Databricks

-Mohammad R Islam

Steps to Create a Databricks Workspace

1) Login to Azure

a) Azure portal: Link

b) Sign in with your Azure credentials.

2) Create Azure Databricks

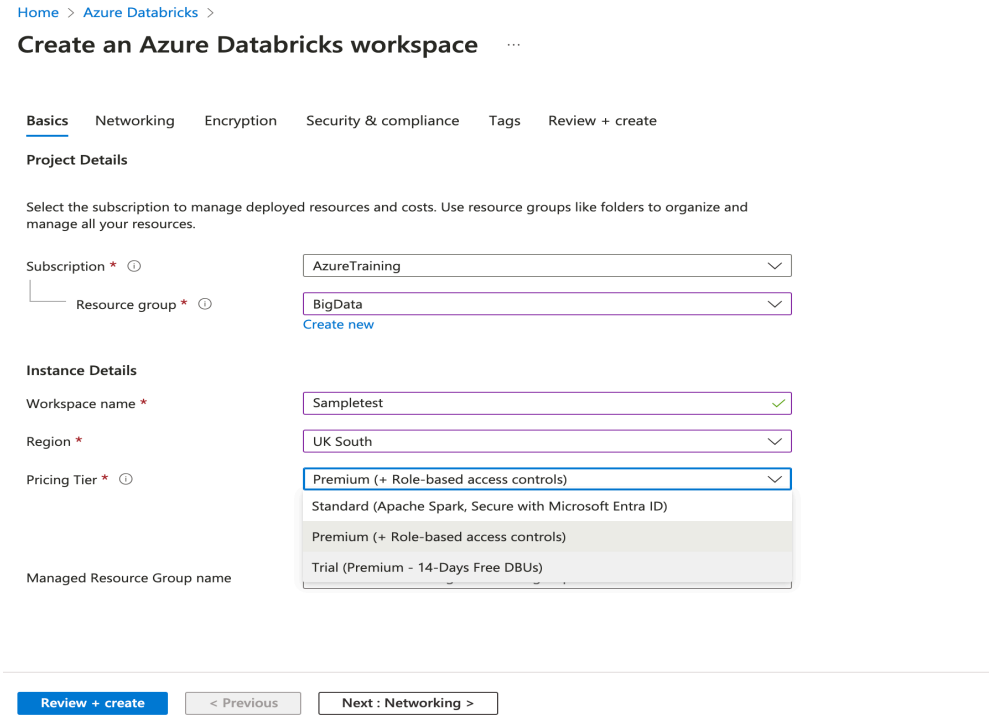
a) Configure Basic Settings

i) **Subscription**: Select the Azure subscription in which you want to create the workspace.

ii) **Resource Group**: Choose an existing resource group or create a new one.

iii) **Workspace Name**: Enter a unique name for your Databricks workspace.

iv) **Region**: Select the Azure region where the workspace will be hosted (choose a region close to your data sources).



