Use Spark in Azure Databricks

Azure Databricks is a Microsoft Azure-based version of the popular open-source Databricks platform. Azure Databricks is built on Apache Spark, and offers a highly scalable solution for data engineering and analysis tasks that involve working with data in files. One of the benefits of Spark is support for a wide range of programming languages, including Java, Scala, Python, and SQL; making Spark a very flexible solution for data processing workloads including data cleansing and manipulation, statistical analysis and machine learning, and data analytics and visualization.

This exercise should take approximately **45** minutes to complete.

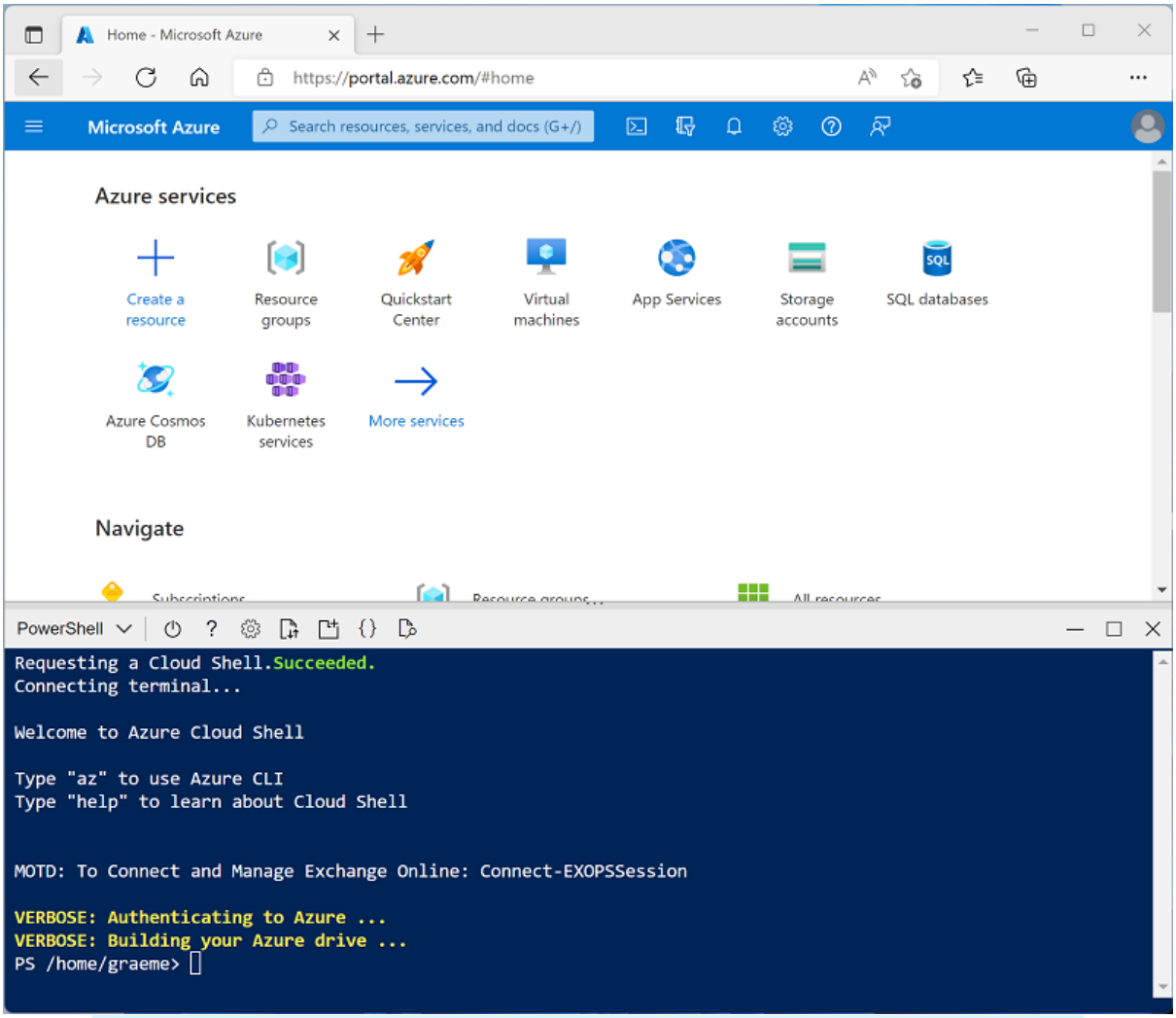
Before you start

You'll need an Azure subscription in which you have administrative-level access.

Provision an Azure Databricks workspace

In this exercise, you'll use a script to provision a new Azure Databricks workspace.

