NAME-MALLU SRAVANA ANDHYA

MAIL.ID-mallusravanasandhya9948@gmail.com

DATA PROCESSING

Project Statement

Implement a data processing pipeline where Azure Data Factory orchestrates data workflows, and Azure Databricks is used as a processing engine for on-demand analytics and transformations.

Project Overview

This project can implement a data processing pipeline using Azure Data Factory (ADF) and Azure Databricks. This setup allows you to orchestrate, schedule, and automate data workflows using ADF while leveraging the power of Databricks for on-demand analytics and transformations.

Project Requirements

Azure Databricks:

Create a Databricks workspace from the Azure portal. This workspace will act as the processing engine for your data pipeline.

Azure Data Factory:

Create a Databricks Linked Service in ADF. This linked service will enable ADF to connect to your Databricks workspace and submit jobs for execution**.**

Data processing:

Define the necessary data inputs and outputs for your data processing pipeline. This could involve creating datasets in ADF.

Execution Overview

- Azure Databricks is responsible for creating a notebook and loading, processing and analysis the data
- Azure Data Factory is used to linked service in ADF.This linked service will enable ADF to connect to Databricks Workspace and submit jobs for execution in Pipelines

Tools/Technology used in project

- ✓ Azure Databricks
- ✓ Azure Data Factory
- ✓ Linked services
- ✓ Pipelines
- ✓ Data Processing
- ✓ Python

Prerequisites

Azure Subscription: Ensure you have an active Azure Subscription to provision the required services.

Access to Azure Portal: You'll need to Azure portal to create and manage resources.

Source Data Files:

We are referring to the source file which is used to run the code in the notebook.

Filename	Type
Diamond	csv

Steps to implement:

To implement this data processing pipeline, you can follow these steps:

Set up Azure Data Factory (ADF) to create a new pipeline to define the data workflow.

Create an Azure Databricks workspace.

Install and configure the Databricks connectors for both Azure Blob Storage and Azure Data Lake Storage Gen2.

Upload the Databricks Notebook to the Azure Databricks workspace.

Create a Databricks linked service in ADF.

Configure the Databricks Notebook activity in ADF.

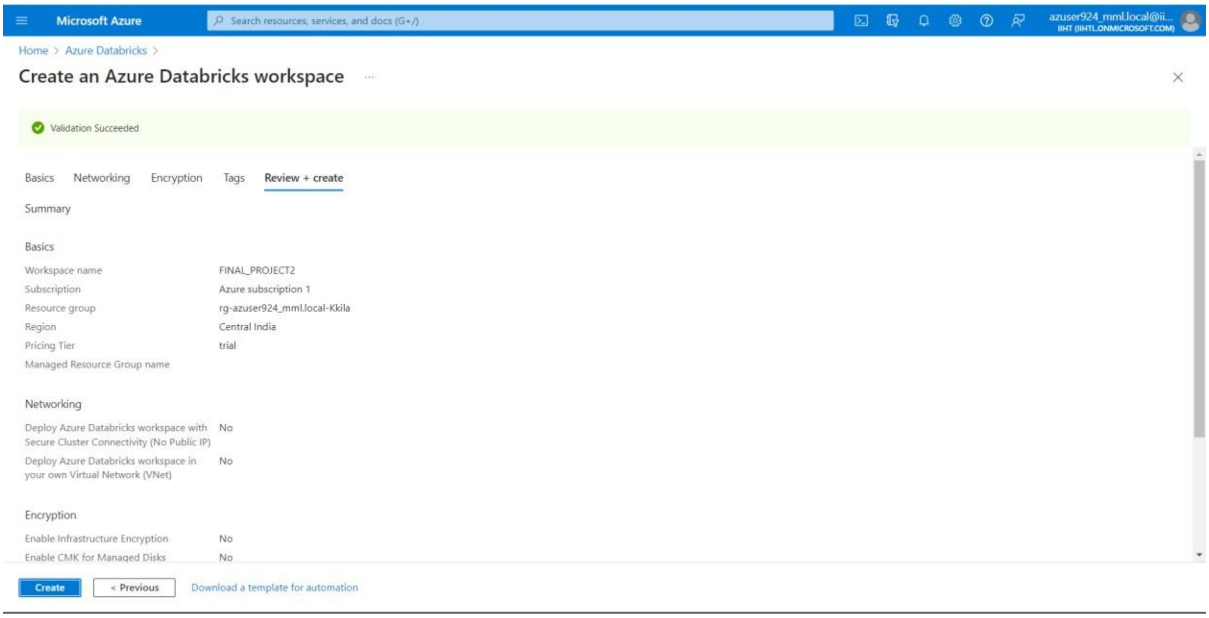
Connect the Copy activity and the Databricks Notebook activity in the ADF pipeline.

Trigger the pipeline execution to initiate the data workflow.

Monitor the progress and view the execution details.

IMPLEMENTING PRACTICALLY IN AZURE PORTAL:

Creating Azure Databricks Workspace:


Microsoft Azure
Search resources, services, and docs (G+/)
azuser924_mml.local@ii...
IIHT (IIHTL.ONMICROSOFT.COM)
Home > Azure Databricks >
Basics
Networking
Encryption
Tags
Review + create
Create
< Previous
Download a template for automation

Create an Azure Databricks workspace

Validation Succeeded

Summary

Basics

Workspace name	FINAL_PROJECT2
Subscription	Azure subscription 1
Resource group	rg-azuser924_mml.local-Kkila
Region	Central India
Pricing Tier	trial
Managed Resource Group name	

