

Name: Rohan Vinayak Chaudhari

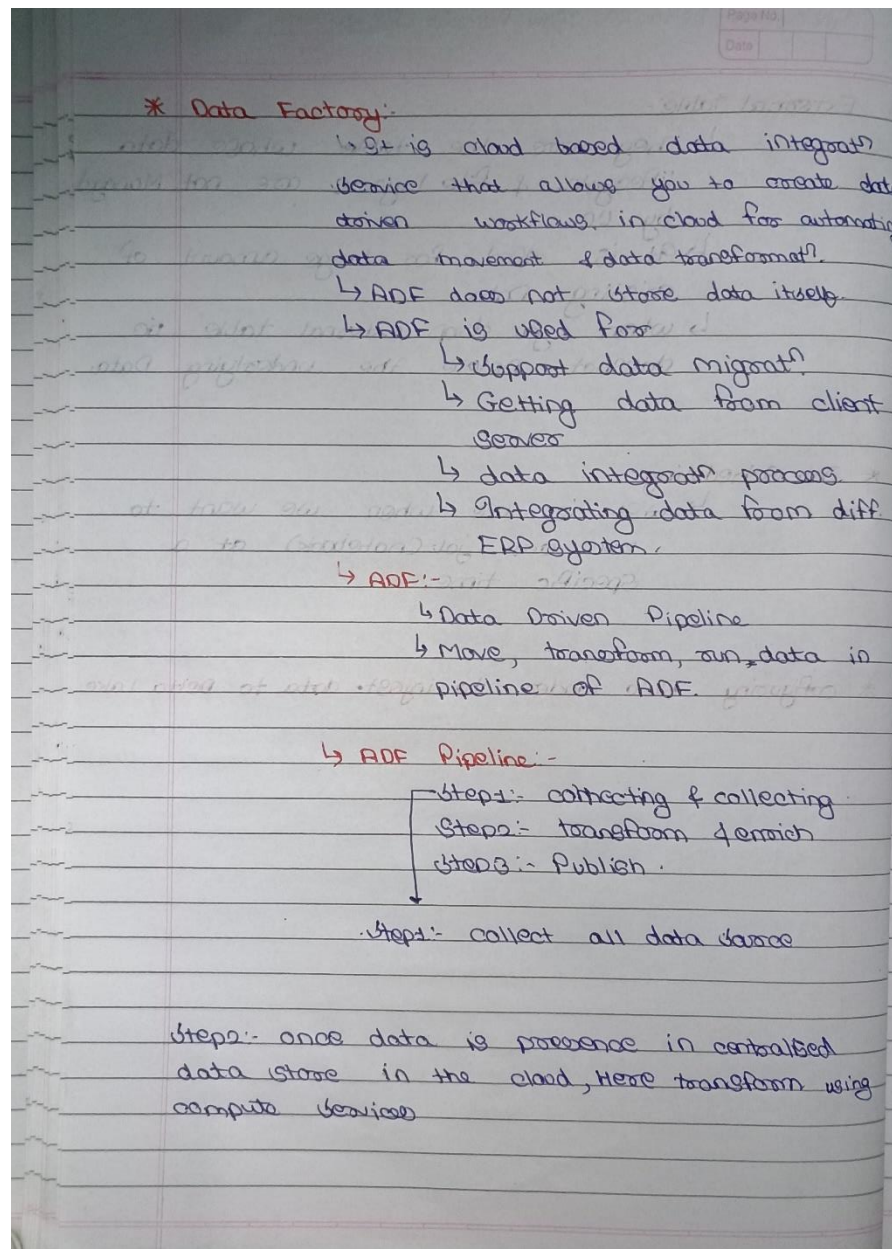
Batch: Data Engineering

Date: 20/02/2024

Topic: Azure DataBricks

Solution:

1. Azure Databricks (Data Factory):



- Data Migration:-

↳ By ADF, data migration occurs betⁿ 2 cloud store & between on premisen data store & a cloud data store.

↳ copy activity in ADF copies data from a Source data to a Sink data store.

- ADF Key component:-

↳ Dataset represent Data structure in ^{Datastore}

↳ Pipeline is a gap of activity

↳ It is gap of activity & ADF can contain 1 or more pipeline.

For eg:- a pipeline could contain a gap of activities that inject data from blob & run on Hive query on an HDInsight cluster.

