Provisioning HDInsight

Data and Analytics Training

Big Data – Self-Study Guide (202)

© Copyright 2014 Avanade Inc. All Rights Reserved. This document contains confidential and proprietary information of Avanade and may be protected by patents, trademarks, copyrights, trade secrets, and/or other relevant state, federal, and foreign laws. Its receipt or possession does not convey any rights to reproduce, disclose its contents, or to manufacture, use or sell anything contained herein. Forwarding, reproducing, disclosing or using without specific written authorization of Avanade is strictly forbidden. The Avanade name and logo are registered trademarks in the US and other countries. Other brand and product names are trademarks of their respective owners.

This study guide serves as a basis for your self-study on the basics of HDInsight. The study guide gives you an overview on the main topics and presents a number of internal and external resources you should use to get a deeper understanding of HDInsight, the related terms and technical details.

Please especially use the external resources provided. They are a main part of this study guide and are required to get a good understanding of the discussed topics.

Estimated Completion Time

4 hours (including time for external resources)

Prerequisites

We recommend completing the course "*Big Data – 201 – HDFS*" to get a basic understanding of the underlying filesystem, as well as completing the level 100 courses (especially the course "*Big Data – 101 – Getting started with Big Data*") before starting with this course.

Objectives

After completing this study guide, you will be able to understand

- Provisioning of HDInsight cluster using
    - Management portal
    - PowerShell
    - Cross-platform command line
    - .NET SDK

© Copyright 2014 Avanade Inc. All Rights Reserved.

Contents

© Copyright 2014 Avanade Inc. All Rights Reserved.

© Copyright 2014 Avanade Inc. All Rights Reserved.

Introduction

Apache Hadoop is a software framework that facilitates big data management and analysis. Apache Hadoop core provides reliable data storage with the Hadoop Distributed File System (HDFS), and a simple MapReduce programming model to process and analyze, in parallel, the data stored in this distributed system. HDFS uses data replication to address hardware failure issues that arise when deploying such highly distributed systems.

Azure HDInsight makes Apache Hadoop available as a service in the cloud. It makes the HDFS/MapReduce software framework and related projects such as Pig, Hive, and Oozie available in a simpler, more scalable, and cost-efficient environment.

HDInsight clusters are deployed in Azure on compute nodes to execute MapReduce tasks and can be dropped by users once these tasks have been completed. Keeping the data in the HDFS clusters after computations have been completed would be an expensive way to store this data. Blob storage is a robust, general purpose Azure storage solution. So storing data in Blob storage enables the clusters used for computation to be safely deleted without losing user data. But Blob storage is not just a low cost solution: It provides a full-featured HDFS file system interface that provides a seamless experience to customers by enabling the full set of components in the Hadoop ecosystem to operate (by default) directly on the data that it manages.[1]

An HDInsight cluster abstracts the Hadoop implementation details so that you don't need to worry about how to communicate with different nodes of a cluster. When you provision an HDInsight cluster, you provision Azure compute resources that contain Hadoop and related applications. The data to be churned is stored in Azure Blob storage, also called *Azure Storage - Blob* (or WASB) in the context of HDInsight.

HDInsight uses Azure PowerShell to configure, run, and post-process Hadoop jobs. HDInsight also provides a Sqoop connector that can be used to import data from an Azure SQL database to HDFS or to export data to an Azure SQL database from HDFS.

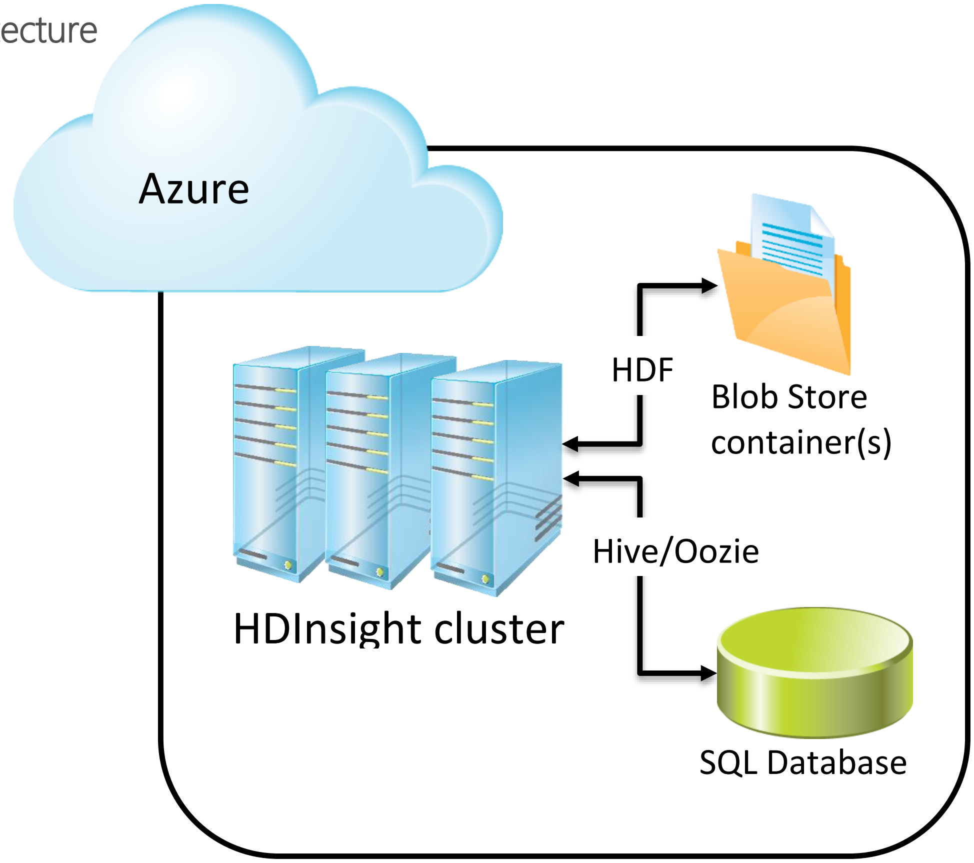
HDInsight has also made YARN available. It is a new, general-purpose, distributed, application management framework that replaces the classic Apache Hadoop MapReduce framework for processing data in Hadoop clusters. It effectively serves as the Hadoop operating system, and takes Hadoop from a single-use data platform for batch processing to a multi-use platform that enables batch, interactive, online and stream processing. This new management framework improves

[1] http://azure.microsoft.com/en-us/documentation/articles/hdinsight-introduction/

© Copyright 2014 Avanade Inc. All Rights Reserved.

scalability and cluster utilization according to criteria such as capacity guarantees, fairness, and service-level agreements.[2]

HDInsight Architecture



In the figure you can see, within Windows Azure, we've got a cluster. These are actually, just virtual machines that are provisioned for us on the cluster. You can use Powershell or you can create a Remote Desktop connection to these node, which allows us to interact with the cluster.

