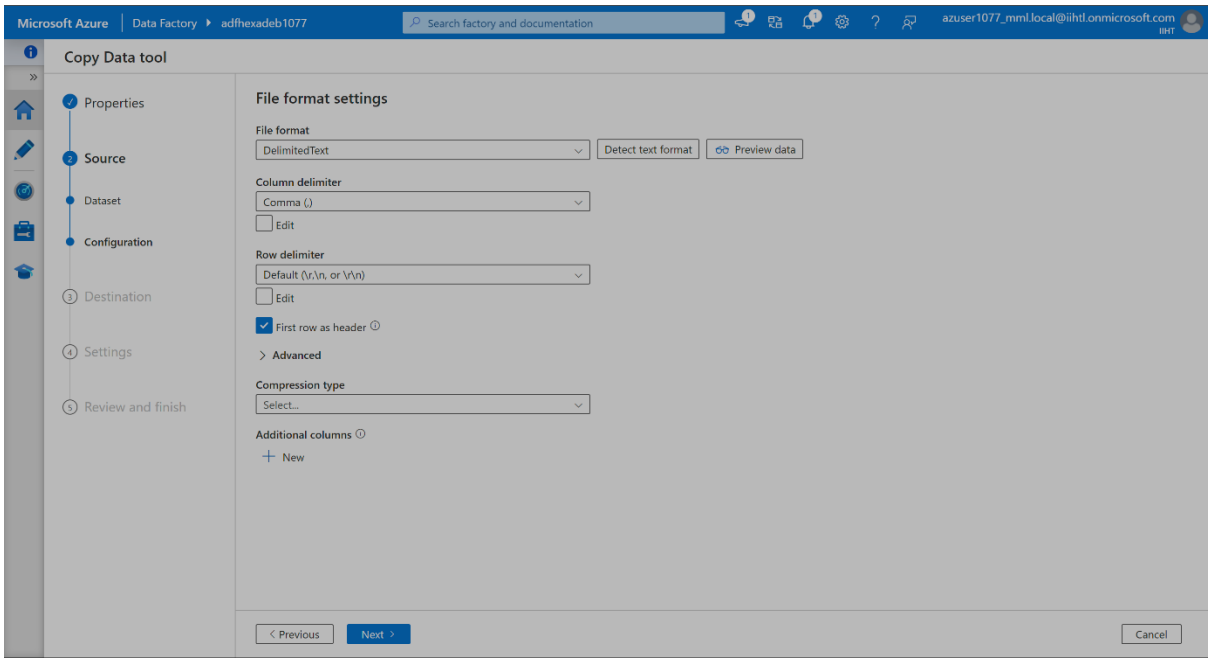


- Making Data Pipeline

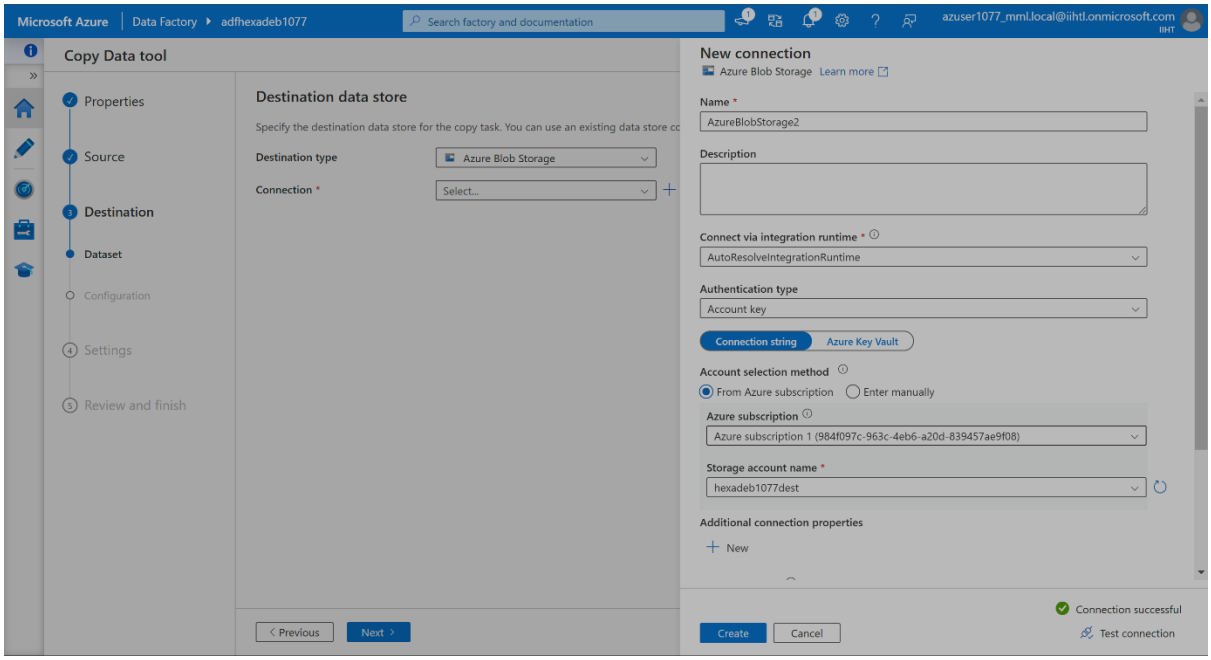


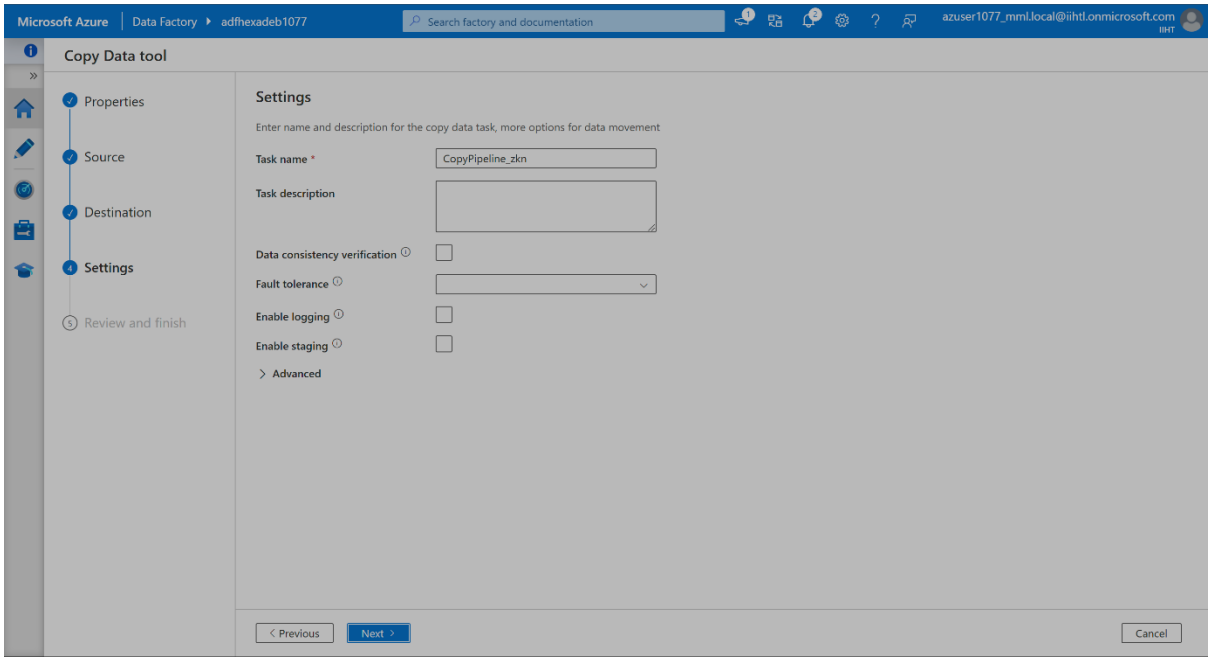
- Source Data



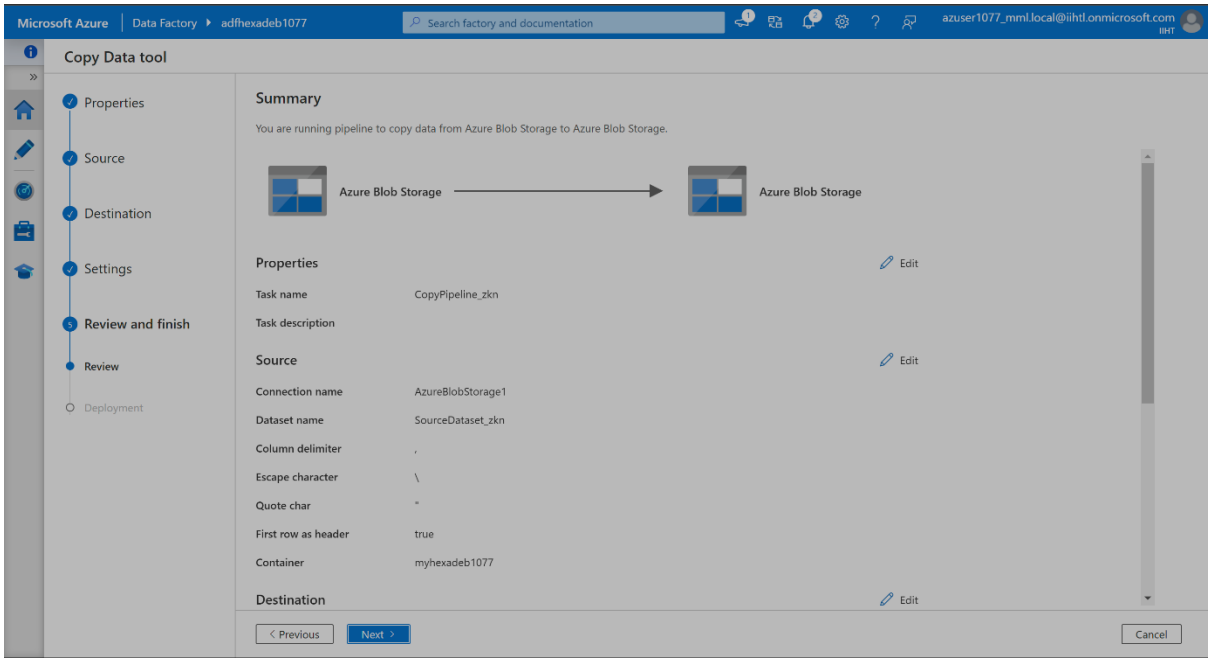


