Creating the Hadoop Cluster on Azure

Course 20773A, Analyzing Big Data with Microsoft R, requires that each student has access to a cluster running Hadoop and Microsoft R server. This cluster runs using HDInsight on Microsoft's Azure cloud platform. To save resources and each student you should create this cluster immediately before module 8, and then delete it again once the labs for module 8 are completed. **Do not leave the cluster running overnight.**

This guide describes the steps for creating the Hadoop cluster. The steps are correct as of the time of publishing.

As Azure is regularly updated and improved, there is a possibility that this guide may be out of date.

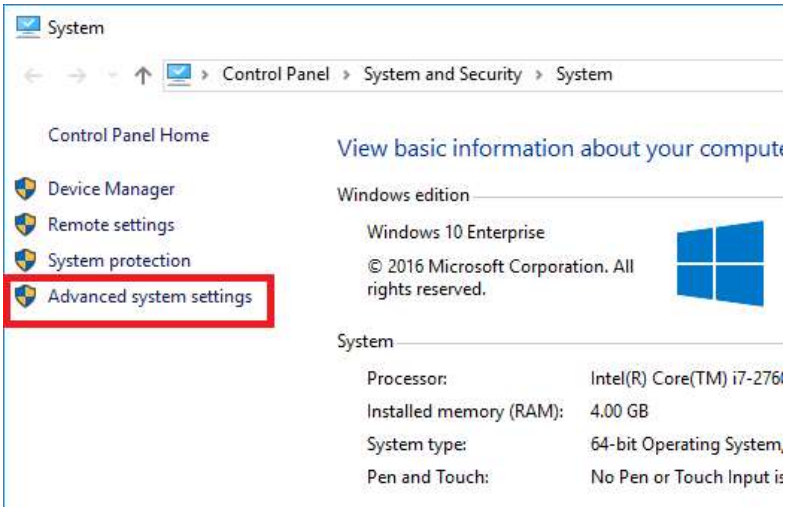
Before following the steps below, please follow the details of how to acquire a Microsoft Azure pass for you and your class here: **http://go.microsoft.com/fwlink/?LinkId=512034**

Install PuTTY

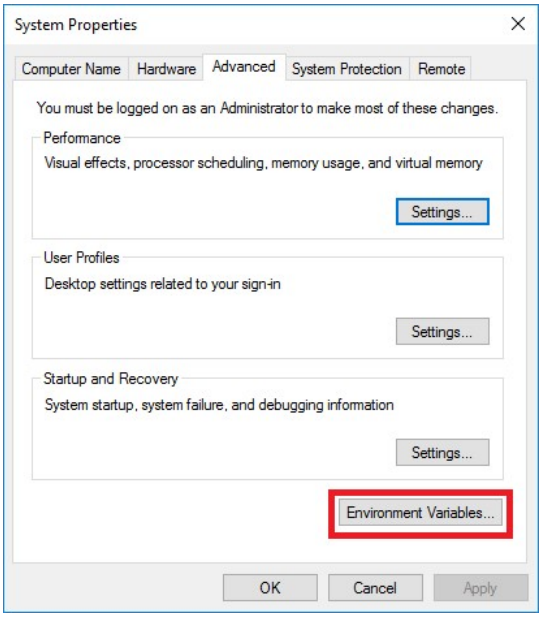
The Hadoop cluster runs Linux. You cannot connect to a Linux VM by using Remote Desktop without installing and configuring additional software, which can be a time-consuming process. Therefore this document uses SSH connections from an SSH client running on the Windows desktop. The simplest SSH client to install and use is PuTTY, a freely available open-source package. Follow these instructions to download and install PuTTY on the desktop machine.

1. In Internet Explorer, browse to https://the.earth.li/~sgtatham/putty/0.68/w64/putty-64bit-0.68-installer.msi.
2. In the Internet Explorer message box, click **Run**.
3. In the **PuTTY Setup** wizard, on the **Welcome** page, click **Next**.
4. On the **Destination Folder** page, click **Next**.
5. On the **Product Features** page, click **Install**.
6. In the **User Account Control** dialog box, click **Yes**.
7. When the wizard has completed, clear **View README file**, and then click **Finish**.
8. Right-click the Windows **Start** button, and then click **System**.

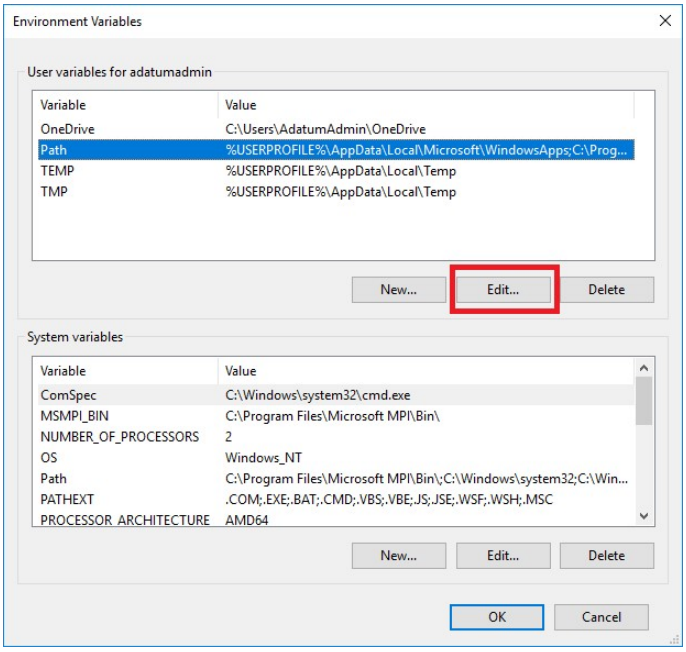
9. In the **System** dialog box, click **Advanced System Settings**.



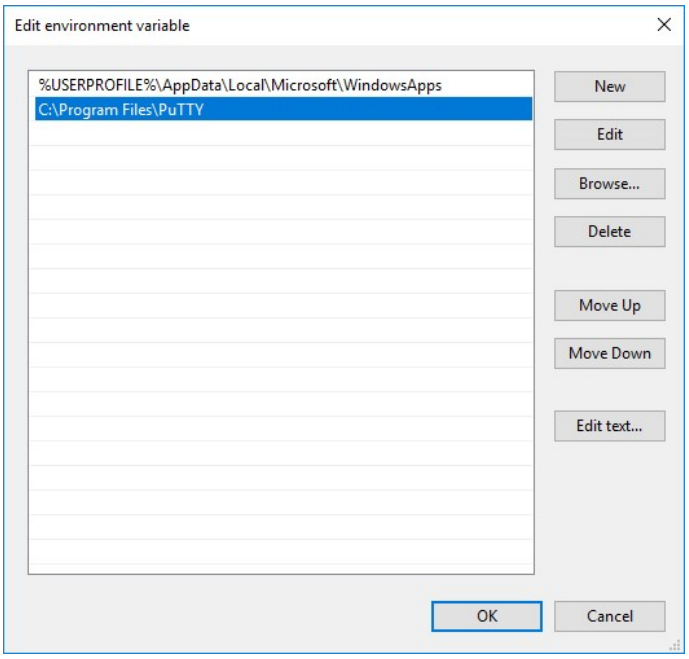
10. In the **System Properties** dialog box, click **Environment Variables**.



11. In the **Environment Variables** dialog box, click **Path**, and then click **Edit**.



12. In the **Edit User Variable** dialog box, append the path **C:\Program Files\PuTTY** to the **Variable value**, and then click **OK**.



13. In the **Environment Variables** dialog box, click **OK**.
14. In the **System Properties** dialog box, click **OK**.
15. Close the **System** dialog box.

Log into Azure

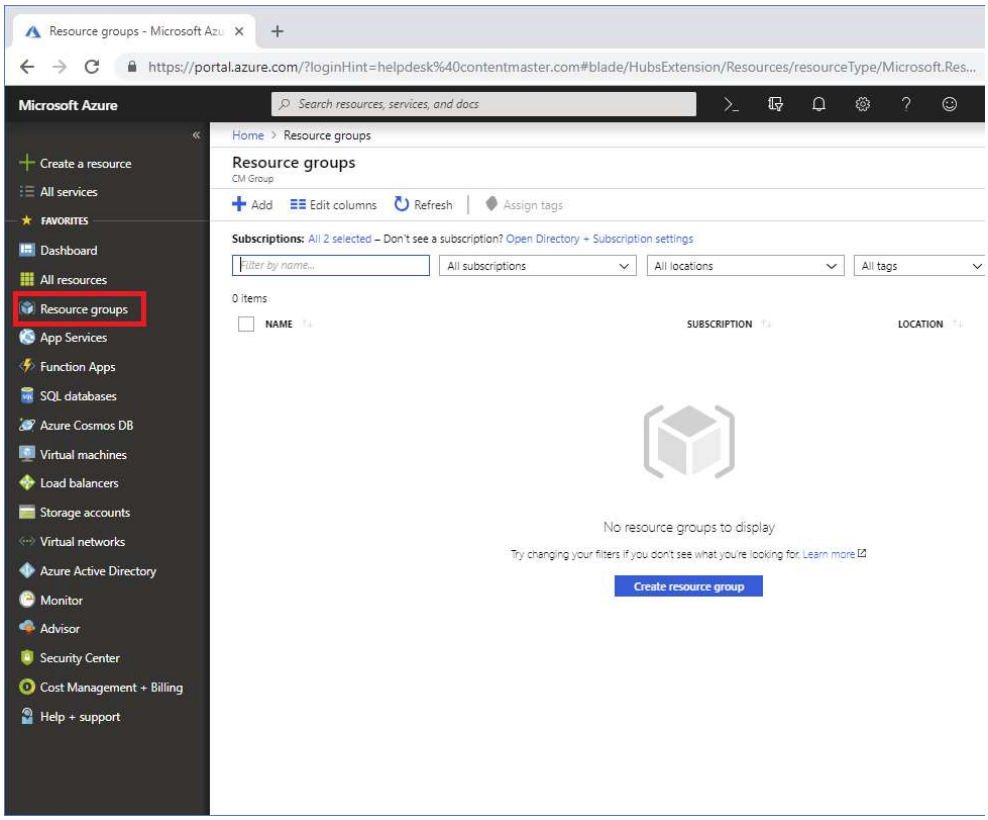
1. You will require a Microsoft account to login to the Azure Portal. The following steps assume you have already created these credentials.
2. On the Start menu, type **Internet Explorer**, and then click **Internet Explorer**.
3. In the address bar, type **portal.azure.com**, and then press Enter.
4. Enter your Microsoft account credentials to log in.

Create the Resource Group for the Virtual Machine

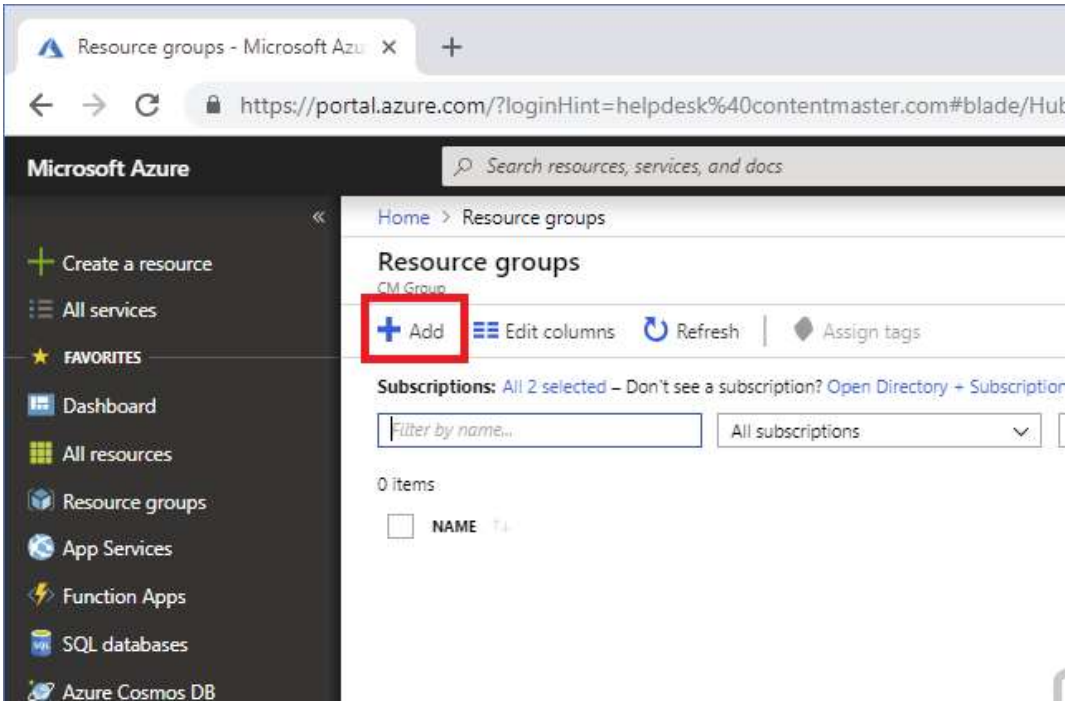
You create the VM and its resources in the same resource group. This helps to make management easier.

