Big Data Orientation

Lab 1 – Working with Data Files in Microsoft Azure

Overview

Before you start to design and build big data processing solutions, you should familiarize yourself with the Microsoft Azure portal and some of the key Azure resources that you will need to provision and use to store and process data.

In this lab, you will explore some commonly used Azure data storage services, and use them to store and manipulate data.

Before You Start

To complete this lab, you will need the following:

- A web browser
- A Windows, Linux, or Mac OS X computer

Before you can perform the lab exercises, you will also need:

- A Microsoft account
- A Microsoft Azure subscription
- The lab files for this course

Sign up for a Microsoft Account

You need a Microsoft account, such as an *outlook.com* or *live.com* address, in order to sign up for a Microsoft Azure subscription. If you already have a Microsoft account that has not already been used to sign up for a free Azure trial subscription, you're ready to get started. If not, don't worry, just create a new Microsoft account at https://signup.live.com.

Sign up for a Microsoft Azure Subscription

If you already have a Microsoft Azure subscription, you can skip this section, download the lab files, and get started.

If you do not have an Azure subscription, you can sign up for a free 30-day trial subscription that includes approximately $200 credit in your local currency at https://aka.ms/edx-dat229x-az. The credit in this trial subscription expires after 30 days (after which time you can purchase a pay-as-you-go Azure subscription.

Note: Some Azure resources consume credit even when not in use, so be careful to delete resources if you don't intend to use them for a while; otherwise you will run out of credit before the month ends.

To sign up for a free Azure subscription, you will need to provide a valid credit card number for identity verification, but you will not be charged for Azure services – for more information, see the frequently asked questions in the Azure sign-up page.

Download the Lab Files

The files you'll need for this lab are provided in a zip file at https://aka.ms/edx-dat229x-labs. Download and extract the lab files to a folder on your local computer.

Exercise 1: Working with Azure Storage

Azure Storage is a general purpose storage service for storing data in Azure. It is widely used by applications and can provide distributed storage services for Big Data technologies. In this exercise, you will create an Azure Storage account and use it to store files.

Provision an Azure Storage Account and a Blob Container

You will use an Azure storage account to store data files. The files will be stored as binary large objects (BLOBs) in a blob store container that is hosted in your storage account.

1. In a web browser, navigate to http://portal.azure.com, and if prompted, sign in using the Microsoft account that is associated with your Azure subscription.
2. In the Microsoft Azure portal, in the Hub Menu (on the left edge of the page), click **New**. Then in the **Storage** menu, click **Storage account**.
3. In the **Create storage account** blade, enter the following settings, and then click **Create**:
    - **Name**: *Enter a unique name for your storage account (and make a note of it!)*
    - **Deployment model:** Resource manager
    - **Account kind**: General purpose
    - **Performance:** Standard
    - **Replication:** *Select **Locally redundant storage (LRS)***
    - **Secure transfer required**: Disabled
    - **Subscription**: *Your subscription*
    - **Resource group**: *Select **Create New** and enter a name for a new resource group (and make a note of it!)*
    - **Location:** *Select any available region*
    - **Pin to dashboard:** *Unselected*
4. At the top of the page, click **Notifications** and verify that deployment has started. Wait until your storage account has been created. This should take a few minutes.
5. In the Hub menu, click **All resources**, and then click your storage account.
6. In the blade for your storage account, click the **Blobs** tile, and note that you don't have any containers yet.
7. In the blob service blade, click **+Container**, and create a new container with the following properties:
    - **Name**: bigdata
    - **Access type**: Private
8. After the **bigdata** blob container has been created, click it and verify that it contains no blobs.
9. In the bigdata blade, click **Properties** and view the **URL** for the blob container, which should be in the form **https://<*your_account_name*>.blob.core.windows.net/bigdata**. This is the URL that client applications can use to access your blob container using HTTP protocol.

> **Note**: Azure blob storage also supports the WASB protocol, which is specific to Azure storage. Some big data processing technologies use this protocol to work with Azure storage.

10. Return to the blade for your storage account, and under **Settings**, click **Access keys**. Then on the **Access Keys** page, note that two keys have been generated. These are used to authenticate client applications that connect to your storage account.

Use the Azure Portal to Upload a File to Azure Storage

The Azure portal includes a rudimentary graphical interface that you can use to work with your Azure storage account. You can use this to transfer files between your local computer and your blob containers, and to browse the data in your storage account.

1. In the blade for your storage account, view the **Overview** page, and then click the **Blobs** tile.
2. Click the **bigdata** container that you created previously, and then click **Upload**.
3. In the **Upload blob** blade, browse to the folder where you extracted the lab files for this course, and select **products.txt**. Then verify the following settings and click **Upload**:
    - **Files**: "products.txt"
    - **Blob type**: Block blob
    - **Block size**: 100 MB

    Note: Azure storage supports three blob types (*block*, *page*, and *append*). Block blobs are formed of one or more blocks of data based on the specified block size, and provide an efficient format for uploading and transferring blobs. For more details about blob types in Azure storage, see https://docs.microsoft.com/en-us/rest/api/storageservices/fileservices/Understanding-Block-Blobs--Append-Blobs--and-Page-Blobs.