Tools to create datapipeline in ADF:-

Azure portal

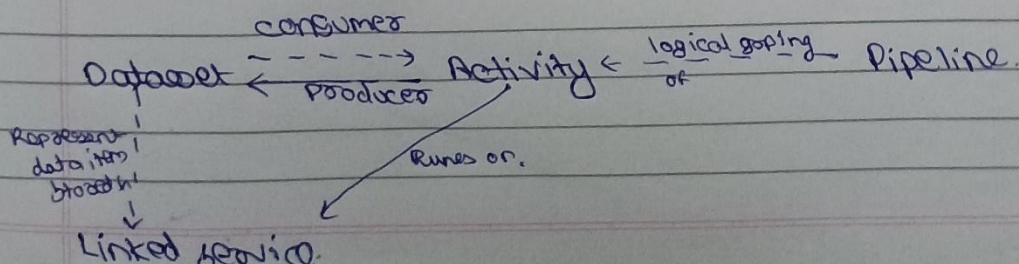
VS code

PowerShell

net API,

↳ Activities define the cluster to perform on data.

↳ Linked Services define the info needed for Azure data Factory to connect to external source.



Job Scheduling:

The screenshot displays the Databricks workspace interface, divided into two main sections: the top notebook editor and the bottom job management view.

Top Section (Notebook Editor):

- Header:** "DataFrame_Practise 2024-02-19 09:45:24" with a Python language selector and a star icon.
- Left Sidebar:** Contains navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Experiments, and Features.
- Main Editor:** Shows a notebook with three commands:
 - Cmd 1:** A simple `spark` command.
 - Cmd 2:** A `df=spark.read.csv("/FileStore/tables/output_file.csv", header=True, inferSchema=True)` command, followed by a display of the first 10 rows of the resulting DataFrame.
 - Cmd 3:** A `df.show()` command.
- Right Panel:** A "Job" configuration panel for the current notebook. It includes a "Run now" button, a "Schedule" button, and a "Share" button. Below these, it shows the job name "Job - At 09:34 AM - DataFrame_Practise 2024-02-19 094524" and a "Last run: No runs" status. There is also an "Add a schedule" button.

Bottom Section (Job Management View):

- Header:** "DataFrame_Practise 2024-02-19 094524" with a star icon and a "Run now" button.
- Left Sidebar:** Similar to the top section, but with "Workflows" and "Jobs" tabs.
- Main View:** A "Runs" section showing a bar chart of "Run total duration" over time. The chart shows a single green bar for the job on Feb 20, 2024, at 09:34 AM, with a duration of 22s. Below the chart is a table of runs:

Start time	Run ID	Launched	Duration	Spark	Status	Run parameters
Feb 20, 2024, 09:34...	560668751...	By scheduler	22s	Spark UI / Logs / Metrics	Running	

Right Panel (Job details):

- Job ID:** 957201426543825
- Creator:** azuser1076_mml.local
- Run as:** azuser1076_mml.local
- Tags:** Add tag
- Description:** Add description
- Git:** Not configured. Add Git settings
- Schedule:** At 09:34 AM (UTC+05:30 — undefined). Buttons: Edit schedule, Pause, Delete.
- Compute:** Section for configuring the compute environment.



Microsoft Azure <azure-noreply@microsoft.com>
to me

9:35AM (0 minutes ago) ☆ 😊 ↩ ⋮

Microsoft Azure

Azure Databricks

Your Azure Databricks job run has started

Run details

Workspace	hexa-deb-1076 (1553062149404298)
Job	DataFrame_Practise 2024-02-19 094524 (957201426543825)
Job run	560668751822764
Started at	2024-02-20 04:04:57 UTC
Launched	By scheduler

[View run in Databricks >](#)

DataFrame_Practise 2024-02-19 09...

Job ID	957201426543825
Job run ID	560668751822764
Launched	By scheduler
Started	02/20/2024, 09:34:57 AM
Ended	02/20/2024, 09:35:18 AM
Duration ⓘ	20s
Queue duration ⓘ	-
Status	Succeeded

[Next >](#)

Feb 20





Microsoft Azure <azure-noreply@microsoft.com>
to me ▾

9:35 AM (11 minutes ago) ☆ 😊 ↶ ⋮

Microsoft Azure

Azure Databricks

Your Azure Databricks job has finished its run

Run details

Workspace	hexa-deb-1076 (1553062149404298)
Job	DataFrame_Practise 2024-02-19 094524 (957201426543825)
Job run	560668751822764
Status	✔ Success
Started at	2024-02-20 04:04:57 UTC
Duration	20s
Launched	By scheduler

Data Factory:

Microsoft Azure

Search resources, services, and docs (G+I)

Home >

Microsoft.DataFactory-20240220225920 | Overview

Deployment

Search

«

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

✔ Your deployment is complete

Deployment name : Microsoft.DataFactory-20240220225920

Subscription : Azure subscription 1

Resource group : rg-azuser1076_mml.local-di1ea

Start time : 20/02/2024, 23:01:10

Correlation ID : 39999c08-728b-45cc-9a94-86db77df13eb

> Deployment details

> Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

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.

