Microsoft

Provision Enterprise Security Package enabled HDInsight - Hadoop Cluster

Provisioning Sequence, and post installation validations

Prepared by

Data SQL Ninja Engineering Team (datasqlninja@microsoft.com**)**

Note: The detail provided in this document has been harvested as part of a customer engagement sponsored through the Data SQL Ninja Engineering.

Table of Contents

1 Introduction 8

2 Setup Azure Active Directory – Domain Services (AAD-DS) 9

2.1 Create AAD-DS Resource 9

2.2 Setup AAD-DS - Scoped Synchronization 10

2.3 Generate & Configure Self-signed Certificates 12

3 Configure Networking 15

3.1 Configure NSG Rules to secure LDAP 16

3.2 Configure DNS Zone for External Access 18

3.3 Configure DNS Server Settings & Enable AAD-DS Password Hash synchronization 20

3.4 Update DNS Server Settings for your Virtual Network 20

3.5 Configure NSG Rules to support HDInsight Management & Health services 23

4 Configure Access Controls 25

4.1 Setup User Assigned Managed Identity 25

4.2 Create & Configure *hadoopadmins* AD Group 27

4.3 Assign HDInsight Domain Services Contributor Role 28

5 Create & Configure ADLS Gen2 Storage Account 31

5.1 Create ADLS Gen2 Storage Account 31

5.2 Validate network restrictions 36

5.3 Configure Storage Blob Data Owner Role 38

5.3.1 Add role to User Assigned Managed Identity 38

5.3.2 Add role to *hadoopadmins* group 40

5.4 Add HDInsight Cluster Admin User 42

5.5 Additional role configurations for group *hadoopadmins* 47

6 Setup HDInsight ESP enabled cluster 49

6.1 Create HDInsight + ESP Resource 50

6.2 Enable NSG rules for Ambari UI (port 443) 62

6.3 Verify Ambari UI HTTPs Port access 63

6.4 Setup SSH-Tunneling to access Ambari URLs 65
6.5 Validate SSH-Tunnel 68
6.6 Known Issues – Web UI 69
7 Post Installation Review 71
7.1 Ambari Landing Page 71
7.2 HDFS Service Summary 72
7.3 YARN Summary 73
7.4 Ranger 73
7.5 Storage explorer – HDInsight Container view 75
8 Appendix – Reference URLs 76
9 Feedback and suggestions 76

Table of Figures

Figure 1: Enable Azure Active Directory - Domain Services 9
Figure 2: Configure AAD-DS Synchronization Scope - Add Groups 10
Figure 3: AAD-DS Add Group - Group Selection 11
Figure 4: Selected AAD-DS Groups for Synchronization scope 12
Figure 5: Azure AD Domain Services - Configure Certificate (PFX file) 14
Figure 6: Azure AD Domain Services - Secure LDAP Enabled 14
Figure 7: Azure AD Domain Services - Secure LDAP Brute-force Warning 15
Figure 8: HDInsight Networking Interdependencies Overview 15
Figure 9: Azure Active Directory - Domain Services Resource (from the listing) 16
Figure 10: Azure AD Domain Services NSG Rules (Inbound/Outbound) 16
Figure 11: Azure AD Domain Services NSG Rules - Add Inbound Security Rule 17
Figure 12: Azure AD Domain Services NSG Rules - Updated 17
Figure 13: ldp.exe client UI 18
Figure 14: ldp.exe - connect error sample 19
Figure 15: Azure AD Domain Services - Required Configuration Steps 20
Figure 16: HDInsight Cluster Deployment Failure - DomainNotFoundInActiveDirectory 21
Figure 17: HDInsight Deployment - Set of Resources Deployed! 21
Figure 18: Azure AD Domain Services - Properties 22
Figure 19: Azure AD Domain Services - VNET Properties - DNS IP address updates 22
Figure 20: Subscription/All Resources View - AAD-DS NSG Resource 23
Figure 21: Configured NSG Rules for HDInsight Management & Health Service IP Addresses 24
Figure 22: Search for User Assigned Managed Identity Resource type 25
Figure 23: Create User Assigned Managed Identity 25
Figure 24: User Assigned Managed Identity - Create Form 26
Figure 25: User Assigned Managed Identity - Blade View/Overview 26
Figure 26: Azure AD - Groups - All Groups view 27
Figure 27: Azure AD Domain Services - Add Group to Synchronization scope 28
Figure 28: Updated Azure AD Domain Services - Synchronization Scope 28
Figure 29: Assign HDInsight Domain Services Contributor role to User Assigned Managed Identity 29
Figure 30: Select Managed Identity for the HDInsight Domain Services Contributor Role assignment 30
Figure 31: Updated HDInsight Domain Services Contributor Role Assignment 30
Figure 32: List of Storage Accounts 31
Figure 33: Create Storage Account - Basics Tab 32

Figure 34: Create Storage Account - Networking Tab 33
Figure 35: Create Storage Account - Advanced Tab 33
Figure 36: Create Storage Account - Review + Create - Validation Passed 34
Figure 37: Storage Account Deployment (underway) 35
Figure 38: Storage Account Deployment Complete 35
Figure 39: ADLS Gen2 - Overview 36
Figure 40: ADLS Gen2 - Validate Network Restrictions - Sample Error Message 37
Figure 41: ADLS Gen2 - Validate Network Restrictions - Unable to Load Data (Sample) 37
Figure 42: ADLS Gen2 - Add Assign User Assigned Managed Identity 38
Figure 43: ADLS Gen2 - Managed Identity selection 39
Figure 44: ADLS Gen2 - Save selected Managed Identity 39
Figure 45: ADLS Gen2 - Updated Storage Blob Data Owner Role view 40
Figure 46: ADLS Gen2 - Add role to *hadoopadmins* group 40
Figure 47: ADLS Gen2 - Updated Role Assignments View 41
Figure 48: Add *hadoopadmin* user to *hadoopadmins* group 42
Figure 49: Add hadoopadmin user to hadoopadmins group (Contd...) 43
Figure 50: Azure Active Directory - All Users view 43
Figure 51: Hadoop Administrator User Profile (Azure Active Directory) 44
Figure 52: Azure Portal - Sign-in with a different account 44
Figure 53: Azure Portal Login 45
Figure 54: Azure Portal First time login - Update Password 46
Figure 55: Azure Portal New User Login - Landing Page 46
Figure 56: RBAC Roles Application Overview 47
Figure 57: *hadoopadmins* group assigned roles - Access Control (IAM) overview 48
Figure 58: Hadoop Admin User - Storage Explorer View - Folder View Sample 48
Figure 59: Create Azure HDInsight - Search 50
Figure 60: Create Azure HDInsight Resource 50
Figure 61: Create HDInsight + ESP - Choose cluster type 51
Figure 62: Create HDInsight + ESP cluster - Details 52
Figure 63: Create HDInsight + ESP cluster - Storage Tab 53
Figure 64: Create HDInsight + ESP cluster - Security + Networking Tab 53
Figure 65: Create HDInsight + ESP cluster - Enable ESP 54
Figure 66: Create HDInsight + ESP cluster - Select Cluster Admin User 55
Figure 67: Create HDInsight + ESP cluster - Price Estimate 57
Figure 68: Create HDInsight + ESP cluster - Configuration Summary 58

