Microsoft

Processing Big Data with Hadoop in Azure HDInsight

Lab 3A – Using Pig

Overview

In this lab, you will use Pig to process data. You will run Pig Latin statements and create Pig Latin scripts that cleanse, shape, and summarize data.

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
    - Microsoft 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, installed and configured Azure PowerShell, and imported the publisher settings for your Azure subscription into PowerShell. You must also have installed Microsoft Power BI Desktop.

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.

Using Pig Interactively

Pig is designed to provide a high-level processing engine over MapReduce that can be used to process structured, semi-structured, and unstructured data by executing a sequence of commands to transform the data. Pig commands are specified in a language called Pig Latin, which is designed to be intuitive and easy to use, while at the same time providing powerful functionality to make complex transformations to data.

Provision an Azure Storage Account and HDInsight Cluster

Note: If you already have an HDInsight cluster and associated storage account, you can skip this task.

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 and Upload the Source Data

1. Use Notepad to view the **Heathrow.txt** file in the C:\HDILabs\Lab03A folder. Note that the file contains monthly weather observation data for Heathrow airport in London from 1948 to 2015. The first few lines of the file contain unstructured text, which is followed by a multiple rows of tab-delimited values.

   Note: The file contains public sector information licensed under the Open Government License.
2. Close Notepad without saving any changes.
3. In the C:\HDILabs\Lab03A folder, rename **Upload Source Data.txt** to **Upload Source Data.ps1** (you may need to modify the *View* options for the folder to see file extensions). Then right-click **Upload Source Data.ps1** and click **Edit** to open the script in the Windows PowerShell ISE.
4. Change the values assigned to the **$clusterName**, **$storageAccountName**, and **$containerName** variables to match the configuration of your HDInsight cluster.
5. Review the code in the script, noting that it uploads the **Heathrow.txt** file to your Azure storage container. Then save the PowerShell script file and then click **Run Script** on the toolbar.
6. Wait for the script to finish and view the status information for the job that is displayed in the console pane. Then close Windows PowerShell ISE.
7. 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.

8. In the Microsoft Azure management portal, click the name of your HDInsight cluster to view its dashboard.
9. Click the **Configuration** tab, and then at the bottom of the page, click **Enable Remote**.
10. On the **Configure Remote Desktop** page, type the following details, and then click the **Complete** icon:
    - **User Name**: *A user name of your choice (but not the same as the existing HTTP user*)
    - **Password** (and **Confirm Password**): *A suitably complex password*.
    - **Expires On**: *Tomorrow's date*.
11. Wait for the operation to complete. Then click **Connect** and open a remote desktop session to your cluster using the remote user credentials you specified in step 4.
12. On the desktop, double-click the **Hadoop Command Line** shortcut.
13. In the Hadoop Command Line window, enter the following command to view the contents of the **/data/weather** folder in your cluster storage account:

    ```
    hadoop fs -ls /data/weather
    ```

14. Note that the folder contains a text file named **heathrow.txt**. To view the contents of this file, enter the following command:

    ```
    hadoop fs -cat /data/weather/heathrow.txt
    ```

15. Note that this file is the same text file you viewed on your local computer – it was uploaded by the PowerShell script.

Use the Grunt Shell to Run Pig Commands

1. In the Hadoop Command Line window, enter the following command to start Pig and open the Grunt command shell:

    ```
    %PIG_HOME%\bin\pig
    ```

2. In the Grunt shell, enter the following Pig Latin statement to load the files in the /data/weather folder into a relation called **Source**. The data is loaded into a tab-delimited schema with seven columns. Lines with no tab characters will result in a row (or *tuple*) containing a single value in the first column and NULL values in the remaining columns.

    ```
    Source = LOAD '/data/weather' USING PigStorage('\t') AS
    (year:chararray, month:chararray, maxtemp:float, mintemp:float,
    frost:int, rainfall:float, sunshine:float);
    ```

3. Ignore any warning messages, and enter the following Pig Latin statement to filter the data so that only tuples with a value in the **maxtemp** and **mintemp** columns are included.

    ```
    Data = FILTER Source BY maxtemp IS NOT NULL AND mintemp IS NOT NULL;
    ```

4. Enter the following Pig Latin statement to further filter the data to remove the header row.

    ```
    Readings = FILTER Data BY year != 'yyyy';
    ```

5. Enter the following Pig Latin statement to group the data by year.

    ```
    YearGroups = GROUP Readings BY year;
    ```

6. Enter the following Pig Latin statement to calculate average **maxtemp** and average **mintemp** for each year.

```
AggTemps = FOREACH YearGroups GENERATE group AS year,
AVG(Readings.maxtemp) AS avghigh, AVG(Readings.mintemp) AS avglow;
```

7. Enter the following Pig Latin statement to sort the temperature data by year.

```
SortedResults = ORDER AggTemps BY year;
```

8. Enter the following Pig Latin statement to display the results in the console. This runs a MapReduce job to perform all of the transformations you have entered so far.

```
DUMP SortedResults;
```

9. View the results, which include the average monthly maximum and minimum temperatures for each year. Then enter the following command to exit Pig.

```
quit;
```

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

Running Pig Scripts with PowerShell

Pig is commonly used to scrub source data by cleaning and reshaping it. Instead of entering Pig Latin commands interactively to do this, you can create a script that includes a sequence of Pig Latin commands to be performed on the data, and then use PowerShell to upload and execute the script as a Hadoop job.

