Azure Databricks Assignment No 7

Pratik Wani

- Notes

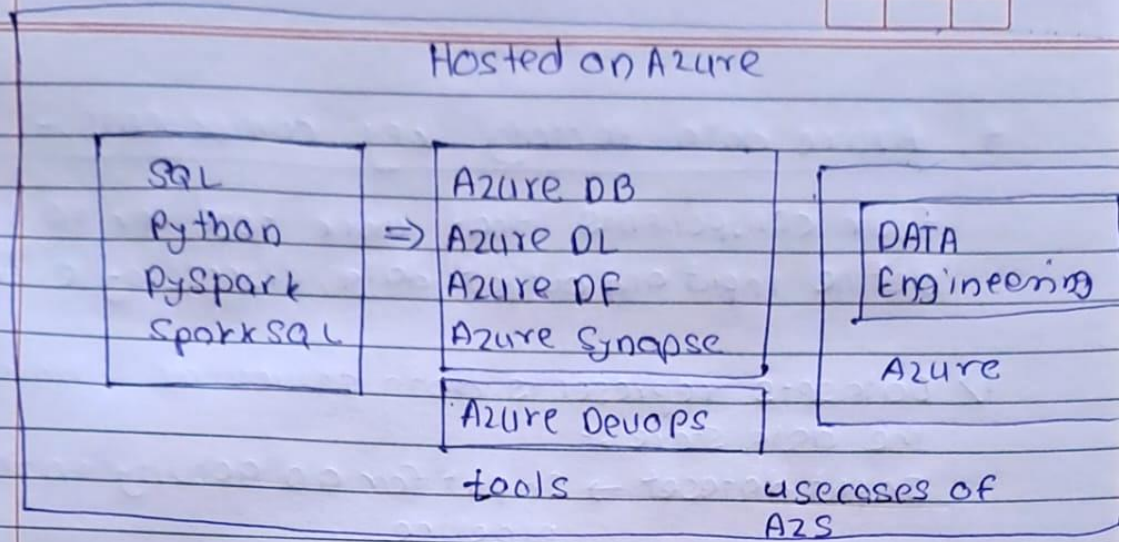## Day 7 Azure Datafactory

* what is Azure data factory?
  - Cloud based data integration service that allows you to create data -driven workflow in the cloud orchestrating and automating data movement and data transformation
  - ADF does not store data itself
  - Also use to monitor and manages workflows using both programmatic and UI mechan'sm

* Uses :-

  ① Supporting data migrations
  ② Getting data from a client's server or online data to an Azure data lake
  ③ Carring out various data integration processes

* How Azure Data factory work?

  - Allows you to create data pipeline that move and transform data and Specified Shedule
  - Data is produced by workflows is time sliced data and we can specify the pipeline made as scheduled.

Hosted on Azure

SQL Python PySpark SparkSQL	⇒	Azure DB Azure DL Azure DF Azure Synapse	DATA Engineering
		Azure Devops	Azure

tools   usecases of
AZS

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 centralized data store in the cloud, it is transformed using Compute services such as HDInsight Hadoop, Spark, Azure data Lake Analytic

Step 3 : Publish :-
- Cloud → on premise sources like SQL server

* **Copy Activity** :- copy data from data store (source) to sink data store

* **Azure data factory key components :-**

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

1) Datasets represents data structures within the data store :-

    Input dataset → Input for an activity in the pipeline

    output dataset → output for the activity

2) A pipeline is a group of Activities :-
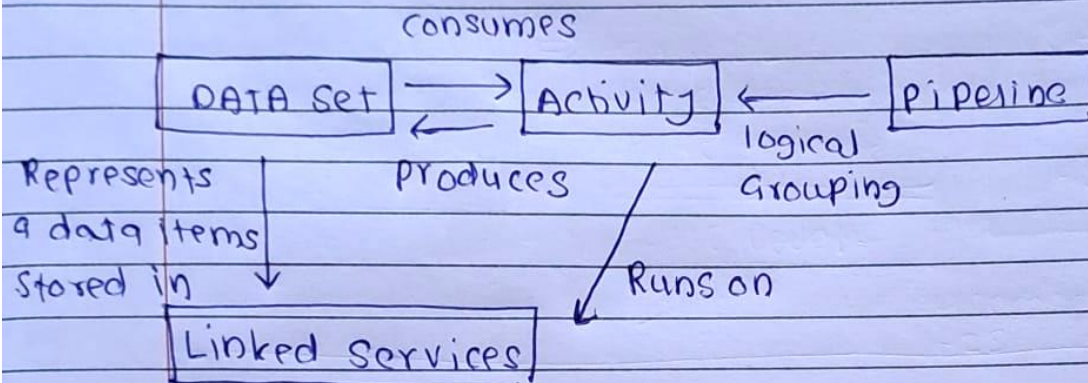- Group activities into a unit that together performs a task
- may have one or more pipelines