The following steps create the resource group.

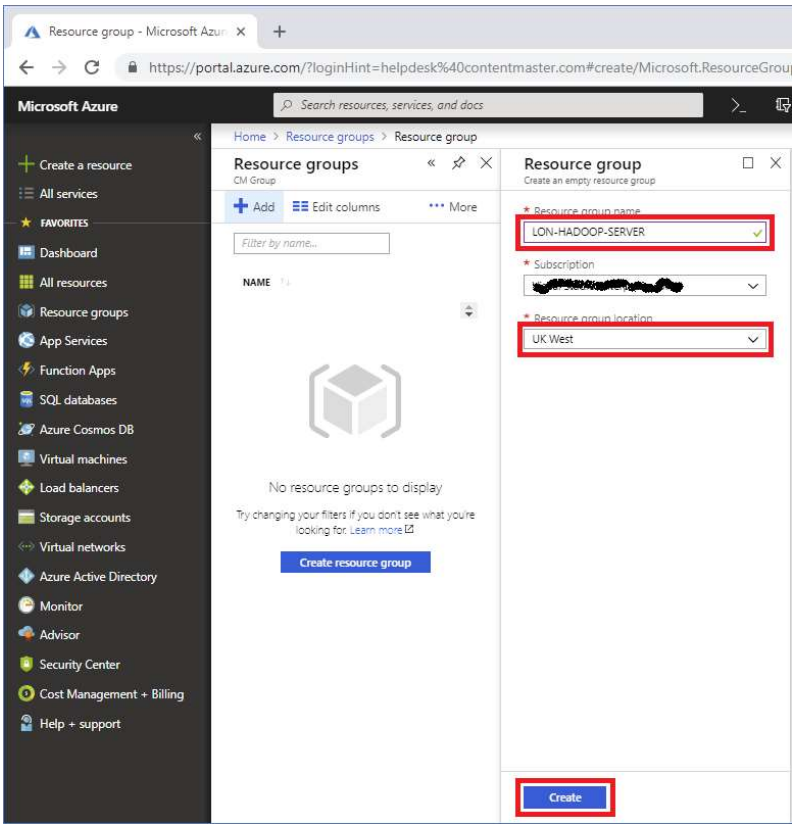
1. On the navigation blade on the left side of the portal, click **Resource groups**.



2. On the **Resource groups** blade, click **Add**.

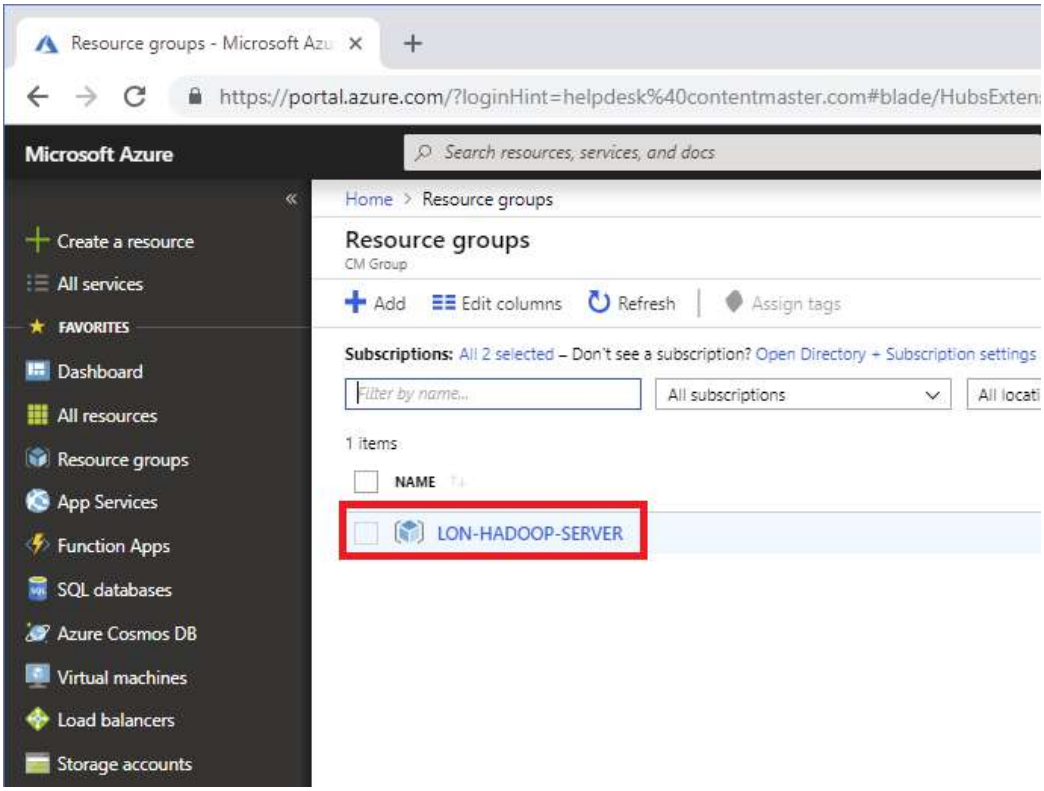


3. On the **Resource group** blade, in the **Resource group name** box, type **LON-HADOOP-SERVER**, select your nearest location, and then click **Create**.

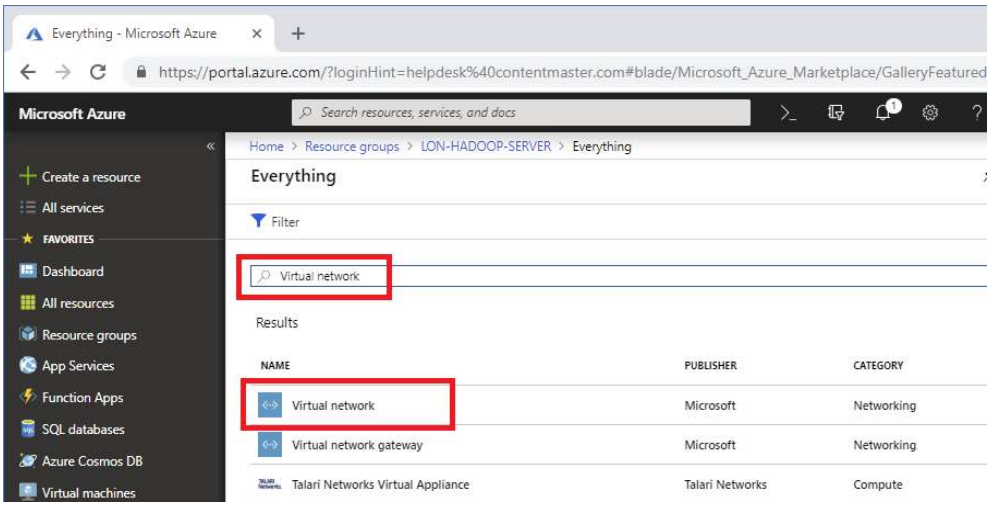


Create a VNet for the LON-HADOOP Cluster

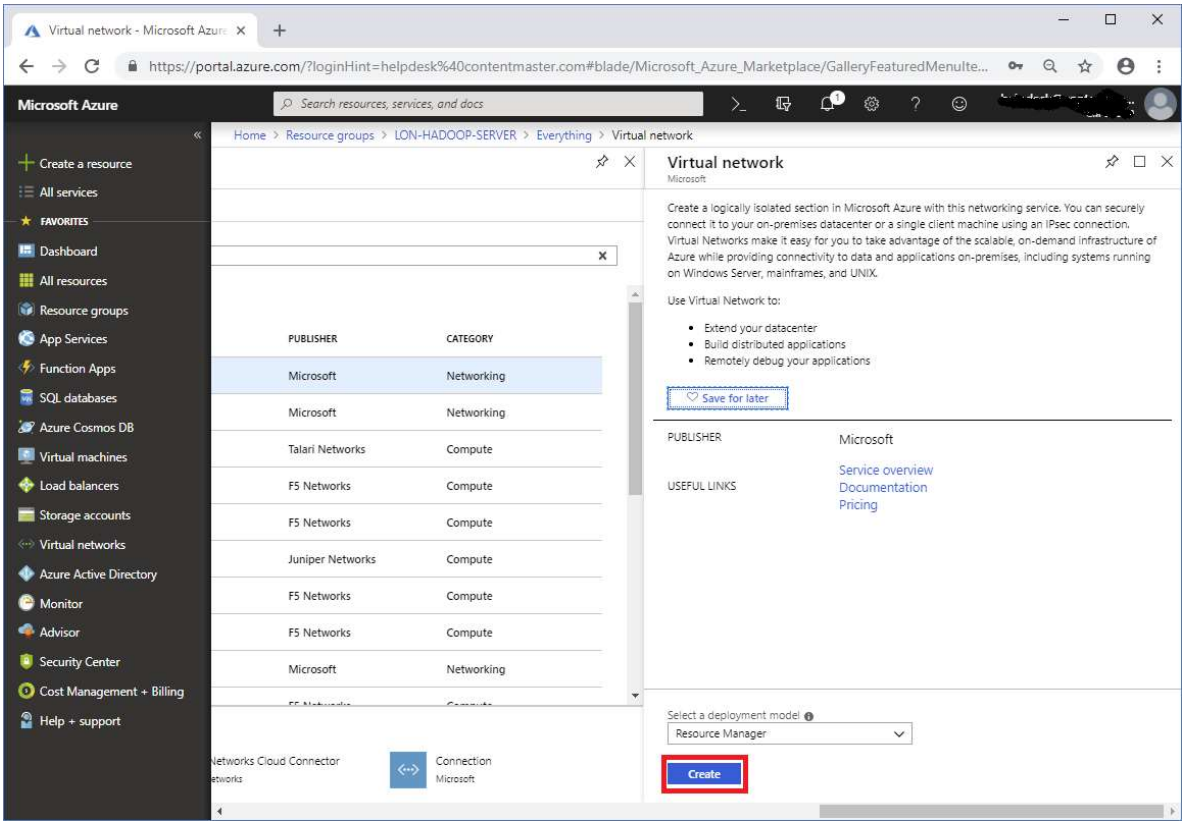
1. On the navigation blade on the left side of the portal, click **Resource groups**.
2. Click the **LON-HADOOP-SERVER** resource group.



3. On the **LON-HADOOP-SERVER** blade, click **Add**.
4. In the search box, type **Virtual network**, and then press Enter.
5. Click **Virtual network**

