

Azure Databricks

Lab: Setup Azure Databricks

The main tasks for this exercise are as follows:

1. Create and Configure Azure Databricks Instance.
2. Create a Spark Cluster in Azure Databricks.
3. Create a Storage Account

Task 1: Create and configure an Azure Databricks instance.

1. In the Azure portal, at the top left of the screen, click on the **Home** hyperlink.
2. In the Azure portal, click on the **+ Create a resource** icon.
3. In the New screen, click in the **Search services and marketplace** text box, and type the word **Azure databricks**. Click **Azure Databricks** in the list that appears.
4. In the **Azure Databricks** blade, click **Create**.
5. In the **Azure Databricks Service** blade, create an Azure Databricks Workspace with the following settings:
 - **Subscription:** the name of the subscription you are using in this lab
 - **Resource group:** **adbwkxx-rg**, where **xx** are your initials.
 - **Workspace name:** **adbwkxx**, where **xx** are your initials.
 - **Region:** the name of the Azure region which is closest to you.
 - **Pricing Tier:** **Standard (Apace Spark, Secure with Azure AD)**.

Microsoft Azure

Search resources, services, and docs (G+)

Home > Azure Databricks >

Create an Azure Databricks workspace

Basics Networking Advanced 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 * ⓘ Azure Pass - Sponsorship

Resource group * ⓘ kpmgadfpc [Create new](#)

Instance Details

Workspace name * adbwkpc ✓

Region * West US

Pricing Tier * ⓘ Standard (Apache Spark, Secure with Azure AD)

Review + create < Previous Next : Networking >

6. In the **Azure Databricks Service** blade, click **Review +create**. Then click **Create**

Note: The provision will take approximately 3 minutes. The Databricks Runtime is built on top of Apache Spark and is natively built for the Azure cloud. Azure Databricks completely abstracts out the infrastructure complexity and the need for specialized expertise to set up and configure your data infrastructure. For data engineers, who care about the performance of production jobs, Azure Databricks provides a Spark engine that is faster and performant through various optimizations at the I/O layer and processing layer (Databricks I/O).

7. Confirm that the Azure Databricks service has been created.
8. In the Azure portal, navigate to the **Resource group** screen.
9. In the Resource groups screen, click on the **adbwkxx-rg** resource group, where **xx** are your initials.

10. In the **adbwkxx-rg** screen, click **adbwkxx**, where **xx** are your initials to open Azure Databricks. This will open your Azure Databricks service.

The screenshot shows the Azure portal interface for the resource group 'kpmgadfpc'. On the left, a list of resource groups includes 'databricks-rg-adbwkpc-hpkldgfawqjfk', 'kpmgadfpc' (highlighted with a red box), and 'NetworkWatcherRG'. A central navigation pane shows options like 'Overview' (highlighted with a red box), 'Activity log', 'Access control (IAM)', 'Tags', 'Resource visualizer', 'Events', 'Settings', and 'Monitoring'. The right pane displays the 'Essentials' section with subscription details and a 'Resources' table. The table lists resources: 'adbwkpc' (highlighted with a red box), 'adf-pc-kpmg', 'datastoragepc1', and 'SQLDB (sqlservicepc/SQLDB)'.

11. Click on **adbwkxx** and click on **Launch Workspace**

The screenshot shows the Azure Databricks workspace launch page. At the top, there's a URL and a pricing tier of 'standard'. Below this, a large red Databricks logo is centered, with a blue 'Launch Workspace' button underneath it. At the bottom, there are four tabs: 'Getting Started' (selected), 'Import Data from File', 'Import Data from Azure Stor...', and another 'Import Data from File' tab.

Task 2: Create a Spark Cluster in Azure Databricks.

1. Once you Launch the Workspace.
2. Under **Compute Tab**, click **Create Cluster**.

3. In the **Create Cluster** screen, under New Cluster, create a Databricks Cluster with the following settings, and then click on **Create Cluster**:
- **Cluster name:** Test Cluster
 - **Cluster Mode:** Single Node
 - **Databricks Runtime Version:** Runtime: 11.x LTS (Scala 2.12, Spark 3.1.2)
 - Make sure you select and set the **Terminate after 30** minutes of inactivity check box. If the cluster isn't being used, provide a duration (in minutes) to terminate the cluster.
 - Leave all the remaining options to their current settings.

The screenshot displays the 'Create Cluster' interface in the Microsoft Azure Databricks portal. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, and a search bar. The main header shows 'Clusters / New Compute' with links for 'UI Preview' and 'Provide feedback'. The cluster name 'Pankaj Choudhary's Cluster' is visible with an edit icon. The configuration section includes: Cluster Mode set to 'Single node'; Access mode set to 'Single user'; Single user access set to 'Pankaj Choudhary (pankajacks@...)'; Performance section with Databricks runtime version set to 'Runtime: 11.3 LTS (Scala 2.12, Spark 3.3.0)'; 'Use Photon Acceleration' is unchecked; Node type set to 'Standard_DS3_v2' (14 GB Memory, 4 Cores); and the 'Terminate after 30 minutes of inactivity' checkbox is checked. The 'Tags' section is partially visible at the bottom.

4. In the **Create Cluster** screen, click on **Create Cluster** and leave the browser screen open.

Note: The creation of the Azure Databricks instance will take approximately 5-8 minutes as the creation of a Spark cluster is simplified through the graphical user interface.