Implementing Predictive Analytics with Spark in Azure Databricks

Lab 1 – Exploring Data with Spark

Overview

In this lab, you will use Spark to explore data and prepare it for predictive analysis.

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 settings*
    - **Auto Termination:** Terminate after 60 minutes.
    - **Spark Config**: Add a key-value pair (separated by a space) 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.

Exploring Data

Now that you have provisioned and configured a Spark cluster, you can use it to explore data.

Upload Source Data to Azure Storage

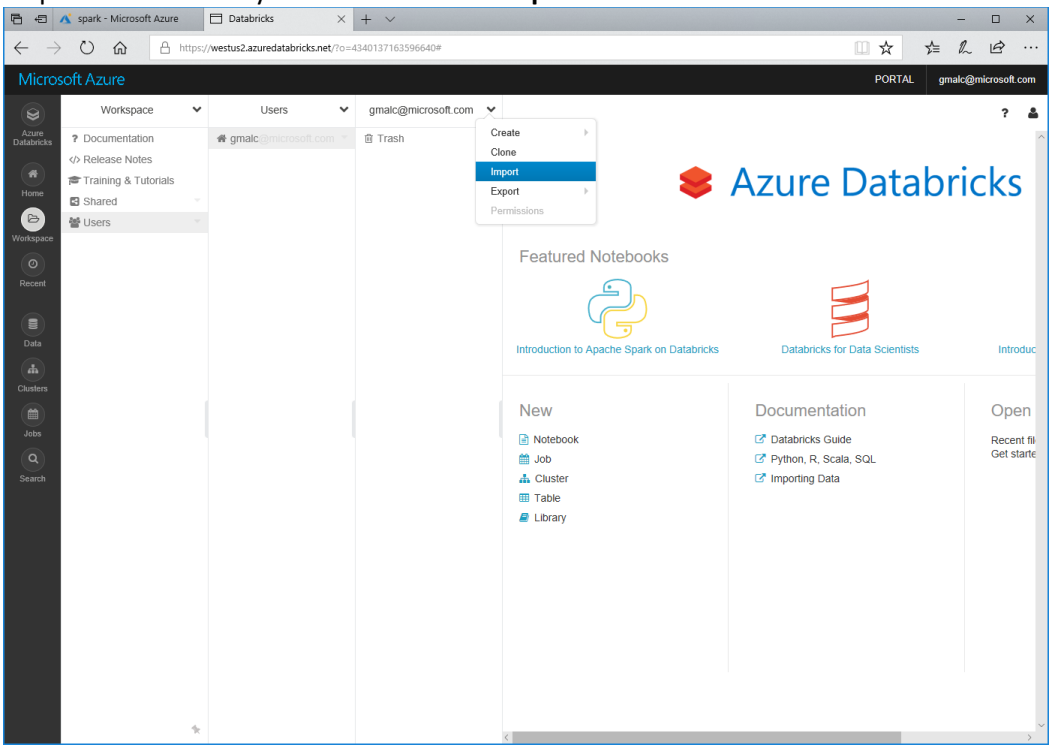
In this lab, you will explore data that contains records of flights. Before you can do this, you must store the flight data files in the shared storage used by your cluster. The instructions here assume you will use Azure Storage Explorer to do this, but you can use any Azure Storage tool you prefer.

1. In the folder where you extracted the lab files for this course on your local computer, in the **data** folder, verify that the **raw-flight-data.csv** and **airports.csv** files exist. These files contain the flight data you will explore
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.
4. In the **Upload** drop-down list, click **Upload Files**. Then upload **raw-flight-data.csv** and **airports.csv** as block blobs to a new folder named **data** in root of the **spark** container.

Upload and Explore a Notebook

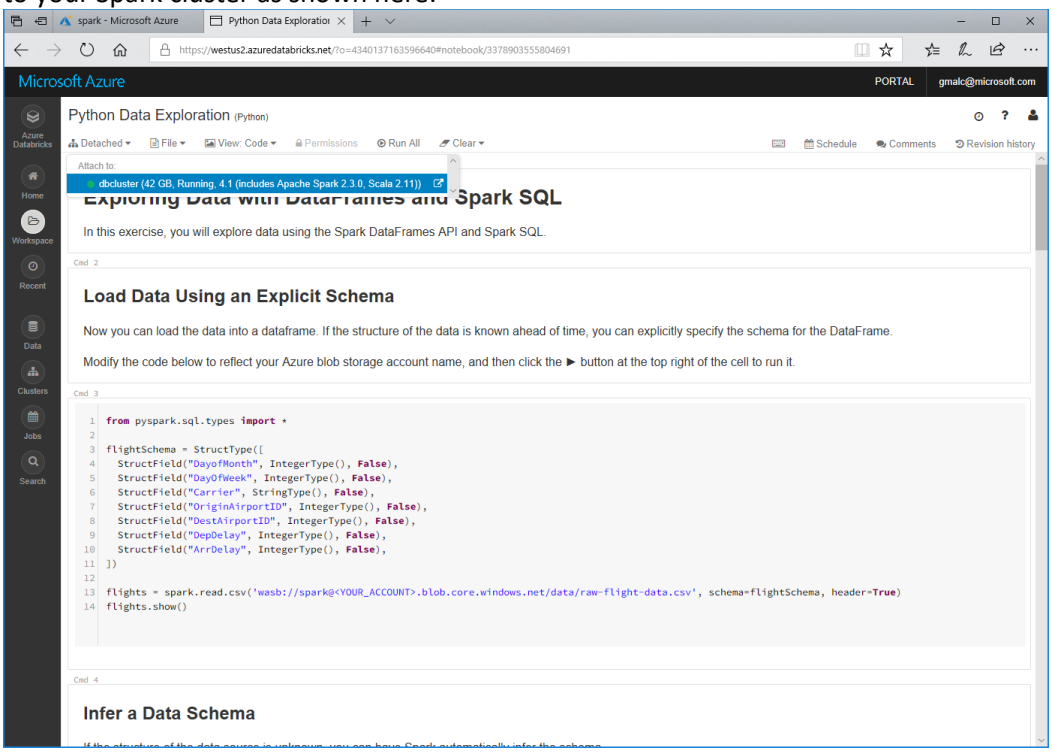
You will use a Notebook to explore the data. You can choose to work with Python or Scala.

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:



2. Browse to the **Lab01** folder in the folder where you extracted the lab files. Then select either **Python Data Exploration.ipynb** or **Scala Data Exploration.ipynb**, depending on your preferred choice of language, and upload it.

3. Open the notebook you uploaded and in the **Detached** drop-down menu, attach the notebook to your Spark cluster as shown here:



4. Read the notes and run the code cells to explore the flight data.

Clean Up

Note: If you intend to proceed straight to the next lab, skip this section. Otherwise, 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.