Processing Big Data with Hadoop in Azure HDInsight

Lab 1B – Using PowerShell with HDInsight

Overview

In this lab, you will provision an HDInsight cluster. You will then use a PowerShell script to upload source data to Azure storage, run a MapReduce job, and download the results. Finally, you will delete your cluster and its storage.

What You'll Need

To complete the labs, you will need the following:

- A web browser
- A Microsoft account
- A Microsoft Azure subscription
- A Microsoft Windows computer with the following software installed:
    - Microsoft Azure PowerShell
- 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, installed and configured Azure PowerShell, and imported the publisher settings for your Azure subscription into PowerShell.

When working with cloud services, transient network errors can occasionally cause scripts to fail. If a script fails, and you believe that you have entered all of the required variables correctly; wait a few minutes and run the script again.

Provision an HDInsight Cluster

If you already have an HDInsight Hadoop cluster, you can skip this procedure.

1. In a web browser, navigate to http://azure.microsoft.com. Then click **Portal**, and if prompted, sign in using the Microsoft account that is associated with your Azure subscription.
2. In the Microsoft Azure portal, view the **HDInsight** page and verify that there are no existing HDInsight clusters in your subscription.

3. Click **NEW** (at the bottom of the page) and then click **CUSTOM CREATE**. Then use the New HDInsight Cluster wizard to create a new cluster with the following settings. Click the arrows to navigate through all of the wizard pages:
    - **Cluster Name**: *Enter a unique name (and make a note of it!)*
    - **Cluster Type**: Hadoop
    - **Operating System**: Windows Server 20012 R2 Datacenter
    - **HDInsight Version**: 3.2 (HDP 2.2, Hadoop 2.6)
    - **Data Nodes**: 2
    - **Region**: *Select any available region*
    - **Head Node Size**: A3 (4 cores, 7 GB memory)
    - **Data Node Size**: A3 (4 cores, 7 GB memory)
    - **HTTP User Name**: *Enter a user name of your choice (and make a note of it!)*
    - **HTTP Password**: *Enter and confirm a strong password (and make a note of it!)*
    - **Enable the remote desktop for cluster:** Selected
    - **RDP User Name:** *Enter another user name of your choice (and make a note of it!)*
    - **RDP Password:** *Enter and confirm a strong password (and make a note of it!)*
    - **Expires on**: Select tomorrow's date
    - **Enter the Hive/Oozie Metastore**: Unselected
    - **Storage Account**: Create New Storage
    - **Account Name**: *Enter a unique name for your storage account (and make a note of it!)*
    - **Default Container**: *Enter a unique name for your container (and make a note of it!)*
    - **Additional Storage Accounts**: 0
    - **Additional scripts to customize the cluster**: *None*
4. Wait for the cluster to be provisioned and the status to change to **Running** (this can take a while.)

Using PowerShell to Run a MapReduce Job

Now that you have provisioned an HDInsight cluster, you can use it to process data. You can use PowerShell to upload the source data you want to process, run a job, and then download the output generated by the job.

View Source Data

1. In the C:\HDILabs\Lab1B\Reviews folder, open **reviews1.txt** in Notepad and note that the file contains some sample text representing product reviews that have been posted on a web site by customers. The files named **reviews2.txt** and **reviews3.txt** include similar data.
2. Close Notepad without saving any changes.

Run a PowerShell Script to Process the Data

1. C:\HDILabs\Lab01B folder, rename **MapReduce.txt** to **MapReduce.ps1** (you may need to modify the *View* options for the folder to see file extensions). Then right-click **MapReduce.ps1** and click **Edit** to open the script file in the PowerShell ISE.
2. Change the values assigned to the **$clusterName**, **$storageAccountName**, and **$containerName** variables to match the names of the HDInsight cluster, storage account, and container you created in the previous exercise.
3. Examine the rest of the script, noting that it performs the following tasks:
    a. Removes any output left over from previous execution of the job.
    b. Uploads the contents of the **Reviews** subfolder to a folder named **reviewprocessing** in your Azure blob container.
    c. Runs the **wordcount** MapReduce code in the **hadoop-mapreduce-examples.jar** executable in the Azure blob store.
    d. Waits for the job to complete, and displays the job status output.

   e. Downloads the results file generated by the job to a local folder.
   f. Uses the **cat** command to display the contents of the downloaded results file.

4. Save the script. Then on the toolbar, click **Run Script**.

   Note: If you installed Azure PowerShell after August 14th 2015, you may see the following error. You can ignore this.

   ```
   Get-AzureHDInsightJobOutput : Could not load file or assembly
   'Microsoft.WindowsAzure.Storage, Version=3.0.3.0, Culture=neutral,
   PublicKeyToken=31bf3856ad364e35' or one of its dependencies. The system cannot find
   the file specified.
   ```

5. Observe the information displayed in the PowerShell command line pane as the script runs. When the script finishes, the words and their counts are displayed.
6. Keep the PowerShell ISE open for the next exercise.

Delete an HDInsight Cluster

Now that you have finished using the HDInsight cluster, you can delete it to reduce costs.

Delete the HDInsight Cluster and its Storage

1. In the Azure portal, click the **HDInsight** tab.
2. Select the row containing your HDInsight cluster, and then at the bottom of the page, click **Delete**. When prompted to confirm the deletion, click **Yes**.
3. Wait for your cluster to be deleted, and then click the **Storage** tab, and if necessary refresh the browser to view the storage account that was created with your cluster.
4. Select the row containing the storage account, and then at the bottom of the page, click **Delete**. When prompted to confirm the deletion, enter the storage account name and click **OK**.
5. Close the browser.