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 Hub menu (on the left edge), click **New** (indicated by a **+**), and in the **Intelligence + 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*
- **Select Cluster Type**:
    - **Cluster Type**: Hadoop
    - **Cluster Operating System**: Linux
    - **Version**: *Choose the latest version of Hadoop available*.
    - **Cluster Tier**: Standard
- **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 Authentication Type**: Password
    - **SSH Password:** *Enter a strong password (and make a note of it!)*
- **Data Source:**
    - **Create a new storage account**: *Enter a unique name consisting of lower-case letters and numbers only (and make a note of it!)*
    - **Choose Default Container**: *Enter the cluster name you specified previously*
    - **Location**: *Select any available region*
- **Pricing:**
    - **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*
- **Optional Configuration:** *None*
- **Resource Group:**
    - **Create a new resource group**: *Enter a unique name (and make a note of it!)*
- **Pin to dashboard**: *Unselected*

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!)

Important: As soon as an HDInsight cluster is running, the credit in your Azure subscription will start to be charged. The free-trial subscription includes a credit limit of approximately $100 (or local equivalent) that you can spend over a period of 30 days, which is 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 in order to avoid using your Azure credit unnecessarily.

View Cluster Configuration in the Azure Portal

1. In the Microsoft Azure portal, on the **HDInsight Cluster** blade, view the summary information for your cluster.
2. On the **HDInsight Cluster** blade, click **Settings**, then click **Properties**, and view the detailed properties of your cluster.
3. On the **HDInsight Cluster** blade, click **Scale Cluster**, and note that you can dynamically scale the number of cluster nodes to meet processing demand.

View the Cluster Dashboard

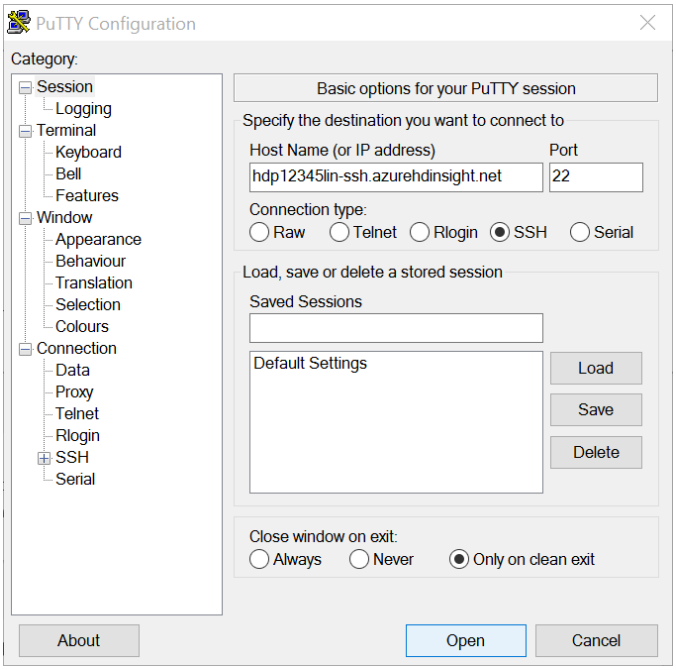
1. On the **HDInsight Cluster** blade, click **Dashboard**, and when prompted, log in using the cluster login username and password you specified when provisioning the cluster.
2. Explore the dashboard for your cluster. The dashboard is an Ambari web application in which you can view and configure settings 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**.



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 (not 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 (not 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.

Using the Azure CLI

The azure command-line interface is a cross-platform tool that you can use to work with Azure services, including HDInsight. In this exercise, you will use the Azure CLI to upload data to the Azure blob store for processing with Hadoop, and then download the results for analysis on your local computer.

View Azure Service Information

1. Open a new command line window.
2. Enter the following command to switch the Azure CLI to *resource manager* mode.

```
azure config mode arm
```

Note: If a *command not found* error is displayed, ensure that you have followed the instructions in the setup guide to install the Azure CLI.

3. Enter the following command to log into your Azure subscription:

```
azure login
```

4. Follow the instructions that are displayed to browse to the Azure device login site and enter the authentication code provided. Then sign into your Azure subscription using your Microsoft account.
5. Enter the following command to view your Azure resources:

```
azure resource list
```

6. Verify that your HDInsight cluster and the related storage account are both listed. Note that the information provided includes the resource group name as well as the individual resource names.
7. Note the resource group and storage account name, you will need them in the next procedure.

Upload a File to Azure Blob Storage

1. Enter a **dir** (Windows) or **ls** (Mac OS X or Linux) command to view the contents of the **HDILabs\Lab01** folder where you extracted the lab files for this course (for example, `dir c:\HDILabs\Lab01` or `ls HDILabs/Lab01`), 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. Enter the following command on a single line to determine the connection string for your Azure storage account, specifying the storage account and resource group names you noted earlier:

```
azure storage account connectionstring show storage_account -g
resource_group
```

3. Note the connection string, copying it to the clipboard if your command line tool supports it.
4. If you are working on a Windows client computer, enter the following command to set a system variable for the connection string:

   ```
   SET AZURE_STORAGE_CONNECTION_STRING=your_connection_string
   ```

   If you are using a Linux or Mac OS X client computer, enter the following command to set a system variable for the connection string (enter the connection string in quotation marks):

   ```
   export AZURE_STORAGE_CONNECTION_STRING="your_connection_string"
   ```

5. Enter the following command on a single line to upload the **reviews.txt** file to a blob named **reviews/reviews.txt** in the container used by your HDInsight cluster. Replace *local_path* with the local path to reviews.txt (for example *c:\HDILabs\Lab01\reviews.txt* or *HDILabs/Lab01/reviews.txt*) and replace *container* with the name of the storage container used by your cluster (which should be the same as the cluster name):

   ```
   azure storage blob upload local_path container reviews/reviews.txt
   ```

6. Wait for the file to be uploaded.

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 the command window for your local computer (in which you logged into Azure and uploaded the reviews.txt file).
2. Enter the following command on a single line to download the output file generated by the MapReduce job, replacing *container* with the name of the storage container used by your cluster (which should be the same as the cluster name), and *Lab01_path* with the path to the Lab01 folder (for example *c:\HDILabs\Lab01* or *HDILabs/Lab01*):

   ```
   azure storage blob download container reviews/results/part-r-00000
   Lab01_path
   ```

3. Use a text editor to open the **part-r-0000** text file that has been downloaded to the **HDILabs\Lab01\reviews\results** folder on your local computer 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).
4. Close the **part-r-00000** file and all remote desktop connections and command windows.

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. Close the browser tab containing the HDInsight Query Console.
2. In the Microsoft Azure portal, click **Resource Groups**.
3. 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**.
4. 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.
5. Close the browser.