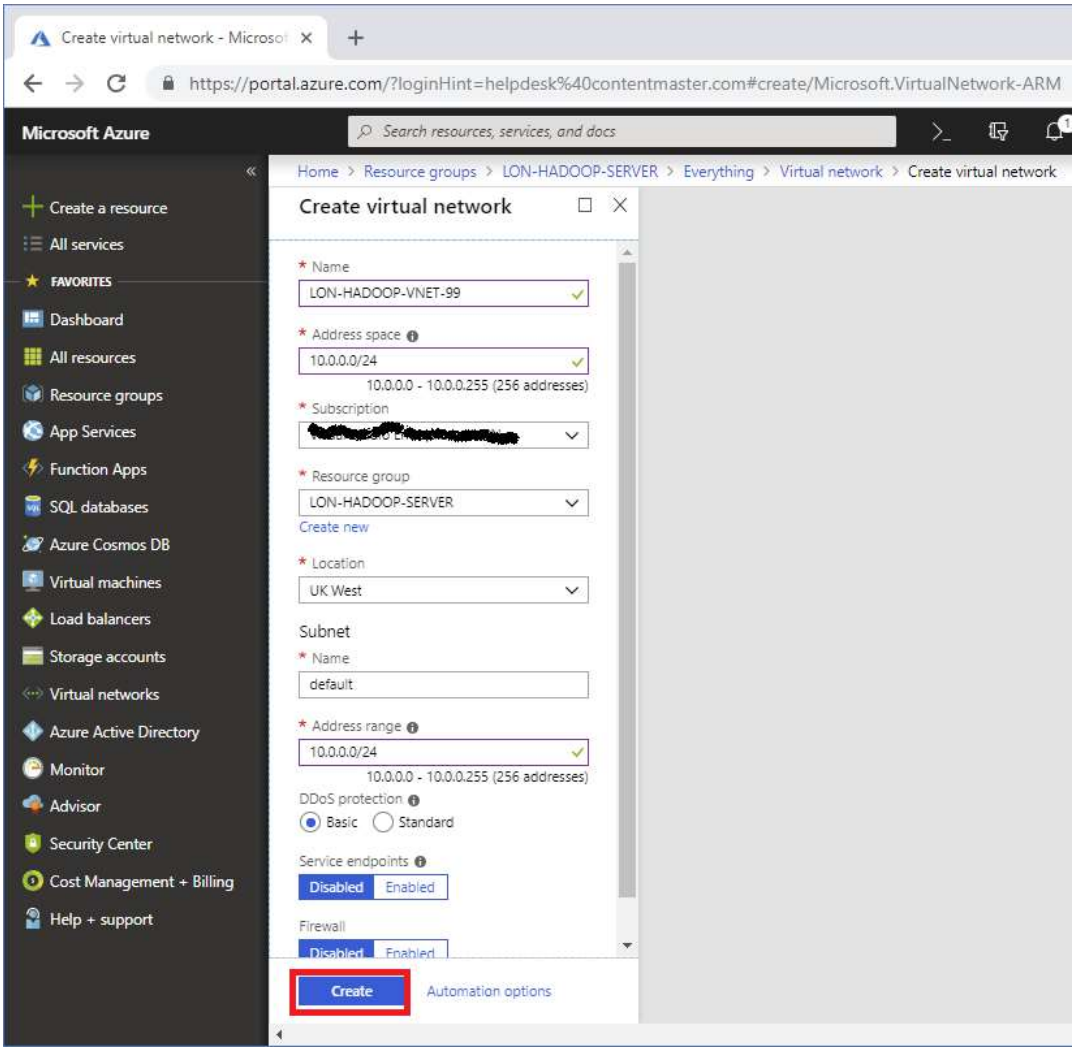


6. On the **Virtual network** blade, click **Create**



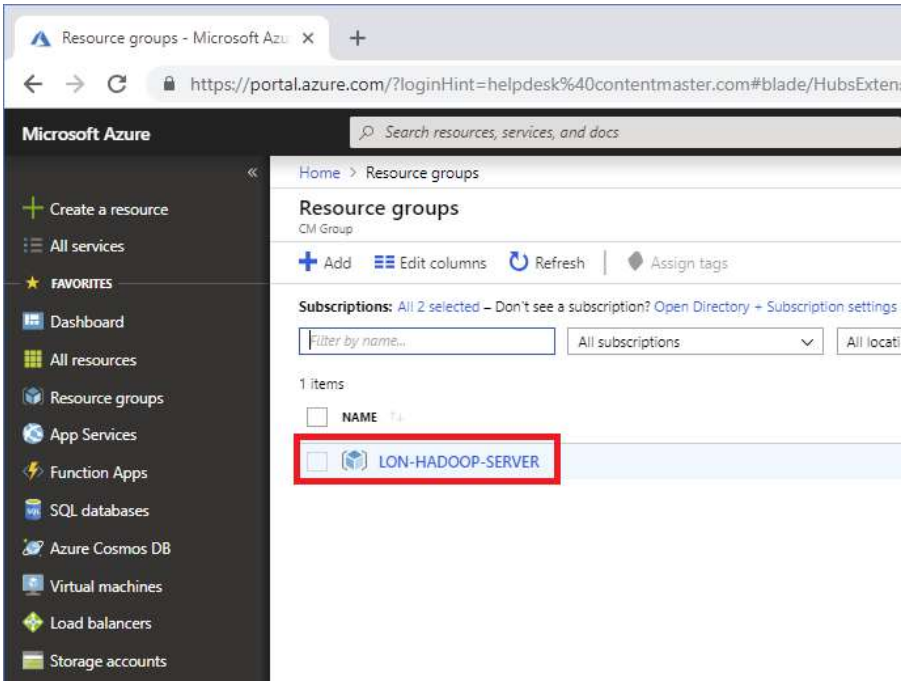
7. On the **Create virtual network** blade, enter the values shown in the following table and then click **Create**.

Property	Value
Name	LON-HADOOP-VNET-*nn* (where *nn* is a unique numeric suffix assigned to each student, such as 01, 02, 03 etc)
Address space	10.0.0.0/24
Subscription	*Specify your subscription*
Resource group	Use existing, LON-HADOOP-SERVER
Location	*Specify the same location that* you used when you created the resource group
Subnet	default
Address range	10.0.0.0/24
DDoS Protection	Basic
Service endpoints	Disabled
Firewall	Disabled

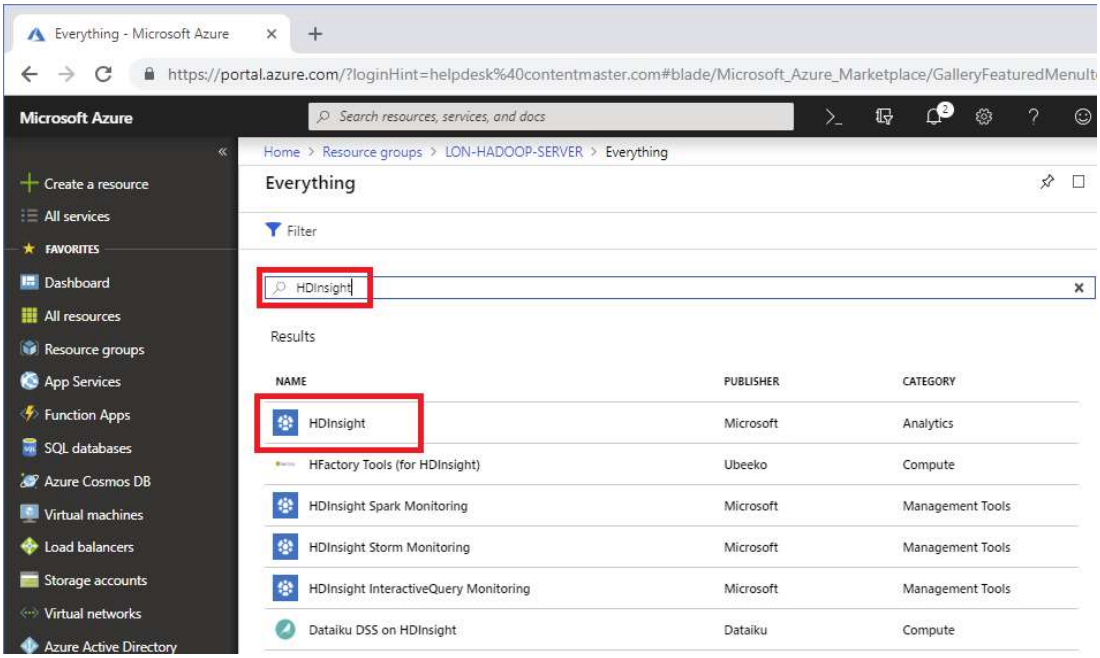


Create the LON-HADOOP Cluster

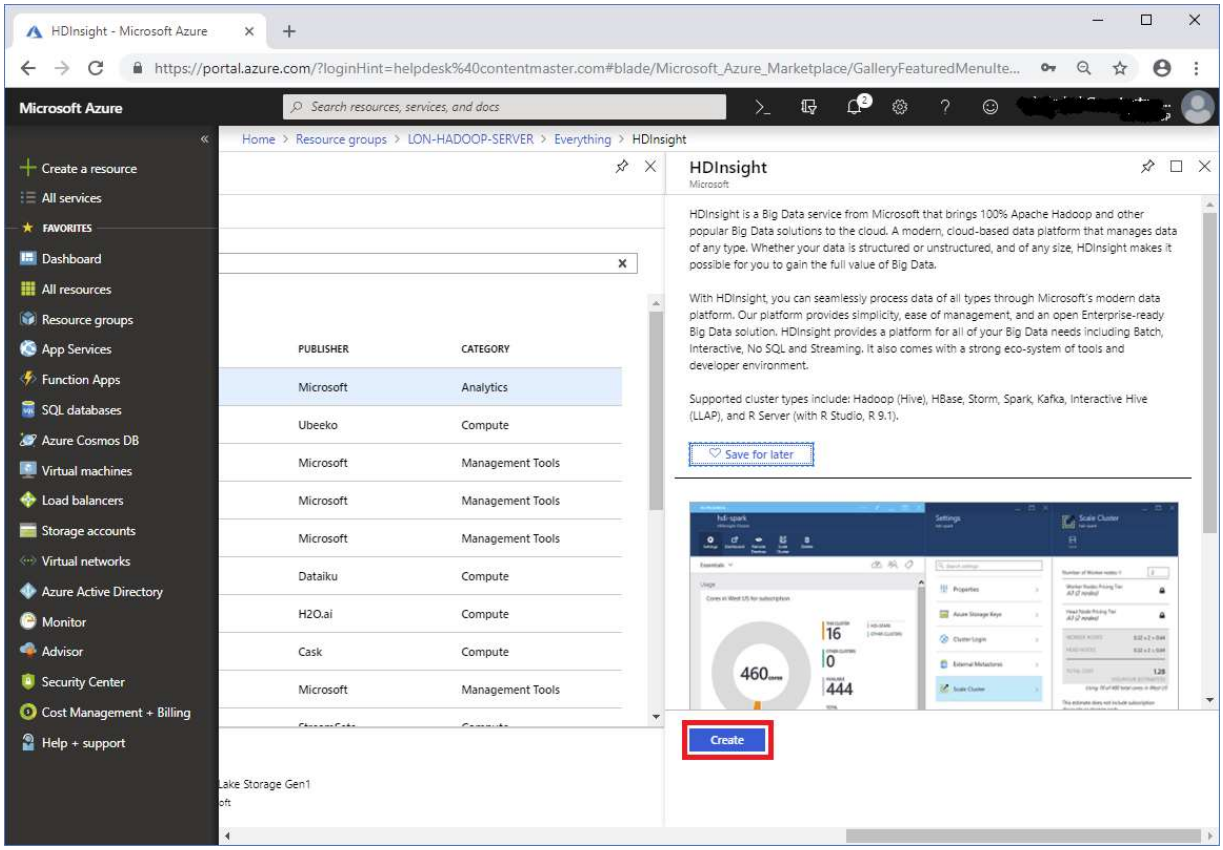
1. In the navigation blade on the left side of the portal, click **Resource groups**.
2. Click the **LON-HADOOP-SERVER** resource group.



3. On the **LON-HADOOP-SERVER** blade, click **Add**.
4. In the search box, type **HDInsight**, and then press Enter.
5. Click **HDInsight**

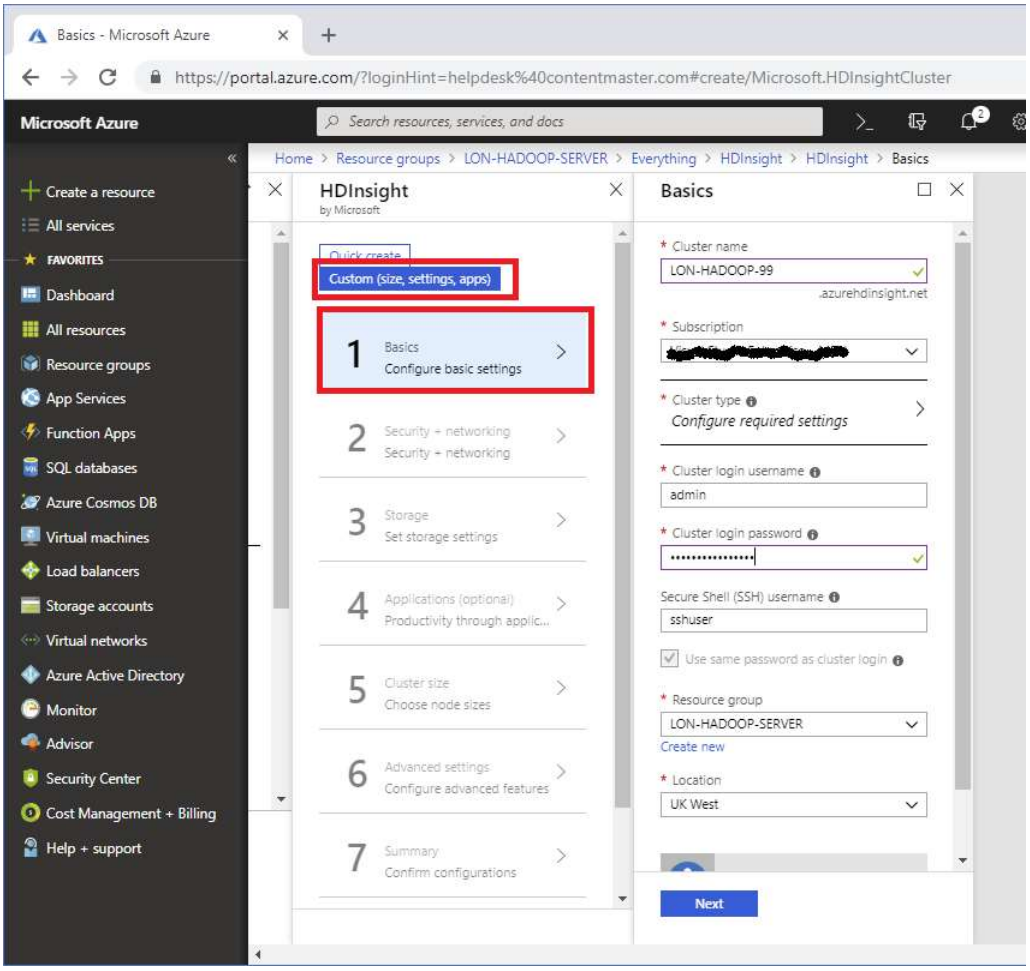


6. On the **HDInsight** blade, Click **Create**.

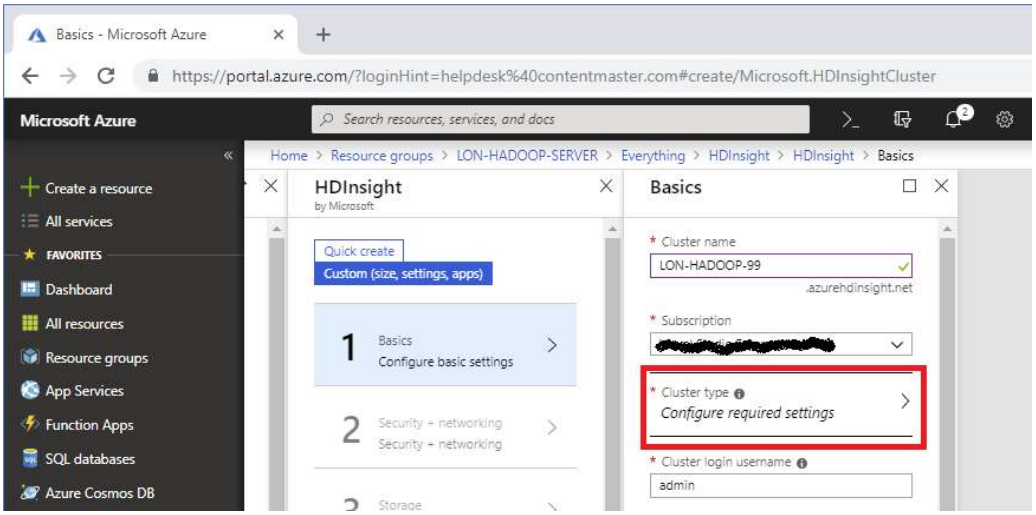


7. On the **HDInsight** blade, click **Custom (size, settings, apps)**, and then in the **Basics** blade, enter the values shown in the following table.