b) Networking

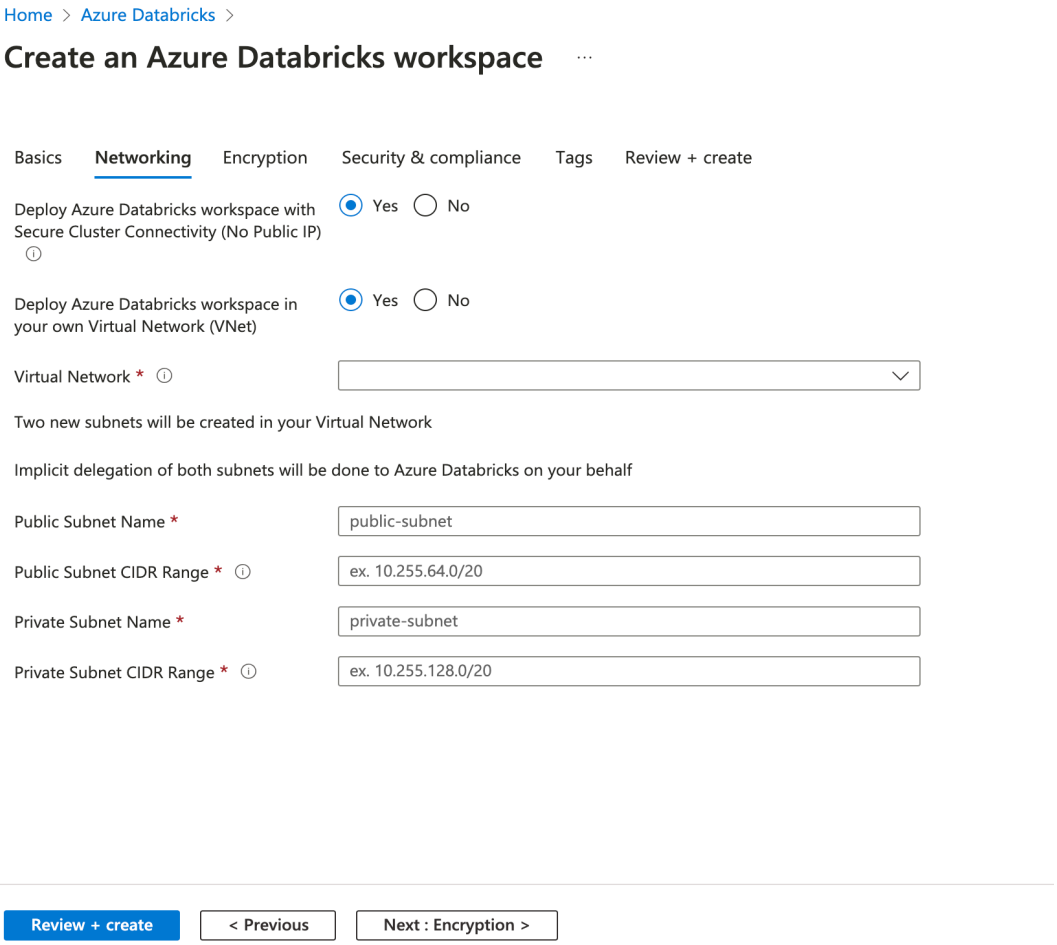
i) Deploy Azure Databricks workspace with secure cluster connectivity (No public IP) –**Yes/No**

ii) Deploy Azure Databricks workspace with you own Virtual network (VNet) – **Yes/No**

iii) If you choose a **VNet,** you will need to specify the following:

(1) **Private Subnet** with CIDR range within the chosen VNet.

(2) **Public Access**: with CIDR range within the chosen VNet.



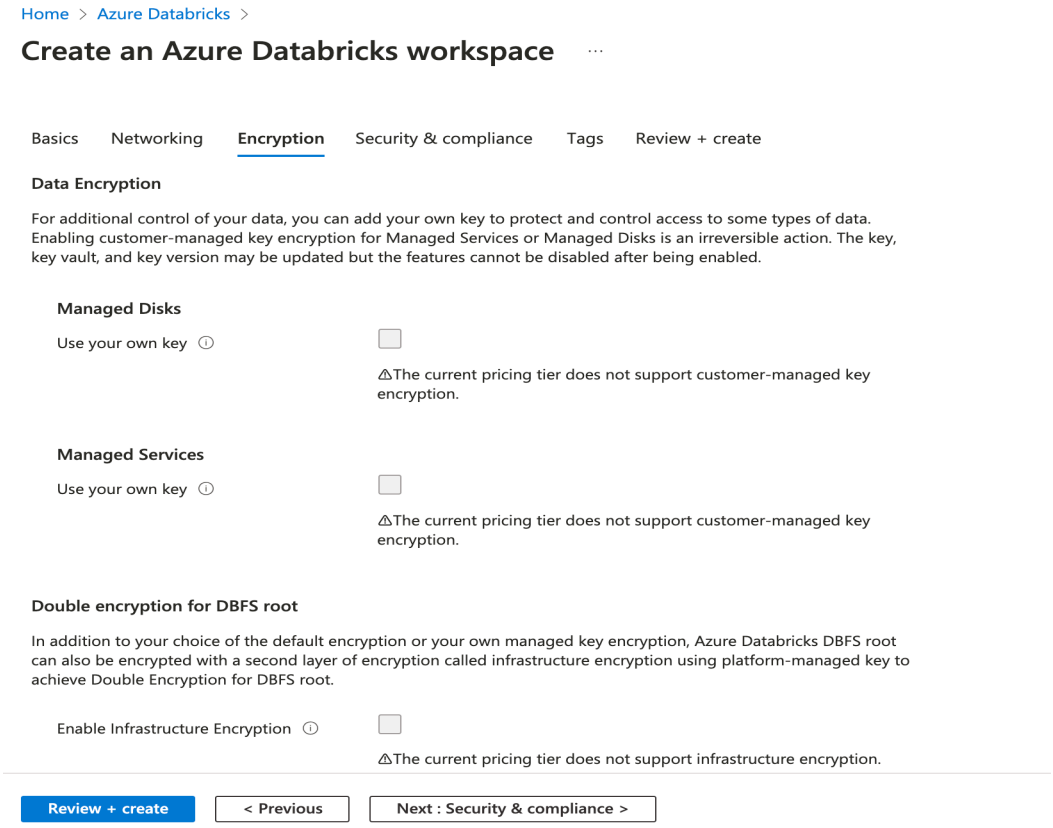
c) Encryption

i) **Encryption at Rest**: Azure Databricks automatically encrypts data at rest. You can choose to use Azure-managed keys or **your own** customer-managed keys (**CMK**).