Microsoft Azure

Search resources, services, and docs (G+V)

azuser1076_mml.local@...
IHIT (IHIT.ONMICROSOFT.COM)

Home > Microsoft.DataFactory-20240220225920 | Overview >

adfhexa1076
Data factory (V2)

Search

Delete

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Networking

Managed identities

Properties

Locks

Getting started

Quick start

Monitoring

Alerts

Metrics

Essentials

Resource group (move) : rg-azuser1076_mml.local-di1ea

Status : Succeeded

Location : East US


Subscription (move) : Azure.subscription.1

Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08

Type : Data factory (V2)

Getting started : [Quick start](#)

JSON View



Azure Data Factory Studio

Launch studio

Quick Starts

Tutorials

Template Gallery

Training Modules

Microsoft Azure

Data Factory > adfhexa1076

Search factory and documentation

azuser1076_mml.local@ihit.onmicrosoft.com
IHIT

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

Microsoft Azure

Data Factory > adfhexa1076

Search factory and documentation

azuser1076_mml.local@ihit.onmicrosoft.com
IHIT

Copy Data tool

1 Properties

2 Source

3 Dataset

4 Configuration

5 Destination

6 Settings

7 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 new data store.

Source type

Connection * [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.

[Browse](#)

Options

☒ Binary copy

Compression type

☒ Recursively

☐ Delete files after completion

Max concurrent connections

Filter by last modified

< Previous Next > Cancel

Microsoft AzureData Factoryadhexa1076Search factory and documentationazuser1076_mml.local@ihl.onmicrosoft.com

Copy Data tool

1 Properties

2 Source

3 Destination

4 Dataset

5 Configuration

6 Settings

7 Review and finish

Destination data store

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

Destination typeAzure Blob Storage

Connection *fordestinationEdit+ New connection

Folder path *
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.
destinationdata/Browse

File name
Filenames are defined by source

Compression type
None

Copy behavior ①
Select...

Max concurrent connections ①

Block size (MB) ①

< PreviousNext >Cancel

Microsoft AzureData Factoryadhexa1076Search factory and documentationazuser1076_mml.local@ihl.onmicrosoft.com

Copy Data tool

1 Properties

2 Source

3 Destination

4 Settings

5 Review and finish

6 Review

7 Deployment

Summary

You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage.

Source

Connection nameforsourceEdit

Dataset nameSourceDataset_g55

Containersourcedata

Destination

Connection namefordestinationEdit

Dataset nameDestinationDataset_g55

Copy settings

Timeout0.12:00:00

Retry0

Retry interval (sec)30

Secure outputfalse

Secure inputfalse

< PreviousNext >Cancel

Microsoft AzureData Factory> adfhexa1076

Search factory and documentation

azuser1076_mm1.local@iitl.onmicrosoft.com

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 AzureData Factory> adfhexa1076

Search factory and documentation

azuser1076_mm1.local@iitl.onmicrosoft.com

Data Factory

Validate all

Publish all

Preview experienceOff

Factory Resources

Filter resources by name

Pipelines1

CopyActivity_1076

Change Data Capture (preview)0

Datasets2

Data flows1

dataflow1

Power Query0

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

CopyActivity_1076

Copy data

Copy_g55

Parameters

Variables

Settings

Output

Pipeline run ID: 027683e4-9b63-4c91-a35d-4f4904afced5

Pipeline statusIn progress

Monitor in Azure Metrics

Export to CSV

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
Copy_g55	Queued	Copy data	2/20/2024, 11:17:19 PM	4s	Us

Source data:

Microsoft Azure

Search resources, services, and docs (G+)

azuser1076 mm1.local@...
BHT (BHTLONMICROSOFT.COM)

Home > forsource | Containers >

sourcedata

Container

Search

«

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Create snapshot

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

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

Location: sourcedata

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/>	output_file.csv	20/02/2024, 17:19:11	Hot (Inferred)		Block blob	61.42 KiB	Available ***
<input type="checkbox"/>	output_file1.json	20/02/2024, 23:12:44	Hot (Inferred)		Block blob	161.8 KiB	Available ***

Destination Data:

Microsoft Azure

Search resources, services, and docs (G+)

azuser1076 mm1.local@...
BHT (BHTLONMICROSOFT.COM)

Home > fordestination | Containers >

destinationdata

Container

Search

«

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Create snapshot

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

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

Location: destinationdata

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/>	output_file.csv	20/02/2024, 23:17:33	Hot (Inferred)		Block blob	61.42 KiB	Available ***
<input type="checkbox"/>	output_file1.json	20/02/2024, 23:17:33	Hot (Inferred)		Block blob	161.8 KiB	Available ***