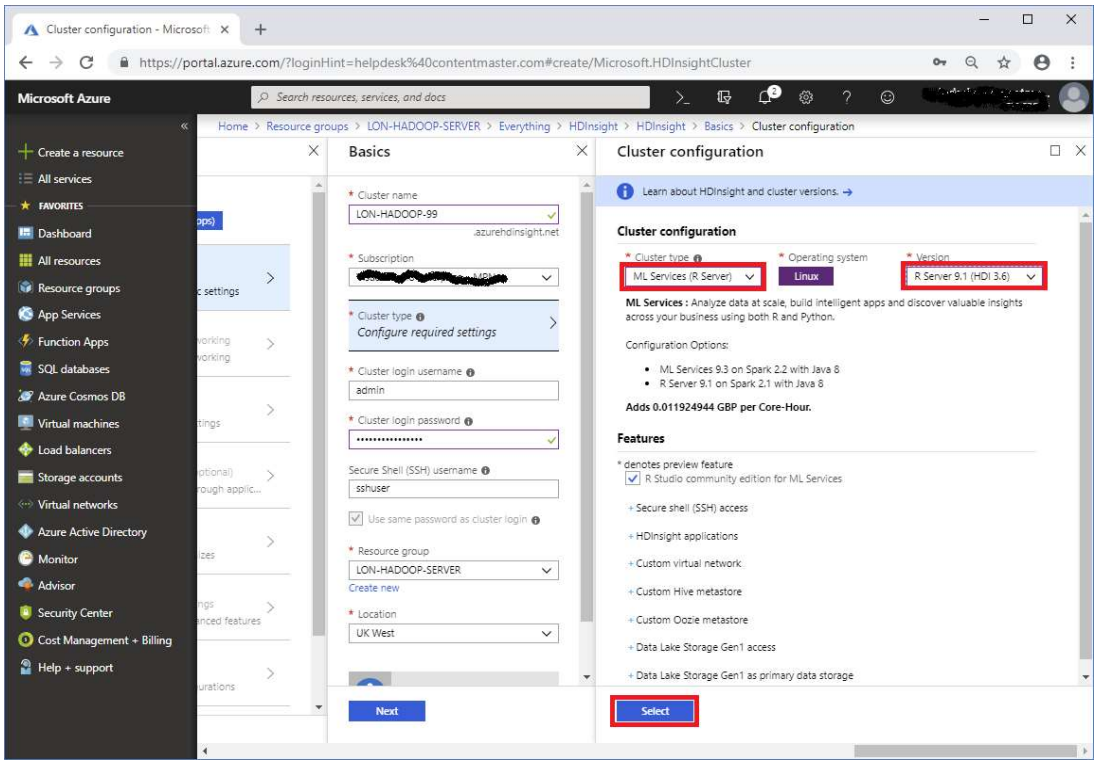
Property	Value
Cluster name	LON-HADOOP-*nn,* where *nn* is the unique suffix for the student
Subscription	*Specify your subscription*
Cluster login username	admin
Cluster login password	Pa55w.rdPa55w.rd *(Note: The repetition is intentional)*
Secure Shell (SSH) username	sshuser
Use same password as cluster login	*checked*
Resource group	Use existing, LON-HADOOP-SERVER
Location	*Specify the same location that* you used when you created the resource group



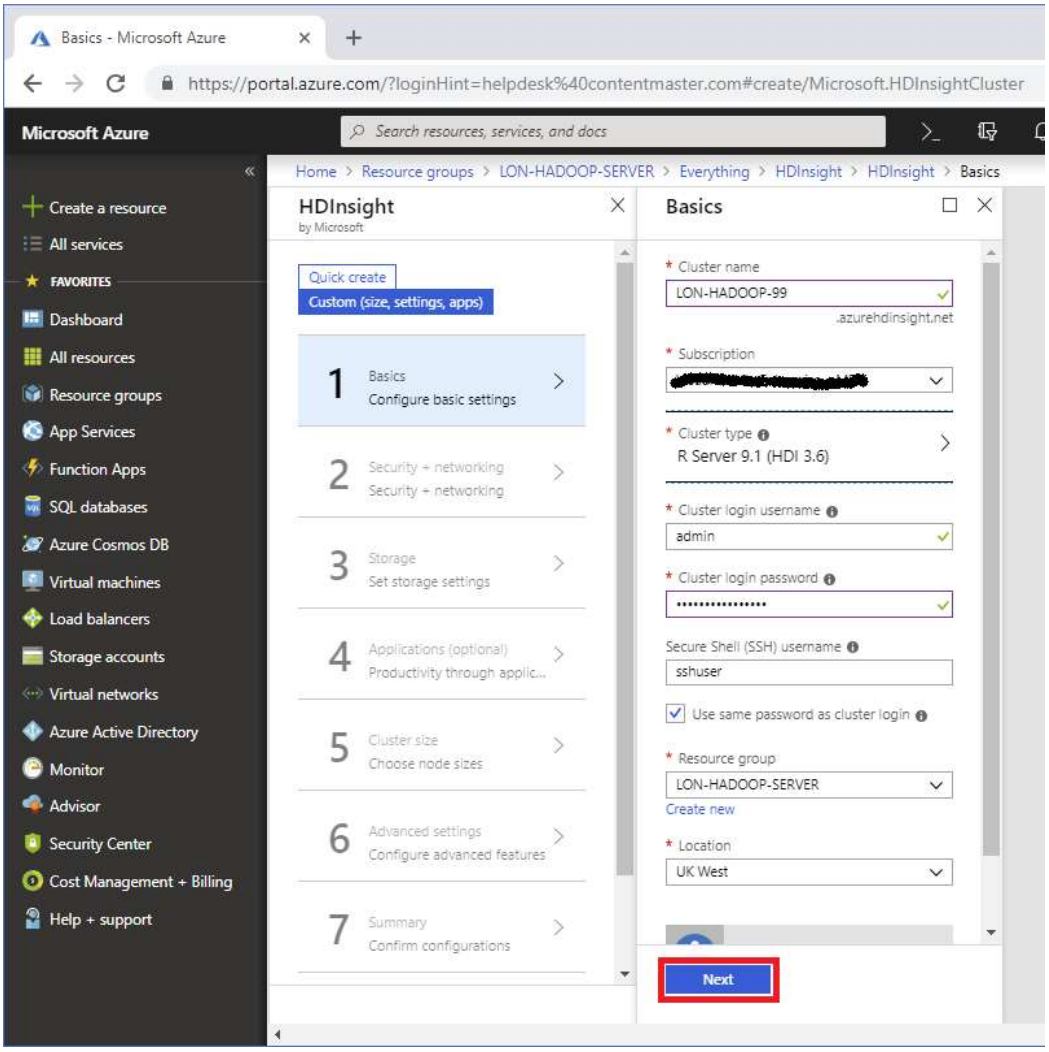
8. On the **Basics** blade, click **Cluster type**



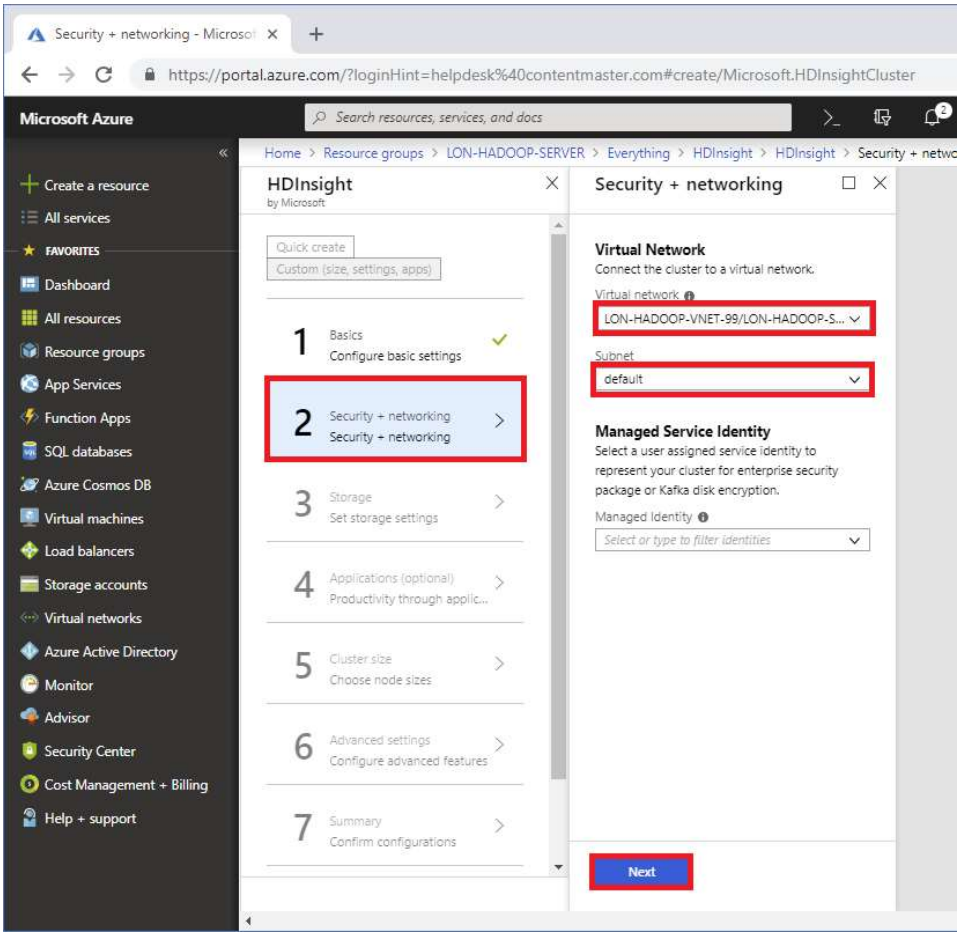
9. On the **Cluster configuration** blade, in the **Cluster type** drop-down list box, click **ML Services (R Server)**. In the **Version** drop-down list box, click **R Server 9.1 (HDI 3.6)**. Accept the default features, and then click **Select**



