Analyzing Data with Spark in Azure Databricks

Lab 4 – Introduction to Machine Learning

Overview

In this lab, you will use Spark in a Databricks cluster to train and test a machine learning model.

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: To set up the required environment for the lab, follow the instructions in the **Setup** document for this course. Specifically, you must have signed up for an Azure subscription.

Provisioning Azure Resources

Note: If you already have an Azure Databricks Spark cluster and an Azure blob storage account, you can skip this section.

Provision a Databricks Workspace

1. In a web browser, navigate to http://portal.azure.com, and if prompted, sign in using the Microsoft account that is associated with your Azure subscription.
2. In the Microsoft Azure portal, click **+ Create a resource**. Then in the **Analytics** section select **Azure Databricks** and create a new Azure Databricks workspace with the following settings:
    - **Workspace name**: *Enter a unique name (and make a note of it!)*
    - **Subscription:** *Select your Azure subscription*
    - **Resource Group:** *Create a new resource group with a unique name (and make a note of it!)*
    - **Location:** *Choose any available data center location.*
    - **Pricing Tier:** Standard
3. In the Azure portal, view **Notifications** to verify that deployment has started. Then wait for the workspace to be deployed (this can take few minutes)

Provision a Storage Account

1. In the Azure portal tab in your browser, and click ✚ **Create a resource**.
2. In the **Storage** category, click **Storage account**.
3. Create a new storage account with the following settings:
    - **Name**: *Specify a unique name (and make a note of it)*
    - **Deployment model**: Resource manager
    - **Account kind**: Storage (general purpose v1)
    - **Location**: *Choose the same location as your Databricks workspace*
    - **Replication:** Locally-redundant storage (LRD)
    - **Performance:** Standard
    - **Secure transfer required:** Disabled
    - **Subscription:** *Choose your Azure subscription*
    - **Resource group:** *Choose the existing resource group for your Databricks workspace*
    - **Virtual networks:** Disabled
4. Wait for the resource to be deployed. Then view the newly deployed storage account.
5. In the blade for your storage account, click **Blobs**.
6. In the **Browse blobs** blade, click ✚ **Container**, and create a new container with the following settings:
    - **Name**: spark
    - **Public access level**: Private (no anonymous access)
7. In the **Settings** section of the blade for your blob store, click **Access keys** and note the **Storage account name** and **key1** values on this blade – you will need these in the next procedure.

Create a Spark Cluster

1. In the Azure portal, browse to the Databricks workspace you created earlier, and click **Launch Workspace** to open it in a new browser tab.
2. In the Azure Databricks workspace home page, under **New**, click **Cluster**.
3. In the **Create Cluster** page, create a new cluster with the following settings:
    - **Cluster Mode**: Standard
    - **Cluster Name**: *Enter a unique cluster name (and make a note of it)*
    - **Databricks Runtime Version**: *Choose the latest available version*
    - **Python Version:** 3
    - **Driver Type**: Same as worker
    - **Worker Type**: *Leave the default type*
    - **Min Workers:** 1
    - **Max Workers:** 2
    - **Auto Termination:** Terminate after 60 minutes.
    - **Spark Config**: Add two key-value pairs for your storage account and key like this:

      fs.azure.account.key.***your_storage_account***.blob.core.windows.net ***your_key1_value***

4. Wait for the cluster to be created.

Creating and Testing A Machine Learning Model

Spark includes an API named Spark MLLib (often referred to as Spark ML), which you can use to create machine learning solutions. Machine learning is a technique in which you train a predictive *model* using a large volume of data so that when new data is submitted to the model it can predict unknown values. The most common types of machine learning are *supervised* learning and *unsupervised* learning. In a supervised learning scenario, you start with a large volume of data that includes both *features* (categorical and numeric values that describe characteristics of the entity you're trying to predict something about) and *labels* (the value your model will predict. Training the model involves applying a statistical algorithm that *fits* the features to the labels. Because your initial data includes known values for the labels, you can train the model and test its accuracy with these known label values – giving you confidence that the model will work accurately with new data for which the label values aren't known. Unsupervised learning is a technique in which there are no known label values, and the model is trained to group (or *cluster*) similar entities together based on their features.

In this lab, we'll focus on supervised learning; and specifically a type of machine learning called *classification* in which you train a model to identify which category, or *class* an entity belongs to.

Upload Source Data to Azure Storage

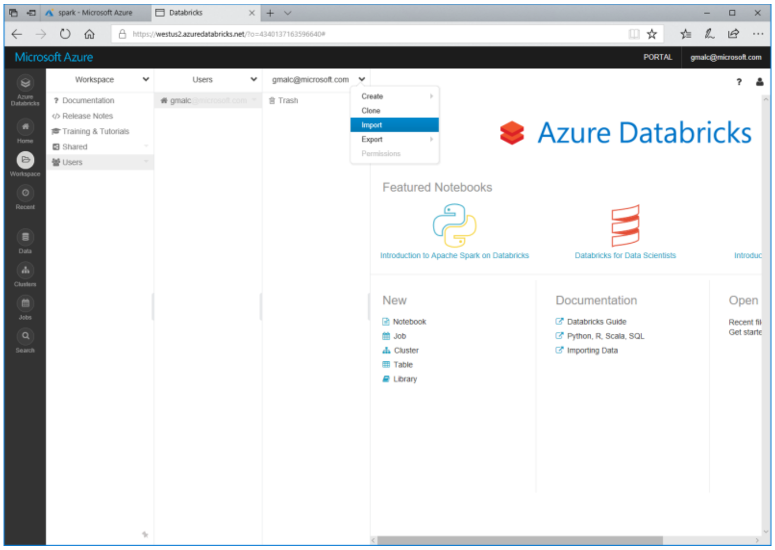
In this lab, you will train a classifier to use features of flights that are enroute to an airport, and predict whether they will be late or on-time. Before you can do this, you must upload the data file containing the data to your blob storage container where it can be accessed by your cluster. The 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 **raw-flight-data.csv** files exists. This file contain historic flight data you will use to train a classification model that predicts whether a flight will be late or not.
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 you created previously in this lab.
4. In the **Upload** drop-down list, click **Upload Files**. Then upload **raw-flight-data.csv** as a block blob to the **data** folder in root of the **spark** container that you created previously.

Train a Machine Learning Model

In this procedure, you will use your choice of Python or Scala to prepare and explore the flight data, before training and testing a classification model.

1. In the Databricks workspace, click **Workspace**. Then click **Users**, click your user name, and in the drop-down menu for your username click **Import** as shown here:



1. Browse to the folder where you extracted the lab files. Then select either **Machine Learning.ipynb** or **Machine Learning.scala**, depending on your preferred choice of language (Python or Scala), and upload it.
2. Open the file you uploaded to view the code it contains, read the comments to understand what it does, and change <*YOUR_ACCOUNT*> to the name of your Azure Storage account in the first code cell.
3. Read the notes and run the code cells to explore the data.

Clean Up

Note: This is the final lab in this course. If you have finished exploring Spark, follow the steps below to delete your Azure resources and avoid being charged for them 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.