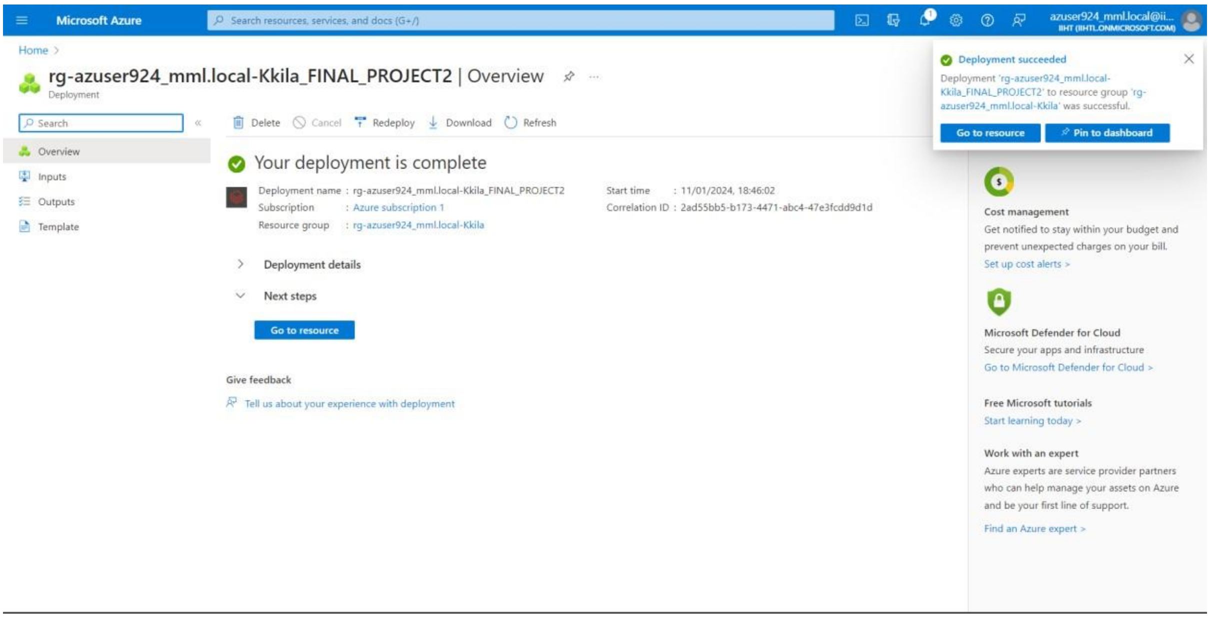
Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)	No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)	No

Encryption

Enable Infrastructure Encryption	No
Enable CMK for Managed Disks	No

Databricks Workspace is successfully created:



rg-azuser924_mml.local-Kkila_FINAL_PROJECT2 | Overview

Deployment

Your deployment is complete

Deployment name	: rg-azuser924_mml.local-Kkila_FINAL_PROJECT2	Start time	: 11/01/2024, 18:46:02
Subscription	: Azure subscription 1	Correlation ID	: 2ad55bb5-b173-4471-abc4-47e3fcdd9d1d
Resource group	: rg-azuser924_mml.local-Kkila		

Deployment details

Next steps

Give feedback

Tell us about your experience with deployment

Deployment succeeded

Deployment 'rg-azuser924_mml.local-Kkila_FINAL_PROJECT2' to resource group 'rg-azuser924_mml.local-Kkila' was successful.

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

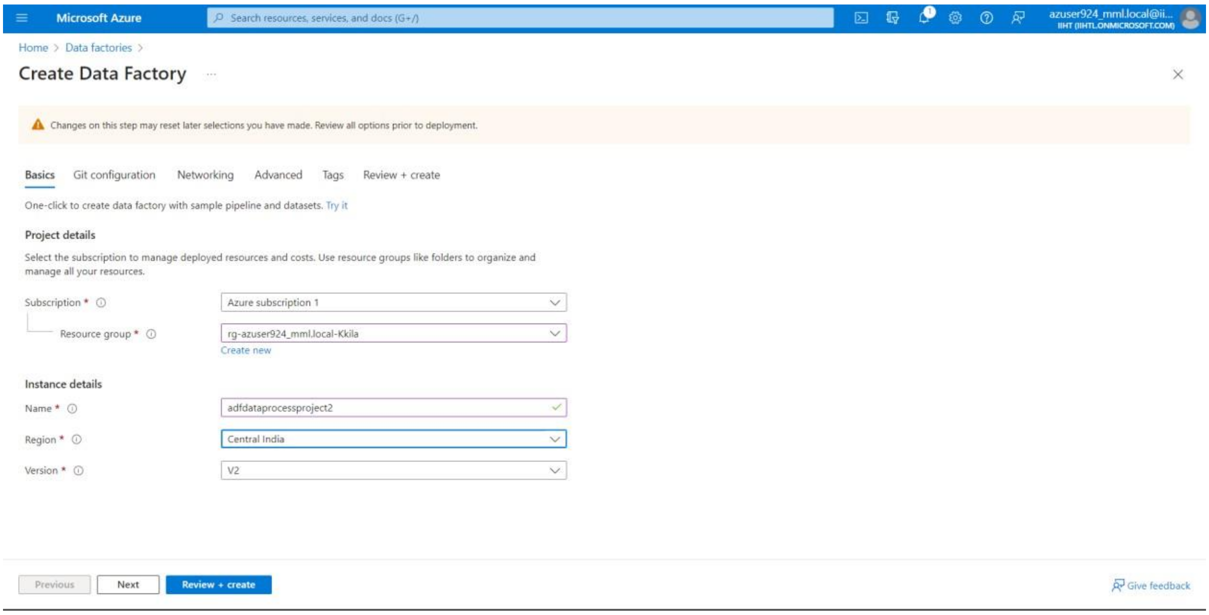
Start learning today >

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.

Find an Azure expert >

Create A Data Factory in azure portal:



Microsoft Azure
Search resources, services, and docs (G+/)
azuser924_mml.local@ii...
IIHT (IIHTL.ONMICROSOFT.COM)
Home > Data factories >

Create Data Factory

Changes on this step may reset later selections you have made. Review all options prior to deployment.

