Implementing Predictive Analytics with Spark in Azure Databricks

Lab 5 – Introduction to MMLSpark

Overview

The Microsoft Machine Learning for Spark (MMLSpark) package provides a number of deep learning and data science tools for Apache Spark, including seamless integration of Spark Machine Learning pipelines with Microsoft Cognitive Toolkit (CNTK) and OpenCV, enabling you to quickly create powerful, highly-scalable predictive and analytical models for large image and text datasets.

In this lab, you'll explore how to install and use the MMLSpark library in a Databricks Spark cluster.

What You'll Need

To complete the labs, you will need the following:

- A web browser
- A Microsoft account
- A Microsoft Azure subscription
- A Windows, Linux, or Mac OS X computer
- Azure Storage Explorer
- The lab files for this course

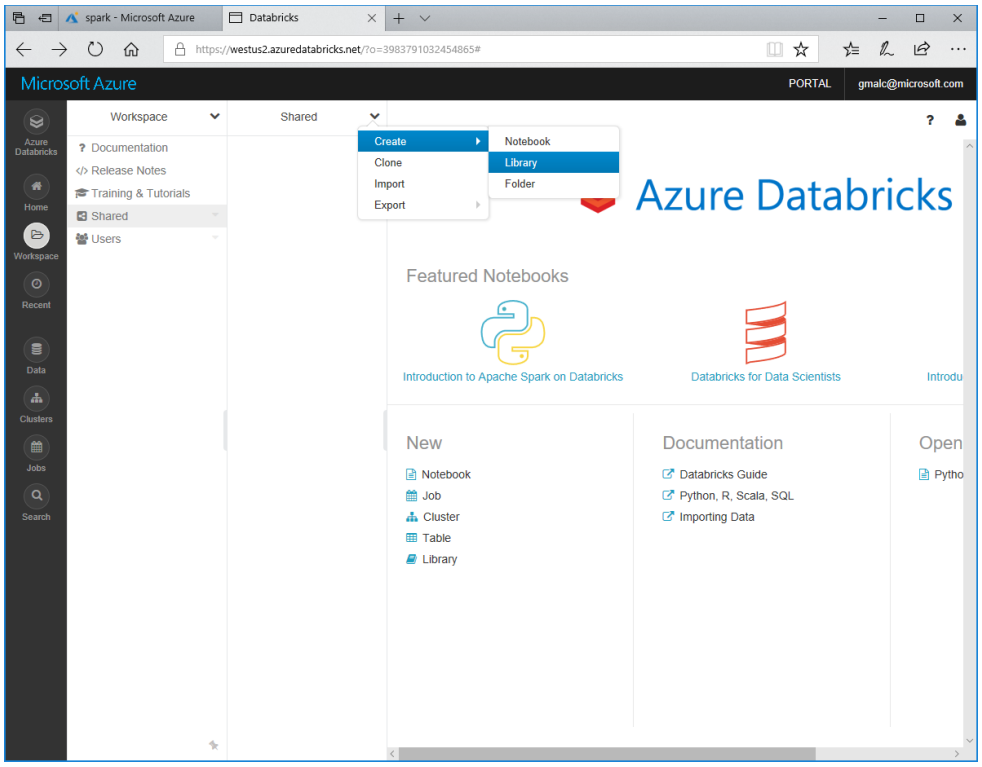
Note: If you have not already done so, set up the required environment for the lab by following the instructions in the Setup document for this course. Then follow the instructions in Lab 1 to provision the required Azure resources.