View a Pig Latin Script

1. Use Notepad to open the scrubweather.pig file in the C:\HDILabs\Lab03A folder.
2. Review the Pig Latin code in this file, and note that the script performs the following tasks:
    a. Loads the contents of the **/data/weather** folder into a schema that includes the columns **year**, **month**, **maxtemp**, **mintemp**, **frostdays**, **rainfall**, and **sunshinehours**.
    b. Filters the data to remove text notes and header rows.
    c. Replaces any "---" values (which indicate missing data) in the **sunshinehours** column with an empty string.
    d. Splits the data into a relation in which **sunshinehours** contains a "'**#**" character denoting a sensor reading (which makes the data dirty) and a relation in which **sunshinehours** is already clean.
    e. Cleans the rows in which **sunshinehours** contains a "**#**" by using the SUBSTRING and INDEXOF functions.
    f. Re-combines the cleaned rows with the rows that were already clean.
    g. Sorts the data by year and month.
    h. Stores the cleaned and sorted data in the **/data/scrubbedweather** folder.
3. When you have finished viewing the script, close Notepad.

Use PowerShell to Upload and Run the Pig Latin Script

1. In the C:\HDILabs\Lab03A folder, rename **Run Pig Script.txt** to **Run Pig Script.ps1** (you may need to modify the *View* options for the folder to see file extensions). Then right-click **Run Pig Script.ps1** and click **Edit** to open the script in the Windows PowerShell ISE.
2. Change the values assigned to the **$clusterName**, **$storageAccountName**, and **$containerName** variables to match the configuration of your HDInsight cluster.
3. Review the code in the script, noting that it performs the following tasks:
    a. Removes any existing files left over from previous executions.
    b. Uploads the **Heathrow.txt** source data file to your Azure storage container.

    c. Uploads the **scrubweather.pig** script file to the **/data** folder in your Azure storage container.
    d. Starts an Azure HDInsight job to run the Pig script.

4. Save the PowerShell script file and then click **Run Script** on the toolbar.
5. Wait for the script to finish and view the status information for the job that is displayed in the console pane. Then close Windows PowerShell ISE.

> **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.
```

Analyze the Pig Output in Power BI Desktop

Unlike Hive, there is no ODBC driver for Pig. However, you can retrieve and analyze the results of Pig jobs (and other Hadoop jobs) using Power BI Desktop.

1. In your web browser, if you are not already logged into the Azure portal, navigate to http://azure.microsoft.com and click **Portal**. Then log into Azure using the Microsoft credentials associated with your Azure subscription.
2. On the **Storage** or **All Items** page, select your storage account. Then at the bottom of the page, click **Manage Access Keys**.
3. Click the icon next to the **Primary Access Key** and copy it to the clipboard. Then click **OK**.
4. Start Power BI Desktop and close the welcome page if it opens.

> **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.

5. On the **Home** tab, in the **Get Data** list, click **More**.
6. In the **Get Data** page, click **Azure**. Then select **Microsoft Azure HDInsight** and click **Connect**.
7. In the **Microsoft Azure HDInsight** dialog box, enter the name of the Azure storage account associated with your HDInsight cluster (for example *hd123456store*), and then click **OK**.
8. In the **Account Key** box, paste the primary access key you copied earlier. Then click **Connect**.
9. In the **Navigator** window, expand your storage account and select the container where your HDInsight files are stored. Then click **Edit**.
10. In the query editor, in the drop-down list for the **Folder Path** column heading, click **Text Filters** and then click **Contains**.
11. In the **Filter Rows** dialog box, in the first row, next to **Contains**, type “scrubbedweather”; then click **OK**.
12. In the query editor, in the row for the file named **part-r-00000**, click **Binary**. The query editor loads the delimited text file generated by Pig.
13. In the query editor, right-click the **Column1** heading and click **Rename**. Then rename the column to **Year**.
14. Repeat the previous step to rename the remaining columns as follows:
    - **Column2**: Month
    - **Column3**: MaxTemp

- **Column4**: MinTemp
- **Column5**: FrostDays
- **Column6**: Rainfall
- **Column7**: Sunshine

15. Right-click the **FrostDays** column, point to **Change Type**, and click **Whole Number**
16. Right-click the **Sunshine** column header, point to **Change Type**, and click **Decimal Number**.
17. On the ribbon, click **Close & Load**, and wait for the data to be loaded into the data model
18. In the **Fields** pane, expand the table for your storage container and select the **Rainfall** and **Sunshine** fields. Then in the **Visualizations** pane, in the **Value** area, for each of the **Rainfall** and **Sunshine** fields, click the drop-down arrow and select **Average**.
19. Drag the **Month** field to the **Axis** area in the **Visualizations** pane, and resize the chart if necessary to see the average rainfall and sunshine for each month.
20. Close Power BI Desktop without saving your report.

Cleaning Up

Now that you have finished this lab, you can delete the HDInsight cluster and storage account.

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 the HDInsight Cluster

If you no longer need the HDInsight cluster used in this lab, you should delete it to avoid incurring unnecessary costs (or using credits in a free trial subscription).

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.