Processing Big Data with Hadoop in Azure HDInsight

Lab 1A - 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 computer containing:
  - Power BI Desktop
  - 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 and installed Microsoft Power BI Desktop.

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://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.)

View Cluster Configuration

1. In the Azure portal, on the **HDInsight** page, click the name of your cluster.
2. On the **Dashboard** tab, view the summary information for your cluster.
3. On the **Monitor** tab, view the activity in your cluster (there should be none so far!)
4. On the **Configuration** tab, view the configuration status and note that this page contains a **Connect** icon at the bottom, with which you can open a remote desktop connection to the cluster.
5. On the **Scale** tab, note that you can dynamically scale the number of cluster nodes to meet processing demand.
6. Switch back to the **Configuration** tab. Then proceed to the next exercise.

Running a MapReduce Job

Now that you have provisioned an HDInsight cluster, you can use it to run a MapReduce job and process data for analysis.

Open a Remote Desktop Connection to the Cluster

1. In the Azure portal, on the **Configuration** tab for your HDInsight cluster, click **Connect** to open a remote desktop session to your HDInsight cluster. When prompted to open or save the .rdp file, click **Open**; then connect to the cluster using the RDP User Name and password you specified when provisioning the cluster.
2. When the remote desktop window opens, on the desktop, double-click the Hadoop Command Line icon and view the syntax documentation for the Hadoop command line tool.

Browse HDFS

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

1. In the Hadoop Command Line console window, enter the following command to see a list of supported HDFS commands.

   ```
   hadoop fs
   ```

2. Enter the following command to view the contents of the root folder in the HDFS file system.

   ```
   hadoop fs –ls /
   ```

3. 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.

   ```
   hadoop fs –ls /example
   ```

4. Enter the following command to view the contents of the **/example/jars** folder, and note that it contains a jar named **hadoop-mapreduce-examples.jar**:

   ```
   hadoop fs –ls /example/jars
   ```

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

   ```
   hadoop fs –ls /example/data/gutenberg
   ```

6. Enter the following command (on a single line) to copy the **davinci.txt** file to the local file system:

   ```
   hadoop fs –copyToLocal /example/data/gutenberg/davinci.txt
   c:\davinci.txt
   ```

7. Enter the following command to view the copied file in Notepad:

   ```
   Notepad.exe c:\davinci.txt
   ```

8. Note that the file contains unstructured text, and then close Notepad.

Run a MapReduce Job

1. In the Hadoop Command Line console window, enter the following command to run a MapReduce job using the **wordcount** class in the **hadoop-mapreduce-examples.jar** Java jar to process the davinci.txt file you viewed earlier and store the results of the job in the **/example/results** folder (enter the command on a single line – there is a space between *wordcount* and */example…*).

   ```
   hadoop jar hadoop-mapreduce-examples.jar wordcount
   /example/data/gutenberg/davinci.txt /example/results
   ```

2. 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.

   ```
   hadoop fs –ls /example/results
   ```

3. Close the Hadoop Command Line console window and sign out of the remote desktop session.

Download the Job Output

1. In the Azure portal, on the **Configuration** tab for your HDInsight cluster, click **Query Console**. Then log into the query console using the HTTP user name and password you specified when provisioning the cluster.
2. In the HDInsight Query Console, on the **File Browser** tab, navigate through your Azure storage account and container to the **\examples\results** folder. Then click **part-r-00000** and save it to a folder on your computer.
3. Start Power BI Desktop.

> **Note**: Power BI Desktop is the released version of the Power BI Designer preview tool used in the demonstrations for this course. The tool has been renamed and updated, and looks cosmetically different from the preview version; but still provides the same functionality as shown in the demonstrations.

4. If the welcome dialog box is displayed, click, **Get Data**; otherwise, in the **Get Data** drop-down list on the toolbar, click **More**.
5. In the **Get Data** dialog box, on the **File** page, click **Text**. Then click **Connect** and browse to the folder where you downloaded the **part-r-00000** file in step 2. Change the file type to **All Files** so you can see the **part-r-00000** file, then open it. The data has been imported as delimited text, but the delimiter has misidentified. You will resolve this issue by redesigning the query in the following steps.
6. Click **Edit** to open the query editor.
7. In the **Query Settings** pane, in the **Applied Steps** list, delete the **Changed Type** step.
8. In the **Applied Steps** list, click the settings icon next to the **Source** step, in the **Open File As** list, select **Text File**, and click **OK**. This combines the data into a single column.
9. Select the **Column1** header, and then on the ribbon, in the **Split Column** list, click **By Delimiter**. Then select the **Tab** delimiter and the option to split at the right-most delimiter, and click **OK**. This splits the data in the column correctly.
10. Right-click the column header for **Column 1.1**, click **Rename**, and rename the column to **Word**. Then rename **Column1.2** to **Count**.
11. With the **Count** column selected, on the toolbar, click the reverse sort (Z-A) button so that the most commonly used words are at the top of the list.
12. On the ribbon, click **Close & Load**. The data is loaded into the data model for the report.
13. Then in the **Fields** pane, expand the **part-r-00000** table. Then select **Word** and **Count**.
14. In the **Visualizations** pane, select **Bar Chart**. Then resize the chart to show the comparative frequency of the words in the source text.
15. Move the mouse to the top of the chart, click the **More Options** ellipse (…), and in the **Sort By** list, select **Count**.
16. After you have reviewed the results of the MapReduce job, close Power BI Desktop without saving any changes.

Clean Up

Now that you have finished using HDInsight to process data, you can delete your cluster. 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.

Delete the HDInsight Cluster

1. Close the browser tab containing the HDInsight Query Console.
2. In the Azure portal, click the **HDInsight** tab.
3. 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**.
4. 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.
5. 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**.
6. Close the browser.