Processing Big Data with Hadoop in Azure HDInsight

Lab 4A – Using Sqoop

Overview

In this lab, you will use Sqoop to transfer the results of data processing in HDInsight to Azure SQL Database. HDInsight provides a powerful platform for transforming and cleansing data; but while Hive offers a SQL-like interface through which to query data, many organizations prefer to store data in a relational database from where it can be accessed by applications and users who need to query it. Sqoop provides support for bi-directional data transfer between a Hadoop data store (in the case of HDInsight, Azure blob storage) and a range of databases that can be accessed using JDBC.

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
    - Microsoft Visual Studio with the Azure SDK
- 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, imported the publisher settings for your Azure subscription into PowerShell, installed Visual Studio and the Azure SDK, and 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.

Provisioning Azure Services

Before you can process data with HDInsight and transfer the results to Azure SQL Database, you will need an HDInsight cluster and an Azure SQL Database server. Azure SQL Database is a platform-as-a-service (PaaS) solution in Azure that offers a cloud-based SQL Server compatible relational database.

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.)

Provision Azure SQL Database

1. In a web browser, navigate to http://azure.microsoft.com and click **Portal**, and if prompted, sign in using the Microsoft account that is associated with your Azure subscription.
2. View the **HDInsight** page and note the **Location** of your HDInsight cluster. Then, in the Azure portal, view the **SQL Databases** page.
3. Click **New**, and click **Custom Create**. Then in the **Name** box, type **AnalysisDB** and in the **Server** list, select **New SQL database server**. Then click **Next**.
4. On the **Create Server** page, enter a login name and password of your choice, make a note of the login name and password you have specified, and select the region in which your HDInsight cluster is defined. Then ensure that the **Allow Windows Azure Services to Access the Server** and **Enable Latest SQL Database Updates (V12)** are both selected, and click **Complete**.

5. Wait a minute or so for the database status to become online (you may need to refresh the web page in the browser), and then note the name in the **Server** column – this is the Azure SQL Database server that has been created to host your database.
6. Click the server name, and then click **Configure**.
7. Under **allowed ip addresses**, next to your current IP address, click **Add to the Allowed IP Addresses**. Then at the bottom of the page, click **Save**.

Create a Table in the Database

1. In the Microsoft Azure portal, view the **SQL Databases** page, and select the **AnalysisDB** database. Then, at the bottom of the page, click **Open in Visual Studio** (you may need to allow pop-ups and repeat this step). In the various confirmation messages that are displayed, confirm that you want to open the database in Visual Studio.
2. When prompted to connect to SQL Server, connect to the Azure SQL database server using the following settings (based on the server name, login name, and password for your Azure SQL Database):
    - **Server name**: *your_azure_sql_server_name*.database.windows.net
    - **Login name**: *your_sql_login_name*
    - **Password**: *your_sql_login_password*.
3. After a few moments, when the Azure SQL database server and its **Databases** folder is displayed in Object Explorer, right-click the **AnalysisDB** database and click **New Query**.
4. In the SQLQuery1sql pane, enter the following Transact-SQL code to create a table named **weather**, and then click the **Execute** button on the SQLQuery1.sql pane toolbar (you can copy and paste this code from **Create Table.txt** in the C:\HDILabs\Lab04A folder.)

```
CREATE TABLE weather
([Year] int PRIMARY KEY CLUSTERED,
 [AvgMaxTemp] float,
 [AvgMinTemp] float,
 [FrostDays] int,
 [Rainfall] float,
 [Sunshine] float);
```

5. Close Visual Studio without saving any files.

Process Data in HDInsight and Transfer Results to Azure SQL Database

Now that you have an Azure SQL database containing an empty table, you are ready to use HDInsight to process your source data and transfer the results to Azure SQL Database.