- Connecting with Source Container

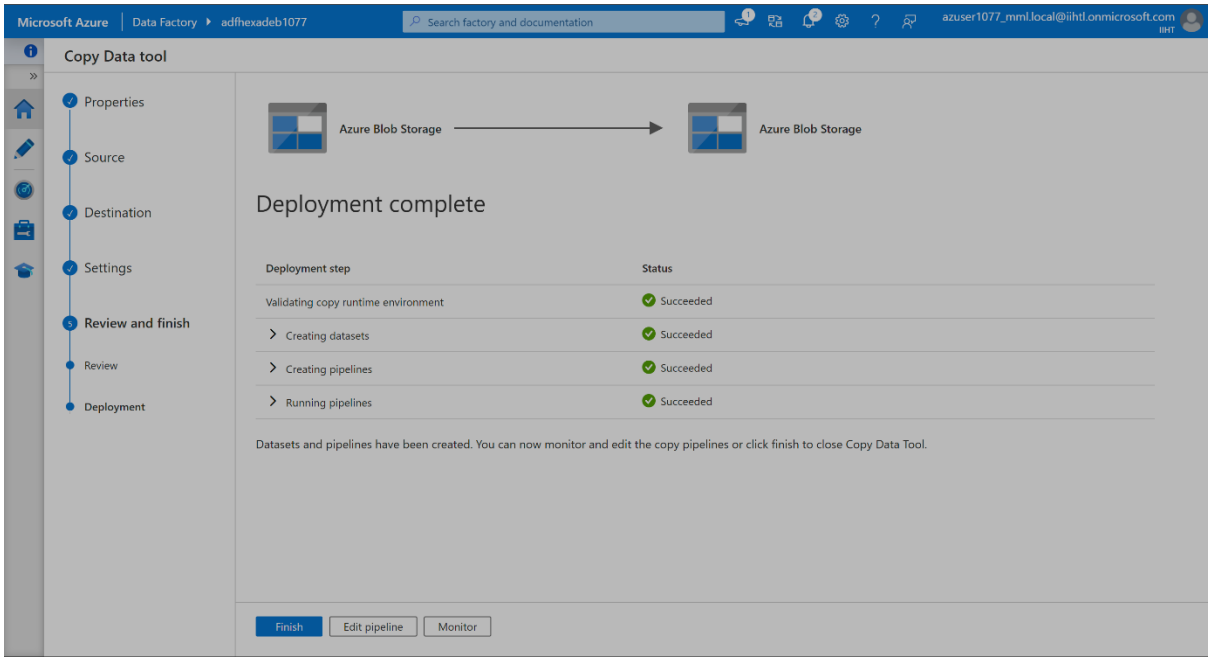


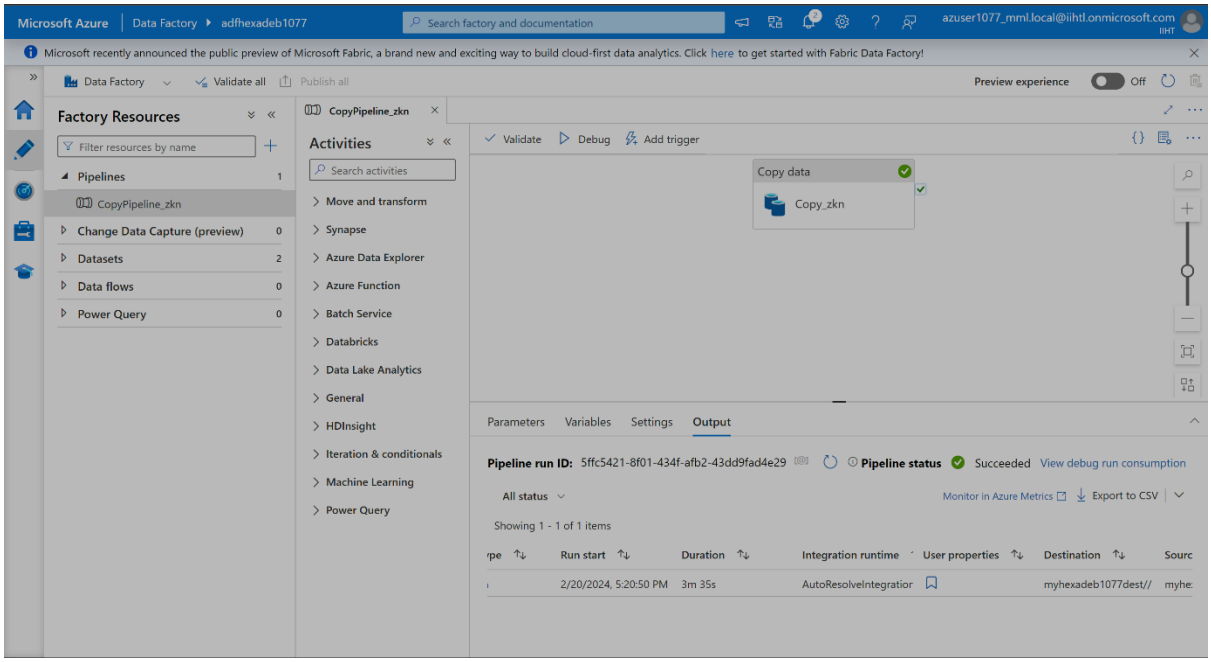

Microsoft Azure
Data Factory
adfhexadeb1077
Search factory and documentation
azuser1077_mml.local@iihtl.onmicrosoft.com
IIHT
Copy Data tool
Properties
Source
Destination
Settings
Review and finish
Settings
Enter name and description for the copy data task, more options for data movement
Task name *
CopyPipeline_zkn
Task description
Data consistency verification
Fault tolerance
Enable logging
Enable staging
Advanced
Previous
Next
Cancel