The HDFS storage is actually implemented using the Windows Azure blob store. So, if you are familiar with Windows Azure, the blob store is generic storage location within Windows Azure for all sorts of things and HDInsight leverages that for its HDFS store. So, we've got a robust replicated storage location there, that what we use for our HDFS.

The other thing you might occasionally use is the Windows Azure SQL Database. You can either explicitly point to a SQL Database that you provisioned and you can use that to store your metadata for Hive and Oozie. Or alternatively, if you don't specify one, when you provision an HDInsight cluster,

[2] http://azure.microsoft.com/en-us/documentation/articles/hdinsight-introduction/

Windows Azure will spin one up anyway. And it will use that. You never get to see it, but it will use that internally to store the data.

Storage account

HDInsight makes Apache Hadoop available as a service in cloud. Hadoop offers a distributed platform to store and manage large volumes of unstructured data. HDInsight uses Azure Blob storage for storing data. A specific Blob container from the account is used as the default file system, just like HDFS.

You can find details on the HDFS blob storage in the course "*Big Data – 201 – HDFS*".

HDInsight uses Azure Blob Storage for storing data. It is called *WASB* or *Windows Azure Storage - Blob*. WASB is Microsoft's implementation of HDFS on Azure Blob storage. When you provision an HDInsight cluster, you specify an Azure Storage account. The HDInsight cluster is by default provisioned in the same data center as the storage account you specify.

In addition to this storage account, you can add additional storage accounts when you custom-configure an HDInsight cluster. This additional storage account can either be from the same Azure subscription or different Azure subscriptions.

To simplify this tutorial, only the default blob container and the default storage account are used. In practice, the data files are usually stored in a designated storage account.

The quick-create option to provision an HDInsight cluster, like the one we use in this tutorial, does not ask for a location while provisioning the cluster. Instead, it by default co-locates the cluster in the same data center as the storage account. You should make sure to create your storage account in the locations supported for the cluster, which are: East Asia, Southeast Asia, North Europe, West Europe, East US, West US, North Central US, South Central US.

Virtual Networking

Azure Virtual Network easily extends your on-premises network through site-to-site VPN, much the way you'd set up and connect to a remote branch office. You control the network topology, including configuration of DNS and IP address ranges, and manage it just like your on-premises infrastructure.[3]

The details of virtual networking are not important for the examples shown in this course, but may get relevant once you think about integration in more difficult productive environments.

[3] http://azure.microsoft.com/en-us/services/virtual-network/

© Copyright 2014 Avanade Inc. All Rights Reserved.

Azure Virtual Network allows you to create a secure, persistent, network containing the resources you need for your solution. A virtual network allows you to:

- Connect cloud resources together in a private network (cloud-only)
- Connect your cloud resources to your local datacenter network (site-to-site or point to site) using a Virtual Private Network (VPN)
    - Site-to-site configuration allows you to connect multiple resources from your data center to the Azure Virtual Network using a hardware VPN or the Routing and Remote Access Service
    - Point-to-site configuration allows you to connect a specific resource to the Azure Virtual Network using software VPN

You must create the Azure Virtual Network before provisioning an HDInsight cluster.

You can find a detailed overview on virtual networking here:

http://msdn.microsoft.com/en-us/library/azure/jj156007.aspx

© Copyright 2014 Avanade Inc. All Rights Reserved.

Using Azure Management Portal

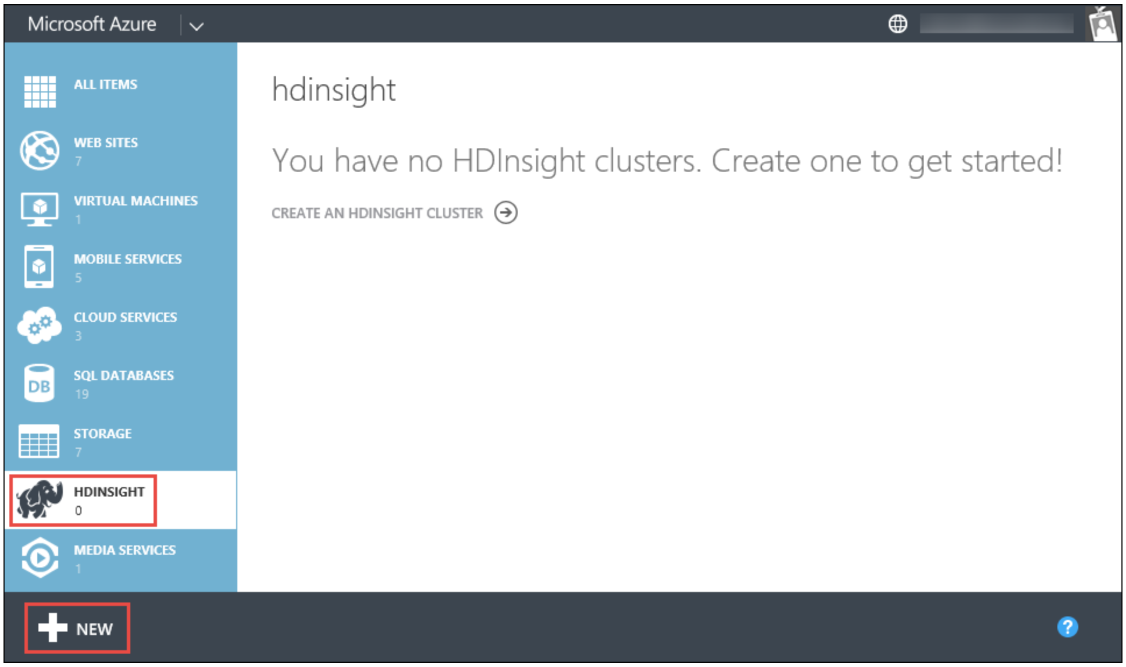
You can use the management portal to configure and control your Windows Azure services and applications.

https://manage.windowsazure.com/

The following step-by-step guide will give you an overview on how to create a cluster out-of-the-box by using the Quick create option and how to create a custom cluster by defining various different parameters.

