Azure Databricks Coding Challenge

Pratik Wani

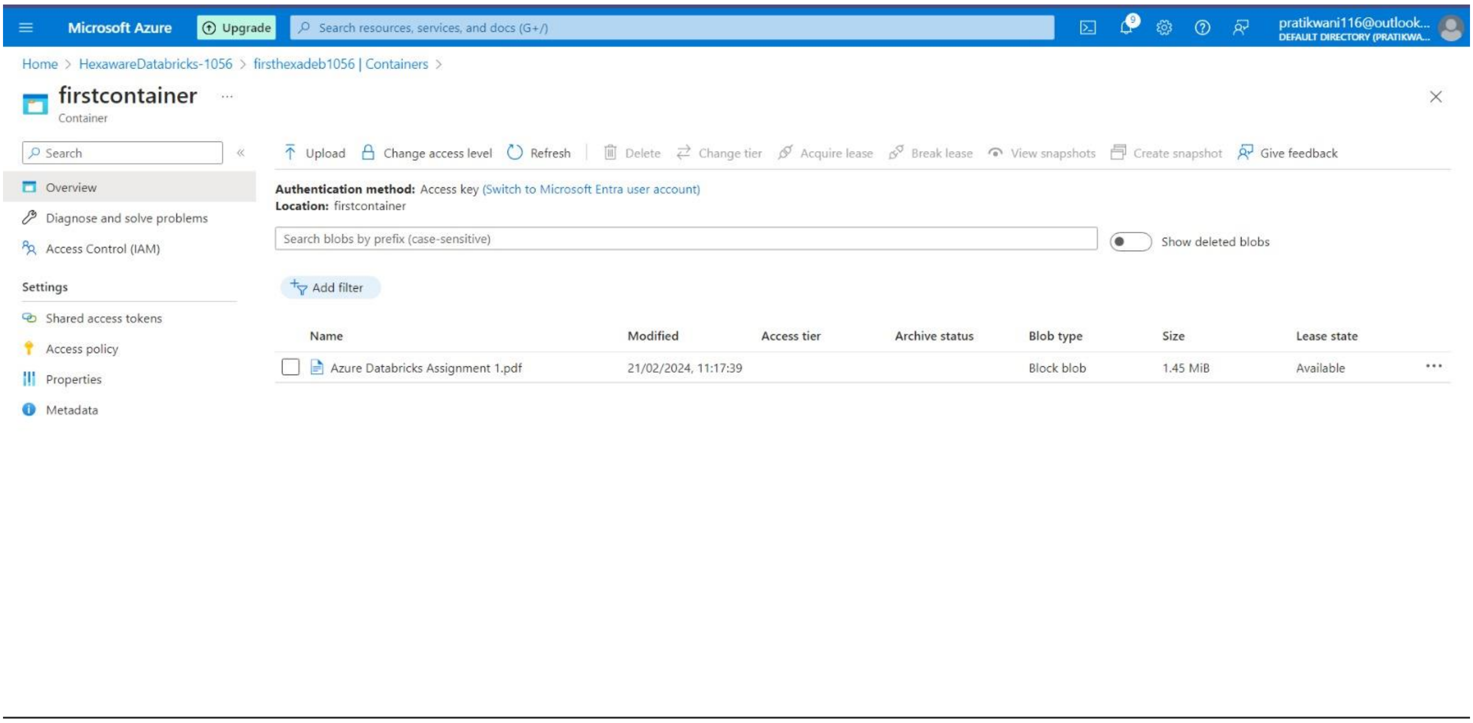
Question 3

❖ Azure Data factory and Copy Activity:

- Azure Data Factory is a cloud-based data integration service that allows us to manage data pipelines for data transformation tasks.
- It enables us to build workflows to transform and publish data from various sources at scale.
- One of the main features of Azure Data Factory is the Copy Activity, which is used to copy data between different data stores.
- We perform copy activity between Source and Sink Data stores
- We can perform copy activity between different data sources like
  - Blob – Blob
  - Blob – ADLS
  - ADLS – Blob
  - ADLS – ADLS
- I perform the copy activity by creating two Blob storage ‘1056hexadeb1’ and ‘1056hexadeb2’
- ‘1056hexadeb1’ is source storage and ‘1056hexadeb2’ is sink storage

- Copy Activity:

