Processing Big Data with Hadoop in Azure HDInsight

Lab 1 - Getting Started with HDInsight

Overview

In this lab, you will provision an HDInsight cluster. You will then run a sample MapReduce job on the cluster and view the results.

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, Linux, or Apple Mac OS X computer on which the Azure CLI has been installed.
- 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.

Provisioning and Configuring an HDInsight Cluster

The first task you must perform is to provision an HDInsight 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

1. In a web browser, navigate to http://portal.azure.com. If prompted, sign in using the Microsoft account that is associated with your Azure subscription.
2. In the Microsoft Azure portal, click **All resources**, and verify that there are no existing HDInsight clusters in your subscription.
3. In the menu (on the left edge), click **New** (indicated by a **+**), and in the **Data + Analytics** category, click **HDInsight**. Then use the **New HDInsight Cluster** blade to create a new 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**: Hadoop
  - **Operating System**: Linux
  - **Version**: *Choose the latest version of Hadoop available*.
  - **Cluster Tier**: Standard
- **Resource Group:**
  - **Create a new resource group**: *Enter a unique name (and make a note of it!)*
- **Credentials:**
  - **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*
- **Storage:**
  - **Primary storage type**: Azure Storage
  - **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**: *View all and choose the smallest available size*
  - **Head Node Size**: *View all and choose the smallest available size*
- **Advanced Settings:** *None*

4. After you have clicked **Create**, wait for the cluster to be provisioned and the status to show as **Running** (this can take a while, so now is a good time for a coffee break!)

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 Cluster Configuration in the Azure Portal

1. In the Microsoft Azure portal, browse your resources and select your cluster. Then on the **HDInsight Cluster** blade, view the summary information for your cluster.
2. On the **HDInsight Cluster** blade, click **Scale Cluster**, and note that you can dynamically scale the number of worker nodes to meet processing demand.

View the Cluster Dashboard

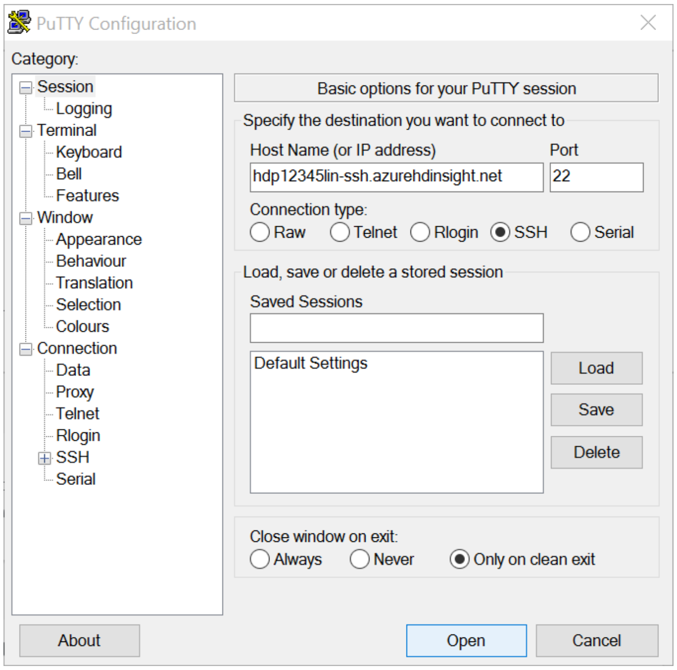
1. On the **HDInsight Cluster** blade, click **Dashboard**, and when prompted, log in using the cluster login username and password you specified when provisioning the cluster.
2. Explore the dashboard for your cluster. The dashboard is an Ambari web application in which you can view and configure settings for the Hadoop services running in the cluster. When you are finished, close its tab and return to the Azure portal tab.

Connecting to an HDInsight Cluster

Now that you have provisioned an HDInsight cluster, you can connect to it and process data for analysis.

If you are using a Windows client computer:

1. In the Microsoft Azure portal, on the **HDInsight Cluster** blade for your HDInsight cluster, click **Secure Shell**, and then in the **Secure Shell** blade, under **Windows users**, copy the **Host name** (which should be ***your_cluster_name*-ssh.azurehdinsight.net**) to the clipboard.
2. Open PuTTY, and in the **Session** page, paste the host name into the **Host Name** box. Then under **Connection type**, select **SSH** and click **Open**.



Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): hdp12345lin-ssh.azurehdinsight.net

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions: Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

3. If a security warning that the host certificate cannot be verified is displayed, click **Yes** to continue.
4. Wen prompted, enter the SSH username and password you specified when provisioning the cluster (<u>not</u> the cluster login).

If you are using a Mac OS X or Linux client computer:

1. In the Microsoft Azure portal, on the **HDInsight Cluster** blade for your HDInsight cluster, click **Secure Shell**, and then in the **Secure Shell** blade, under **Linux, Unix, and OS X users**, note the command used to connect to the head node.
2. Open a new terminal session, and enter the following command, specifying your SSH user name (<u>not</u> the cluster login) and cluster name as necessary:

```
ssh your_ssh_user_name@your_cluster_name-ssh.azurehdinsight.net
```

3. If you are prompted to connect even though the certificate can't be verified, enter **yes**.
4. When prompted, enter the password for the SSH username.