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.

consumes

DATA set → Activity ← pipeline

logical
Grouping

Represents   Produces

a data items

Stored in

Runs on

Linked Services

* 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

- Copy Activity

1056hexadeb-datafactory

Data factory (V2)

Search

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

Settings
- Networking
- Managed identities
- Properties
- Locks

Getting started
- Quick start

Monitoring
- Alerts
- Metrics
- Diagnostic settings
- Logs

Delete

∧ Essentials

JSON View

Resource group (move)	: HexawareDatabricks-1056	Type	: Data factory (V2)
Status	: Succeeded	Getting started	: Quick start
Location	: Central India		
Subscription (move)	: Free Trial		
Subscription ID	: 60cd428c-ab8a-474b-b8cc-2aeb893e8d3e		

Azure Data Factory Studio

Launch studio

Quick Starts	Tutorials	Template Gallery	Training Modules

Monitoring

Microsoft Azure | Data Factory ▸ 1056hexadeb-datafactory   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

< Previous   Next >   Cancel

Copy Data tool

New connection

- ✓ 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

🔍 Search

All | Azure | Database | File | Generic protocol | NoSQL | Services and apps
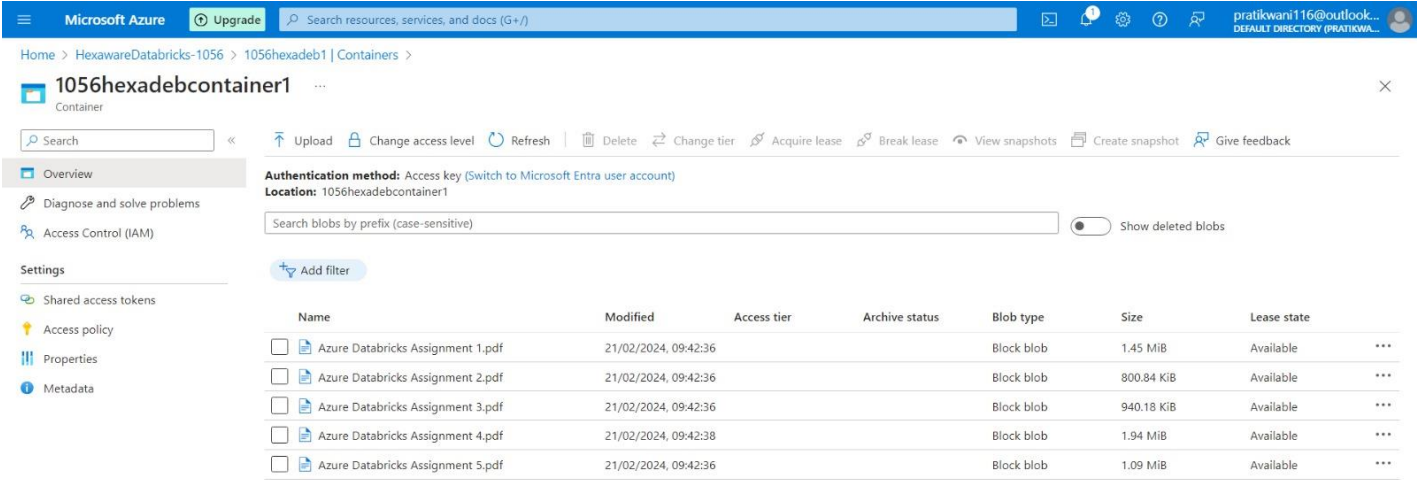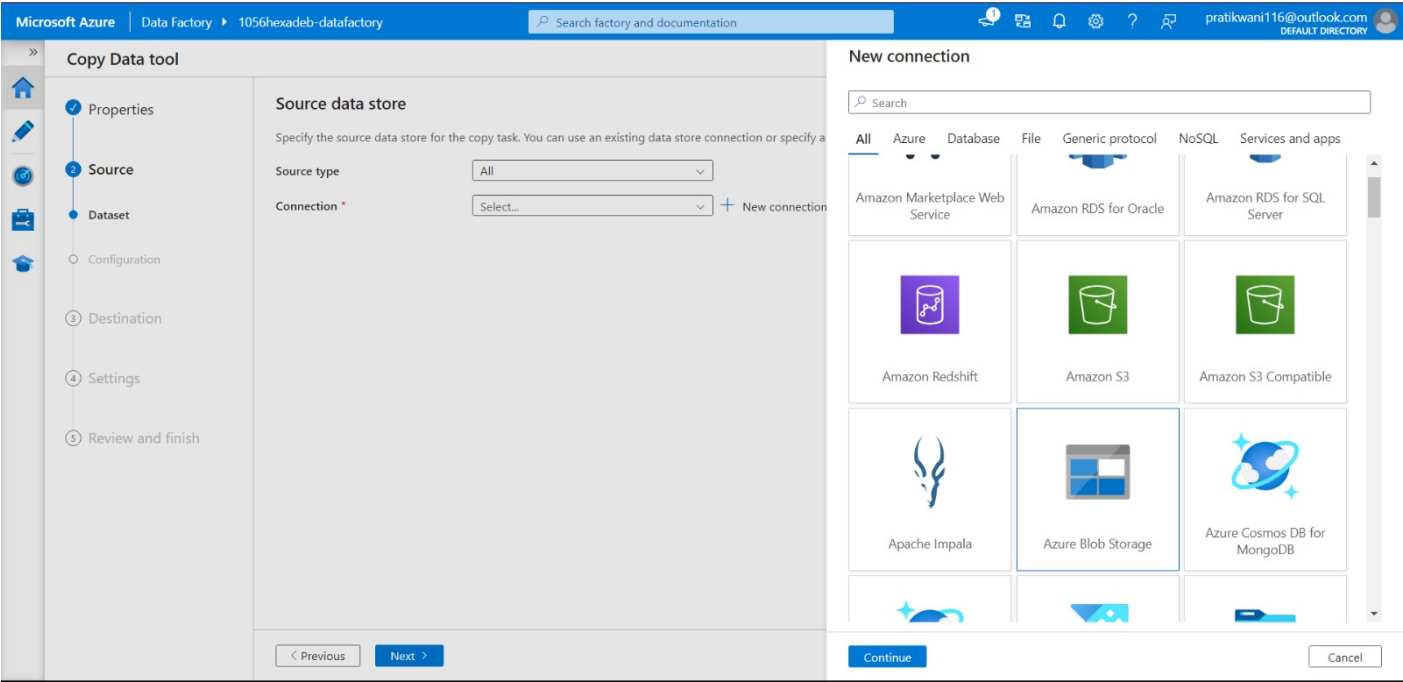
Amazon Marketplace Web Service	Amazon RDS for Oracle	Amazon RDS for SQL Server
Amazon Redshift	Amazon S3	Amazon S3 Compatible
Apache Impala	Azure Blob Storage	Azure Cosmos DB for MongoDB