One-click to create data factory with sample pipeline and datasets. Try it

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : Azure subscription 1

Resource group * : rg-azuser924_mml.local-Kkila

Create new

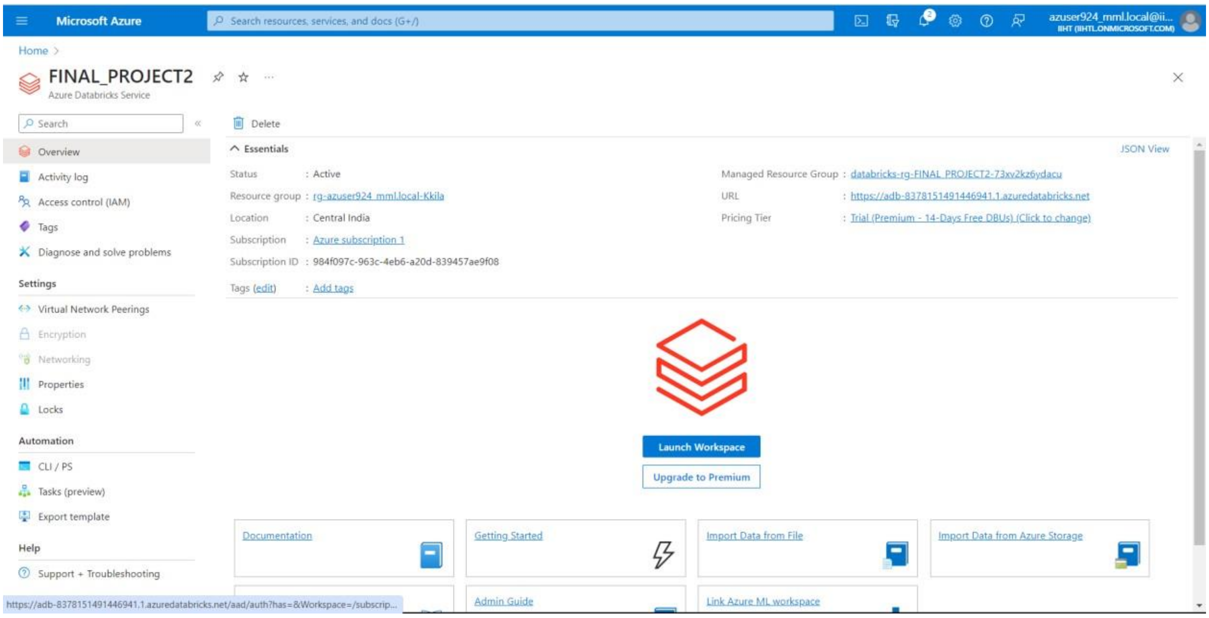
Instance details

Name * : adfdataprocessproject2

Region * : Central India

Version * : V2

Lanching Azure databricks workspace:



Microsoft Azure
Search resources, services, and docs (G+/)
azuser924_mml.local@ii...
IIHT (IIHTL.ONMICROSOFT.COM)
Home >

FINAL_PROJECT2

Azure Databricks Service

Status : Active

Resource group : rg-azuser924_mml.local-Kkila

Location : Central India

Subscription : Azure subscription 1

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

Tags (edit) : Add tags

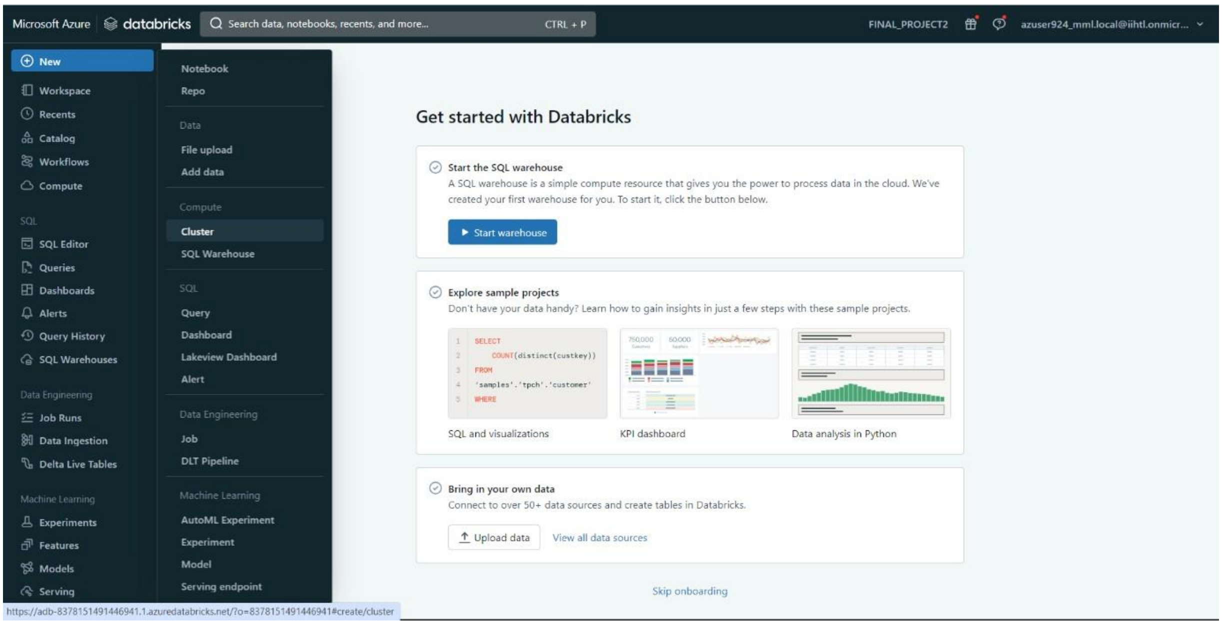
Managed Resource Group : databricks-rg-FINAL_PROJECT2-73xv2kz6ydacu

URL : https://adb-8378151491446941.1.azuredatabricks.net

Pricing Tier : Trial (Premium - 14-Days Free DBUs) (Click to change)

https://adb-8378151491446941.1.azuredatabricks.net/aad/auth?has=&Workspace=/subscrip...

To create a cluster go to new and select cluster:



Get started with Databricks

Start the SQL warehouse

A SQL warehouse is a simple compute resource that gives you the power to process data in the cloud. We've created your first warehouse for you. To start it, click the button below.

Explore sample projects

Don't have your data handy? Learn how to gain insights in just a few steps with these sample projects.