Provisioning a Quick Cluster

When you provision an HDInsight cluster, you provision an Azure compute resource that contains Hadoop and the related applications. In this section you provision a HDInsight cluster of version 3.1, which is based on Hadoop version 2.4. You can also create Hadoop clusters for other versions using the Azure portal, HDInsight PowerShell cmdlets, or the HDInsight .NET SDK.



hdinsight

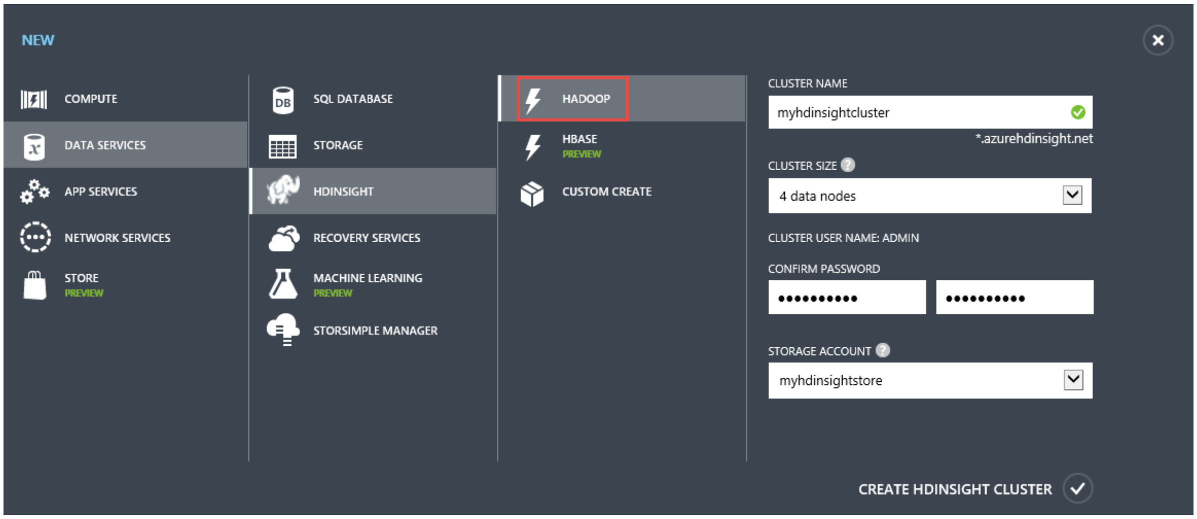
You have no HDInsight clusters. Create one to get started!

CREATE AN HDINSIGHT CLUSTER

Click HDInsight on the left to list the status of the clusters in your account. In the screenshot, there are no existing HDInsight clusters.

Click NEW on the lower left side.

© Copyright 2014 Avanade Inc. All Rights Reserved.



CLUSTER NAME: myhdinsightcluster *.azurehdinsight.net

CLUSTER SIZE: 4 data nodes

CLUSTER USER NAME: ADMIN

CONFIRM PASSWORD: •••••••••• ••••••••••

STORAGE ACCOUNT: myhdinsightstore

Click Data Services, click HDInsight, and then click Hadoop.

Enter or select the following values:

Cluster Name: Name of the cluster

Cluster Size: Number of data nodes you want to deploy. The default value is 4. But the option to use 1 or 2 data nodes is also available from the drop-down. Any number of cluster nodes can be specified by using the Custom Create option. Pricing details on the billing rates for various cluster sizes are available. Click the ? symbol just above the dropdown box and follow the link on the pop up.

Password: The password for the *admin* account. The cluster user name "admin" is specified when you are not using the Custom Create option. Note that this is NOT the Windows Administrator account for the VMs on which the clusters are provisioned. The account name can be changed by using the Custom Create wizard.

Storage Account: If you have already created a storage account, you can select this storage account from the dropdown box. Once a storage account is chosen, it cannot be changed. If the storage account is removed, the cluster will no longer be available for use. The HDInsight cluster is co-located in the same datacenter as the storage account.

NOTE: You will find detailed information on how to create a new storage account in a below chapter.

Configuration Options

Additional storage

© Copyright 2014 Avanade Inc. All Rights Reserved.

During configuration, you must specify an Azure Blob Storage account, and a default container. This is used as the default storage location by the cluster. Optionally, you can specify additional blobs that will also be associated with your cluster.

Metastore

The Metastore contains information about Hive tables, partitions, schemas, columns, etc. This information is used by Hive to locate where data is stored on HDFS (or WASB for HDInsight.) By default, Hive uses an embedded database to store this information.

When provisioning an HDInsight cluster, you can specify a SQL Database that will contain the Metastore for Hive. This allows the metadata information to be preserved when you delete a cluster, as it is stored externally in SQL Database.

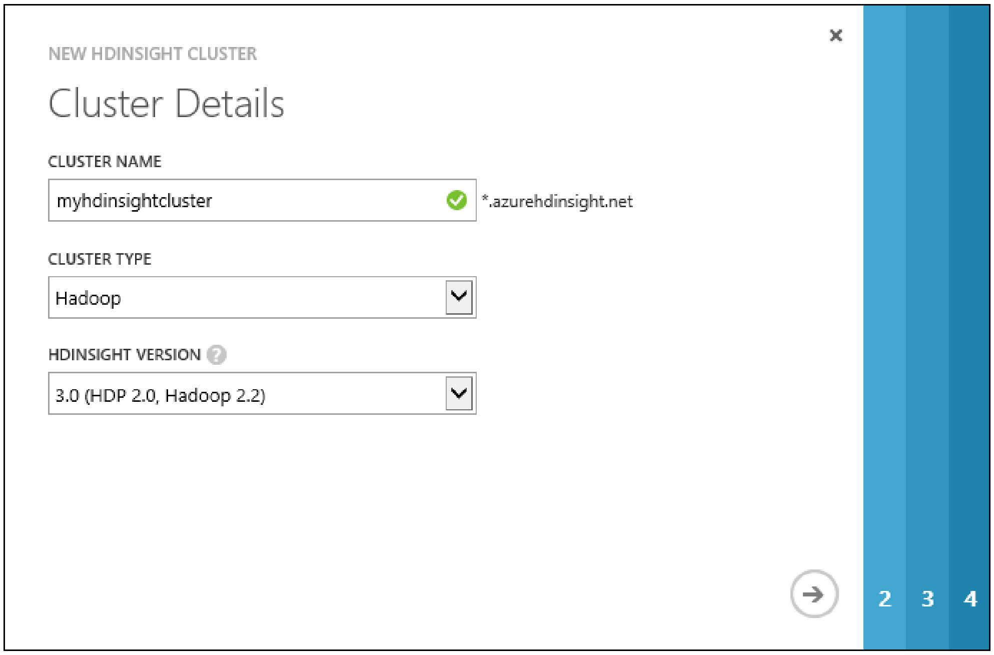
Provisioning a custom cluster

HDInsight clusters use an Azure Blob Storage container as the default file system. An Azure storage account located on the same data center is required before you can create a HDInsight cluster.

To create a cluster, you have to sign in to the Azure Management Portal.

Click + NEW on the bottom of the page, click DATA SERVICES, click HDINSIGHT, and then click CUSTOM CREATE.

On the Cluster Details page, type or choose the values shown in the figure.



NEW HDINSIGHT CLUSTER

Cluster Details

CLUSTER NAME

myhdinsightcluster ✓ *.azurehdinsight.net

CLUSTER TYPE

Hadoop

HDINSIGHT VERSION

3.0 (HDP 2.0, Hadoop 2.2)

© Copyright 2014 Avanade Inc. All Rights Reserved.

Fill in the following information:

Cluster name: Name the cluster.