- Summary

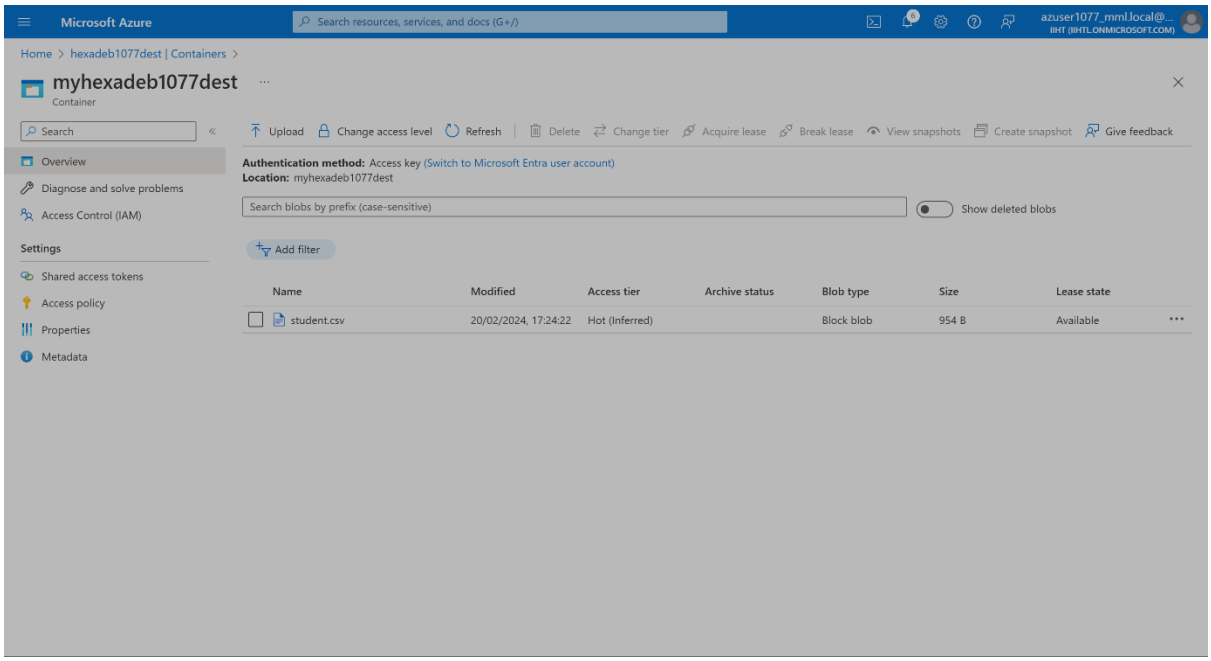

Microsoft Azure
Data Factory
adfhexadeb1077
Search factory and documentation
azuser1077_mml.local@iihtl.onmicrosoft.com
IIHT
Copy Data tool
Properties
Source
Destination
Settings
Review and finish
Review
Deployment
Summary
You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage.
Azure Blob Storage
Azure Blob Storage
Properties
Edit
Task name
CopyPipeline_zkn
Task description
Source
Edit
Connection name
AzureBlobStorage1
Dataset name
SourceDataset_zkn
Column delimiter
,
Escape character
\
Quote char
"
First row as header
true
Container
myhexadeb1077
Destination
Edit
Previous
Next
Cancel

- Deployment


Microsoft Azure
Data Factory
adfhexadeb1077
Search factory and documentation
azuser1077_mml.local@iihtl.onmicrosoft.com
IIHT
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


Microsoft Azure
Data Factory
adfhexadeb1077
Search factory and documentation
azuser1077_mml.local@iihtl.onmicrosoft.com
IIHT
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
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 1
CopyPipeline_zkn
Change Data Capture (preview) 0
Datasets 2
Data flows 0
Power Query 0
CopyPipeline_zkn
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Copy data
Copy_zkn
Parameters
Variables
Settings
Output
Pipeline run ID: 5ffc5421-8f01-434f-afb2-43dd9fad4e29
Pipeline status
Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Run start
Duration
Integration runtime
User properties
Destination
2/20/2024, 5:20:50 PM
3m 35s
AutoResolveIntegratior
myhexadeb1077dest//

- Destination Container



- Source Container