  - Created Storage Account and Container



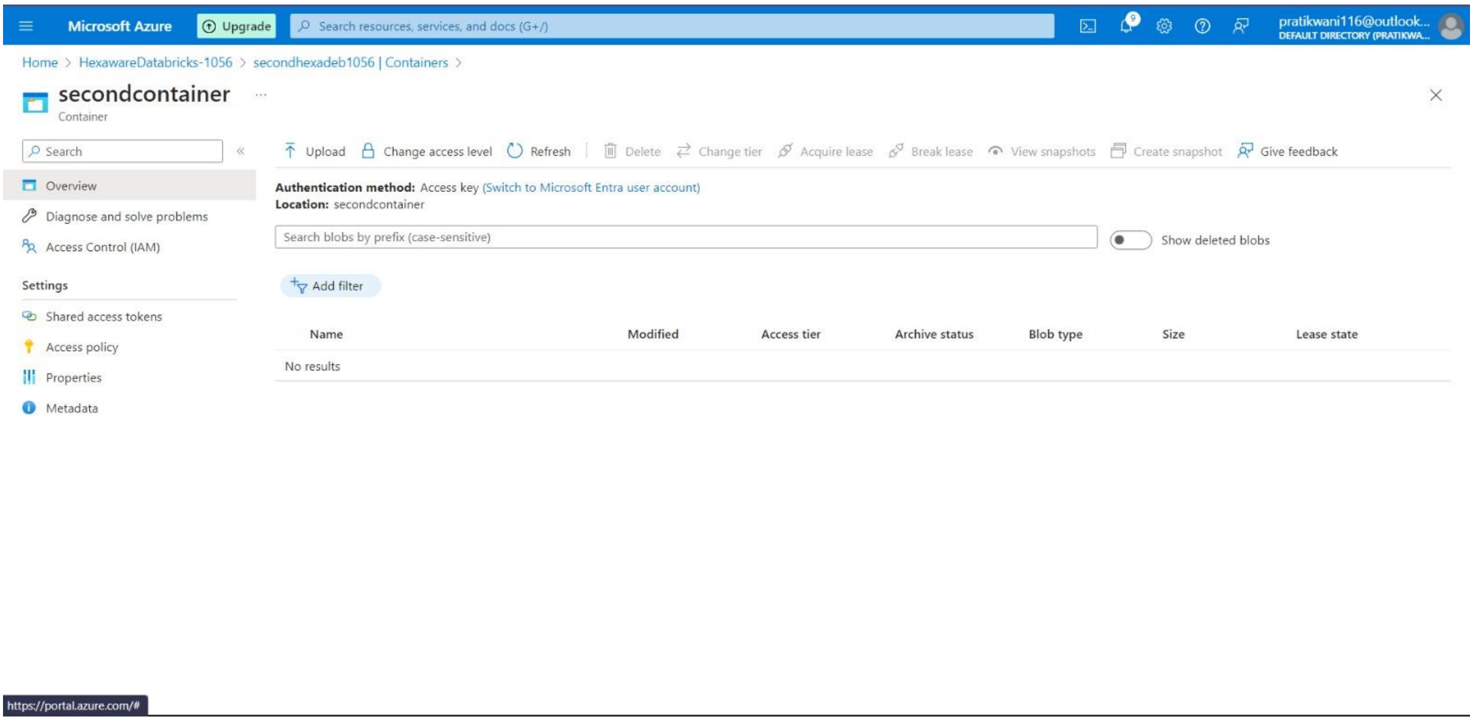
Home > HexawareDatabricks-1056 > firsthexadeb1056 | Containers >

firstcontainer

Container

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

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
Azure Databricks Assignment 1.pdf	21/02/2024, 11:17:39			Block blob	1.45 MiB	Available