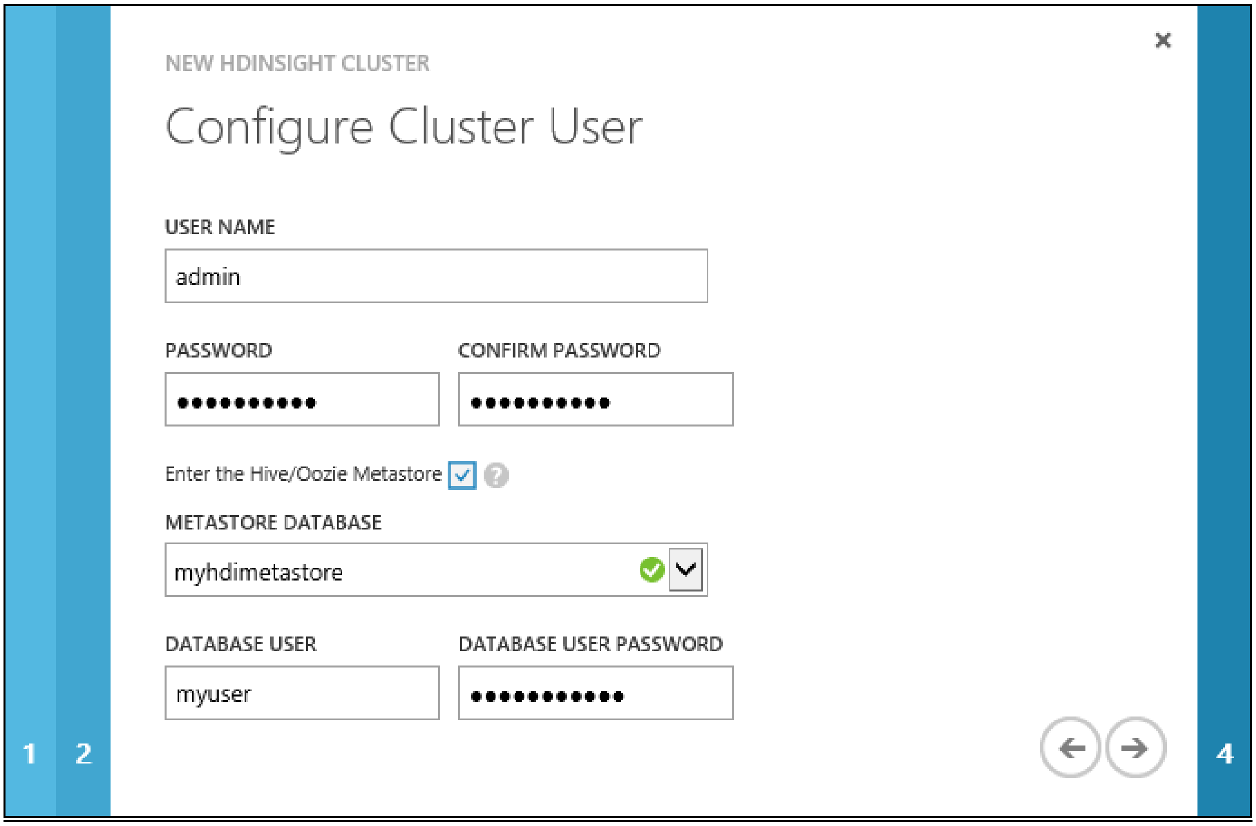
- DNS name must start and end with alpha numeric, may contain dashes.
- The field must be a string between 3 to 63 characters.

Cluster Type: For cluster type, select **Hadoop**.

HDInsight version: Choose the version. For Hadoop, the default is HDInsight version 3.1, which uses Hadoop 2.4.

Configure Cluster User

Furthermore you have to specify the credentials and a database that can be used to storage metadata information for project such as Hive or Oozie:



NEW HDINSIGHT CLUSTER

Configure Cluster User

USER NAME: admin

PASSWORD: ●●●●●●●●●●

CONFIRM PASSWORD: ●●●●●●●●●●

Enter the Hive/Oozie Metastore ☑

METASTORE DATABASE: myhdimetastore

DATABASE USER: myuser

DATABASE USER PASSWORD: ●●●●●●●●●●●

User name: Specify the HDInsight cluster user name.

Password/Confirm Password: Specify the HDInsight cluster user password.

© Copyright 2014 Avanade Inc. All Rights Reserved.

Enter Hive/Oozie Metastore: Select this checkbox to specify a SQL database on the same data center as the cluster, to be used as the Hive/Oozie metastore. This is useful if you want to retain the metadata about Hive/Oozie jobs even after a cluster has been deleted.

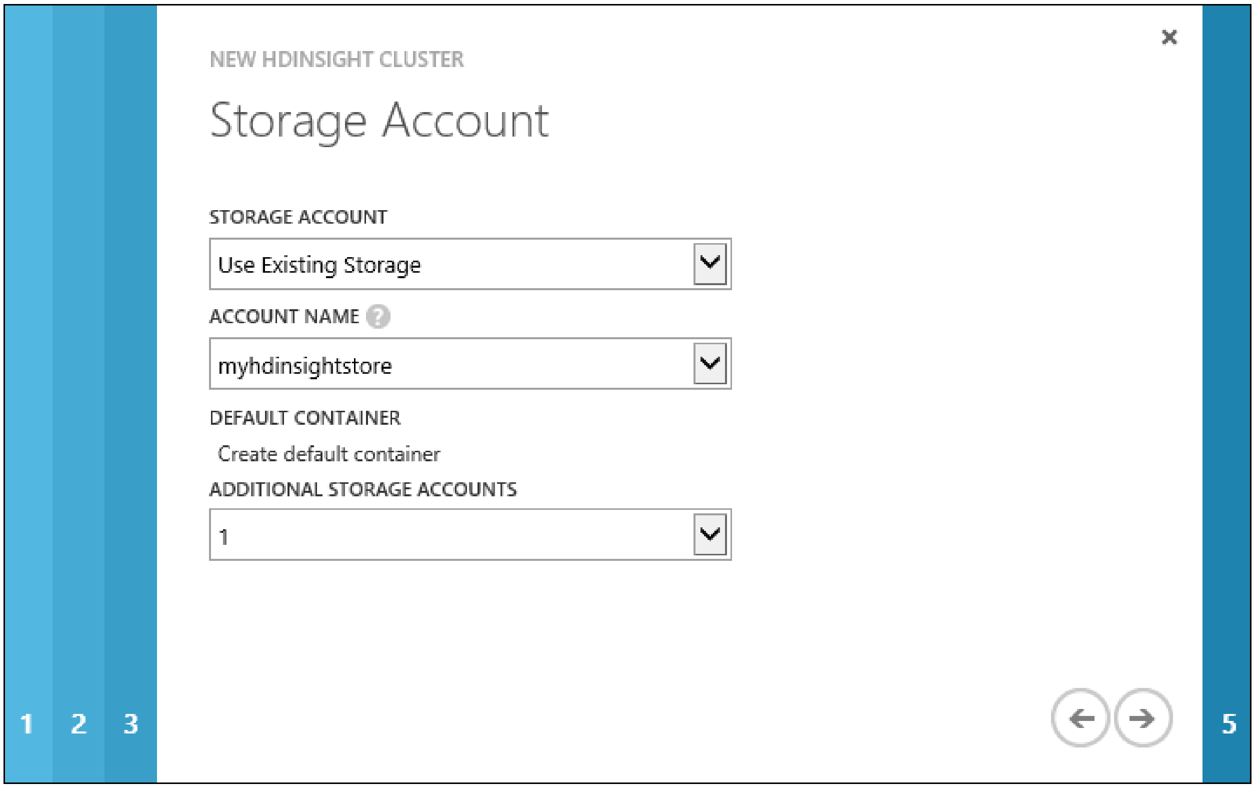
Metastore Database: Specify the Azure SQL database that will be used as the metastore for Hive/Oozie. This SQL database must be in the same data center as the HDInsight cluster. The list box only lists the SQL databases in the same data center as you specified on the Cluster Details page.