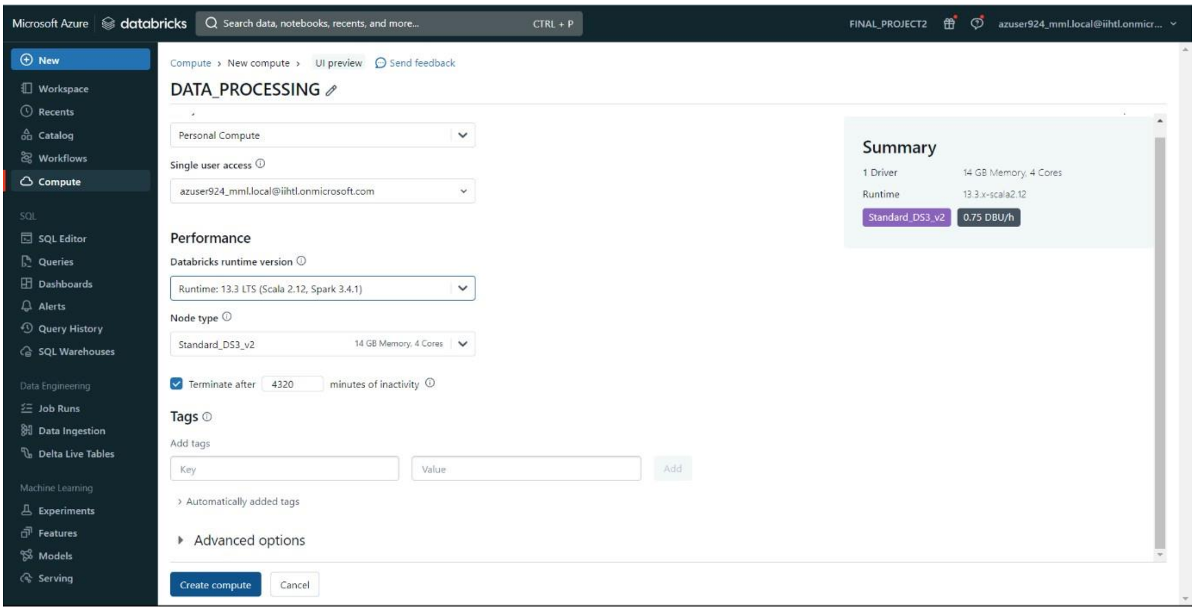
```
SELECT
    COUNT(distinct(custkey))
FROM
'samples'.'tpch'.'customer'
WHERE
```

Bring in your own data

Connect to over 50+ data sources and create tables in Databricks.

https://adb-8378151491446941.1.azuredatabricks.net/?o=8378151491446941#create/cluster

Give a name to the cluster and select the policy as personal compute, select theruntime version from the dropdown then click on create compute



DATA_PROCESSING

Personal Compute

Single user access: azuser924_mml.local@iihtl.onmicrosoft.com

Performance

Databricks runtime version: Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1)

Node type: Standard_DS3_v2 14 GB Memory, 4 Cores

☑ Terminate after 4320 minutes of inactivity

Tags

Add tags

Key | Value

› Automatically added tags

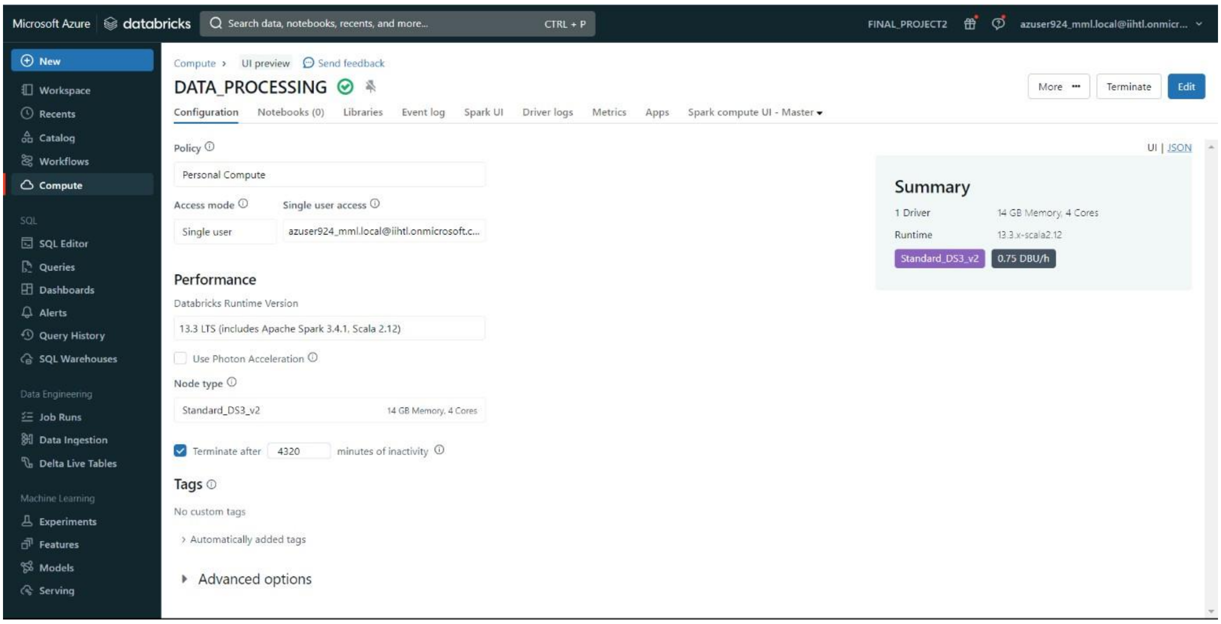
Advanced options

Summary

1 Driver	14 GB Memory, 4 Cores
Runtime	13.3.x-scala2.12

Standard_DS3_v2 0.75 DBU/h

Cluster is successfully created:



DATA_PROCESSING

Policy: Personal Compute

Access mode: Single user

Single user access: azuser924_mml.local@iihtl.onmicrosoft.c...