(1) If using a **CMK**, specify the Key Vault and the key you want to use for encryption.

ii) **Configure DBFS Encryption:** Databricks will use the specified CMK for encrypting the data in DBFS. You do not need to make additional configurations to enable this, as Databricks will automatically encrypt the data stored in DBFS using the specified key.

iii) **Encryption in Transit**: Databricks uses TLS/SSL for encrypting data in transit by default.



d) Workspace Security and compliance Settings

i) Enable Compliance security profile

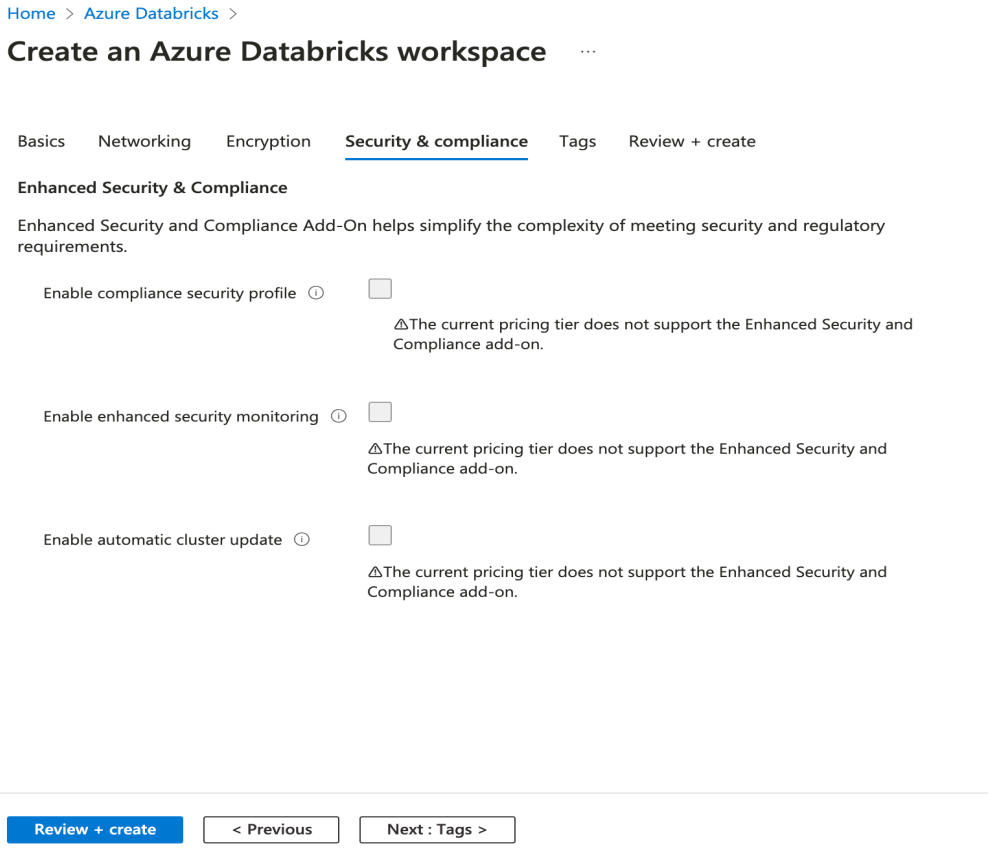
Databricks does not have a specific "Compliance Security Provider," you can enhance compliance in your Databricks environment by enabling audit logging, integrating with external compliance and security solutions, configuring access controls, setting up encryption, enabling IP access lists, and monitoring compliance metrics. This multi-layered approach will help you meet various regulatory requirements and maintain a secure environment.

ii) Enable Enhanced security monitoring

Enabling Enhanced Security Monitoring in Databricks involves configuring audit logs, access control, IP access lists, encryption, user activity monitoring, and integration with external security solutions. These steps help ensure that your Databricks environment is secure and compliant with best practices for data protection and security monitoring.

iii) Enable automatic cluster update

Enabling automatic cluster updates in Databricks helps ensure that your clusters are always running the latest version of Databricks Runtime, which includes performance improvements, security updates, and new features.