Database user: Specify the SQL database user that will be used to connect to the database.

Database user password: Specify the SQL database user password.

NOTE: The Azure SQL database used for the metastore must allow connectivity to other Azure services, including Azure HDInsight. On the Azure SQL database dashboard, on the right side click the server name. This is the server on which the SQL database instance is running. Once you are on the server view, click Configure, and then for Windows Azure Services, click Yes, and then click Save.

Storage Account


NEW HDINSIGHT CLUSTER

Storage Account

STORAGE ACCOUNT: Use Existing Storage

ACCOUNT NAME: myhdinsightstore

DEFAULT CONTAINER: Create default container

ADDITIONAL STORAGE ACCOUNTS: 1

Storage Account: Specify the Azure storage account that will be used as the default file system for the HDInsight cluster. You can choose one of the three options:

© Copyright 2014 Avanade Inc. All Rights Reserved.

- Use Existing Storage
- Create New Storage
- Use Storage From Another Subscription

Account Name:

- If you chose to use existing storage, for **Account name**, select an exising storage account. The drop-down only lists the storage accounts located in the same data center where you chose to provision the cluster.
- If you chose **Create new storage** or **Use storage from another subscription** option, you must provide the storage account name.

Account Key: If you chose the **Use Storage From Another Subscription** option, specify the account key for that storage account.

Default container: Specifies the default container on the storage account that is used as the default file system for the HDInsight cluster. If you chose **Use Existing Storage** for the **Storage Account** field, and there are no existing containers in that account, the container is created by default with the same name as the cluster name. If a container with the name of the cluster already exists, a sequence number will be appended to the container name. For example, mycontainer1, mycontainer2, and so on. However, if the existing storage account has a container with a name different from the cluster name you specified, you can use that container as well.

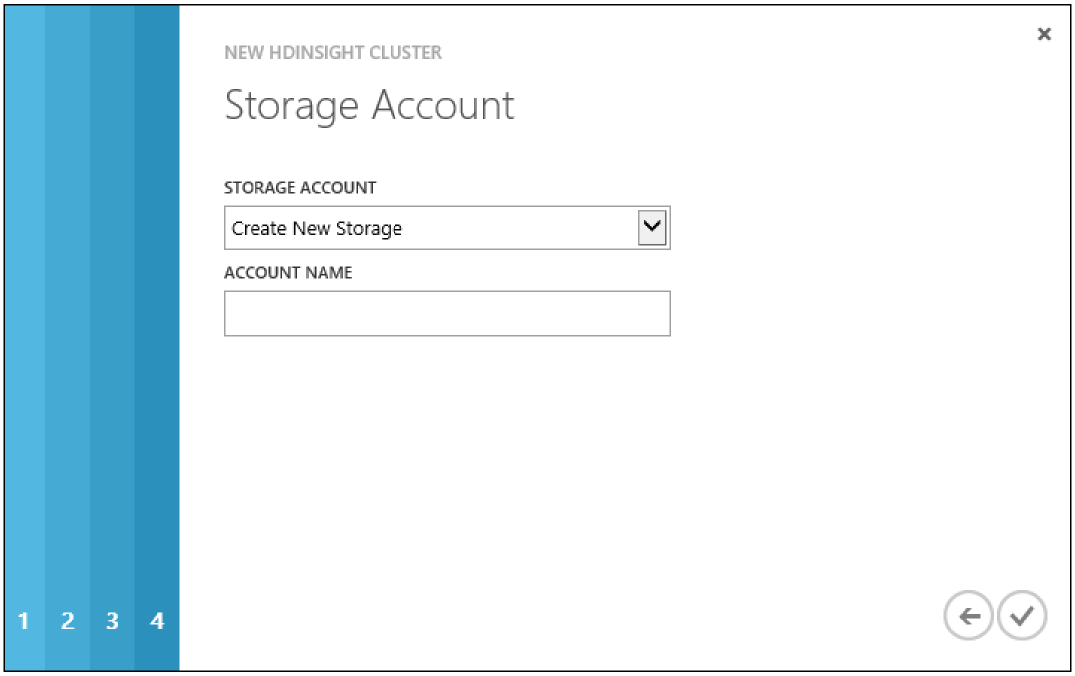
If you chose to create a new storage or use storage from another Azure subscription, you must specify the default container name

Additional Storage Accounts: HDInsight supports multiple storage accounts. There is no limit on the additional storage account that can be used by a cluster. However, if you create a cluster using the Management Portal, you have a limit of seven due to the UI constraints. Each additional storage account you specify adds an extra Storage Account page to the wizard where you can specify the account information. For example, in the screenshot here, one additional storage account is selected, and hence page 5 is added to the dialog.

Additional Storage Account

If required by your environment, you can specifiy an additional storage account. In this example no additional storage account is needed.

© Copyright 2014 Avanade Inc. All Rights Reserved.



STORAGE ACCOUNT: Create New Storage

ACCOUNT NAME:

On the **Storage Account** page, enter the account information for the additional storage account.

Here again, you have the option to choose from existing storage, create new storage, or use storage from another Azure subscription. The procedure to provide the values is similar to the previous step.

Click the check mark to start provisioning the cluster. When the provisioning completes, the status column shows **Running**.

NOTE:

Once an Azure storage account is chosen for your HDInsight cluster, you can neither delete the account, nor change the account to a different account.

© Copyright 2014 Avanade Inc. All Rights Reserved.

Using Azure PowerShell

Azure PowerShell is a powerful scripting environment that you can use to control and automate the deployment and management of your workloads in Azure. This section provides instructions on how to provision an HDInsight cluster using PowerShell.