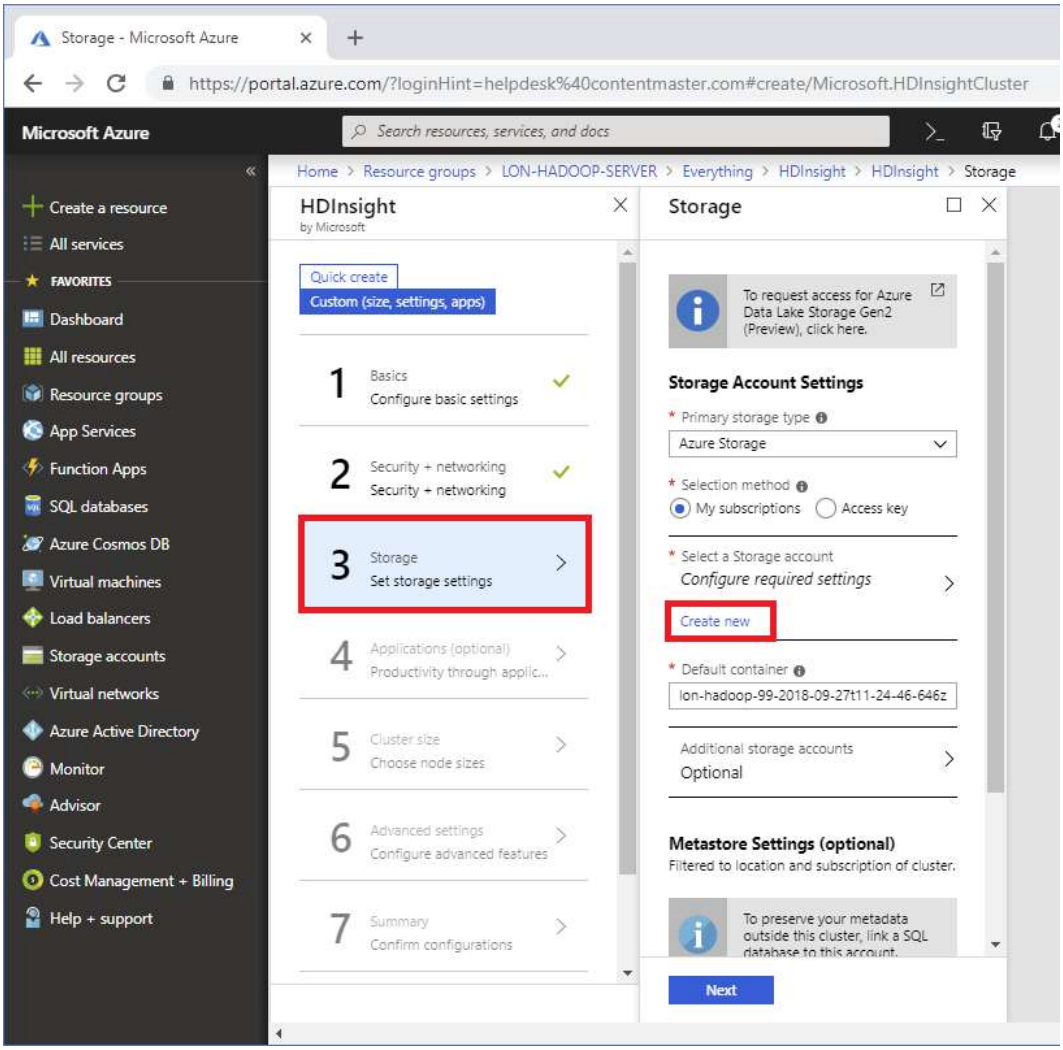
10. On the **Basics** blade, click **Next**



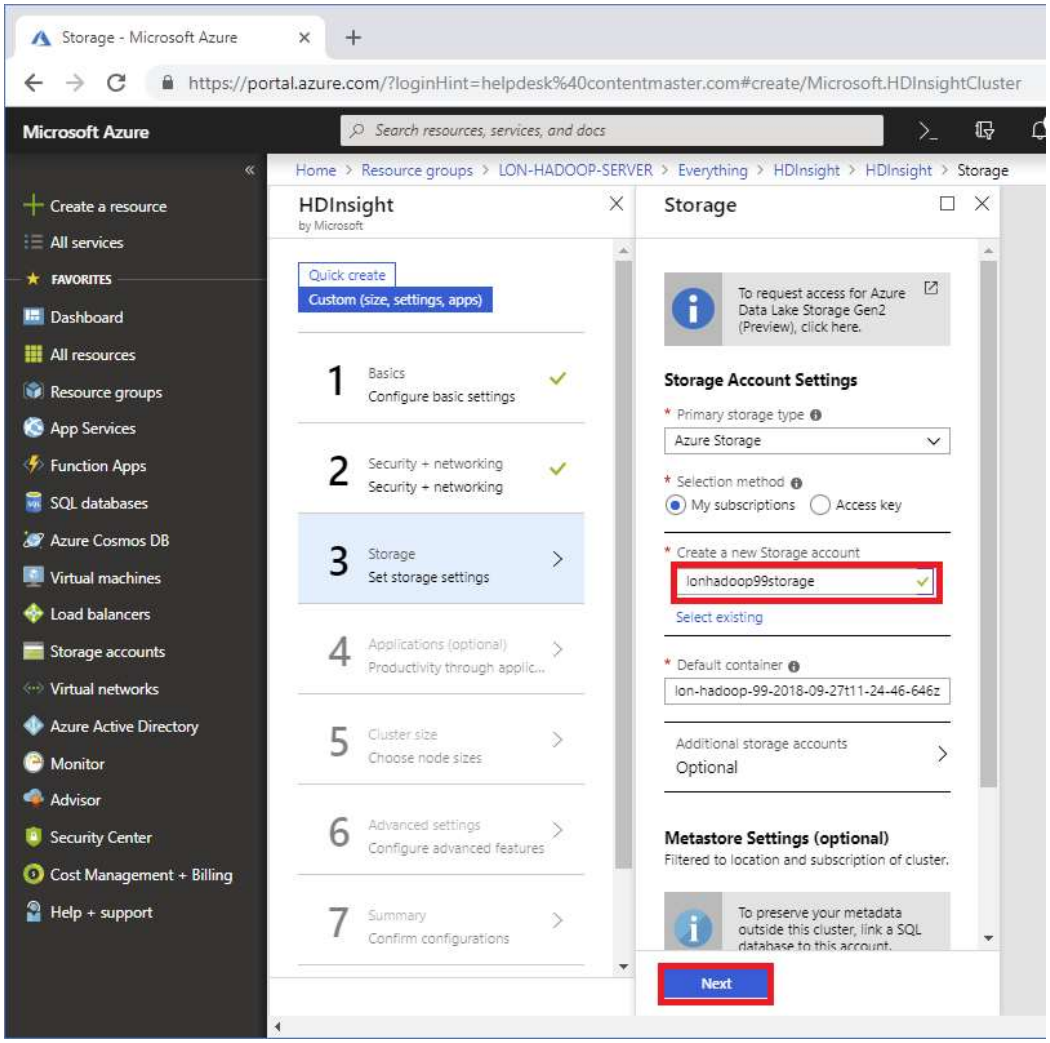
11. On the **Security + networking** blade, in the **Virtual network** drop-down list box, select the **LON-HADOOP-VNET-*NN*/LON-HADOOP-SERVER** network, select the **default** subnet, and then click **Next**



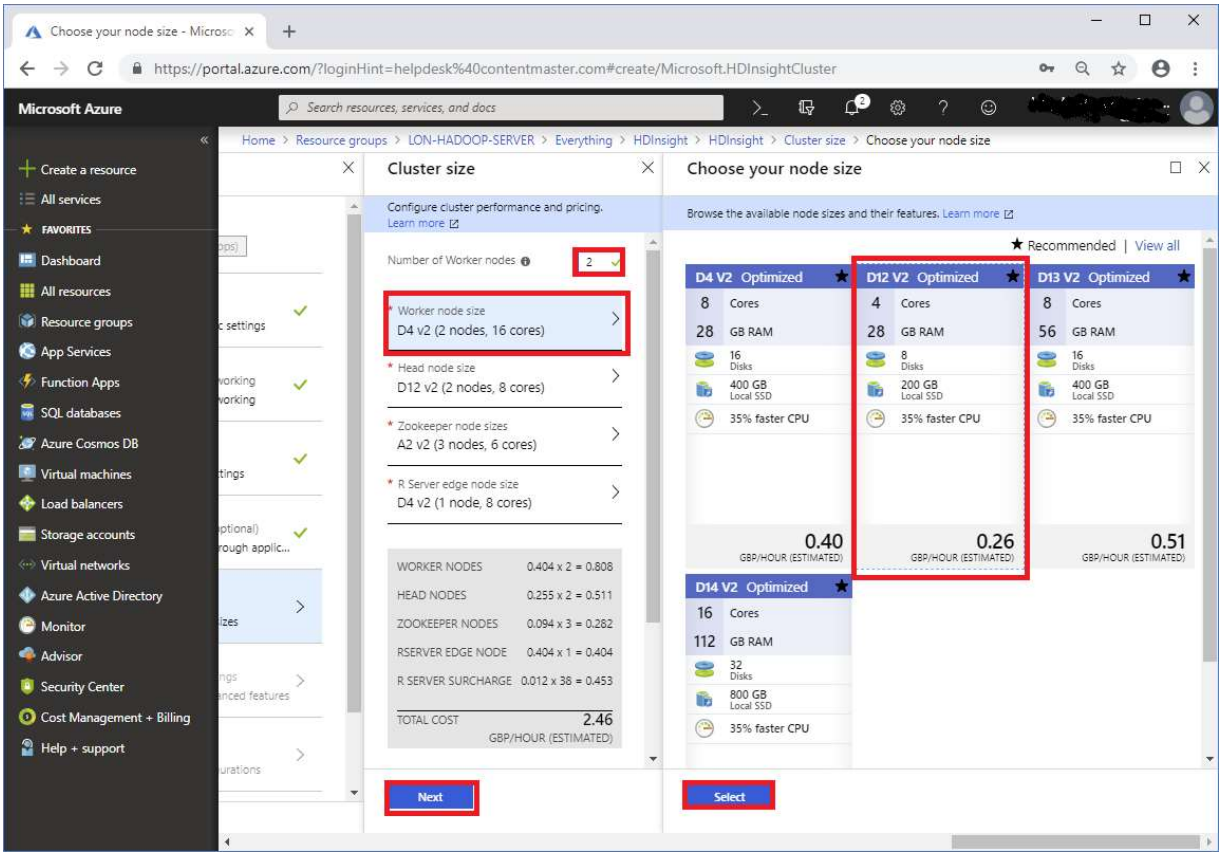
12. On the **Storage** blade, under **Select a Storage account**, click **Create new**



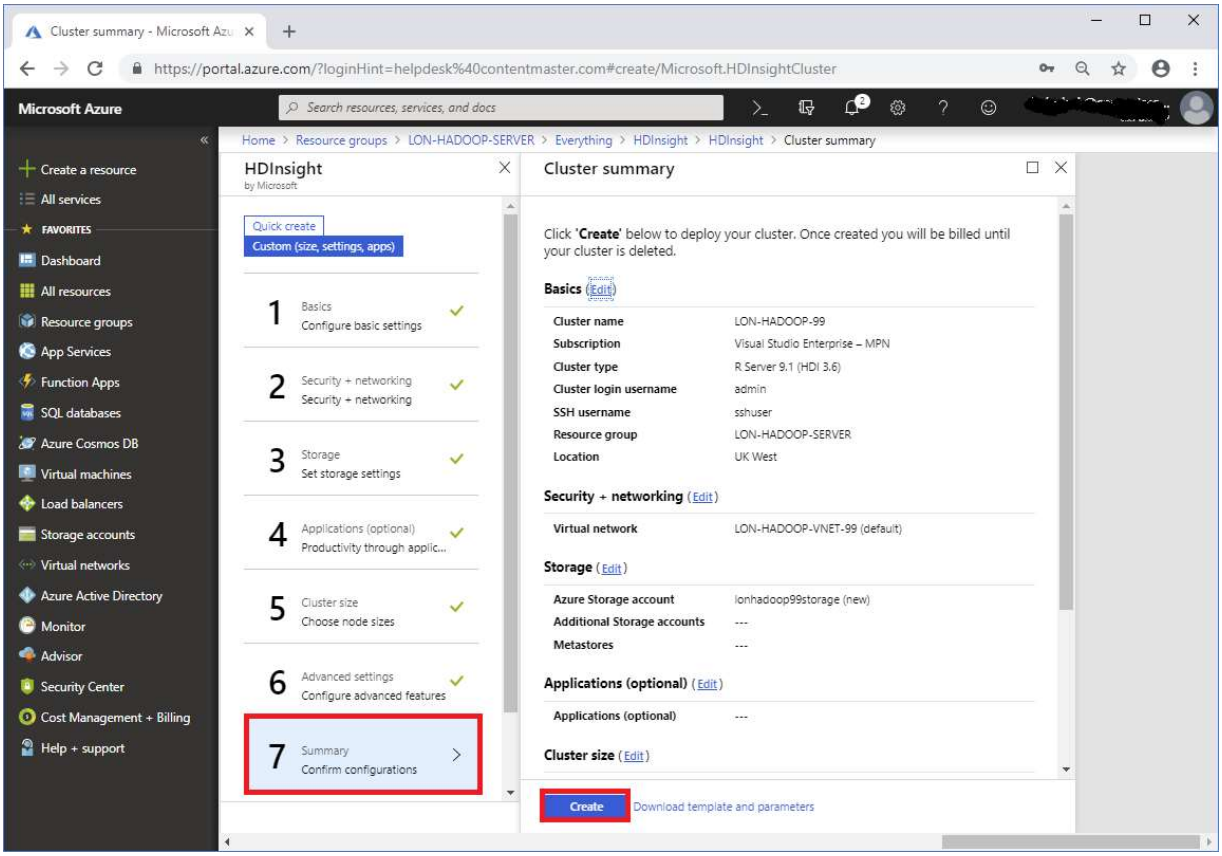
13. In the **Create a new Storage account** box type **lonhadoop*nn*storage** (where *nn* is the unique suffix assigned to the student) and then click **Next**.