Performance

Databricks Runtime Version: 13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)

□ Use Photon Acceleration

Node type: Standard_DS3_v2 14 GB Memory, 4 Cores

☑ Terminate after 4320 minutes of inactivity

Tags

No custom tags

› Automatically added tags

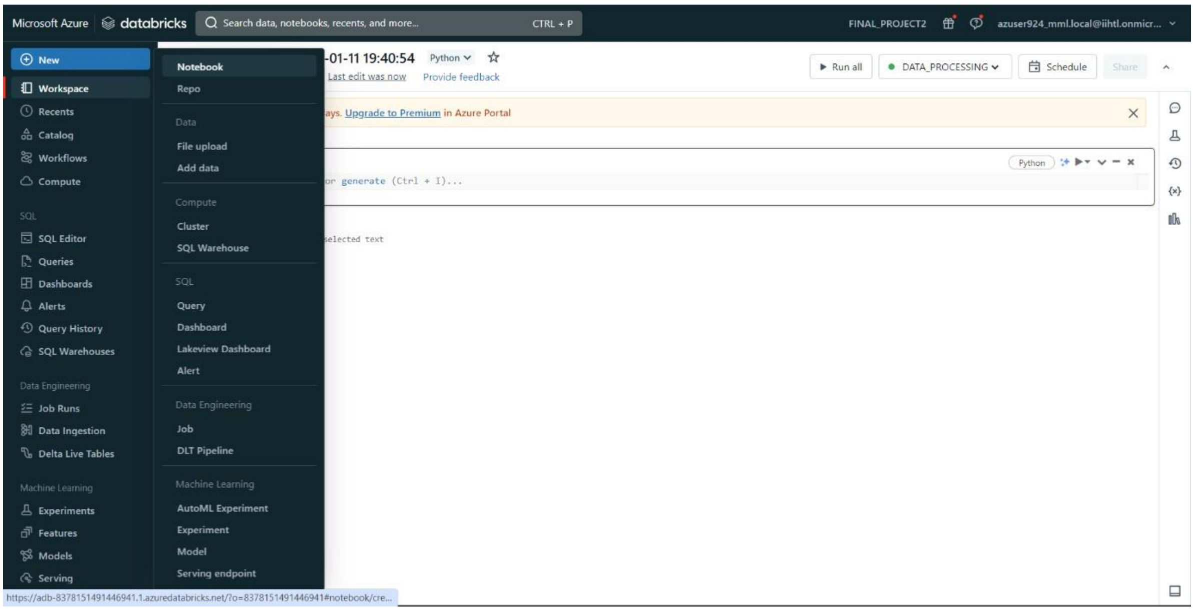
Summary

1 Driver 14 GB Memory, 4 Cores

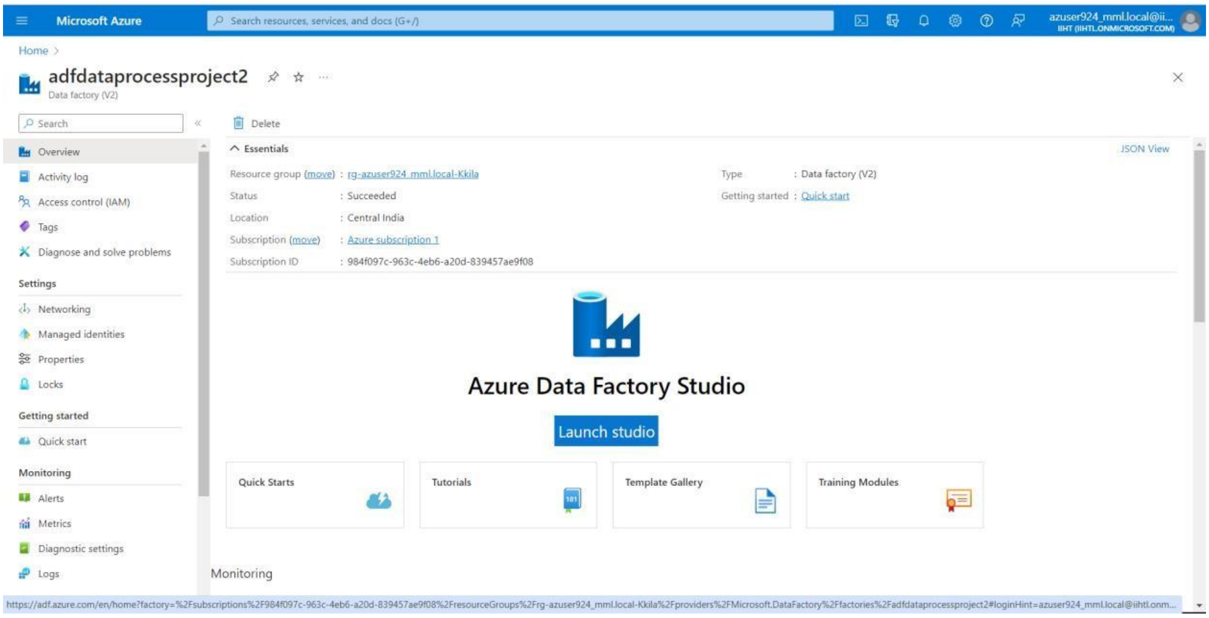
Runtime 13.3.x-scala2.12

Standard_DS3_v2 0.75 DBU/h

Create a notebook as shown in the figure:



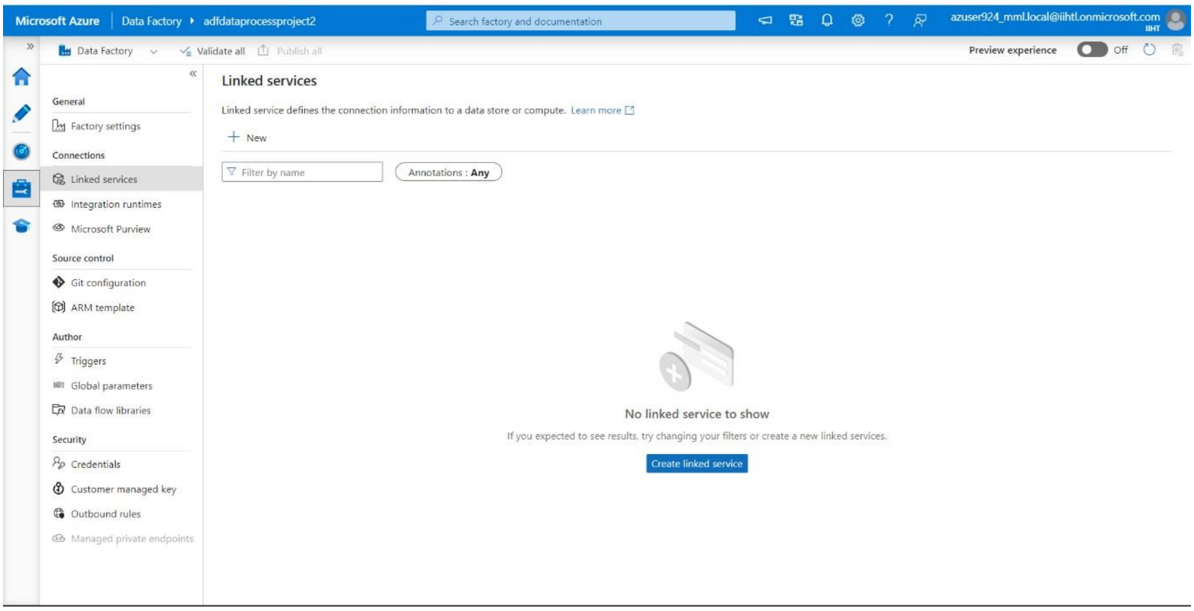
Launch studio in azure data factory:



Resource group (move) : rg-azuser924_mml.local-Kkila
Status : Succeeded
Location : Central India
Subscription (move) : Azure subscription 1
Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08
Type : Data factory (V2)
Getting started : Quick start

https://adf.azure.com/en/home?factory=%2Fsubscriptions%2F984f097c-963c-4eb6-a20d-839457ae9f08%2FresourceGroups%2Frg-azuser924_mml.local-Kkila%2Fproviders%2FMicrosoft.DataFactory%2Ffactories%2Fadfdataprocessproject2#loginHint=azuser924_mml.local@iihtl.onm...

Click on Create a Linked service:

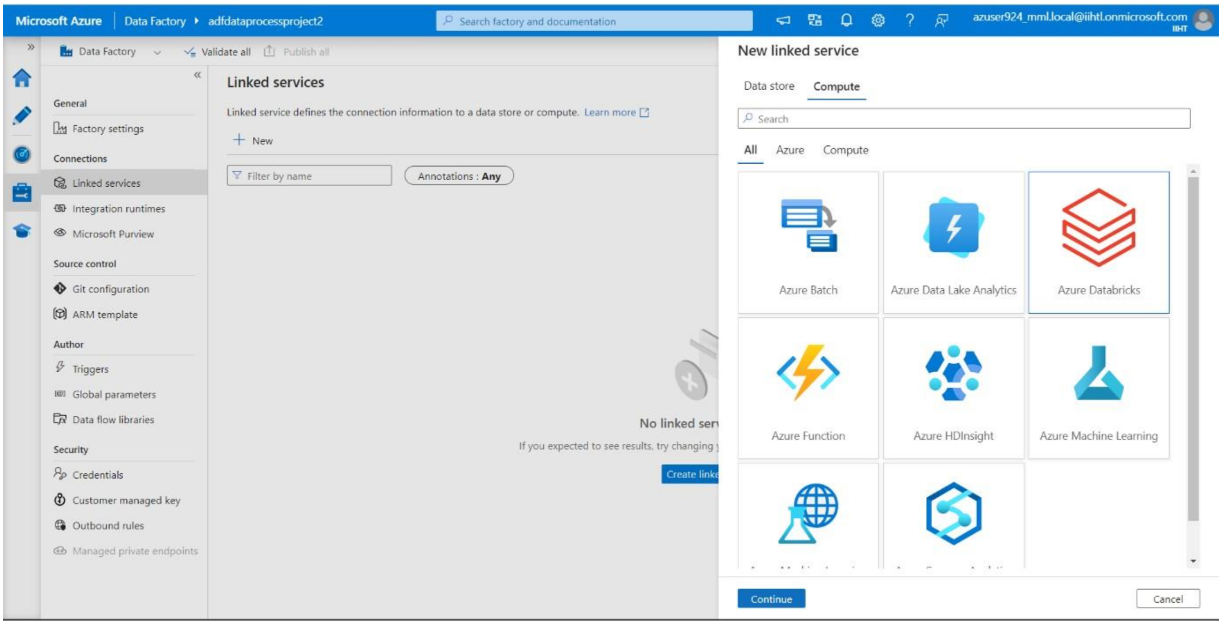


Linked service defines the connection information to a data store or compute. Learn more

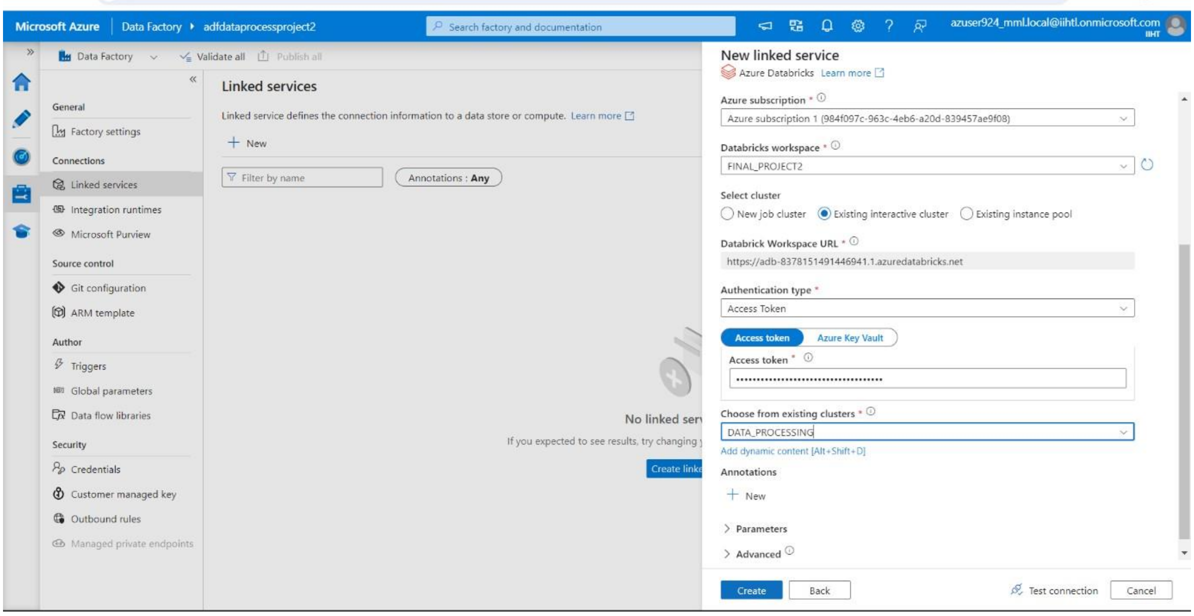
No linked service to show

If you expected to see results, try changing your filters or create a new linked services.