Figure 69: Create HDInsight + ESP cluster - Deployment Underway 59
Figure 70: Create HDInsight + ESP cluster - Deployment Complete 60
Figure 71: Create HDInsight + ESP cluster - Deployment Operation Details 60
Figure 72: HDInsight + ESP cluster overview 61
Figure 73: Ambari UI - Cannot reach the page 61
Figure 74: Enable Port 443 on NSG for Ambari UI 62
Figure 75: NSG Inbound Rule Updated - Ambari UI - HTTPS port enabled 63
Figure 76: Retry accessing Ambari Web UI 63
Figure 77: Ambari Dashboard View 64
Figure 78: Ambari Web UI - HDFS Summary 64
Figure 79: Enable SSH access on NSG 65
Figure 80: Firefox - SOCKS5 Configuration 67
Figure 81: SOCKS5 Proxy Enabled - Whatismyip.com verification 68
Figure 82: Ambari Web UI access using SSH-tunneling (port-forwarding) 69
Figure 83: HDInsight - Ambari - Namenode UI (Known Issue) 69
Figure 84: Ambari Web UI - Using SSH-tunnel - HTTP Error 401 70
Figure 85: HDInsight YARN Summary 73
Figure 86: HDInsight - Ranger Overview 73
Figure 87: HDInsight - Ranger - Hive Overview 74
Figure 88: HDInsight - Ranger - Hive Sample Policy 75
Figure 89: HDInsight Storage Container View using Storage Explorer for *hadoopadmin* 75

List of Tables

Table 1: hadoopadmins group RBAC roles Overview 47

1 Introduction

Enterprise readiness for Azure HDInsight clusters is supported by means of enabling Enterprise Security Package (ESP) as part of the cluster deployment procedure. This document presents a detailed process of setting up interdependent connected components such as enabling AAD-DS, setting up Storage Account, Configuring Access Controls, through to successful creation and validation of ESP enabled HDInsight cluster. References to additional information and relevant public product documentation are made available in the Appendix section.

Please note that the Azure Active Directory (AAD) tenant used here has the Microsoft controlled domain suffix - *.onmicrosoft.com. The domain name will be different depending on your AAD setup, where you can choose to have your fully owned routable public DNS Domain Name.

2 Setup Azure Active Directory – Domain Services (AAD-DS)

A secure HDInsight cluster will depend on Azure Active Directory – Domain Services (AAD-DS) for identity services. For example – authentication. Once enabled there will be one AAD-DS instance per region, and it runs inside a VM instance that offers DNS services.

2.1 Create AAD-DS Resource

Follow the below steps to setup Azure Active Directory – Domain Services resource:

- Select Azure AD Domain Services resource from the marketplace.
- Sample create configuration below (note, the Name attribute value is your AAD-DS domain name, that'd be referenced in all the subsequent instruction set; search for *youraaddsdomainname* string in the following sections of this notes.):

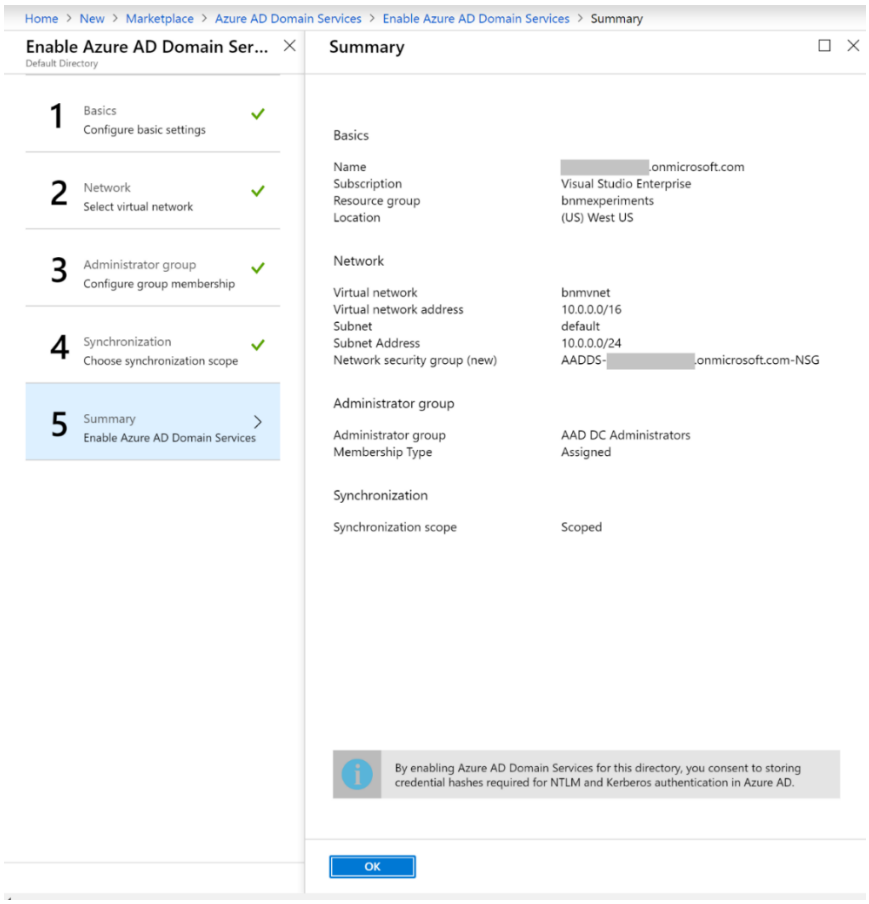


Figure 1: Enable Azure Active Directory - Domain Services

2.2 Setup AAD-DS - Scoped Synchronization

AAD-DS synchronizes its identity database with Azure AD. You can choose to synchronize just about everything (the default process) in the AD or choose scoped synchronization. Difference is that if you opt for the defaults the synchronization process may potentially consume significant amount of time. Rather if you are only interested in certain groups and users you can limit the synchronization to just to those entities.