14. On the **Applications (optional)** blade, click **Next**
15. On the **Cluster size** blade, in the **Number of Worker nodes** box, type **2**, click **Worker node size**, select **D12 V2 (Optimized)**, click **Select**, and then click **Next**



16. On the **Advanced** settings blade, click **Next**
17. On the **Cluster summary** blade, click **Create**

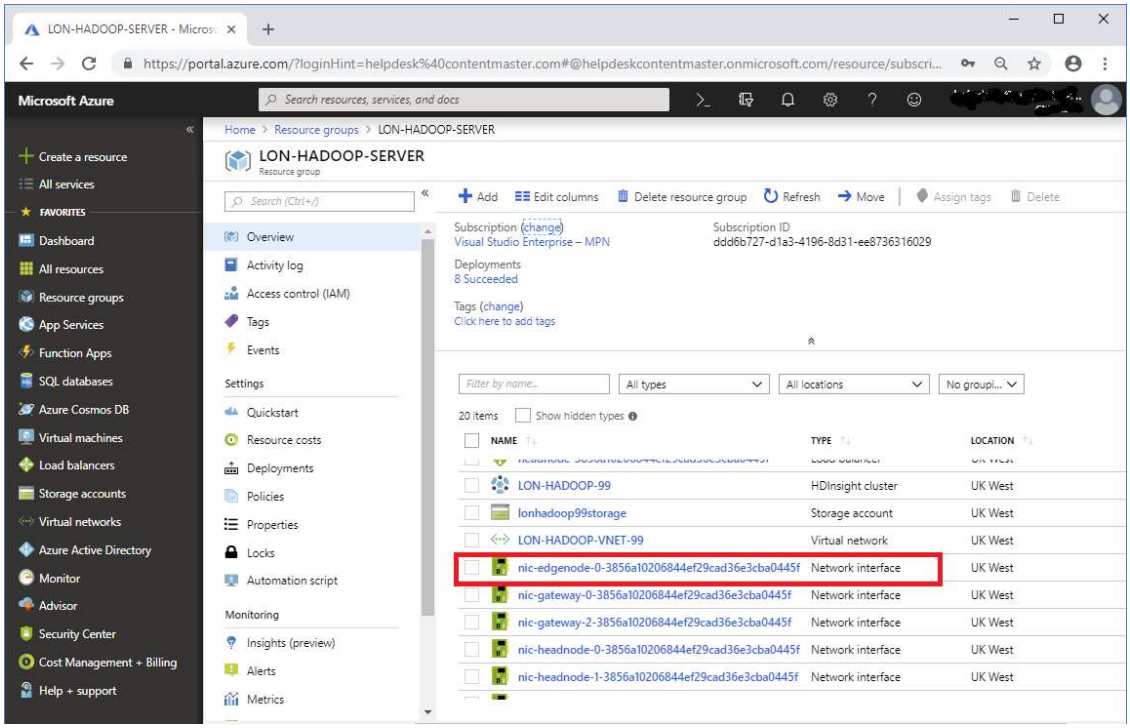


18. Wait while the cluster is created. Note that this can take up to thirty minutes.

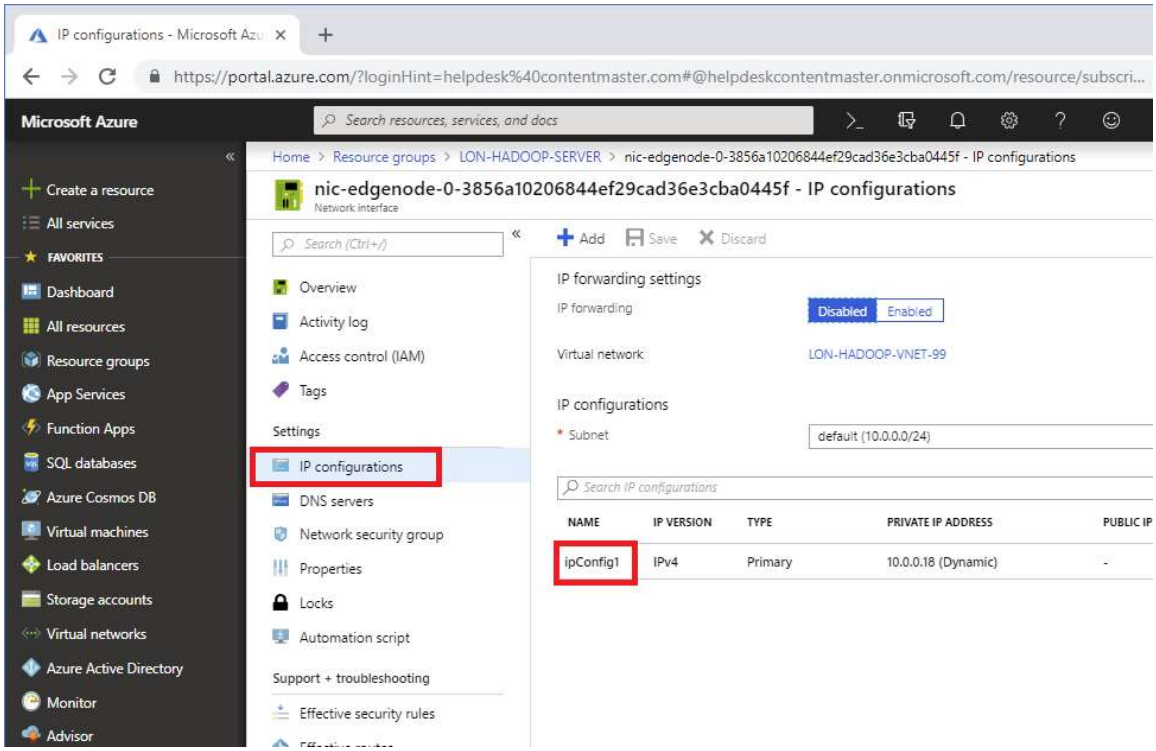
Configure the Edge Node of the Hadoop Cluster

The default configuration of the edge node blocks most IP traffic from the public Internet. To enable remote operations for the R server hosted by the cluster, you must allow access to port 12800. The following procedure adds another public IP address to the edge node with a network security rule that permits traffic for port 12900.

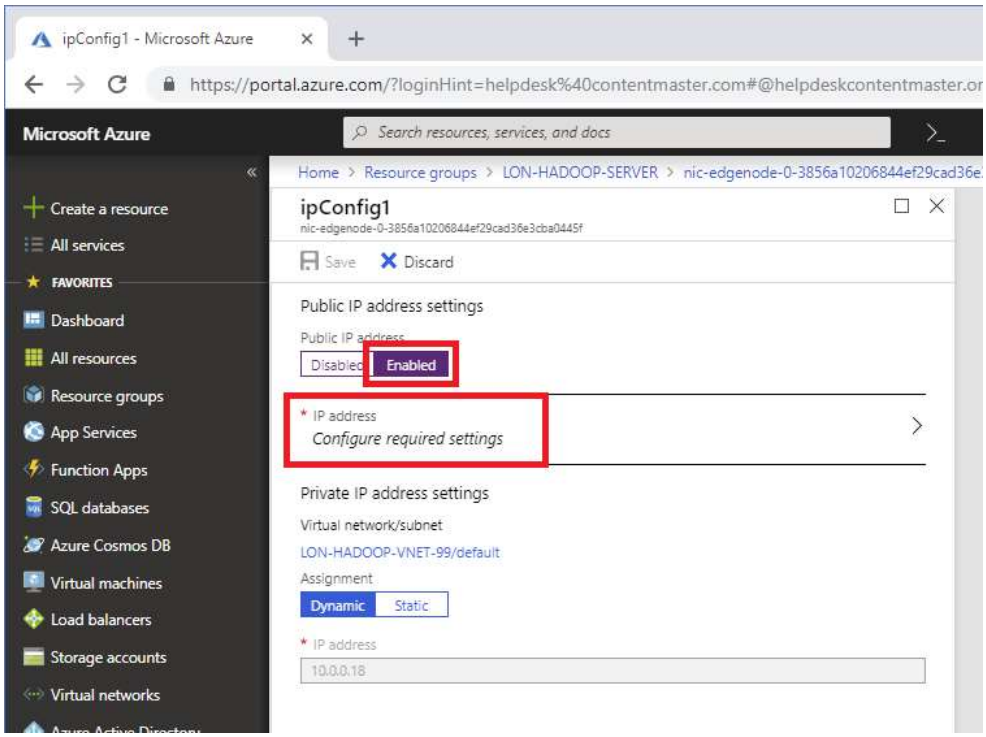
1. On the navigation blade on the left side of the portal, click **Resource groups**.
2. Click the **LON-HADOOP-SERVER** resource group.
3. On the **LON-HADOOP-SERVER** blade, click the **Network interface** for the edge node of the cluster. The name of the network interface will start with **nic-edgenode**



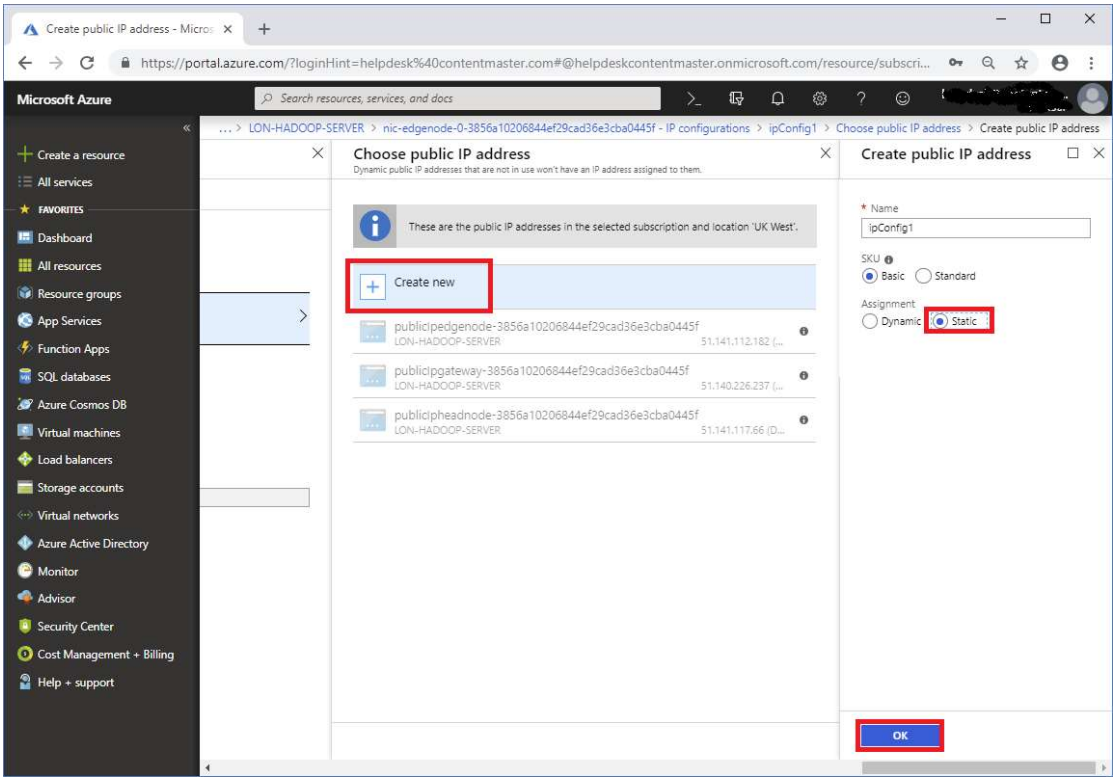
4. On the network interface blade, click **IP configurations**, and then click the **ipConfig1** configuration



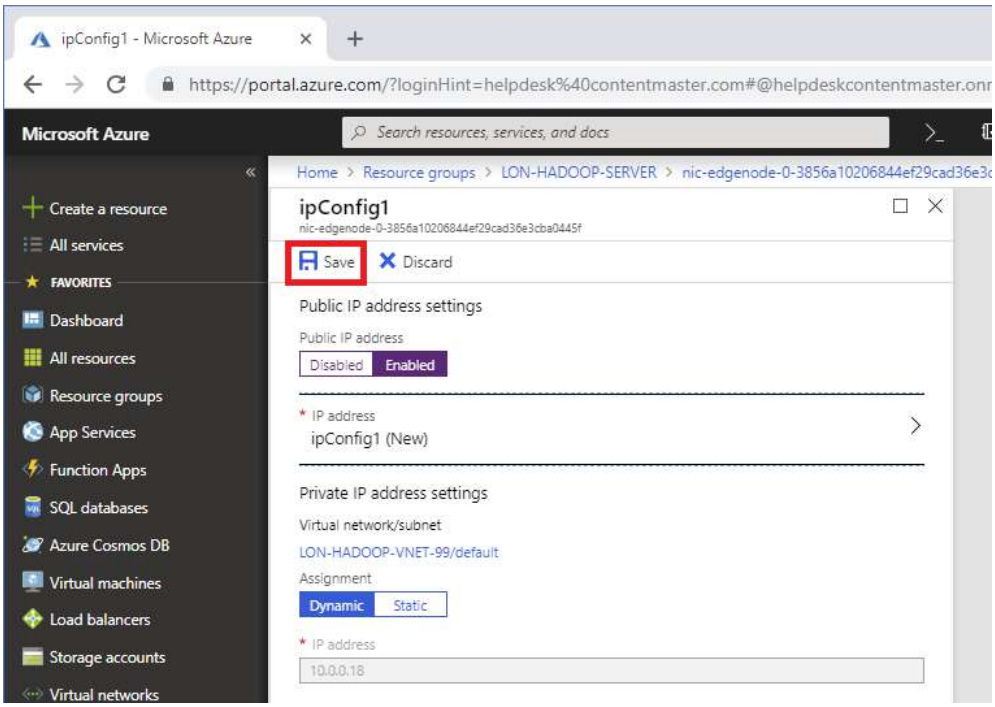
5. On the **ipConfig1** blade, under **Pubic IP address settings**, click **Enabled**, and then click **Configure required settings**