Creating a Storage account

The following procedures are needed to provision an HDInsight cluster using PowerShell:

- Create an Azure Storage account
- Create an Azure Blob container
- Create a HDInsight cluster

HDInsight uses an Azure Blob Storage container as the default file system. An Azure storage account and storage container are required before you can create an HDInsight cluster. The storage account must be located in the same data center as the HDInsight Cluster.

```
$storageAccountName = "mystorage"   # Provide a storage account name
$location = "Southeast Asia"        # For example, "West US"

# Create an Azure storage account
New-AzureStorageAccount -StorageAccountName $storageAccountName -
Location $location
```

To retrieve account name and key:

```
# List storage accounts for the current subscription
Get-AzureStorageAccount

# List the keys for a storage account
Get-AzureStorageKey "mystorage"
```

Create Azure Blob storage container

```
$storageAccountName = "mystorage"
$storageAccountKey = "<StorageAccountKey>"
$containerName="mycontainer"

# Create a storage context object
$destContext = New-AzureStorageContext
     -StorageAccountName $storageAccountName
     -StorageAccountKey $storageAccountKey

# Create a Blob storage container
New-AzureStorageContainer -Name $containerName -Context $destContext
```

© Copyright 2014 Avanade Inc. All Rights Reserved.

Provision an HDInsight cluster

After you have successfully created the Storage account, you are now ready to start the provisioning of the cluster:

```
$subscriptionName = "<SubscriptionName>"
$storageAccountName = "mystorage"
$containerName = "mycontainer"
$clusterName = "mycluster"
$location = "Southeast Asia“
$clusterNodes = 4

Select-AzureSubscription $subscriptionName

$storageAccountKey = Get-AzureStorageKey $storageAccountName | %{
$_.Primary }

New-AzureHDInsightCluster
     -Name $clusterName `
     -Location $location `
     -DefaultStorageAccountName
"$storageAccountName.blob.core.windows.net" `
     -DefaultStorageAccountKey $storageAccountKey `
     -DefaultStorageContainerName $containerName  `
     -ClusterSizeInNodes $clusterNodes
```

When prompted, enter the credentials for the cluster. It can take several minutes before the cluster provision completes.

NOTE: The PowerShell cmdlets are the only recommended way to change configuration variables in an HDInsight cluster. Changes made to Hadoop configuration files while connected to the cluster via Remote Desktop may be overwritten in the event of cluster patching. Configuration values set via PowerShell will be preserved if the cluster is patched.

While provisioning a cluster, you can use the other configuration options such as connecting to more than one Azure Blob storage, using a Virtual Network, or using an Azure SQL database for Hive and Oozie metastores. This allows you to separate lifetime of your data and metadata from the lifetime of the cluster.

```
$subscriptionName = "mysubscription"
$clusterName = "mycluster"
$location = "Southeast Asia"
$clusterNodes = 4
$storageAccountName_Default = "mystorage1"
$containerName_Default = "mycontainer"
$storageAccountName_Add1 = "mystorage2"
$hiveSQLDatabaseServerName = "SQLDatabaseServerNameForHiveMetastore"
$hiveSQLDatabaseName = "SQLDatabaseDatabaseNameForHiveMetastore"
$oozieSQLDatabaseServerName = "SQLDatabaseServerNameForOozieMetastore"
```

© Copyright 2014 Avanade Inc. All Rights Reserved.

```
$oozieSQLDatabaseName = "SQLDatabaseDatabaseNameForOozieMetastore"

# Get the virtual network ID and subnet name
$vnetID = "<AzureVirtualNetworkID>"
$subNetName = "<AzureVirtualNetworkSubNetName>"


# Get the storage account keys
Select-AzureSubscription $subscriptionName

$storageAccountKey_Default = Get-AzureStorageKey
$storageAccountName_Default | %{ $_.Primary }

$storageAccountKey_Add1 = Get-AzureStorageKey $storageAccountName_Add1
| %{ $_.Primary }

$oozieCreds = Get-Credential -Message "Oozie metastore"
$hiveCreds = Get-Credential -Message "Hive metastore"

# Create a Blob storage container
$dest1Context = New-AzureStorageContext
     -StorageAccountName $storageAccountName_Default `
     -StorageAccountKey $storageAccountKey_Default

New-AzureStorageContainer -Name $containerName_Default `
                          -Context $dest1Context


$config =
New-AzureHDInsightClusterConfig -ClusterSizeInNodes $clusterNodes |
          Set-AzureHDInsightDefaultStorage
     -StorageAccountName
"$storageAccountName_Default.blob.core.windows.net" `
     -StorageAccountKey $storageAccountKey_Default `
     -StorageContainerName $containerName_Default |
    Add-AzureHDInsightStorage
     -StorageAccountName
"$storageAccountName_Add1.blob.core.windows.net" `
     -StorageAccountKey $storageAccountKey_Add1 |
    Add-AzureHDInsightMetastore -SqlAzureServerName
"$hiveSQLDatabaseServerName.database.windows.net" `
                                -DatabaseName $hiveSQLDatabaseName `
                                -Credential $hiveCreds `
                                -MetastoreType HiveMetastore |
    Add-AzureHDInsightMetastore -SqlAzureServerName
"$oozieSQLDatabaseServerName.database.windows.net" `
                                -DatabaseName $oozieSQLDatabaseName `
                                -Credential $oozieCreds `
                                -MetastoreType OozieMetastore |
    New-AzureHDInsightCluster -Name $clusterName `
                              -Location $location `
                              -VirtualNetworkId $vnetID `
                              -SubnetName $subNetName
```

© Copyright 2014 Avanade Inc. All Rights Reserved.

Listing HDInsight clusters

After you have successfully created the HDInsight cluster, you can run the command from an Azure PowerShell console window to list the HDInsight cluster and verify that the cluster was successfully provisioned.

```
Get-AzureHDInsightCluster -Name "mycluster"
```

© Copyright 2014 Avanade Inc. All Rights Reserved.

Using Cross-platform Command Line

Another option for provisioning an HDInsight cluster is the Cross-platform Command-line Interface. The command-line tool is implemented in Node.js. It can be used on any platform that supports Node.js including Windows, Mac and Linux. The command-line tool is open source.

The following procedures are needed to provision an HDInsight cluster using Cross-platform command line:

- Install cross-platform command line
- Download and import Azure account publish settings
- Create an Azure Storage account
- Provision a cluster

The command-line interface can be installed using *Node.js Package Manager (NPM)* or Windows Installer. Microsoft recommends that you install using only one of the two options.

Install using NPM

1. Browse to www.nodejs.org.
2. Click INSTALL and following the instructions using the default settings.
3. Open Command Prompt (or *Azure Command Prompt*, or *Developer Command Prompt for VS2012*) from your workstation.
4. Run the following command in the command prompt window.

```
npm install -g azure-cli
```

NOTE:

If you get an error saying the NPM command is not found, verify the following paths are in the PATH environment variable: *C:\Program Files (x86)\nodejs;C:\Users[username]\AppData\Roaming\npm* or *C:\Program Files\nodejs;C:\Users[username]\AppData\Roaming\npm*

5. Run the following command to verify the installation:

```
azure hdinsight -h
```

© Copyright 2014 Avanade Inc. All Rights Reserved.