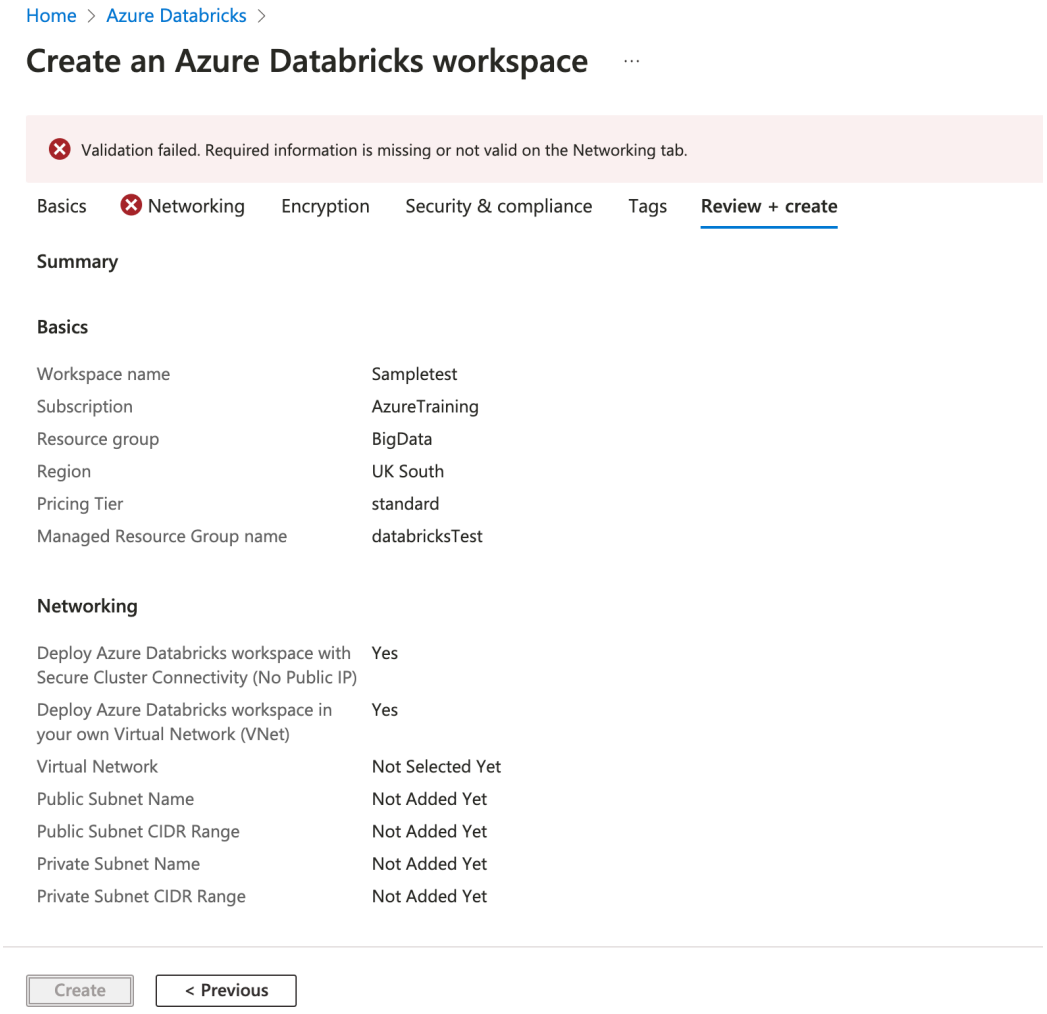


e) Review and Create

i) Once you have filled in all the necessary fields, review your settings on the **Review + create** tab.

ii) Check the relevant summary for any errors or required fields.

iii) Click the **Create** button to deploy your Databricks workspace.

