Setting up Azure Databricks Workspace & Cluster

Exercise 0 – Setup Azure Data Lake Gen2 Account

1. Complete Lab 1 – Working with Azure Data Lake Gen2 account.
2. Upload Files if not already uploaded.

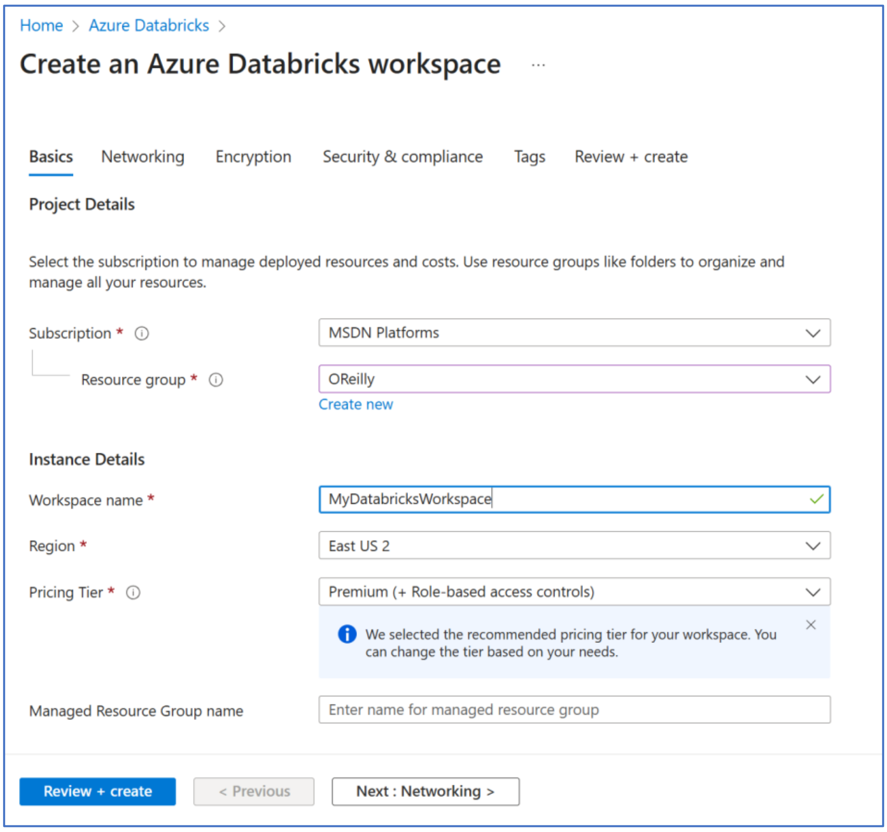
Exercise 1 – Setup Azure Databricks Workspace

1. Go to Azure portal (portal.azure.com).

2. In the search bar, search for Azure Databricks. And select it.

3. Click on **Create**.

4. Fill up the properties to create account
   a. [Basics Tab]
      i. Select subscription
      ii. Select resource group that you created in Exercise 0.
      iii. Provide a unique name
      iv. Select region of your choice (example – East US 2)
      v. Select pricing tier as Premium
      vi. Click **Review + Create**.



Home > Azure Databricks >
Create an Azure Databricks workspace ...
Basics | Networking | Encryption | Security & compliance | Tags | Review + create

Project Details

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

Subscription *: MSDN Platforms

Resource group *: OReilly

Create new

Instance Details

Workspace name *: MyDatabricksWorkspace

Region *: East US 2

Pricing Tier *: Premium (+ Role-based access controls)

We selected the recommended pricing tier for your workspace. You can change the tier based on your needs.

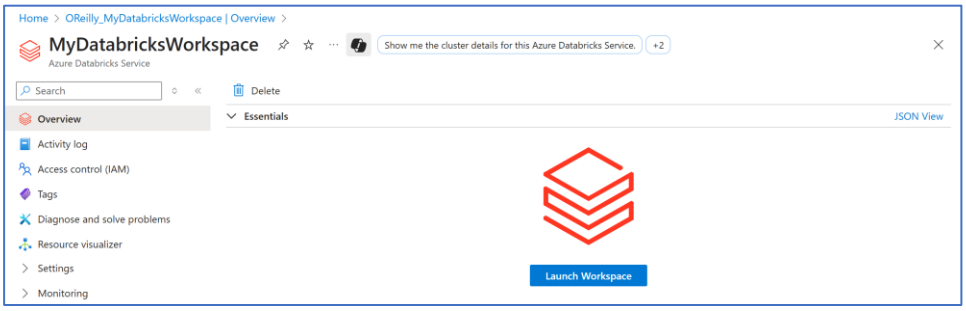
Managed Resource Group name: Enter name for managed resource group

   b. Click **Create**.
      - This will take a few minutes to deploy.