Home > HexawareDatabricks-1056 > secondhexadeb1056 | Containers >

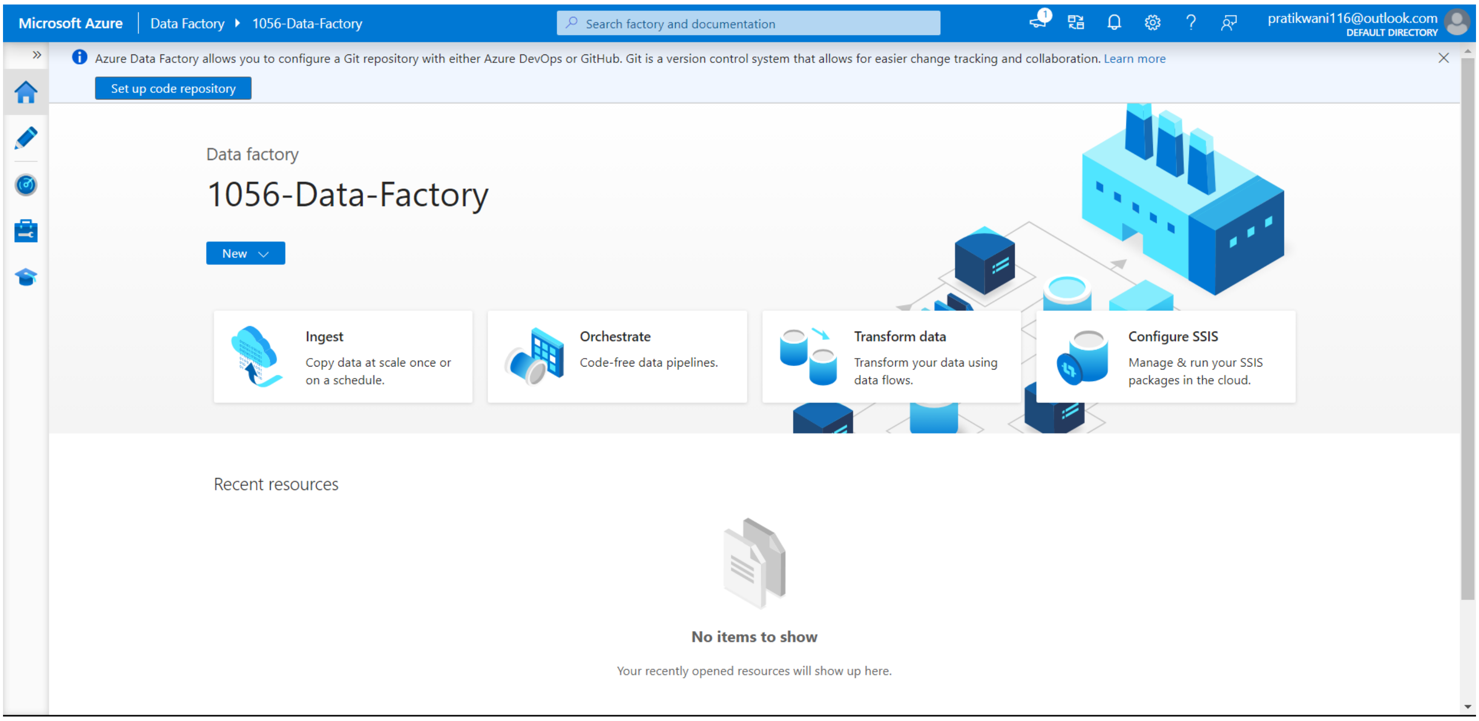
secondcontainer

Container

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

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
No results						

- Create Data Factory



Microsoft Azure | Data Factory ▸ 1056-Data-Factory

Search factory and documentation

pratikwani116@outlook.com
DEFAULT DIRECTORY

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration. Learn more

Set up code repository

Data factory

1056-Data-Factory

New

Ingest
Copy data at scale once or on a schedule.

Orchestrate
Code-free data pipelines.

Transform data
Transform your data using data flows.

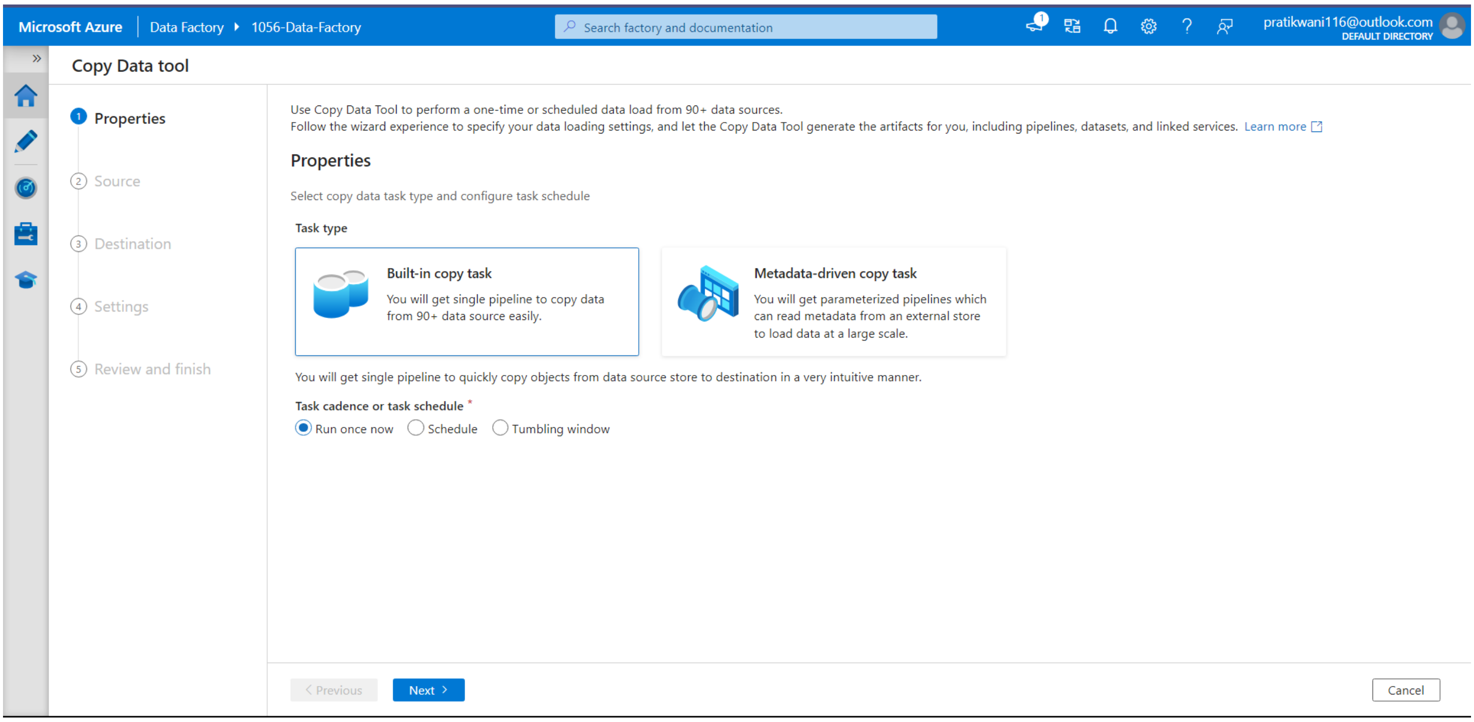
Configure SSIS
Manage & run your SSIS packages in the cloud.