Installing MMLSpark

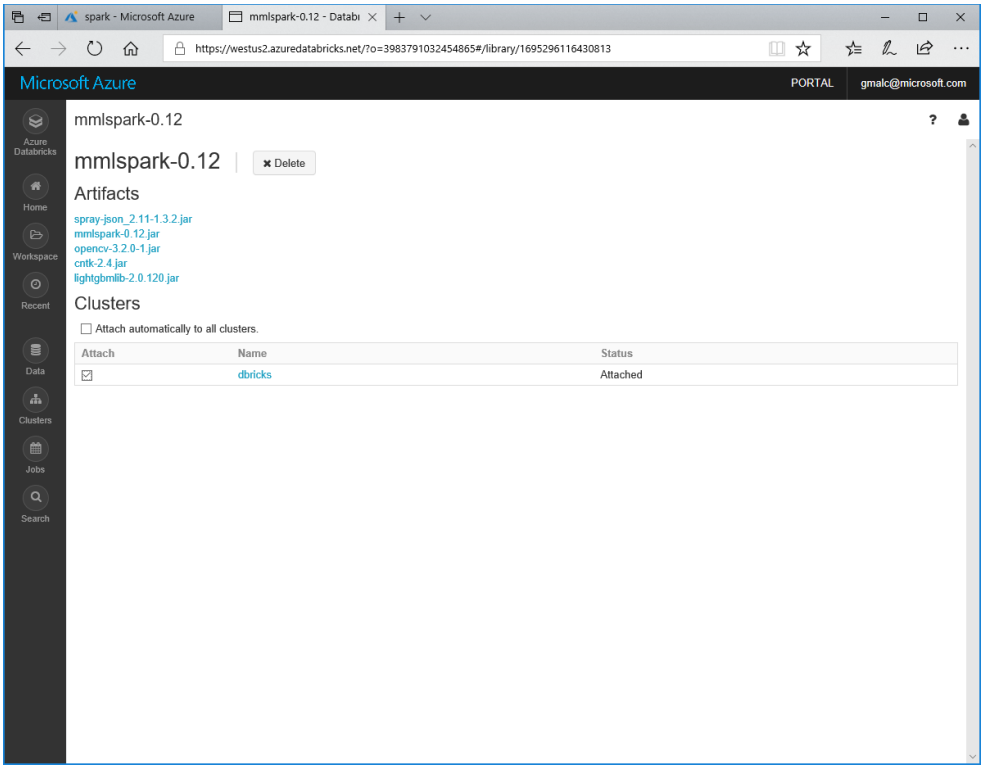
Before you can use the MMLSpark library, you must install it in your Databricks workspace and attach it to your cluster.

Create a New MMLSpark Library

1. In the Databricks workspace, click **Workspace**. Then click **Shared**, and in the drop-down menu for the **Shared** folder, point to **Create** and click **Library** as shown here:



2. In the **New Library** page, in the **Source** drop-down list, select **Maven Coordinate**.
3. In the **Coordinate** textbox, type **Azure:mmlspark:0.12** and then click **Create Library**. After a few minutes, the library and its dependencies will be installed.
4. In the **mmlspark-0.12** page, in the **Clusters** list, select the **Attach** checkbox for your cluster and wait for the **Status** value to indicate that the library has been attached, like this:



Build a Classification Model

Now that you have added the MMLSpark library to your Databricks workspace and attached it to your cluster, you can use it to build a machine learning model.

Upload Source Data to Azure Storage

Note: If you have already uploaded the flights.csv data file to your Azure storage container, you can skip this procedure.

In this lab, you will build a model based on data about flights. Before you can do this, you must store the flight data files in the shared storage used by your cluster. The instructions here assume you will use Azure Storage Explorer to do this, but you can use any Azure Storage tool you prefer.

1. In the folder where you extracted the lab files for this course on your local computer, in the **data** folder, verify that the **flights.csv** file exists. This file contains flight data that has been cleaned and prepared for modeling.
2. Start Azure Storage Explorer, and if you are not already signed in, sign into your Azure subscription.
3. Expand your storage account and the **Blob Containers** folder, and then double-click the **spark** blob container.
4. In the **Upload** drop-down list, click **Upload Files**. Then upload **flights.csv** as a block blob to a folder named **data** in root of the **spark** container.

Upload a Notebook

You will use a Notebook to create a classification model. You can choose to work with Python or Scala.

1. From the **Lab05** folder in the folder where you extracted the lab files, upload **MMLSpark Classifier.ipynb** or **MMLSpark Classifier.scala**, depending on your preferred choice of language, to your Databricks workspace.
2. Open the notebook you uploaded, attach it to your cluster, and then read the notes and run the code it contains to build a classification model for text features.

Clean Up

Note: Follow the steps below to delete your cluster and avoid being charged for cluster resources when you are not using them.

Delete the Resource Group

1. Close the browser tab containing the databricks workspace if it is open.
2. In the Azure portal, view your **Resource groups** and select the resource group you created for your databricks workspace. This resource group contains your databricks workspace and your storage account.
3. In the blade for your resource group, click **Delete**. When prompted to confirm the deletion, enter the resource group name and click **Delete**.
4. Wait for a notification that your resource group has been deleted.
5. After a few minutes, a second resource group containing the resources for your cluster will automatically be deleted.
6. Close the browser.