Exercise 2 – Launch Databricks Workspace & Create Cluster

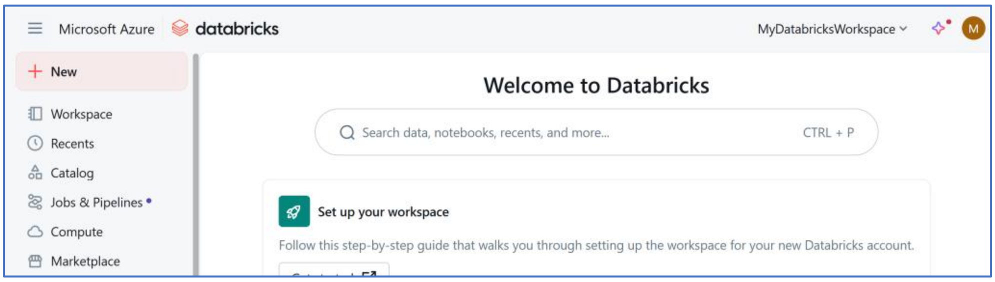
1. Open Azure Databricks instance created in the previous step.

2. Click on Launch workspace, to open Databricks UI.



Show me the cluster details for this Azure Databricks Service.

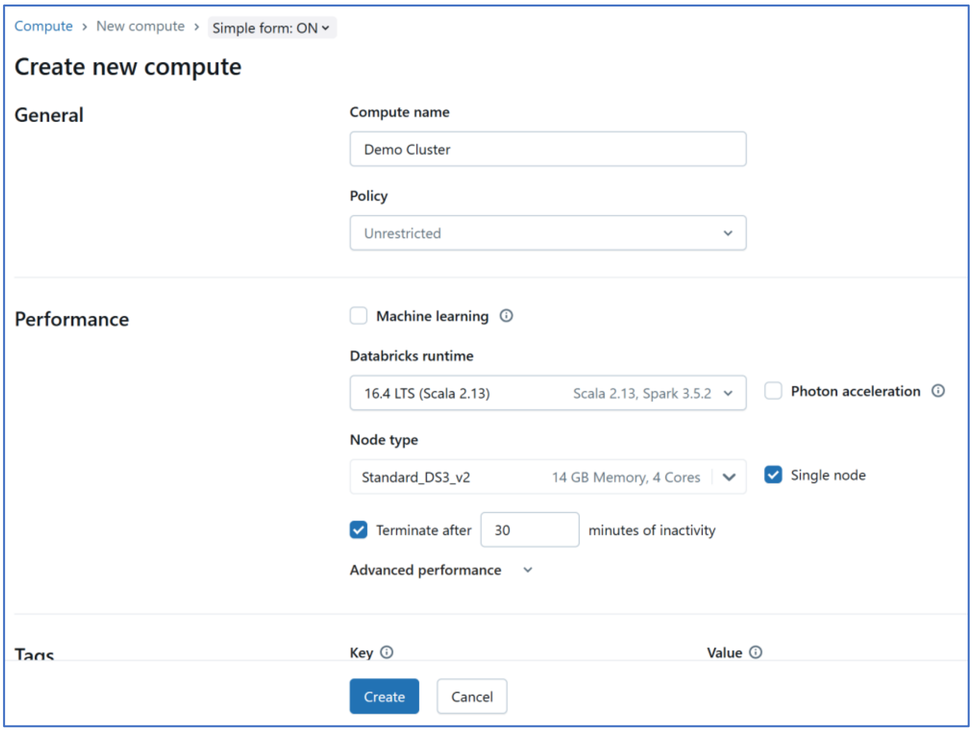
3. In the workspace, from left pane, go to **Compute** tab.



Follow this step-by-step guide that walks you through setting up the workspace for your new Databricks account.

4. Under **All-purpose compute**, Click on **Create Compute** to create a cluster.

5. Fill up cluster properties and click on **Create**. This will take few minutes to setup a single node cluster.
    a. Compute name: Demo Cluster
    b. Databricks Runtime: Select the latest runtime with LTS (long-term support).
    c. Photon acceleration: Disable
    d. Node type: Standard_DS3_v2 (if this type is not available, select any other)
    e. Single node: Enable
    f. Terminate minutes: 30 minutes



Compute › New compute › Simple form: ON

Create new compute

General

Compute name: Demo Cluster

Policy: Unrestricted

Performance

☐ Machine learning

Databricks runtime: 16.4 LTS (Scala 2.13) Scala 2.13, Spark 3.5.2

☐ Photon acceleration

Node type: Standard_DS3_v2 14 GB Memory, 4 Cores

☑ Single node

☑ Terminate after 30 minutes of inactivity

Advanced performance

Tags

Key

Value

[Note]: If you want to setup multi-node cluster, deselect single node option from UI.