Recent resources

No items to show

Your recently opened resources will show up here.

- Copy Data From one Blob Storage to Another Blob Storage



Microsoft Azure
Data Factory ▸ 1056-Data-Factory
Search factory and documentation
pratikwani116@outlook.com
DEFAULT DIRECTORY

Copy Data tool

1 Properties
2 Source
3 Destination
4 Settings
5 Review and finish

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources.
Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines, datasets, and linked services. Learn more

Properties

Select copy data task type and configure task schedule

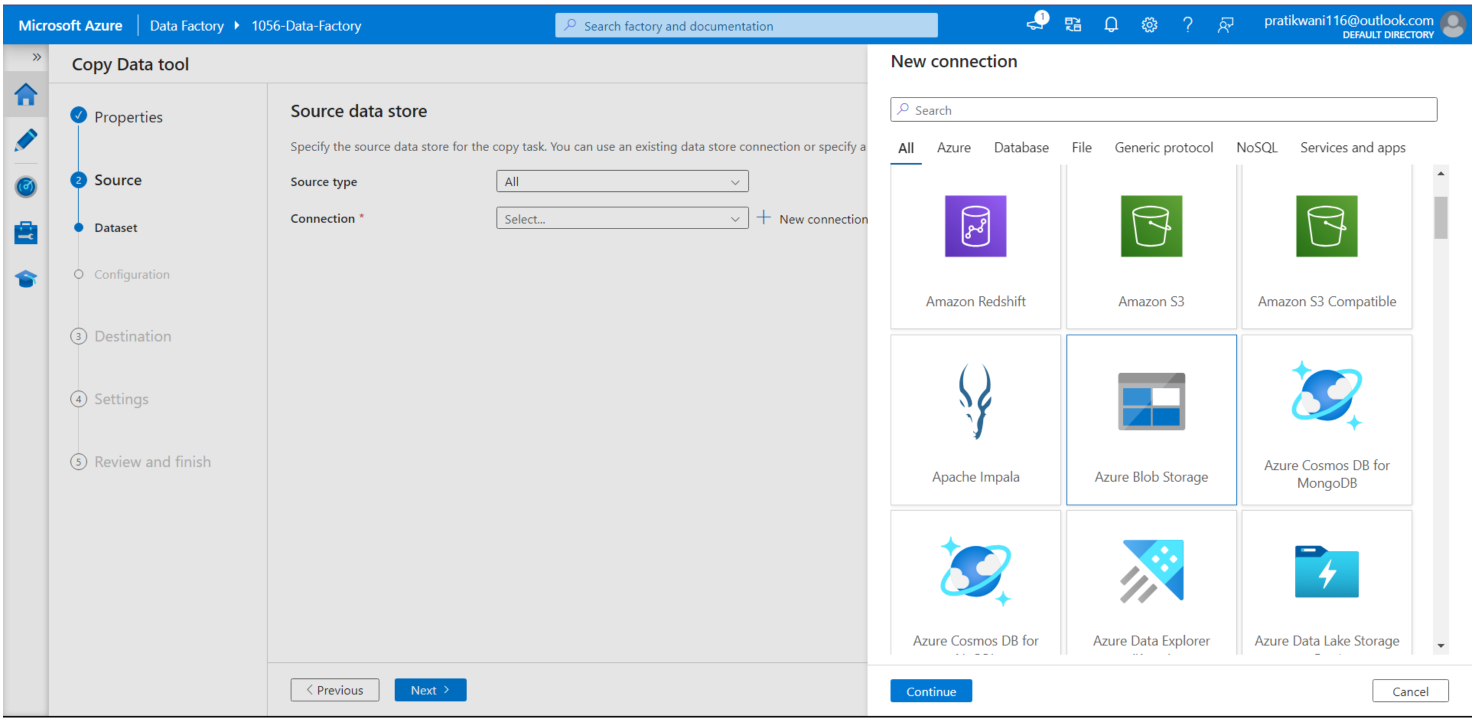
Task type

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.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *
● Run once now ○ Schedule ○ Tumbling window