Note: If you have previously connected to a cluster with the same name, the certificate for the older cluster will still be stored and a connection may be denied because the new certificate does not match the stored certificate. You can delete the old certificate by using the **ssh-keygen** command, specifying the path of your certificate file (**f**) and the host record to be removed (**R**) - for example:

```
ssh-keygen -f "/home/usr/.ssh/known_hosts" -R clstr-ssh.azurehdinsight.net
```

Browse Cluster Storage

Now that you have opened an SSH console for your cluster, you can use it to work with the cluster shared storage system. Hadoop uses a file system named HDFS, which in Azure HDInsight clusters is implemented as a blob container in Azure Storage.

Note: The commands in this procedure are case-sensitive.

1. In the SSH console, enter the following command to view the contents of the root folder in the HDFS file system.

   ```
   hdfs dfs -ls /
   ```

2. Enter the following command to view the contents of the **/example** folder in the HDFS file system. This folder contains subfolders for sample apps, data, and JAR components.

   ```
   hdfs dfs -ls /example
   ```

3. Enter the following command to view the contents of the **/example/data/gutenberg** folder, which contains sample text files:

   ```
   hdfs dfs -ls /example/data/gutenberg
   ```

4. Enter the following command to view the text in the **davinci.txt** file:

   ```
   hdfs dfs -text /example/data/gutenberg/davinci.txt
   ```

5. Note that the file contains a large volume of unstructured text.

Run a MapReduce Job

Hadoop uses MapReduce jobs to distribute the processing of data across nodes in the cluster. Each job is divided into a map phase during which one or more mappers splits the data into key/value pairs, and a reduce phase, during which one or more reducers process the values for each key.

1. Enter the following command to view the sample Java jars stored in the cluster head node:

   ```
   ls /usr/hdp/current/hadoop-mapreduce-client
   ```

2. Enter the following command on a single line to get a list of MapReduce functions in the **hadoop-mapreduce-examples.jar**:

   ```
   hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-
   examples.jar
   ```

3. Enter the following command on a single line to get help for the **wordcount** function in the **hadoop-mapreduce-examples.jar** that is stored in the cluster head:

   ```
   hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-
   examples.jar wordcount
   ```

4. Enter the following command on a single line to run a MapReduce job using the **wordcount** function in the **hadoop-mapreduce-examples.jar** jar to process the davinci.txt file you viewed earlier and store the results of the job in the **/example/results** folder:

   ```
   hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-
   examples.jar wordcount /example/data/gutenberg/davinci.txt
   /example/results
   ```

5. Wait for the MapReduce job to complete, and then enter the following command to view the output folder, and note that a file named **part-r-00000** has been created by the job.

   ```
   hdfs dfs -ls /example/results
   ```

6. Enter the following command to view the results in the output file:

   ```
   hdfs dfs -text /example/results/part-r-00000
   ```

7. Minimize the SSH console window. Then proceed to the next exercise.

Uploading and Processing Data Files

In the previous exercise, you ran a Map Reduce job on a sample file that is provided with HDInsight. In this exercise, you will use Azure Storage Explorer to upload data to the Azure blob store for processing with Hadoop, and then download the results for analysis on your local computer.

Upload a File to Azure Blob Storage

1. View the contents of the **HDILabs\Lab01\reviews** folder where you extracted the lab files for this course, and verify that this folder contains a file named **reviews.txt**. This file contains product review text from a hypothetical web site on which cycles and cycling accessories are sold.
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 **folder**. Then upload the **reviews** folder as a block blob to a new folder named **reviews** in root of the container.

Process the Uploaded Data

1. Switch to the SSH console for your HDInsight cluster, and enter the following command on a single line to run a MapReduce job using the **wordcount** function in the **hadoop-mapreduce-examples.jar** jar to process the reviews.txt file you uploaded and store the results of the job in the **/reviews/results** folder:

   ```
   hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-
   examples.jar wordcount /reviews/reviews.txt /reviews/results
   ```

2. Wait for the MapReduce job to complete, and then enter the following command to view the output folder, and verify that a file named **part-r-00000** has been created by the job.

   ```
   hdfs dfs -ls /reviews/results
   ```

Download the Results

1. Switch back to Azure Storage Explorer, and browse to the **reviews/results** folder (you may need to refresh the root folder to see the **reviews** folder)
2. Double-click the **part-r-0000** text file to download it and open it in a text editor and view the word counts for the review data (the file is tab-delimited, and if you prefer, you can open it using a spreadsheet application such as Microsoft Excel).
3. Close the **part-r-00000** file, all command windows, and Azure Storage Explorer.

Clean Up

Now that you have finished this lab, you can delete the HDInsight cluster and storage account. This ensures that you avoid being charged for cluster resources when you are not using them. If you are using a trial Azure subscription that includes a limited free credit value, deleting the cluster maximizes your credit and helps to prevent using it all before the free trial period has ended.

Note: If you are proceeding straight to the next lab, omit this task and use the same cluster in the next lab. Otherwise, follow the steps below to delete your cluster and storage account.

Delete Cluster Resources

1. In the Microsoft Azure portal, click **Resource Groups**.
2. On the **Resource groups** blade, click the resource group that contains your HDInsight cluster, and then on the **Resource group** blade, click **Delete**. On the confirmation blade, type the name of your resource group, and click **Delete**.
3. Wait for your resource group to be deleted, and then click **All Resources**, and verify that the cluster, and the storage account that was created with your cluster, have both been removed.
4. Close the browser.