1. In a web browser, sign into the Azure portal at `https://portal.azure.com`.
2. Use the **[>_]** button to the right of the search bar at the top of the page to create a new Cloud Shell in the Azure portal, selecting a ***PowerShell*** environment and creating storage if prompted. The cloud shell provides a command line interface in a pane at the bottom of the Azure portal, as shown here:



Note: If you have previously created a cloud shell that uses a *Bash* environment, use the the drop-down menu at the top left of the cloud shell pane to change it to ***PowerShell***.

3. Note that you can resize the cloud shell by dragging the separator bar at the top of the pane, or by using the —, ◻, and **X** icons at the top right of the pane to minimize, maximize, and close the pane. For more information about using the Azure Cloud Shell, see the Azure Cloud Shell documentation.
4. In the PowerShell pane, enter the following commands to clone this repo:

CodeCopy

```
 rm -r dp-203 -f
 git clone https://github.com/MicrosoftLearning/dp-203-azure-data-engineer
dp-203
```

5. After the repo has been cloned, enter the following commands to change to the folder for this lab and run the **setup.ps1** script it contains:

CodeCopy

```
cd dp-203/Allfiles/labs/24
./setup.ps1
```

6. If prompted, choose which subscription you want to use (this will only happen if you have access to multiple Azure subscriptions).
7. Wait for the script to complete - this typically takes around 5 minutes, but in some cases may take longer. While you are waiting, review the Exploratory data analysis on Azure Databricks article in the Azure Databricks documentation.

Create a cluster

Azure Databricks is a distributed processing platform that uses Apache Spark *clusters* to process data in parallel on multiple nodes. Each cluster consists of a driver node to coordinate the work, and worker nodes to perform processing tasks.

Note: In this exercise, you'll create a *single-node* cluster to minimize the compute resources used in the lab environment (in which resources may be constrained). In a production environment, you'd typically create a cluster with multiple worker nodes.

1. In the Azure portal, browse to the **dp203-*xxxxxxx*** resource group that was created by the script you ran.
2. Select the **databricks*xxxxxxx*** Azure Databricks Service resource.
3. In the **Overview** page for **databricks*xxxxxxx***, use the **Launch Workspace** button to open your Azure Databricks workspace in a new browser tab; signing in if prompted.
4. If a **What's your current data project?** message is displayed, select **Finish** to close it. Then view the Azure Databricks workspace portal and note that the sidebar on the left side contains icons for the various tasks you can perform.

   Tip: As you use the Databricks Workspace portal, various tips and notifications may be displayed. Dismiss these and follow the instructions provided to complete the tasks in this exercise.

5. Select the **(+) New** task, and then select **Cluster**.
6. In the **New Cluster** page, create a new cluster with the following settings:
   - **Cluster name**: *User Name's* cluster (the default cluster name)
   - **Cluster mode**: Single Node
   - **Access mode**: Single user (*with your user account selected*)
   - **Databricks runtime version**: 12.2 LTS (Scala 2.12, Spark 3.2.2)
   - **Use Photon Acceleration**: Selected
   - **Node type**: Standard_DS3_v2
   - **Terminate after** *30* **minutes of inactivity**

7. Wait for the cluster to be created. It may take a minute or two.

Note: If your cluster fails to start, your subscription may have insufficient quota in the region where your Azure Databricks workspace is provisioned. See CPU core limit prevents cluster creation for details. If this happens, you can try deleting your workspace and creating a new one in a different region. You can specify a region as a parameter for the setup script like this: `./setup.ps1 eastus`

Explore data using a notebook

As in many Spark environments, Databricks supports the use of notebooks to combine notes and interactive code cells that you can use to explore data.

1. In the sidebar on the left, select **Workspace**. Then select the ⌂ **Home** folder.
2. Create a new notebook from the Workspace.
3. Connect the notebook to your cluster, and follow the queries shared on OneDrive; try running the cells it contains to explore data in files.

Delete Azure Databricks resources

Now you've finished exploring Azure Databricks, you must delete the resources you've created to avoid unnecessary Azure costs and free up capacity in your subscription.

1. Close the Azure Databricks workspace browser tab and return to the Azure portal.
2. On the Azure portal, on the **Home** page, select **Resource groups**.
3. Select the **dp203-*xxxxxxx*** resource group (not the managed resource group), and verify that it contains your Azure Databricks workspace.
4. At the top of the **Overview** page for your resource group, select **Delete resource group**.
5. Enter the **dp203-*xxxxxxx*** resource group name to confirm you want to delete it, and select **Delete**.

   After a few minutes, your resource group and the managed workspace resource groups associated with it will be deleted.