You can use the *-h* switch at different levels to display the help information. For example:

```
azure –h

azure hdinsight -h

azure hdinsight cluster -h

azure hdinsight cluster create -h
```

Install using windows installer

1. Browse to http://azure.microsoft.com/en-us/downloads/.

2. Scroll down to the Command line tools section, and then click Windows install under Azure command-line interface and follow the Web Platform Installer wizard

Publish Settings

Before using the command-line interface, you must configure connectivity between your workstation and Azure. Your Azure subscription information is used by the command-line interface to connect to your account. This information can be obtained from Azure in a publish settings file. The publish settings file can then be imported as a persistent local config setting that the command-line interface will use for subsequent operations. You only need to import your publish settings once.

NOTE:

The publish settings file contains sensitive information. Microsoft recommends that you delete the file or take additional steps to encrypt the user folder that contains the file. On Windows, modify the folder properties or use BitLocker.

1. Open a Command Prompt.
2. Run the following command to download the publish settings file.

   ```
   azure account download
   ```

The command launches the Web page to download the publish settings file from.

1. At the prompt to save the file, click Save and provide a location where the file must be saved.
2. From the command prompt window, run the following command to import the publish settings file:

© Copyright 2014 Avanade Inc. All Rights Reserved.

```
    azure account import <file>
```

Create storage account

HDInsight uses an Azure Blob Storage container as the default file system. An Azure storage account is required before you can create an HDInsight cluster. The storage account must be located in the same data center.

From the command prompt window, run the following command to create an Azure storage account:

```
azure storage account create -l "Southeast Asia" mystorage
```

When prompted for a location, select a location where an HDINsight cluster can be provisioned. The storage must be in the same location as the HDInsight cluster. Currently, only the East Asia, Southeast Asia, North Europe, West Europe, East US, West US, North Central US, and South Central US regions can host HDInsight clusters.

If you already have a storage account but do not know the account name and account key, you can use the following commands to retrieve the information:

```
-- lists storage accounts
azure storage account list

-- Shows information for a storage account
azure storage account show <StorageAccountName>

-- Lists the keys for a storage account
azure storage account keys list <StorageAccountName>
```

An HDInsight cluster also requires a container within a storage account. If the storage account you provide does not already have a container, the *azure hdinsight cluster create* prompts you for a container name and creates it as well. However, if you want to create the container beforehand, you can use the following command:

```
azure storage container create -a mystorage -k <StorageAccountKey>
mycontainer
```

Once you have the storage account and the blob container prepared, you are ready to create a cluster.

Provision a cluster

From the command prompt window, run the following command:

© Copyright 2014 Avanade Inc. All Rights Reserved.

```
azure hdinsight cluster create `
      --clusterName "mycluster" `
      --storageAccountName "mystorage.blob.core.windows.net" `
      --storageAccountKey <storageAccountKey> `
      --storageContainer "mycontainer" `
      --nodes 1 `
      --location "Southeast Asia" `
      --username <HDInsightClusterUsername> `
      --clusterPassword <HDInsightClusterPassword>
```

Provision a cluster using config file

Typically, you provision an HDInsight cluster, run the jobs, and then delete the cluster to cut down the cost. The command-line interface gives you the option to save the configurations into a file, so that you can reuse it every time you provision a cluster.

From the command prompt window, run the following commands:

```
$file = "C:\config.txt"
$storage = "mystorage.blob.core.windows.net"
$key = "<StorageAccountKey>“

#Create the config file
azure hdinsight cluster config create $file

#Add commands to create a basic cluster
azure hdinsight cluster config set $file
      --clusterName "mycluster“
      --nodes 1
      --location "Southeast Asia" `
      --storageAccountName $storage `
      --storageAccountKey $key `
      --storageContainer "mycontainer“
      --username "<Username>"
      --clusterPassword "<UserPassword>"

#If requred, include commands to use additional blob storage with the
cluster
azure hdinsight cluster config storage add $file --storageAccountName
"<StorageAccountName>.blob.core.windows.net"
       --storageAccountKey "<StorageAccountKey>"


#If required, include commands to use a SQL database as a Hive
metastore
azure hdinsight cluster config metastore set $file --type "hive" --
server "<SQLDatabaseName>.database.windows.net"
       --database "<HiveDatabaseName>" --user "<Username>" --
metastorePassword "<UserPassword>"

#If required, include commands to use a SQL database as an Oozie
metastore
azure hdinsight cluster config metastore set <file> --type "oozie" --
server "<SQLDatabaseName>.database.windows.net"
```

```
       --database "<OozieDatabaseName>" --user "<SQLUsername>" --
metastorePassword "<SQLPassword>"

#Run this command to create a cluster using the config file
azure hdinsight cluster create --config $file
```

To list and show cluster details

Use the following commands to list and show cluster details:

```
azure hdinsight cluster list
azure hdinsight cluster show "<ClusterName>"
```

To delete a cluster

Use the following command to delete a cluster:

```
azure hdinsight cluster delete "<ClusterName>"
```

© Copyright 2014 Avanade Inc. All Rights Reserved.

Using HDInsight .NET SDK

The HDInsight .NET SDK provides .NET client libraries that makes it easier to work with HDInsight from a .NET application.

The following procedures must be performed to provision an HDInsight cluster using the SDK:

- Install HDInsight .NET SDK
- Create a self-signed certificate
- Create a console application
- Run the application

Install HDInsight .NET SDK

You can install latest published build of the SDK from NuGet. The instructions will be shown in the next procedure.