Click on Azure Databricks and click continue:



Create a new linked service by clicking create



New linked service

Azure Databricks Learn more

Azure subscription *

Azure subscription 1 (984f097c-963c-4eb6-a20d-839457ae9f08)

Databricks workspace *

FINAL_PROJECT2

Select cluster

○ New job cluster ● Existing interactive cluster ○ Existing instance pool

Databrick Workspace URL *

https://adb-8378151491446941.1.azuredatabricks.net

Authentication type *

Access Token

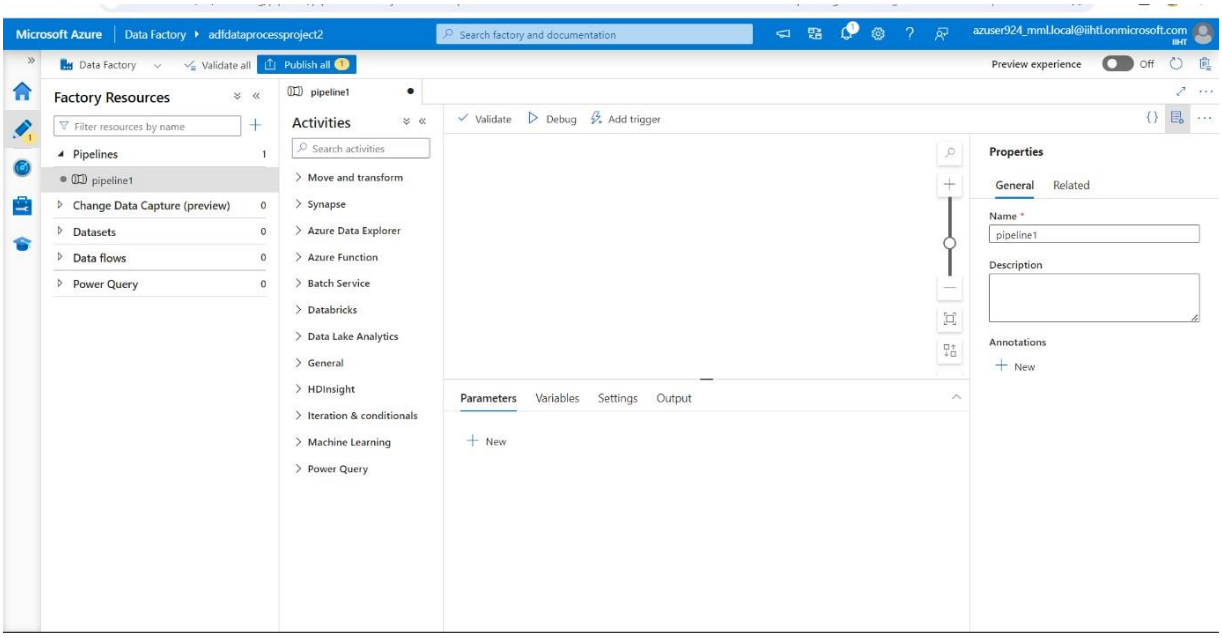
Access token *

••••••••••••••••••••••••••••••••••••••

Choose from existing clusters *

DATA_PROCESSING

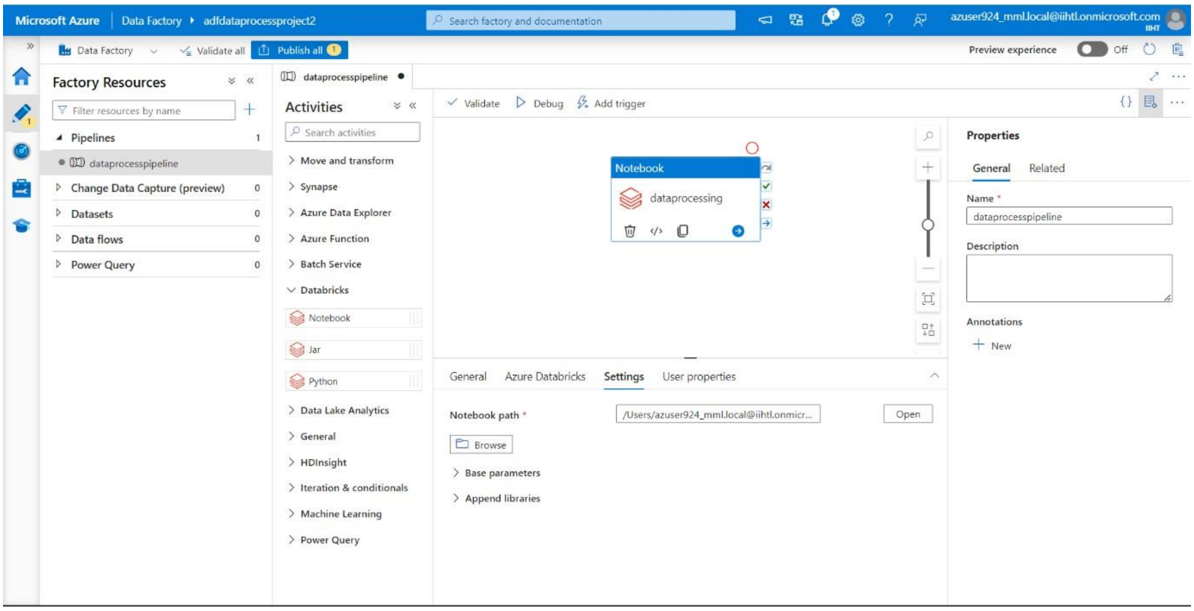
Create a new pipline in Azure data factory:



Name * pipeline1

Description

In pipeline trigger the notebook created in the azuredatabricks:

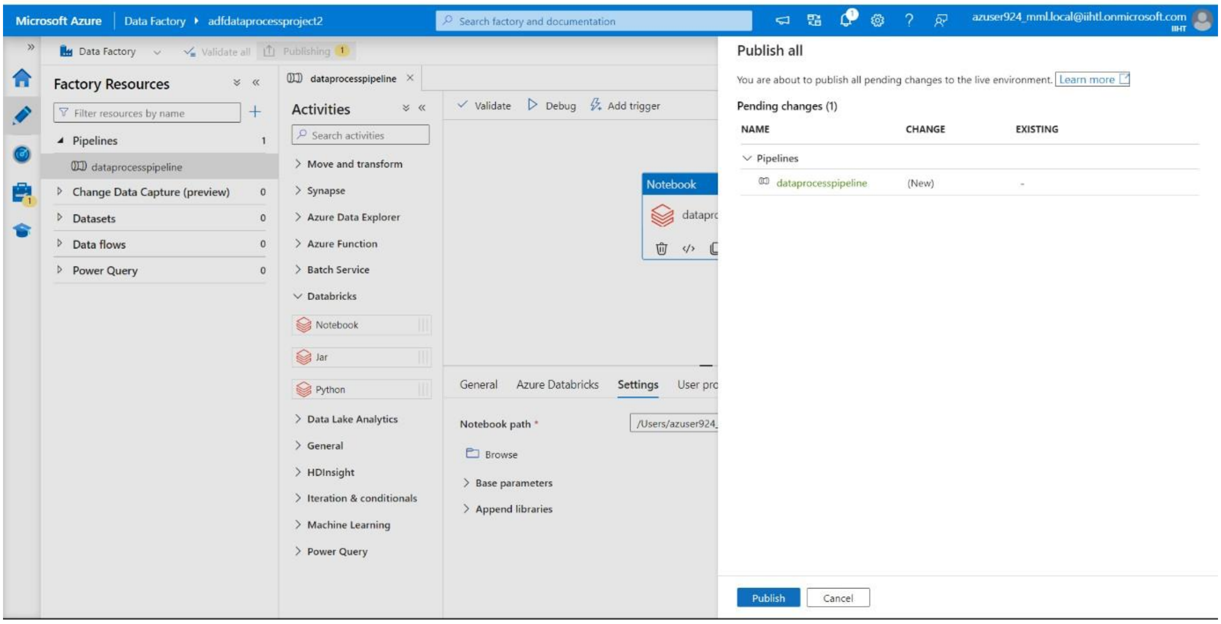


Notebook path * /Users/azuser924_mml.local@iihtl.onmicr...

Name * dataprocesspipeline

Description

Publish pipeline and trigger the notebook



Notebook path * /Users/azuser924_

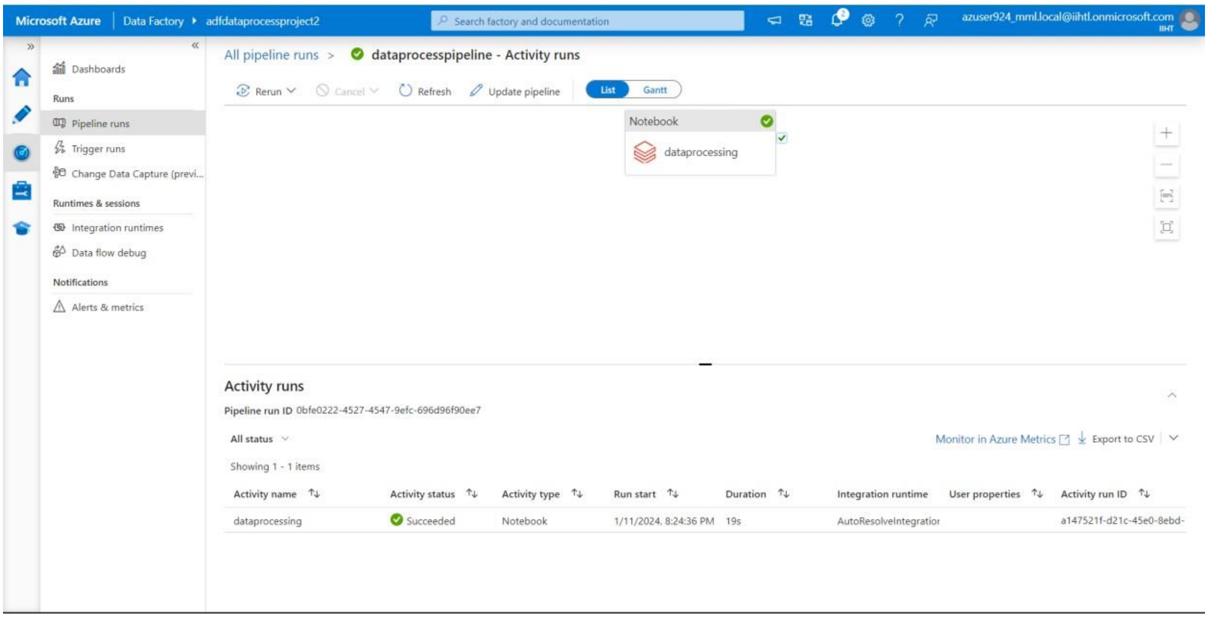
Publish all

You are about to publish all pending changes to the live environment. Learn more

Pending changes (1)

NAME	CHANGE	EXISTING
Pipelines		
dataprocesspipeline	(New)	-

Check the notebook is successfully run in the azure data factory



Activity runs

Pipeline run ID 0bfe0222-4527-4547-9efc-696d96f90ee7

Showing 1 - 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
dataprocessing	✓ Succeeded	Notebook	1/11/2024, 8:24:36 PM	19s	AutoResolveIntegratior		a147521f-d21c-45e0-8ebd-

By following these steps, you will have implemented a data processing pipeline where Azure Data Factory orchestrates data workflows, and Azure Databricks is used as a processing engine for on-demand analytics and transformations.