Microsoft Azure
Data Factory ▸ 1056-Data-Factory
Search factory and documentation
pratikwani116@outlook.com
DEFAULT DIRECTORY

Copy Data tool

✓ Properties
2 Source
Dataset
Configuration
3 Destination
4 Settings
5 Review and finish

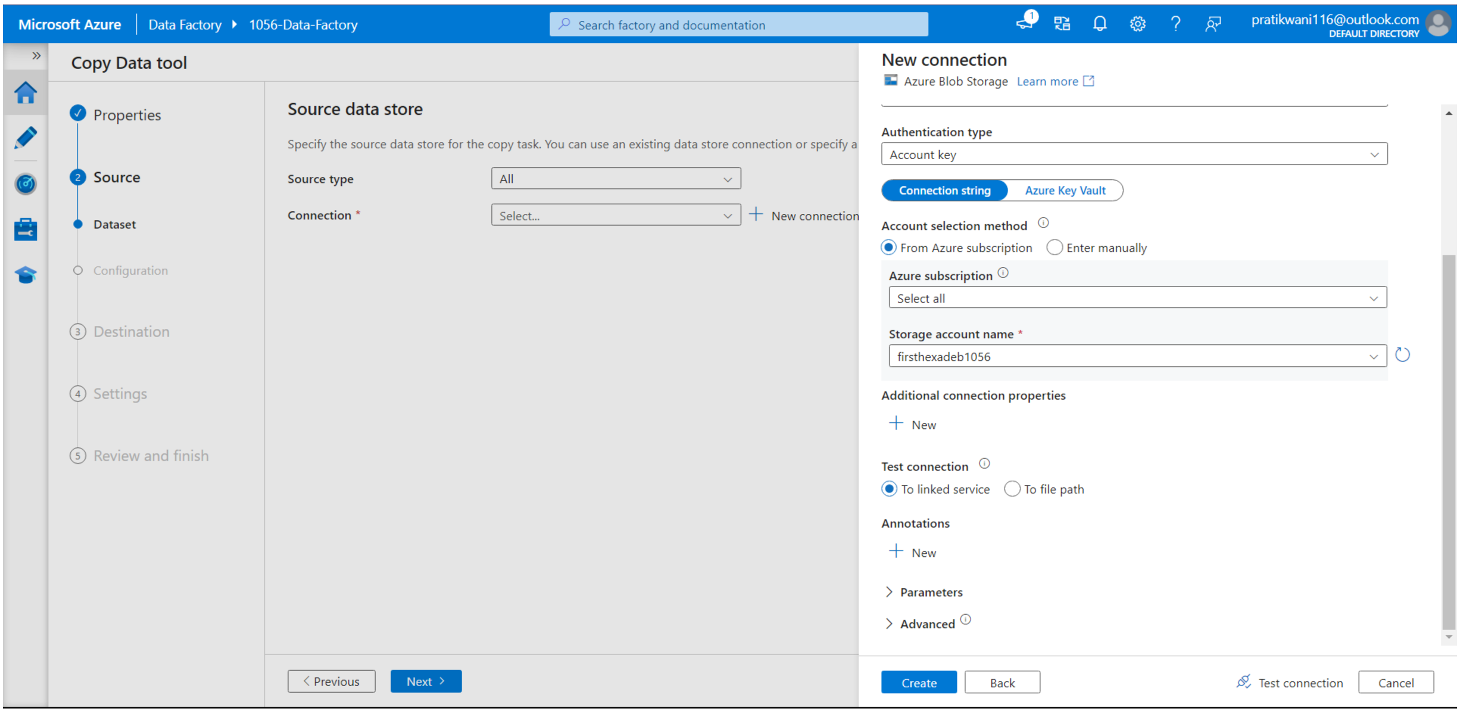
Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a

Source type: All

Connection *: Select... + New connection

New connection



Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a

Source type: All

Connection *: Select... + New connection

New connection

Azure Blob Storage Learn more

Authentication type: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Select all