4. After the blob has been uploaded, note that it is listed in the **bigdata** container blade. Then click the **Products.txt** blob and in the **Blob** properties blade, note its **URL**, which should be similar to **https://*<your_account_name>*.blob.core.windows.net/bigdata/products.txt**.

Use Azure Storage Explorer to Upload files to Azure Storage

The blob container interface in the Azure portal enables you to upload, browse, and download blobs; but it lacks many of the features expected in a modern file management tool. There are various graphical Azure storage management tools available, including support for exploring your Azure storage in Microsoft Visual Studio. However, if you do not need the full Visual Studio environment, you can install Azure Storage Explorer, which is available for Windows, Mac OSX, and Linux.

1. Open a new browser tab and browse to http://storageexplorer.com.
2. Download and install the latest version of Azure Storage Explorer for your operating system (Windows, Mac OSX, or Linux).
3. When the application is installed, launch it. Then add your Azure account, signing in with your Azure credentials when prompted, and configure Storage Explorer to show resources from the Azure subscription in which you created your storage account.
4. After your subscription has been added to the **Explorer** pane expand your storage account, expand **Blob Containers**, and select the **bigdata** container. Note that the **products.txt** file you uploaded previously is listed.
5. In the **Upload** drop-down menu, note that you can choose to upload individual files or folders. Then select **Upload Folder** and browse to the folder where you extracted the lab files for this course and select the **data** folder, and upload it as a block blob.

6. After the upload operation is complete, double-click the **data** folder in your blob container to open it, and verify that it contains files named **customers.txt** and **reviews.txt**.
7. Click the ↑ button to navigate back up to the root of the **bigdata** container, and select the **products.txt** file. Then click **Copy**.
8. Open the **data** folder, and then click **Paste** to copy the **product.txt** file to this folder.
9. Navigate back up to the root of the **bigdata** container, and select the **products.txt** file. Then click **Delete**, and when prompted to confirm the deletion, click **Yes**.
10. Verify that the **bigdata** container now contains only a folder named **data**, which in turn contains files named **customers.txt**, **products.txt**, and **reviews.txt**.

Exercise 2: Working with Azure Data Lake Store

Azure Data Lake Store is a storage service in Azure that is optimized for big data workloads. It supports unlimited numbers of files of unlimited size and can be used to organize and secure files in folder hierarchies. In this exercise, you will provision Azure Data Lake Store and upload some files to it.

Note: For a detailed comparison of Azure Storage and Azure Data Lake Store, see https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-comparison-with-blob-storage.

Provision Azure Data Lake Store

To get started, you must provision Azure Data Lake Store.

1. In the Microsoft Azure portal, in the Hub Menu (on the left edge of the page), click **New**. Then in the **Storage** menu, click **Data Lake Store**.
2. In the **New Data Lake Store** blade, enter the following settings, and then click **Create**:
    - **Name**: *Enter a unique name for your storage account (and make a note of it!)*
    - **Resource group**: *Select **Use existing** and select the resource group you created previously.*
    - **Location:** *Select any available region*
    - **Pricing:** *Pay-as-you-go*
    - **Encryption Settings:** Enabled
    - **Pin to dashboard:** *Unselected*
3. At the top of the page, click **Notifications** and verify that deployment has started. Wait until your storage account has been created. This should take a few minutes.

Upload Files to Azure Data Lake Store

Now that you have provisioned Azure Data Lake Store, you can use the Data Explorer in the Azure Portal to upload files.

1. In the Hub menu, click **All resources**, and then click your Data Lake Store.
2. In the blade for your Data Lake Store, click **Data Explorer**.
3. In the Data Explorer blade, click **New Folder**. Then create a folder named **data**.
4. Click the **data** folder to open it.
5. Click **Folder Properties**; and in the **Properties** blade, view the PATH property. This is the folder URL, and should be similar to **adl://*<your_account_name>*.azuredatalakestore.net/data/**
6. In the blade for the **data** folder, click **Upload**.
7. In the **Upload files** blade, browse to the folder where you extracted the lab files for this course and view the contents of the **data** folder. Then select all the files it contains (hold the CTRL key to select multiple files) and click **Add selected files** to upload them to the **data** folder in your Azure Data Lake Store.

8. After the files have been uploaded, close the **Upload files** blade and verify that the data folder in your Azure Data Lake Store now contains **customers.txt** and **reviews.txt**.
9. Click **reviews.txt** and view the preview of the data. Note that the tab-delimiter has been detected and the data is shown as a table with multiple columns.