Process Data and Transfer the Results

1. In the C:\HDILabs\Lab04A folder, rename **Process Data.txt** to **Process Data.ps1** (you may need to modify the *View* options for the folder to see file extensions). Then right-click **Process Data.ps1** and click **Edit** to open the script in the Windows PowerShell interactive script environment (ISE).
2. Change the values assigned to the **$clusterName**, **$storageAccountName**, and **$containerName** variables to match the configuration of your HDInsight cluster.
3. Change the values assigned to the following variables based on the configuration of your Azure SQL Database server:
    - **$sqlDatabaseServerName**: *Your Azure SQL server name*
    - **$sqlDatabaseUserName**: *Your SQL login name*
    - **$sqlDatabasePassword:** *Your SQL Login Password* (note that you must prefix reserved PowerShell characters such as **$** and **#** with a **`** character, so the value *SecurePa`$`$w0rd* assigns the password *SecurePa$$w0rd*).
4. Save the script and then review the rest of the code, noting that it performs the following actions:

a. Removes leftover files from any previous executions.
b. Uploads the **Source.txt** file in the local folder containing the script to **/data/temp/source** in your Azure storage container.
c. Uploads the **Pig.txt** file to **/data/temp** in your Azure storage account, and then starts a Pig job to run the Pig Latin script that the file contains.
d. Uploads the **Hive.txt** file to **/data/temp** in your Azure storage account, and then starts a Hive job to run the HiveQL script that the file contains.
e. Creates and runs a Sqoop job that exports the data files in the **/data/temp/hivetable** folder in your Azure storage account to the **weather** table in your Azure SQL Database server.

5. Keep the PowerShell ISE window open, and use Notepad to open **Source.txt**. This file contains the source data that you will use HDInsight to process.
   Note: The file contains public sector information licensed under the Open Government License.
6. In Notepad, open **Pig.txt** and review the code it contains. This simple Pig Latin script loads the source data into a specified schema, and then saves it in the **/data/temp/pigoutput** folder.
7. In Notepad, open **Hive.txt** and review the HiveQL script it contains. This script creates a Hive table named **tempdata** based on the **/data/temp/pigoutput** folder (which contains the result of the Pig script), and then creates a table called **tempsummary** based on the **/data/temp/hivetable** folder that is populated by the results of a query against the **tempdata** table.
8. Close notepad without saving any changes. Then in the PowerShell ISE window, on the toolbar, click **Run Script**.
9. Wait for the script to finish (which can take several minutes) and 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.

Access the Processed Data in Azure SQL Database

1. 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.
2. On the **Home** tab, in the **Get Data** list, click **SQL Server**.
3. In the **Microsoft SQL Database** page, in the **Server** box, type the fully-qualified name of your Azure SQL Database server in the format *your_azure_sql_server_name*.database.windows.net, and in the Database box, type **AnalysisDB**. Then click **OK**.
4. In the **Access a Microsoft SQL Database** dialog box, click the **Database** page. Then in the **Username** box type your Azure SQL Database login name, in the **Password** box type your Azure SQL Database login password, and click **Connect**.
5. In the **Navigator** window, select the **weather** table. Then click **Load**.
6. In the **Fields** pane, expand the **weather** table and select **AvgMaxTemp** and **AvgMinTemp**. Then drag **Year** to the **Axis** area in the **Visualizations** pane.

7. In the Visualizations pane, select **Area Chart**, and then view the chart to see average minimum and maximum temperatures for over time.
8. Close Power BI Desktop without saving the 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 and SQL database in the next lab. Otherwise, follow the steps below to delete your cluster and storage account.

Delete the HDInsight Cluster and Azure SQL Database Server

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

1. In the Azure portal, click the **HDInsight** page.
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** page, 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. In the Azure portal, click the **SQL Databases** page.
6. Click the **Servers** tab.
7. Select the row containing your SQL Database server, and then at the bottom of the page, click **Delete**. When prompted to confirm the deletion, enter the server name and click **OK**.
8. Close the browser.