Please note that Users and User Groups that you desire to have access to your HDInsight cluster, must be selected as part of this AAD-DS synchronization process. Else, HDInsight cluster deployment will fail with terminal state error – GroupNotFoundInActiveDirectory. Following is a sample error message for immediate reference:

```
"{\"status\":\"Failed\",\"error\":{\"code\":\"ResourceDeploymentFailure\",\"message\":\"The
resource operation completed with terminal provisioning state
'Failed'.\",\"details\":[{\"code\":\"GroupNotFoundInActiveDirectory\",\"message\":\"Group
IngressUsers not found in the Active Directory.\"}]}}"
```

Following visual provides a reference to the portal pane from where additional users and user-groups can be added to the AAD-DS synchronization scope:

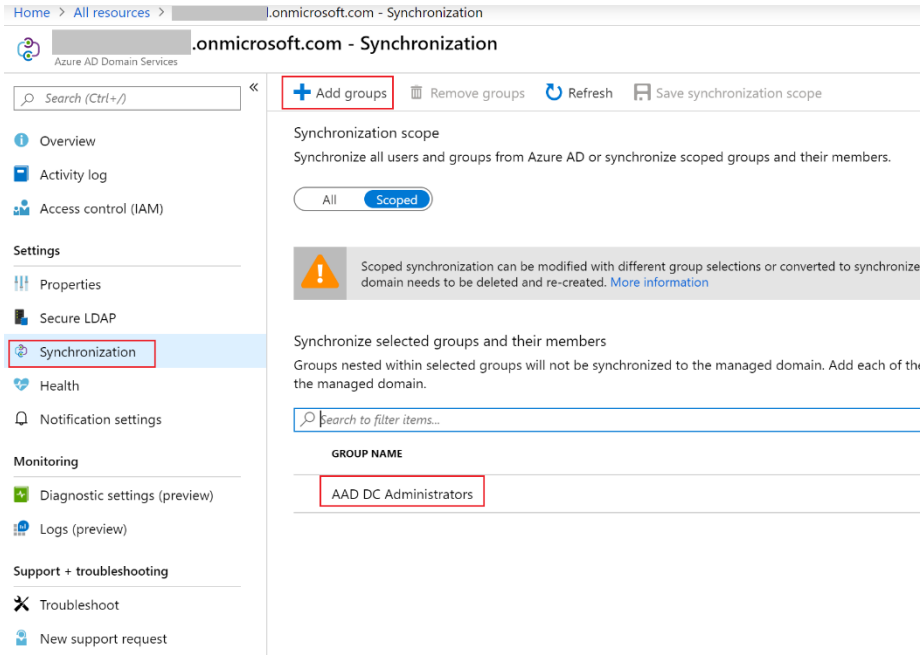


Figure 2: Configure AAD-DS Synchronization Scope - Add Groups

On the similar screen of yours, click on the + *Add groups* link as highlighted in the visual above.

Choose the groups and or users that must be granted the access, as shown in the following reference visual:

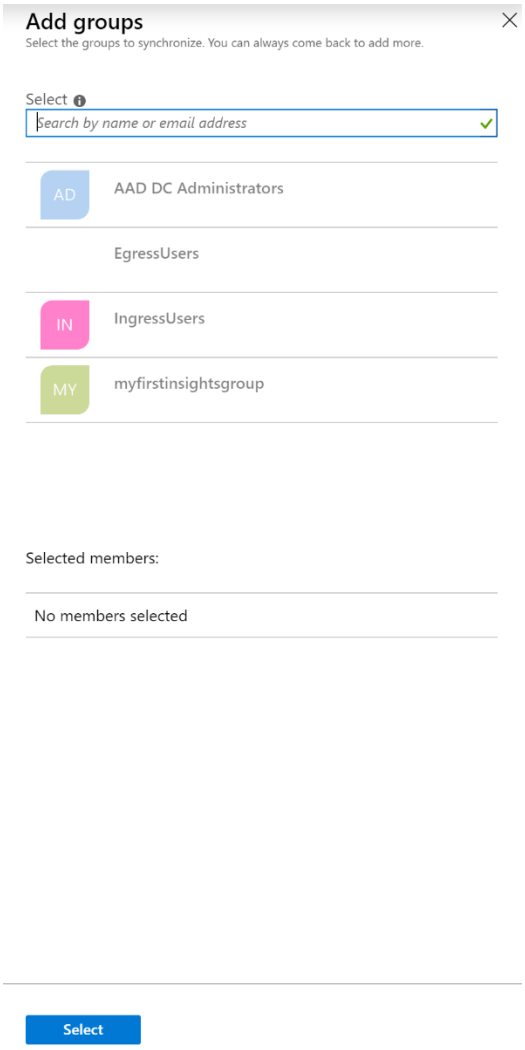


Figure 3: AAD-DS Add Group - Group Selection

Once selected, saved the selections will get added to the synchronization scope as shown below:

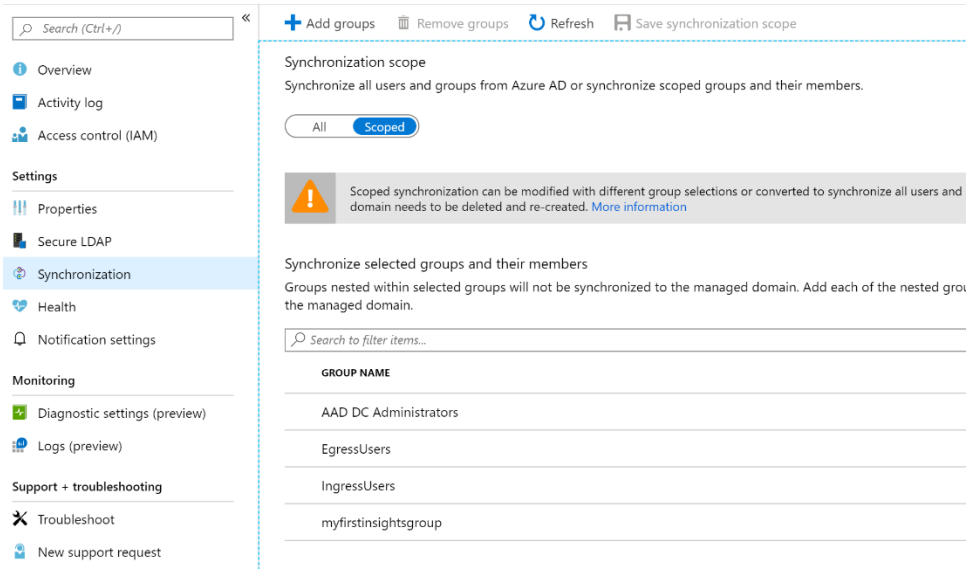


Figure 4: Selected AAD-DS Groups for Synchronization scope

2.3 Generate & Configure Self-signed Certificates

- First step is to define a self-signed certificate using the powershell command below:

```
$lifetime=Get-Date
New-SelfSignedCertificate -Subject youraaddsdomainname.onmicrosoft.com `
  -NotAfter $lifetime.AddDays(365) -KeyUsage DigitalSignature,
KeyEncipherment `
  -Type SSLServerAuthentication -DnsName
*.youraaddsdomainname.onmicrosoft.com, youraaddsdomainname.onmicrosoft.com
```

- Warning note about how you'd launch Powershell, if you run as a normal user i.e., non-administrator and attempt to run above command, you'd end up with following error:

```
PS C:\Users\yourid> $lifetime=Get-Date
PS C:\Users\yourid> New-SelfSignedCertificate -Subject
youraaddsdomainname.onmicrosoft.com `
>>   -NotAfter $lifetime.AddDays(365) -KeyUsage DigitalSignature,
KeyEncipherment `
>>   -Type SSLServerAuthentication -DnsName
*.youraaddsdomainname.onmicrosoft.com, youraaddsdomainname.onmicrosoft.com
New-SelfSignedCertificate : CertEnroll::CX509Enrollment::_CreateRequest:
Access denied. 0x80090010 (-2146893808
NTE_PERM)
At line:1 char:1
+ New-SelfSignedCertificate -Subject youraaddsdomainname.onmicrosoft.com `
+ ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
    + CategoryInfo          : NotSpecified: (:) [New-SelfSignedCertificate],
Exception
    + FullyQualifiedErrorId :
System.Exception,Microsoft.CertificateServices.Commands.NewSelfSignedCertific
ateCommand
```

- Launch Powershell with the option Run as Administrator. Here's the sample command response on the console output:

```
PS C:\WINDOWS\system32> $lifetime=Get-Date
PS C:\WINDOWS\system32> New-SelfSignedCertificate -Subject
youraaddsdomainname.onmicrosoft.com `
>>   -NotAfter $lifetime.AddDays(365) -KeyUsage DigitalSignature,
KeyEncipherment `
>>   -Type SSLServerAuthentication -DnsName
*.youraaddsdomainname.onmicrosoft.com, youraaddsdomainname.onmicrosoft.com

   PSParentPath: Microsoft.PowerShell.Security\Certificate::LocalMachine\MY

Thumbprint                                Subject
----------                                -------
3F6F8XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX  CN=youraaddsdomainname.onmicrosoft.com
```

- Subsequently, follow the instructions from here to export the certificate files – PFX and CER. While you export the certificate (PFX) you'd be prompted to password protect the file. Please take a note of the password. It will be used when configuring the same in the portal.
  - Complete necessary steps through to exporting the certificate file (.PFX extension) for Azure Active Directory Domain Services and .CER file for client computers.
- Upload and configure the PFX file to the Secure LDAP settings as shown in the following visual. Click here for official documentation.

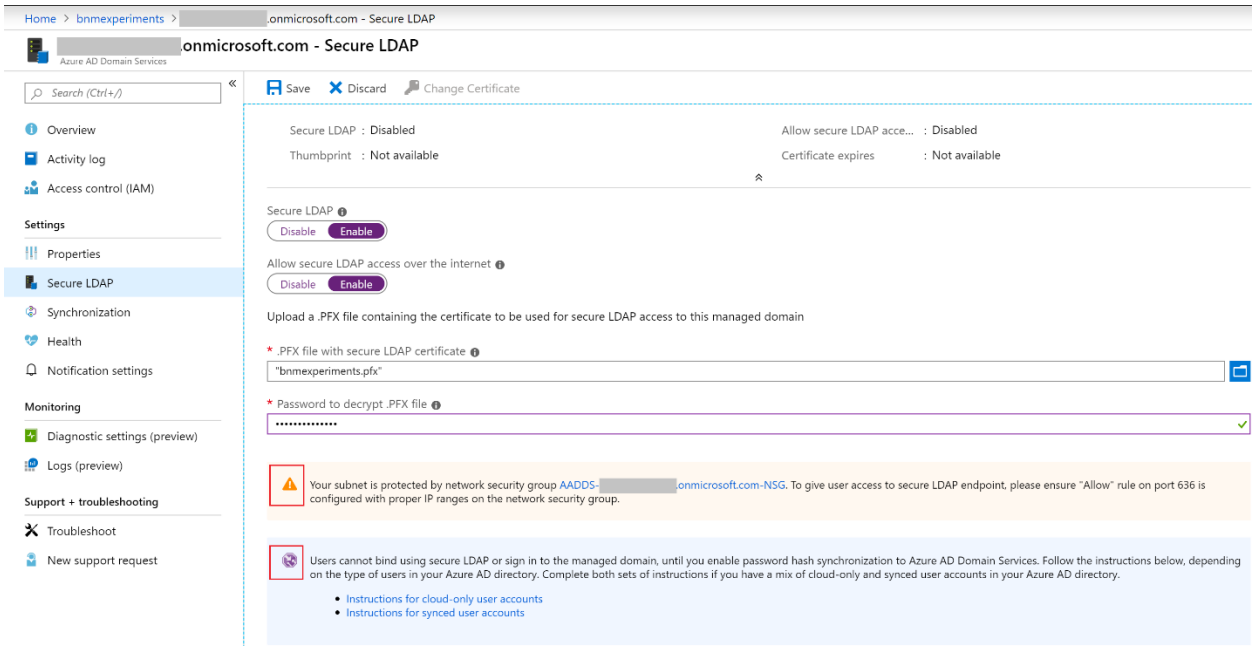


Figure 5: Azure AD Domain Services - Configure Certificate (PFX file)