Storage account name *: firsthexadeb1056

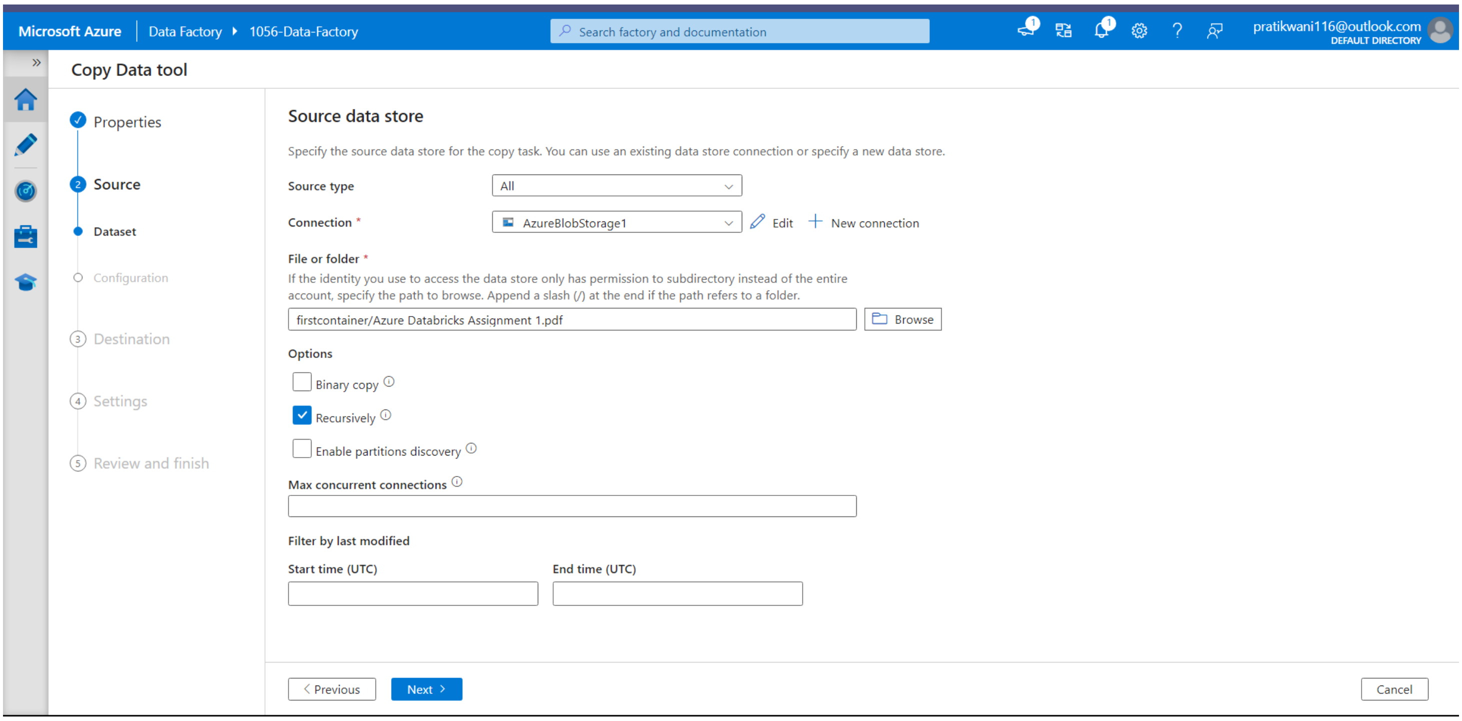
Additional connection properties: + New

Test connection: ● To linked service ○ To file path

Annotations: + New

> Parameters

> Advanced



Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: All

Connection *: AzureBlobStorage1 Edit + New connection

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.