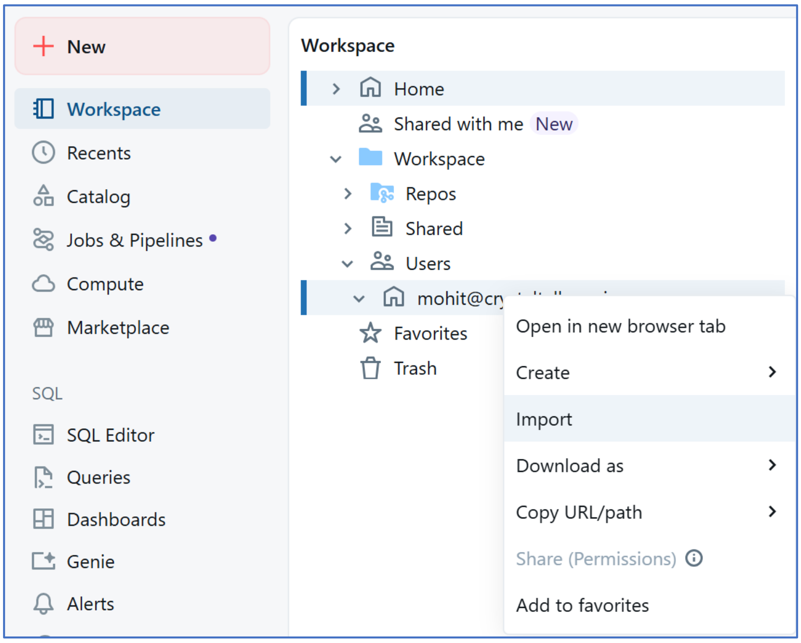
Exercise 3 – Import Notebook & Run Commands

1. Download notebook – **Working with Spark.py** from GitHub repository.

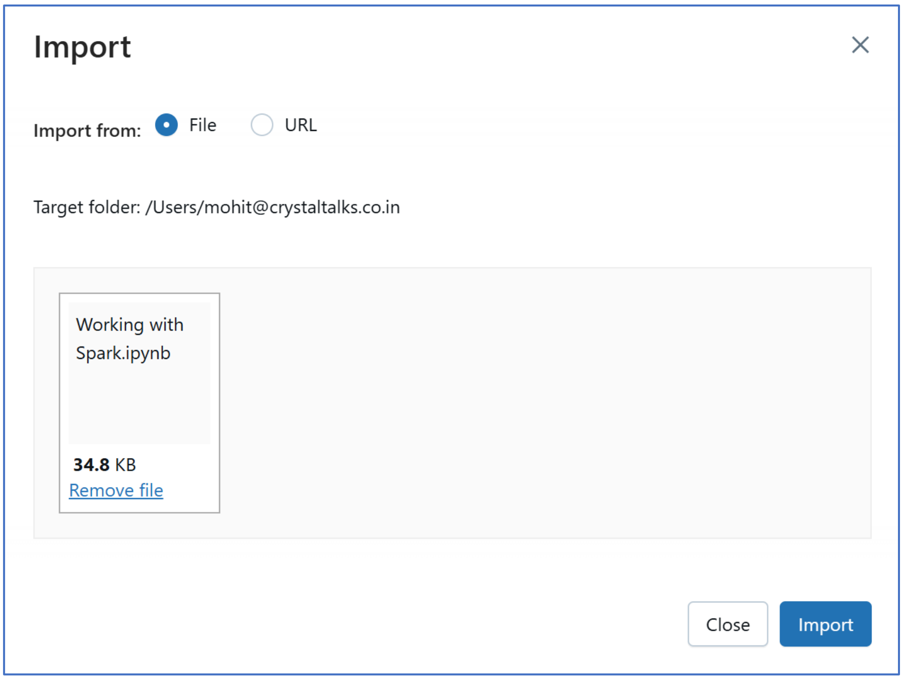
2. Once the cluster is ready, from left pane, go to **Workspace** tab.

3. Click on Workspace folder → Users folder → Your user account.

4. Right-click on the account folder and click **Import**.



5. Upload the notebook - **Working with Spark.py** and click **Import**.



Import

Import from: ● File ○ URL

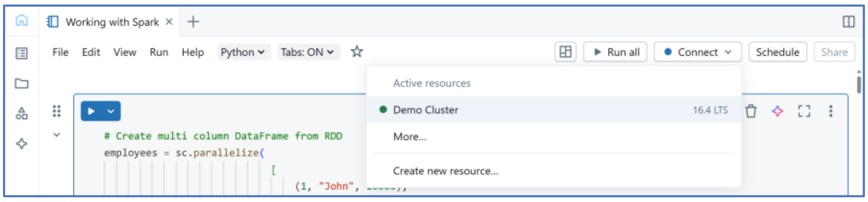
Target folder: /Users/mohit@crystaltalks.co.in

Working with Spark.ipynb

34.8 KB

Remove file

6. Open the notebook and connect to cluster (Demo Cluster) that you previously created.



```
# Create multi column DataFrame from RDD
employees = sc.parallelize(
                                [
                                    (1, "John",
```

7. Run the commands 1 to 12.