- Pay attention to those in red boxes. Look forward to section on configuring NSG rules further down in this article. For now, hit Save to secure LDAP.
- Once enabled… here's what it should look like

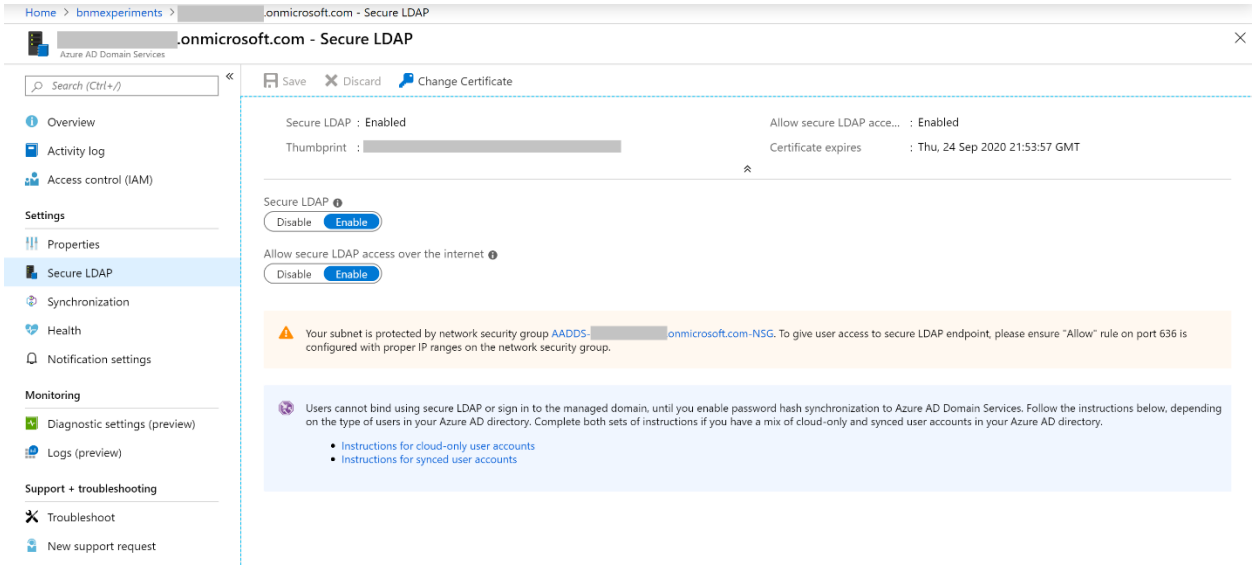


Figure 6: Azure AD Domain Services - Secure LDAP Enabled

- As a cautionary measure, verify the health of AAD-DS. Note the warning below. It will go away once you configure the NSG rules in the next section.

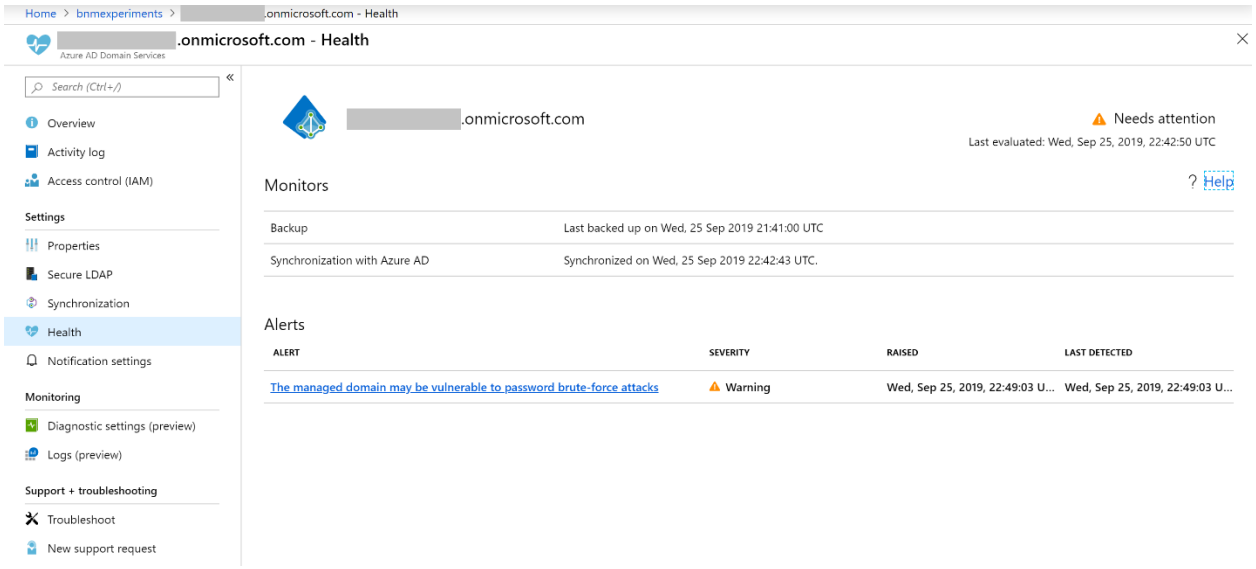


Figure 7: Azure AD Domain Services - Secure LDAP Brute-force Warning

3 Configure Networking

Following visual offers networking inter-dependencies between various moving parts of the HDInsight cluster. It helps in planning the HDInsight virtual network configurations. Click here to access the official documentation reference.