firstcontainer/Azure Databricks Assignment 1.pdf Browse

Options

☐ Binary copy

☑ Recursively

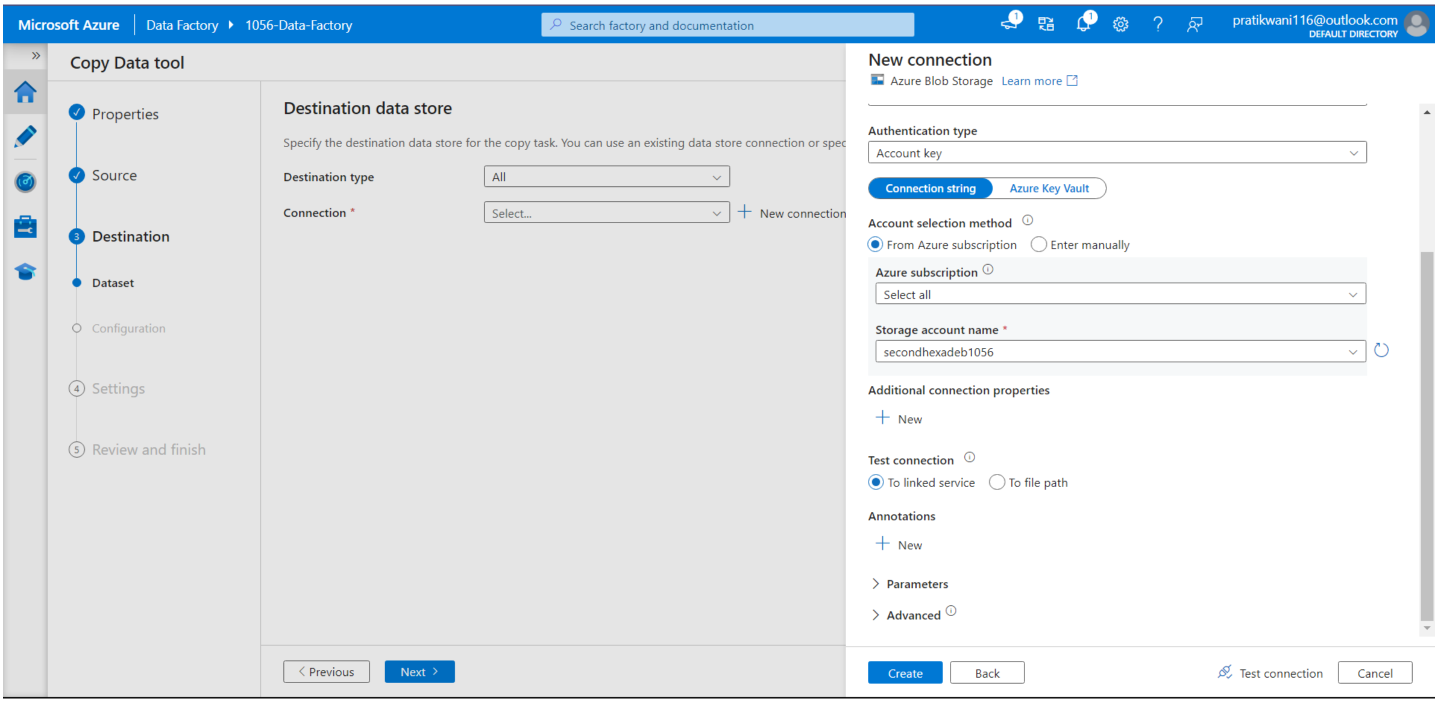
☐ Enable partitions discovery

Max concurrent connections

Filter by last modified

Start time (UTC)

End time (UTC)



Copy Data tool

- Properties
- Source
- 3 Destination
  - Dataset
  - Configuration
- 4 Settings
- 5 Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or spec

Destination type: All

Connection *: Select... + New connection

< Previous   Next >

New connection

Azure Blob Storage Learn more

Authentication type: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Select all

Storage account name *: secondhexadeb1056

Additional connection properties: + New

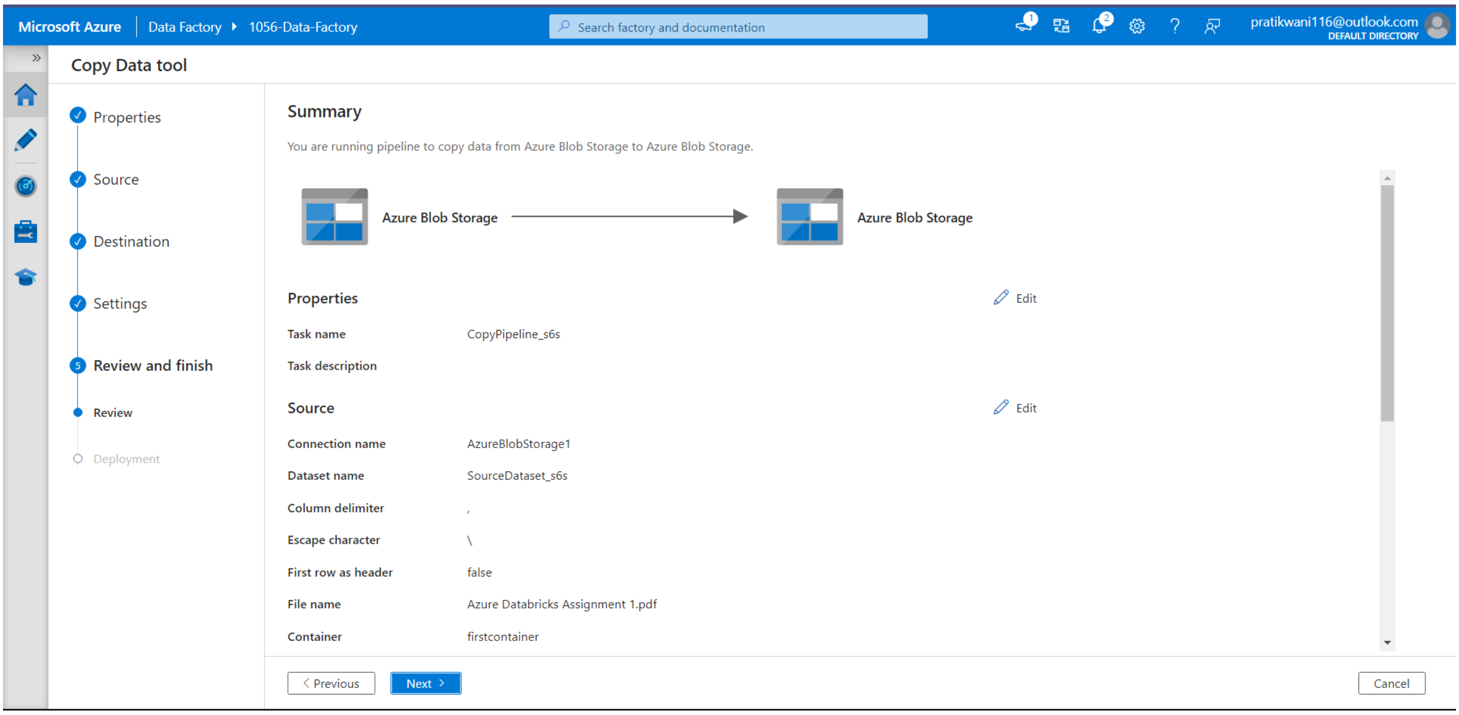
Test connection: ● To linked service ○ To file path