[Previous] [**Next >**]

[**Continue**] [Cancel]

1056hexadebcontainer1 …
Container

🔍 Search «

- □ Overview
- 🔍 Diagnose and solve problems
- 🔑 Access Control (IAM)

Settings
- 🔗 Shared access tokens
- 🔑 Access policy
- ⬛ Properties
- ⓘ Metadata

⬆ Upload 🔓 Change access level ↻ Refresh | 🗑 Delete ⇄ Change tier 🖉 Acquire lease 🖉 Break lease 👁 View snapshots 🗍 Create snapshot 🗨 Give feedback

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: 1056hexadebcontainer1

Search blobs by prefix (case-sensitive)   ⚪ Show deleted blobs

+ Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
☐ 📄 Azure Databricks Assignment 1.pdf	21/02/2024, 09:42:36			Block blob	1.45 MiB	Available	⋯
☐ 📄 Azure Databricks Assignment 2.pdf	21/02/2024, 09:42:36			Block blob	800.84 KiB	Available	⋯
☐ 📄 Azure Databricks Assignment 3.pdf	21/02/2024, 09:42:36			Block blob	940.18 KiB	Available	⋯
☐ 📄 Azure Databricks Assignment 4.pdf	21/02/2024, 09:42:38			Block blob	1.94 MiB	Available	⋯
☐ 📄 Azure Databricks Assignment 5.pdf	21/02/2024, 09:42:36			Block blob	1.09 MiB	Available	⋯

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_9bs
Task description	

Source   ✏ Edit

Connection name	AzureBlobStorage1
Dataset name	SourceDataset_9bs
Column delimiter	,
Escape character	\
First row as header	false
Container	1056hexadebcontainer1

Destination   ✏ Edit

[< Previous] [Next >]   [Cancel]

Copy Data tool

- ✓ Properties
- ✓ Source
- ✓ Destination
- ✓ Settings
- ⑤ 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]