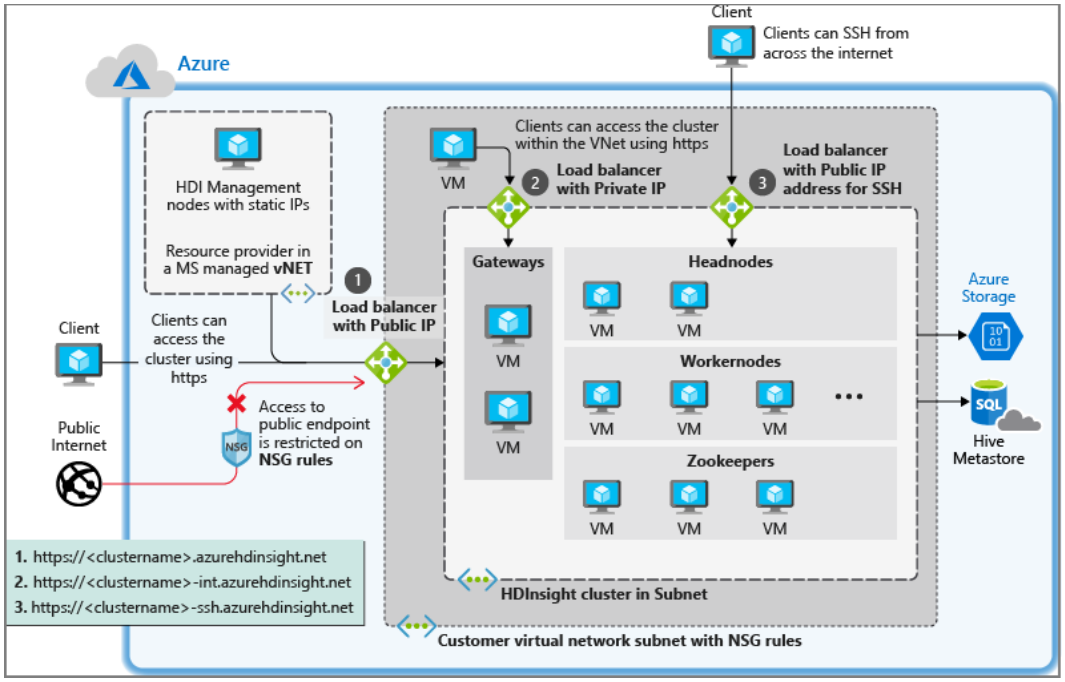
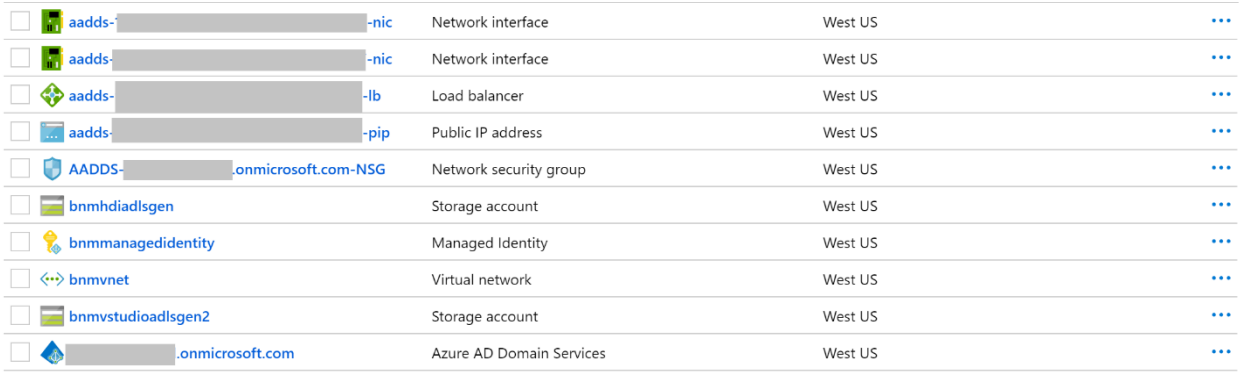


Figure 8: HDInsight Networking Interdependencies Overview

3.1 Configure NSG Rules to secure LDAP

To lock down and secure LDAP access over the Internet, in this section we will go through the sequence of steps to configure NSG rules and secure access to LDAP. Once these steps are successfully completed, users seeking access to your LDAP services on port 636 will be enabled. The warning notification that is displayed at the time of certification configuration will be gone. Follow the sequence below:

- Go to NSG resource from the Resource Group's resources list. Naming convention for the NSG resource will have a prefix AADS and suffix NSG as shown in the visual below:



aadds-[redacted]-nic	Network interface	West US
aadds-[redacted]-nic	Network interface	West US
aadds-[redacted]-lb	Load balancer	West US
aadds-[redacted]-pip	Public IP address	West US
AADDS-[redacted].onmicrosoft.com-NSG	Network security group	West US
bnmhdiadlsgen	Storage account	West US
bnmmanagedidentity	Managed Identity	West US
bnmvnet	Virtual network	West US
bnmvstudioadlsgen2	Storage account	West US
[redacted].onmicrosoft.com	Azure AD Domain Services	West US

Figure 9: Azure Active Directory - Domain Services Resource (from the listing)

- Select the NSG resource. Here's the reference visual of the overview section:

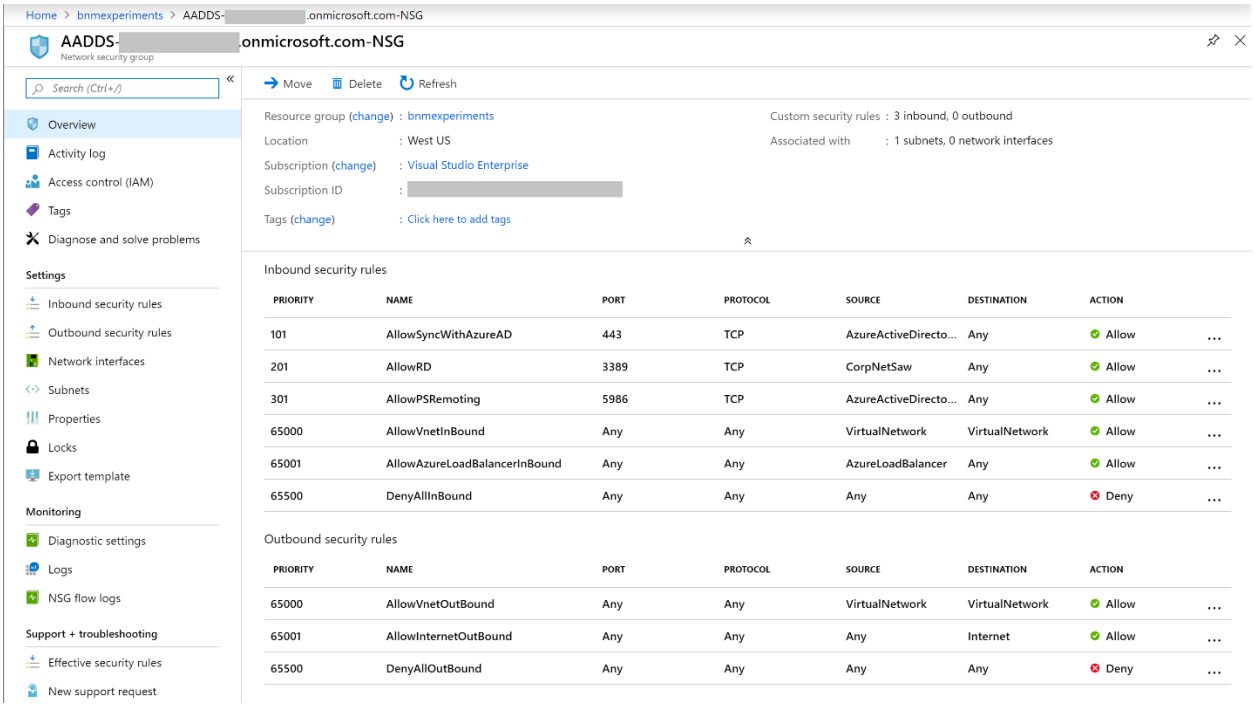


Figure 10: Azure AD Domain Services NSG Rules (Inbound/Outbound)

- Add Inbound security rule to enable traffic on port 636 for secure LDAP

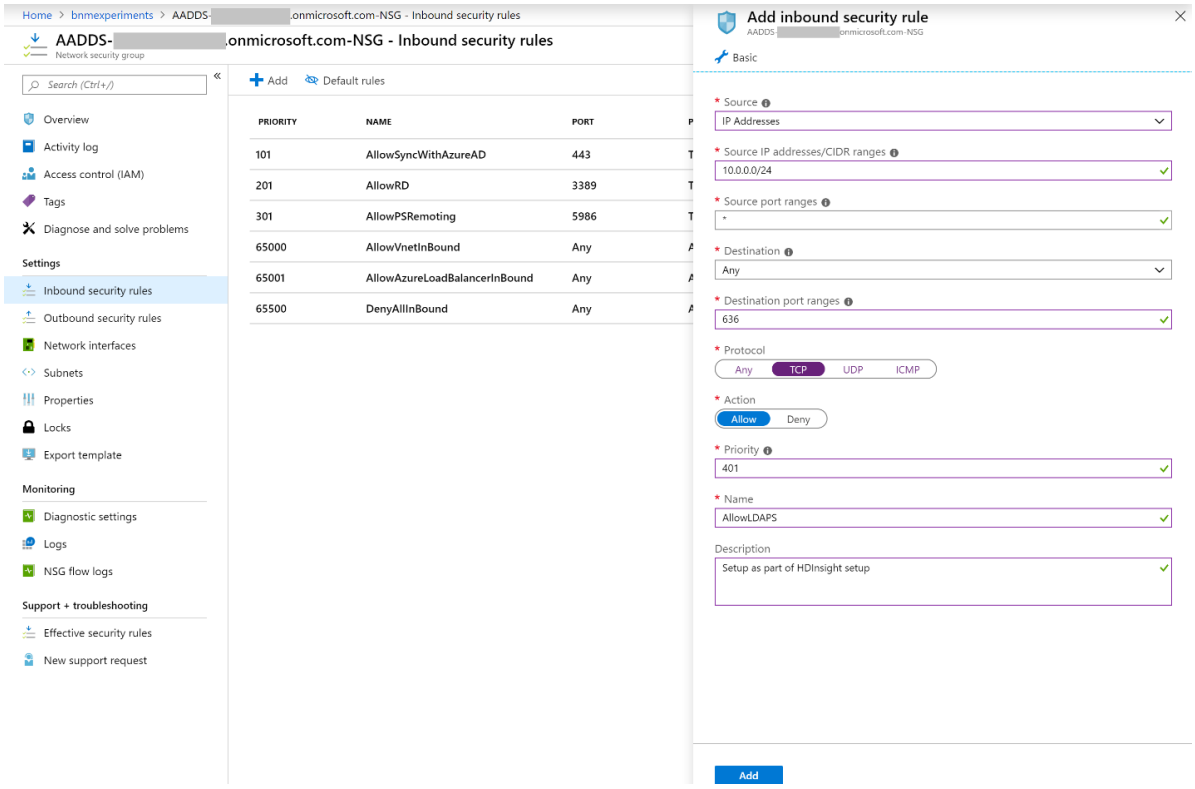


Figure 11: Azure AD Domain Services NSG Rules - Add Inbound Security Rule

- Once you successfully have added the rule, the rules list should appear like below