6. On the **Choose public IP address** blade, click **Create new**, click **Static**, and then click **OK**

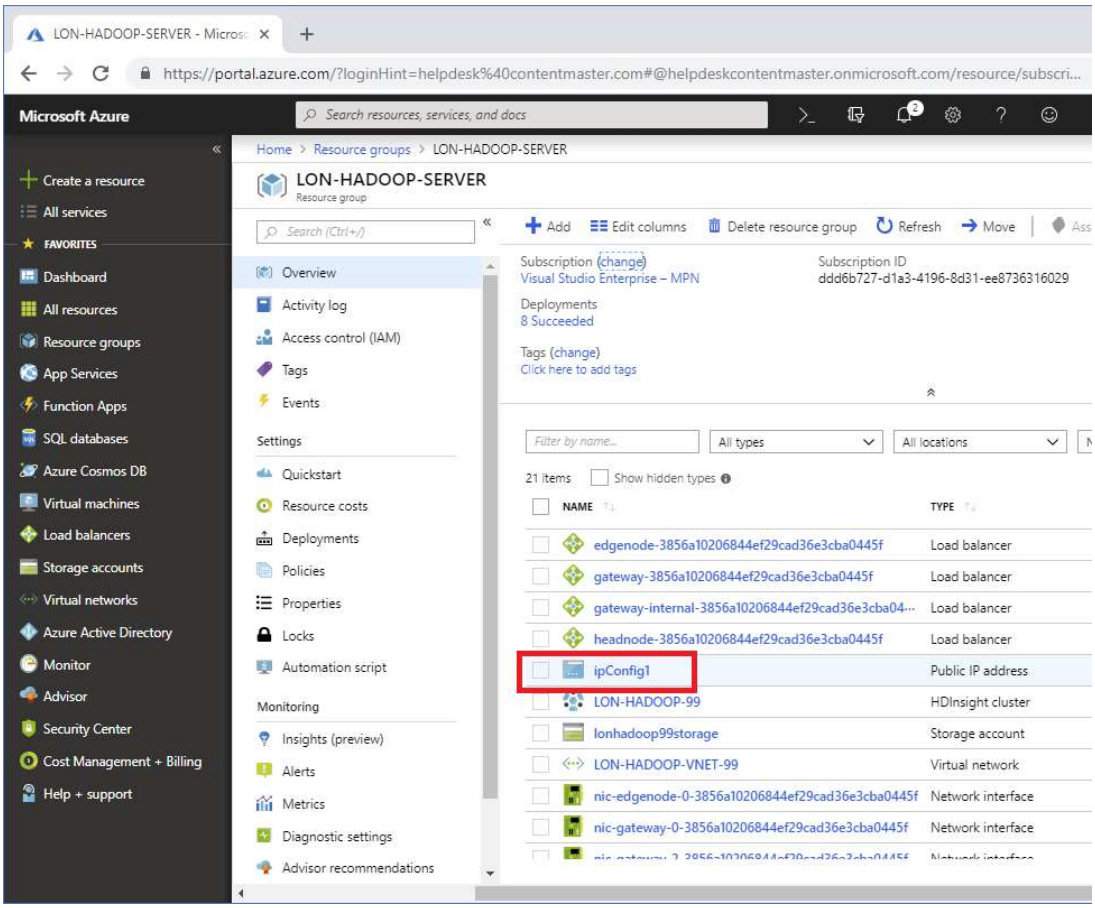


7. On the **ipConfig1** blade, click **Save**

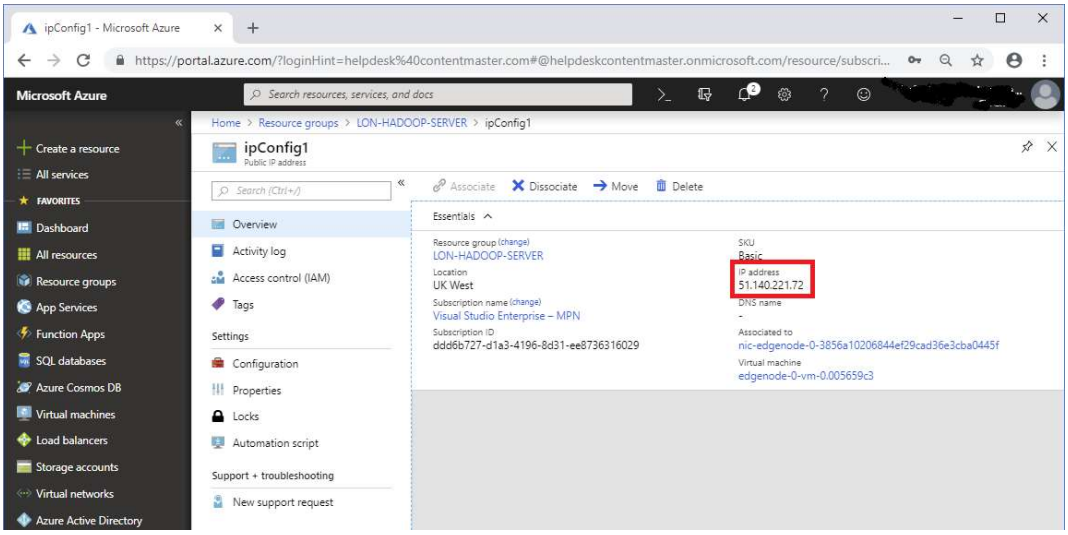


8. On the navigation blade on the left side of the portal, click **Resource groups**.

9. Click the **LON-HADOOP-SERVER** resource group.
10. On the **LON-HADOOP-SERVER** blade, click the **ipConfig1** Public IP address.



11. On the **ipConfig1** blade, make a note of the IP address. You will need this for the demonstrations and lab exercises



Configure PuTTY on the LON-DEV VM to connect to the Hadoop Cluster

1. Log on to the LON-DEV VM as **Adatum\AdatumAdmin** with password **Pa55w.rd**
2. On the LON-DEV VM, open a command prompt.
3. In the command prompt window, run the **putty** command. The putty utility should start and the **PuTTY Configuration** window should appear.
4. In the **PuTTY Configuration** window, in the **Host Name** box, enter **sshuser@*ipaddress*** where ***ipaddress*** is the public IP address of **ipConfig1** (you recorded this earlier).
5. In the **Saved Sessions** box, type **LON-HADOOP**, click **Save**, and then click **Open**.
6. If a **PuTTY Security Alert** dialog box appears, click **Yes**.
7. In the PuTTY terminal window that appears, at the **password** prompt, enter **Pa55w.rdPa55w.rd**.
8. Run the following command to create SSH keys for performing password-less authentication:

```
ssh-keygen
```

9. At the prompt **Enter file in which to save the key (/home/sshuser/.ssh/id_rsa)**, press Enter.
10. At the prompt **Enter passphrase (empty for no passphrase)**, press Enter.
11. At the prompt **Enter same passphrase again**, press Enter.
12. In the PuTTY terminal window, run the following command:

```
cat .ssh/id_rsa.pub >> .ssh/authorized_keys
```

13. In the PuTTY terminal window, run the following commands:

```
chmod 700 .ssh
chmod 600 .ssh/authorized_keys
```

14. Close the PuTTY terminal window.
15. In the **PuTTY Exit Confirmation** dialog box, click **OK**.
16. On the LON-DEV VM, in the command prompt window, move to the **E:** folder.

17. Run the following command to copy the key file for your account on the Hadoop VM to the LON-DEV VM. Replace *ipaddress* with the value of the **ipConfig1** public IP address:

```
pscp sshuser@ipaddress:.ssh/id_rsa id_rsa
```

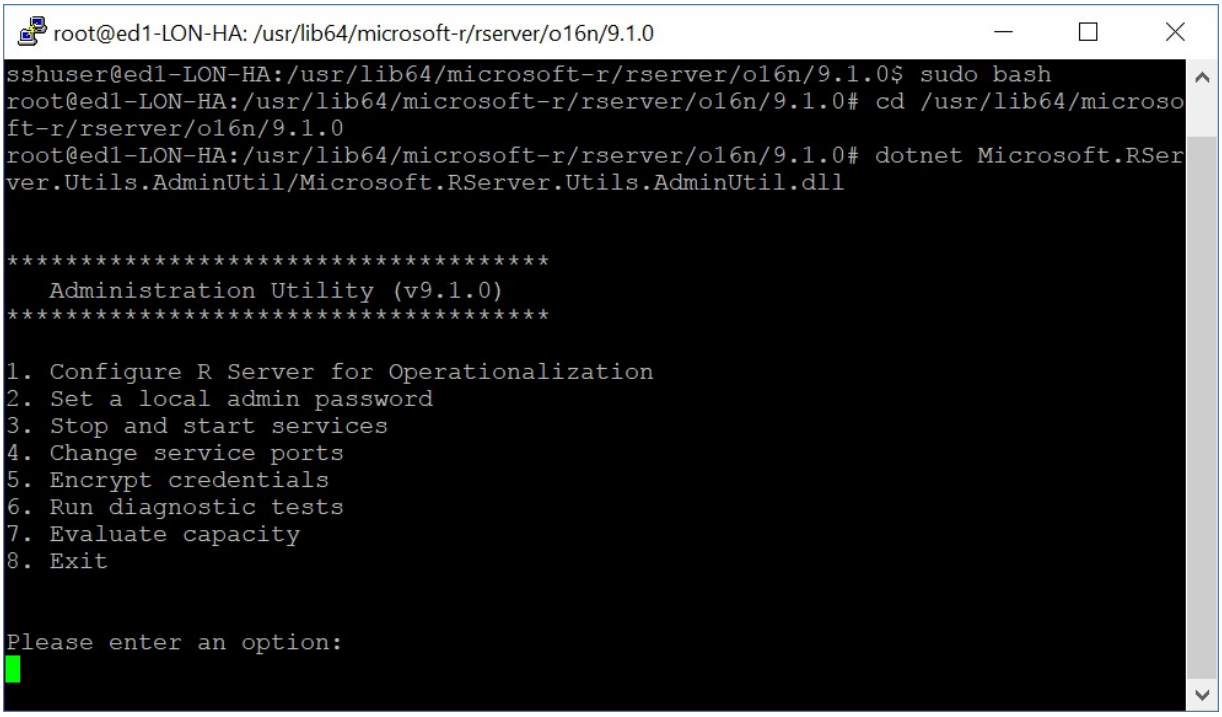
18. At the **Password** prompt, type **Pa55w.rdPa55w.rd**, and then press Enter.
19. In the command prompt window, run the **puttygen** command. The **PuTTY Key Generator** window should appear.
20. In the **PuTTY Key Generator** window, click **Load**.
21. In the **Load private key** dialog box, move to the **E:** folder, in the file selector drop-down list box, click **All Files(*.*)**, click **id_rsa**, and then click **Open**.
22. In the **PuTTYgen Notice** dialog box, verify that the key was imported successfully, and then click **OK**.
23. In the **PuTTY Key Generator** window, click **Save private key**.
24. In the **PuTTYgen Warning** dialog box, click **Yes**.
25. In the **Save private key as** dialog box, in the **File name** box, type **HadoopVM**, and then click **Save**.
26. Close the **PuTTY Key Generator** window.
27. Run the **putty** command again.
28. In the **PuTTY Configuration** window, in the **Saved Sessions** box, click **LON-HADOOP**, and then click **Load.**
29. In the **Category** pane of the **PuTTY Configuration** window, under **Connection**, expand **SSH**, and then click **Auth**.
30. In the **Options controlling SSH authentication** pane, next to the **Private key file for authentication** box, click **Browse**.
31. In the **Select private key file** dialog box, move to the **E:** folder, click **HadoopVM.ppk**, and then click **Open**.
32. In the **Category** pane of the **PuTTY Configuration** window, under **Connection**, click **Data**.
33. In the **Data to send to the server** pane, in the **Auto-login username** box, type **sshuser**
34. In the **Category** pane of the **PuTTY Configuration** window, click **Session**.
35. Click **Save**, and then close the **PuTTY Configuration** window.
36. Close the Command Prompt window.

