Implementing Predictive Analytics with Spark in Azure HDInsight

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 an HDInsight Spark Cluster

The first task you must perform is to provision an HDInsight Spark cluster.

Note: The Microsoft Azure portal is continually improved in response to customer feedback. The steps in this exercise reflect the user interface of the Microsoft Azure portal at the time of writing, but may not match the latest design of the portal exactly.

Provision an HDInsight Cluster

Note: If you already have a Spark HDInsight cluster running, you can skip this procedure.

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, in the Hub Menu, click **New**. Then in the **Data + Analytics** section select **HDInsight** and create a new HDInsight cluster with the following settings:
   - **Cluster Name**: *Enter a unique name (and make a note of it!)*
   - **Subscription**: *Select your Azure subscription*

- **Cluster type**
    - **Cluster Type**: Spark
    - **Cluster Operating System**: Linux
    - **HDInsight Version**: *Choose the latest version of Spark*
    - **Cluster Tier**: Standard
  - **Cluster Login Username:** *Enter a user name of your choice (and make a note of it!)*
  - **Cluster Login Password:** *Enter a strong password (and make a note of it!)*
  - **SSH Username:** *Enter another user name of your choice (and make a note of it!)*
  - **SSH Password:** *Use the same password as the cluster login password*
  - **Resource Group:**
      - **Create a new resource group**: *Enter a unique name (and make a note of it!)*
  - **Location**: *Choose any available data center location.*
  - **Storage:**
      - **Primary storage type**: Azure Storage
      - **Selection Method**: My Subscriptions
      - **Create a new storage account**: *Enter a unique name consisting of lower-case letters and numbers only (and make a note of it!)*
      - **Default Container**: *Enter the cluster name you specified previously*
  - **Applications**: *None*
- **Cluster Size**
    - **Number of Worker nodes:** 1
    - **Worker node size:** *Leave the default size selected*
    - **Head node size:** *Leave the default size selected*
- **Advanced Settings**: *None*

3. In the Azure portal, view **Notifications** to verify that deployment has started. Then wait for the cluster to be deployed (this can take a long time – often 30 minutes or more. Feel free to catch up on your social media networks while you wait!)

Note: As soon as an HDInsight cluster is running, the credit in your Azure subscription will start to be charged. Free-trial subscriptions include a limited amount of credit limit that you can spend over a period of 30 days, which should be enough to complete the labs in this course as long as clusters are deleted when not in use. If you decide not to complete this lab, follow the instructions in the *Clean Up* procedure at the end of the lab to delete your cluster to avoid using your Azure credit unnecessarily.

View the HDInsight Cluster in the Azure Portal

1. In the Azure portal, browse to the Spark cluster you just created.
2. In the blade for your cluster, under **Quick Links**, click **Cluster Dashboards**.
3. In the **Cluster Dashboards** blade, note the dashboards that are available. These include a Jupyter Notebook dashboard that you will use later in this course.

Install the Python Pandas Library

Note: Some Spark configuration and management is best accomplished through a remote secure shell (SSH) session in a console such as Bash. In this case, you will use an SSH session to install the latest version of the Python Pandas library, which is used in the labs. If you have a locally installed console, you can use it to connect to your Spark cluster; but you can also follow the instructions below to create a cloud-based console in the Azure portal.

1. At the top of the Azure portal page, click the CloudShell icon (**>_**).

2. The first time you do this, you will be prompted to choose a shell. Click **Bash (Linux)**. You will then be prompted to provision a storage account for the shell, so select your subscription and click **Create storage**. After a while, the bash prompt (*username*@azure:~$) will be displayed.
3. Enter the following command, replacing ***sshuser*** with the SSH user name you specified when provisioning your cluster (not the cluster login name), and replacing ***cluster*** with the name of your cluster:

   ```
   ssh sshuser@cluster-ssh.azurehdinsight.net
   ```

4. When asked to confirm the connection, enter **yes**.
5. When prompted, enter the password you specified for the SSH user when provisioning the cluster. After you have been authenticated, the prompt will change to show that you are connected to the head node (***hn0***) of your cluster.
6. Enter the following Bash command to install the Pandas library:

   ```
   sudo -HE /usr/bin/anaconda/bin/conda install pandas
   ```

7. Ignore any warnings about a new version of conda being available, and when prompted to proceed (after a minute or so), enter **y**.
8. Wait for the command to finish and the prompt to be displayed. Then close the cloud shell pane.

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 blob container for your HDInsight cluster.
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 container.

Upload and Explore a Jupyter Notebook

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

1. In the Azure portal, in the blade for your HDInsight cluster, under **Quick Links**, click **Cluster Dashboards**.
2. Click **Jupyter Notebook**, and if prompted, log in using the cluster login name you specified when provisioning your cluster (be sure you use login name for HTTP connections and not the SSH user name.)
3. Click **Upload**, and 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.

4. Open the notebook you uploaded and then read the notes and run the code it contains 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 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 Jupyter Notebooks dashboard if it is open.
2. In the Azure portal, view your **Resource groups** and select the resource group you created for your cluster. This resource group contains your cluster and the associated 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. If you have completed the course and no longer require the cloud shell you created, you can also delete the **cloud-shell-storage-*region*** resource group to remove the storage account used for the shell.
6. Close the browser.