Create a self-signed certificate

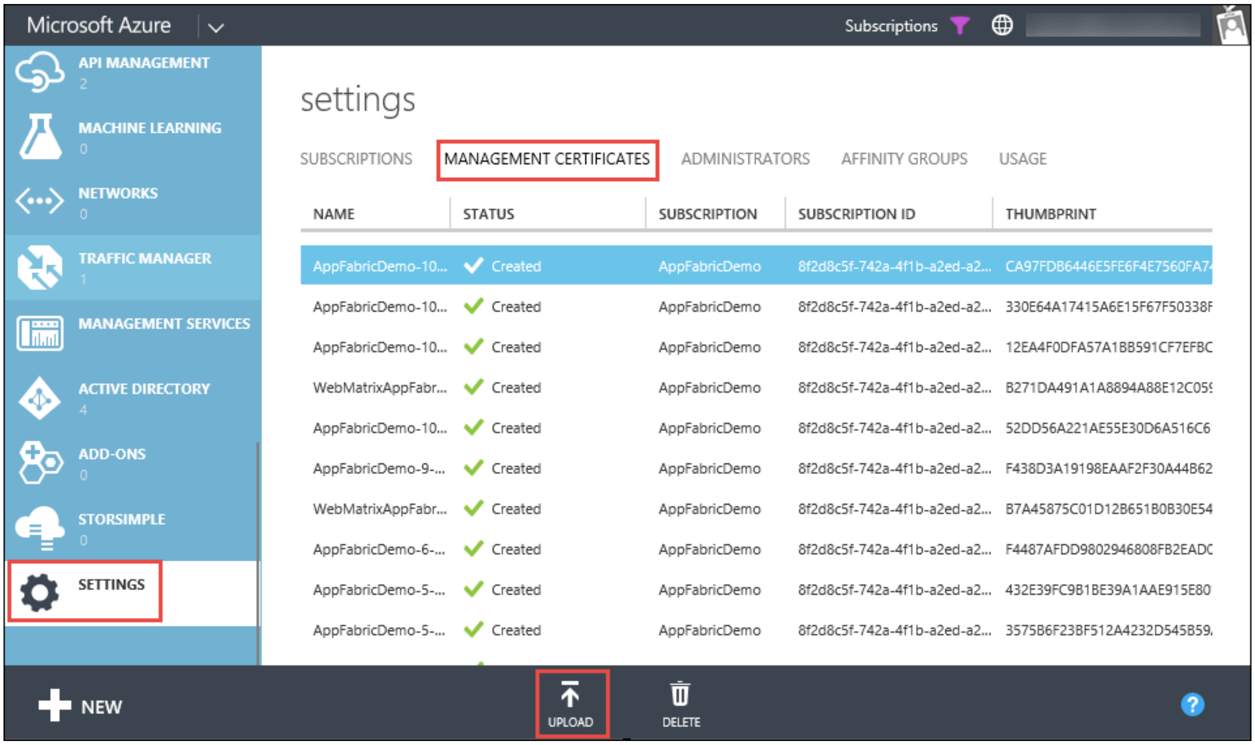
Create a self-signed certificate, install it on your workstation, and upload it to your Azure subscription.

1. Create a self-signed certificate that is used to authenticate the requests. You can use IIS or makecert to create the certificate.

2. Browse to the location of the certificate, right-click the certificate, click Install Certificate, and install the certificate to the computer's personal store. Edit the certificate properties to assign it a friendly name.

3. Import the certificate into the Azure Management Portal. From the portal, click Settings on the bottom-left of the page, and then click Management Certificates. From the bottom of the page, click Upload and follow the instructions to upload the .cer file you created in the previous step.

© Copyright 2014 Avanade Inc. All Rights Reserved.



settings

SUBSCRIPTIONS | MANAGEMENT CERTIFICATES | ADMINISTRATORS | AFFINITY GROUPS | USAGE

NAME	STATUS	SUBSCRIPTION	SUBSCRIPTION ID	THUMBPRINT
AppFabricDemo-10...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	CA97FDB6446E5FE6F4E7560FA7
AppFabricDemo-10...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	330E64A17415A6E15F67F50338F
AppFabricDemo-10...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	12EA4F0DFA57A1BB591CF7EFBC
WebMatrixAppFabr...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	B271DA491A1A8894A88E12C05
AppFabricDemo-10...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	52DD56A221AE55E30D6A516C6
AppFabricDemo-9-...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	F438D3A19198EAAF2F30A44B62
WebMatrixAppFabr...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	B7A45875C01D12B651B0B30E54
AppFabricDemo-6-...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	F4487AFDD9802946808FB2EADC
AppFabricDemo-5-...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	432E39FC9B1BE39A1AAE915E80
AppFabricDemo-5-...	✓ Created	AppFabricDemo	8f2d8c5f-742a-4f1b-a2ed-a2...	3575B6F23BF512A4232D545B59

Create a console application

1. Open Visual Studio 2013.

2. From the File menu, click New, and then click Project.

3. From New Project, type or select the following values:

PROPERTY	VALUE
Category	Templates/Visual C#/Windows
Template	Console Application
Name	CreateHDICluster

4. Click OK to create the project.

5. From the Tools menu, click Library Package Manager, and then click Package Manager Console.

© Copyright 2014 Avanade Inc. All Rights Reserved.

6. Run the following commands in the console to install the packages.

```
Install-Package Microsoft.WindowsAzure.Management.HDInsight
```

   This command adds .NET libraries and references to them to the current Visual Studio project.

7. From Solution Explorer, double-click Program.cs to open it.

8. Add the following using statements to the top of the file:

```
using Microsoft.WindowsAzure.Management.HDInsight;
using System.Security.Cryptography.X509Certificates;
```

9. In the Main() function, copy and paste the following code:

© Copyright 2014 Avanade Inc. All Rights Reserved.

```
string certfriendlyname = "<CertFriendlyName>";// Friendly name for the certificate created earlier
string subscriptionid = "<AzureSubscriptionID>";
string clustername = "<HDInsightClusterName>";
string location = "<MicrosoftDataCenter>";
string storageaccountname = "<AzureStorageAccountName>.blob.core.windows.net";
string storageaccountkey = "<AzureStorageAccountKey>";
string containername = "<HDInsightDefaultContainerName>";
string username = "<HDInsightUsername>";
string password = "<HDInsightUserPassword>";
int clustersize = <NumberOfNodesInTheCluster>;

// Get the certificate object from certificate store using the friendly name to identify it
X509Store store = new X509Store();
store.Open(OpenFlags.ReadOnly);

var certificates = store.Certificates.Cast<X509Certificate2>();
var cert = certificates.First(item => item.FriendlyName == certfriendlyname);

// Create the storage account if it doesn't exist.

// Create the container if it doesn't exist.

// Create an HDInsightClient object
var creds = new HDInsightCertificateCredential(new Guid(subscriptionid), cert);
var client = HDInsightClient.Connect(creds);

// Supply the cluster information
var clusterInfo = new ClusterCreateParameters
{
    Name = clustername,
    Location = location,
    DefaultStorageAccountName = storageaccountname,
    DefaultStorageAccountKey = storageaccountkey,
    DefaultStorageContainer = containername,
    UserName = username,
    Password = password,
    ClusterSizeInNodes = clustersize
};

// Create the cluster
Console.WriteLine("Creating the HDInsight cluster ...");

ClusterDetails cluster = client.CreateCluster(clusterInfo);

Console.WriteLine("Created cluster: {0}.", cluster.ConnectionUrl);
Console.WriteLine("Press ENTER to continue.");
Console.ReadKey();
```

10. Replace the variables at the beginning of the main() function.

© Copyright 2014 Avanade Inc. All Rights Reserved.

Run the application

While the application is open in Visual Studio, press F5 to run the application. A console window should open and display the status of the application. It can take several minutes to create a HDInsight cluster.

© Copyright 2014 Avanade Inc. All Rights Reserved.