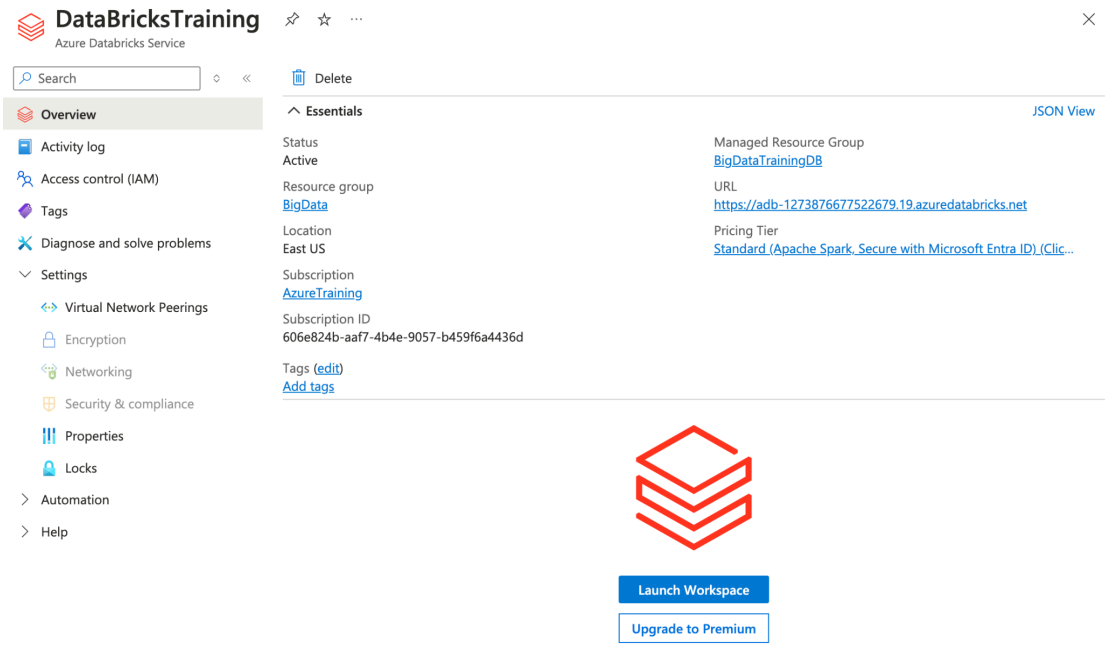


f) Accessing the Workspace

i) After the deployment is complete, go to the resource group where your Databricks workspace is located.

ii) Click on the Databricks workspace resource.

iii) You can now launch the workspace by clicking on the **Launch Workspace** button.



Databricks workspace

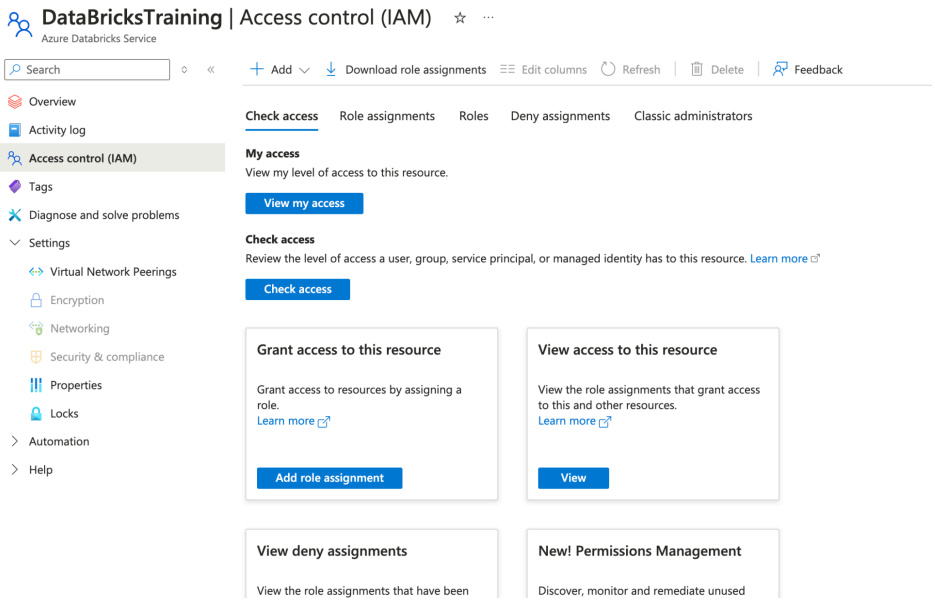
1. Activity logs

Databricks activity logs are crucial for tracking and auditing various events and activities occurring within your Databricks workspace. These logs help with monitoring for security, operational efficiency, and compliance purposes. You can configure activity logging and access the logs via Azure, AWS, or your preferred cloud provider, depending on your Databricks environment.

2. Access Control (IAM)

Access control in Databricks is essential for managing who can access your workspace, clusters, data, and other resources. Databricks provides various levels of access control mechanisms to ensure security and compliance, including

- Workspace access control
- cluster access control
- Job access control
- Notebook permissions, and
- Data access control.



3. Diagnose and solve problems

Diagnosing and solving problems in Databricks involves a combination of monitoring logs, debugging code, optimizing performance, and managing resources effectively.

- Cluster-Related Issues
- Job and Notebook Issues
- Performance Tuning
- Data Access and Storage Issues

- Security Issues
- Monitoring and Logging
- Network and Connectivity Issues