Annotations: + New

> Parameters

> Advanced

Create   Back   Test connection   Cancel



Copy Data tool

- Properties
- Source
- Destination
- Settings
- 5 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

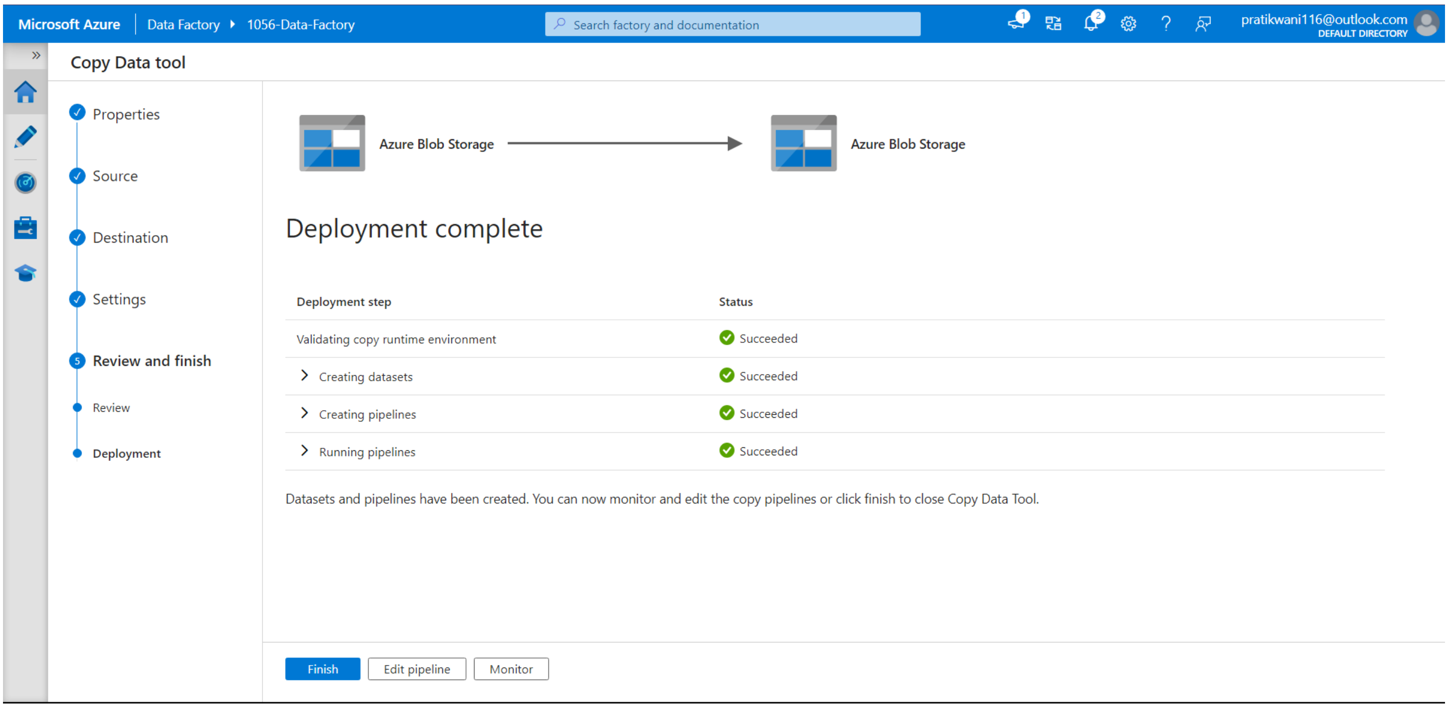
Task name	CopyPipeline_s6s
Task description	

Source Edit

Connection name	AzureBlobStorage1
Dataset name	SourceDataset_s6s
Column delimiter	,
Escape character	\
First row as header	false
File name	Azure Databricks Assignment 1.pdf
Container	firstcontainer

< Previous   Next >   Cancel

- Data Copied Successfully



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.