Operationalize R Server on the LON-HADOOP Cluster

1. On the desktop machine, open a command prompt.
2. In the command prompt window, run the **putty** command. The putty utility should start and the **PuTTY Configuration** window should appear.
3. In the **Saved Sessions** box, click **LON-HADOOP**, click **Load**, and then click **Open**.
4. In the PuTTY terminal window, run the following commands to start the Microsoft R Administrator Utility:

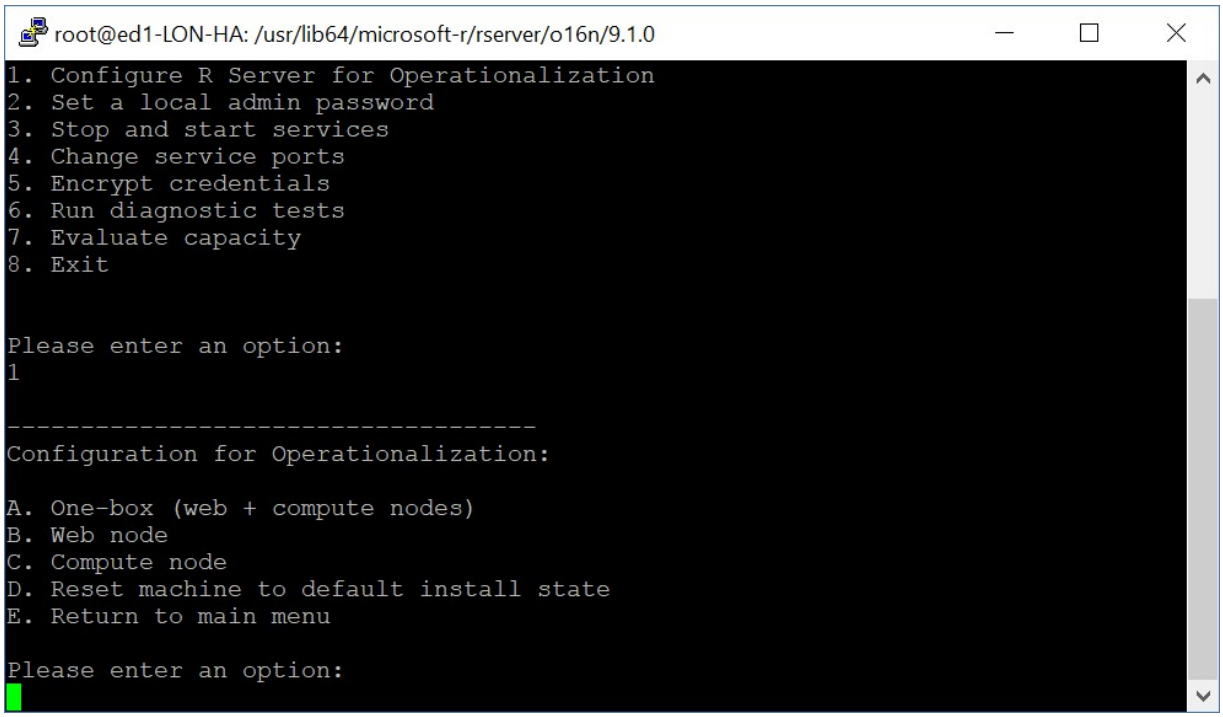
```
sudo bash
cd /usr/lib64/microsoft-r/rserver/o16n/9.1.0
```

```
dotnet
Microsoft.RServer.Utils.AdminUtil/Microsoft.RServer.Utils.AdminUtil.dll
```



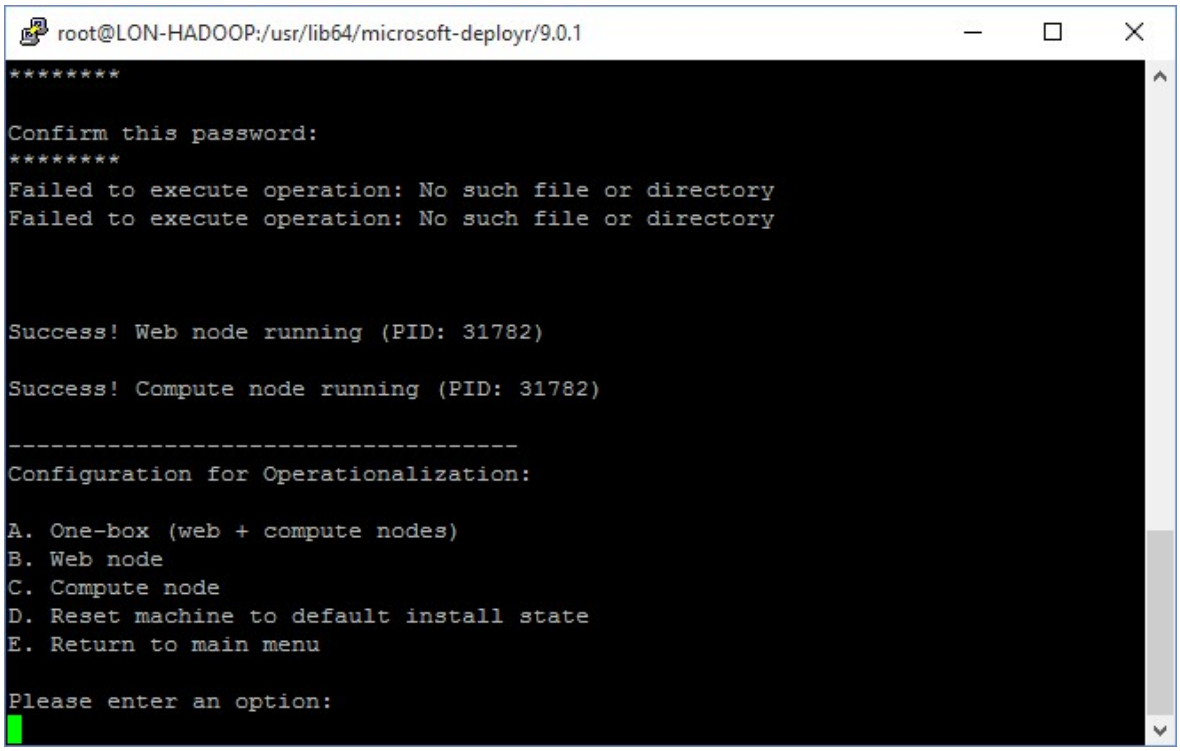
5. In the **Administration Utility** menu, type **1**, and then press Enter.

6. In the **Configuration for Operationalization** menu, type **A**, and then press Enter.



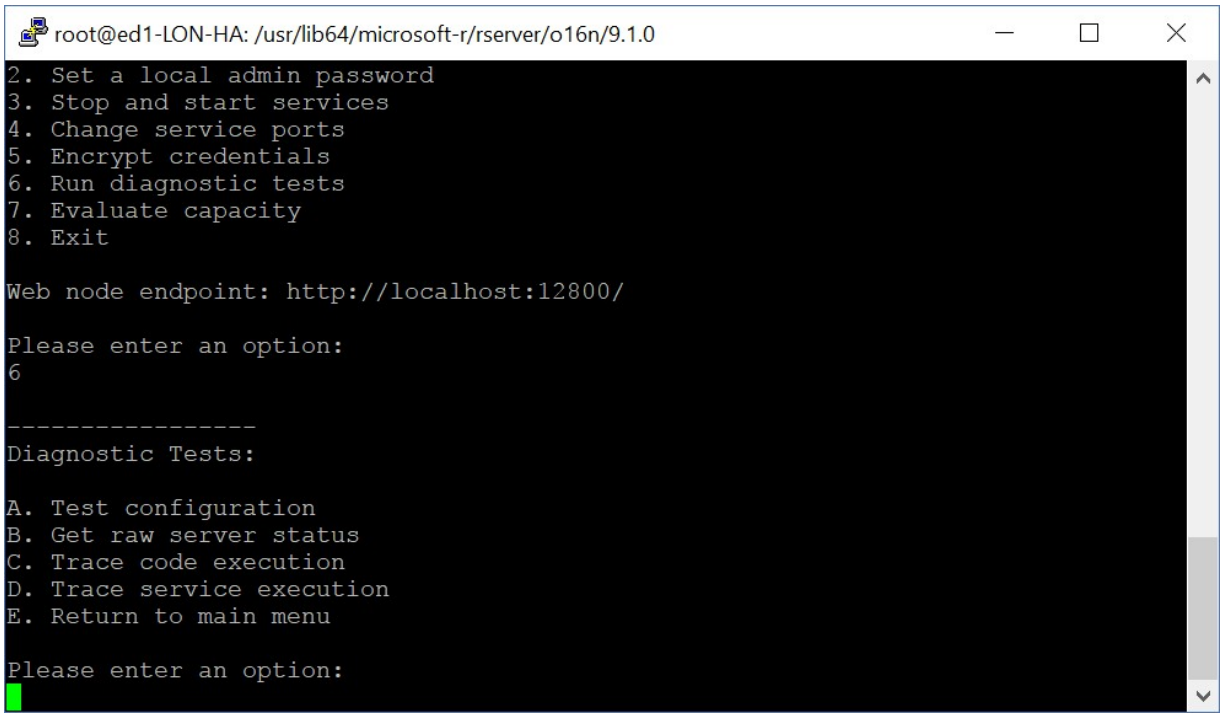
7. At the **Set the admin password** prompt, type **Pa55w.rd**, and then press Enter.

8. At the **Confirm this password** prompt, type **Pa55w.rd**, and then press Enter.

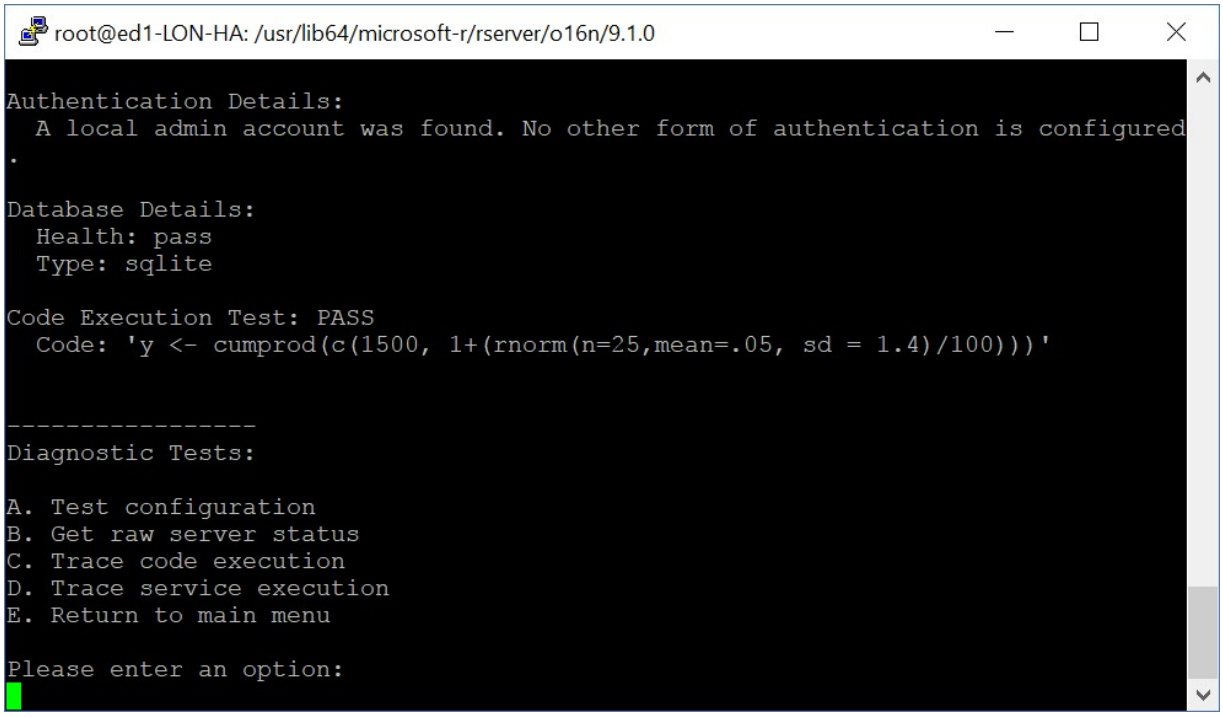


9. In the **Configuration for Operationalization** menu, type **E**, and then press Enter.
10. In the **Administration Utility** menu, type **6**, and then press Enter.

11. In the **Diagnostic Tests** menu, type **A**, and then press Enter.



12. At the **Username** prompt, type **admin**, and then press Enter.
13. At the **Password** prompt, type **Pa55w.rd**, and then press Enter.
14. Verify that the diagnostic results show that the server is healthy:



15. In the **Diagnostic Tests** menu, type **E**, and then press Enter.
16. In the **Administration Utility** menu, type **8**, and then press Enter.
17. In the PuTTY terminal window, run the following commands to connect as the **hdfs** user (this user has admin privileges over the HDFS file system):

```
su - hdfs
```

18. In the PuTTY terminal window, run the following commands to create the HDFS folders required by R server for the sshuser user:

```
hadoop fs -mkdir /user/RevoShare/sshuser
hadoop fs -chmod 777 /user/RevoShare/sshuser
hadoop fs -mkdir /user/sshuser
hadoop fs -chmod 777 /user/sshuser
```

19. In the PuTTY terminal window, run the following command to return to running as the root user:

```
exit
```

20. In the PuTTY terminal window, run the following commands to create the file system folders required by R server for each user:

```
mkdir -p /var/RevoShare/sshuser
chmod 777 /var/RevoShare/sshuser
```

Ignore any warnings that occur if these directories already exist

21. Close the PuTTY terminal window. When prompted, click **OK** to confirm that you wish to exit the session.