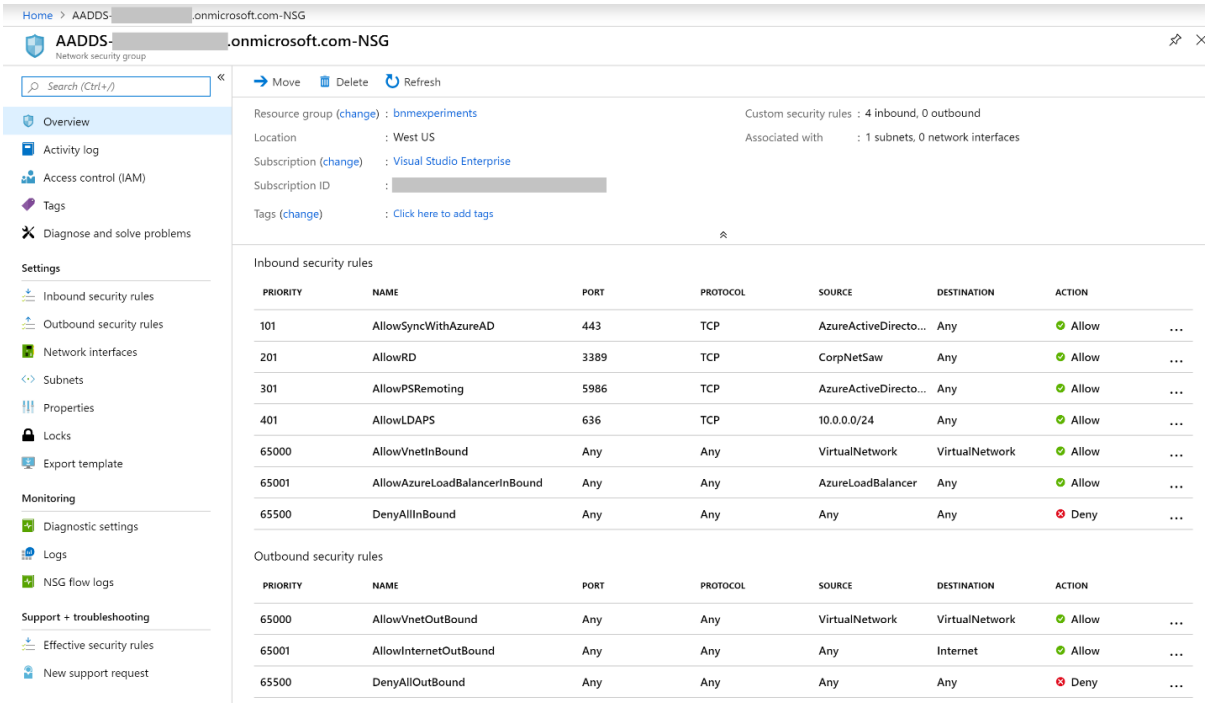


Figure 12: Azure AD Domain Services NSG Rules - Updated

3.2 Configure DNS Zone for External Access

- DNS Zone entry must be configured on the client resource that intends to interact with the AAD-DS resource. Here's the link to the official documentation.
- Configure your external DNS provider to create a host record, such as ldaps, to resolve to the secure external IP address available from the AAD-DS properties blade entry - Secure LDAP external IP address. In this case the IP address resolves to xxx.xxx.xxx.xxx.

```
xxx.xxx.xxx.xxx        youraaddsdomainname.onmicrosoft.com
```

- Test Queries for the Managed Domain (Note - Work in progress sequence below; needs further NSG configurations to allow remote ldp.exe to be able to reach our LDAPS on the AAD-DS):
  - Install the Remote Server Administrator tools on your local windows machine, following the instruction detail from here.
  - Launch Windows Power-shell as an administrator (i.e., Run as Administrator).
  - And, execute the ldp.exe command as shown in the following visual:

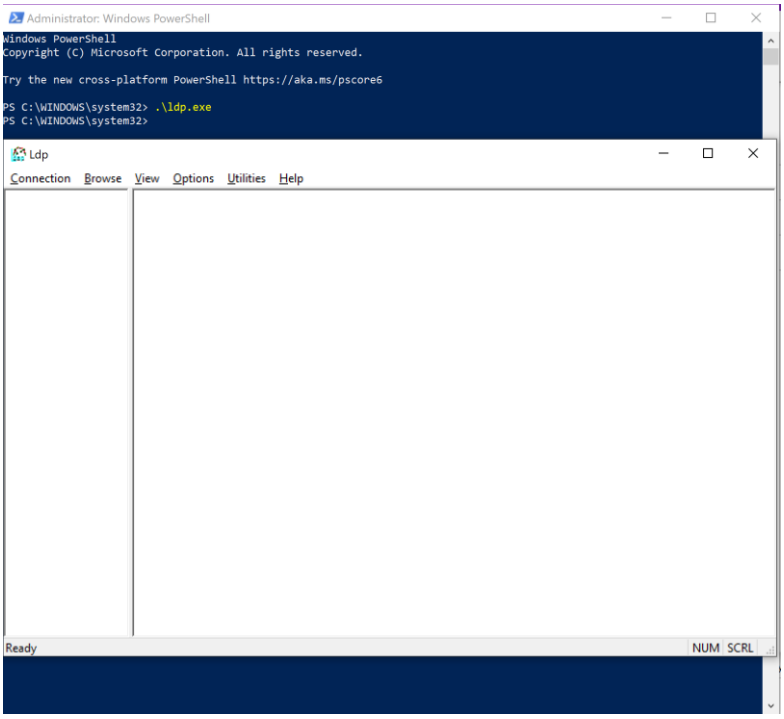


Figure 13: ldp.exe client UI

- It is possible that you'd encounter similar error when attempting to connect and bind with the DNS server using the external IP address as found in the AAD-DS properties.

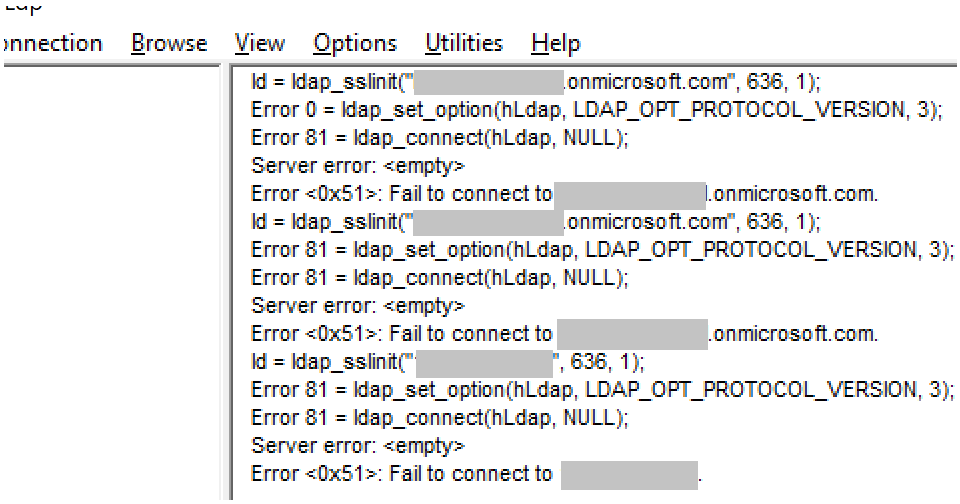


Figure 14: ldp.exe - connect error sample

- Detailed instructions of how to accomplish the steps is accessible from the official documentation section here.

3.3 Configure DNS Server Settings & Enable AAD-DS Password Hash synchronization

For password has synchronization click here for the official documentation. This part of the process is essential for hybrid environment setups and is out of scope for this document.

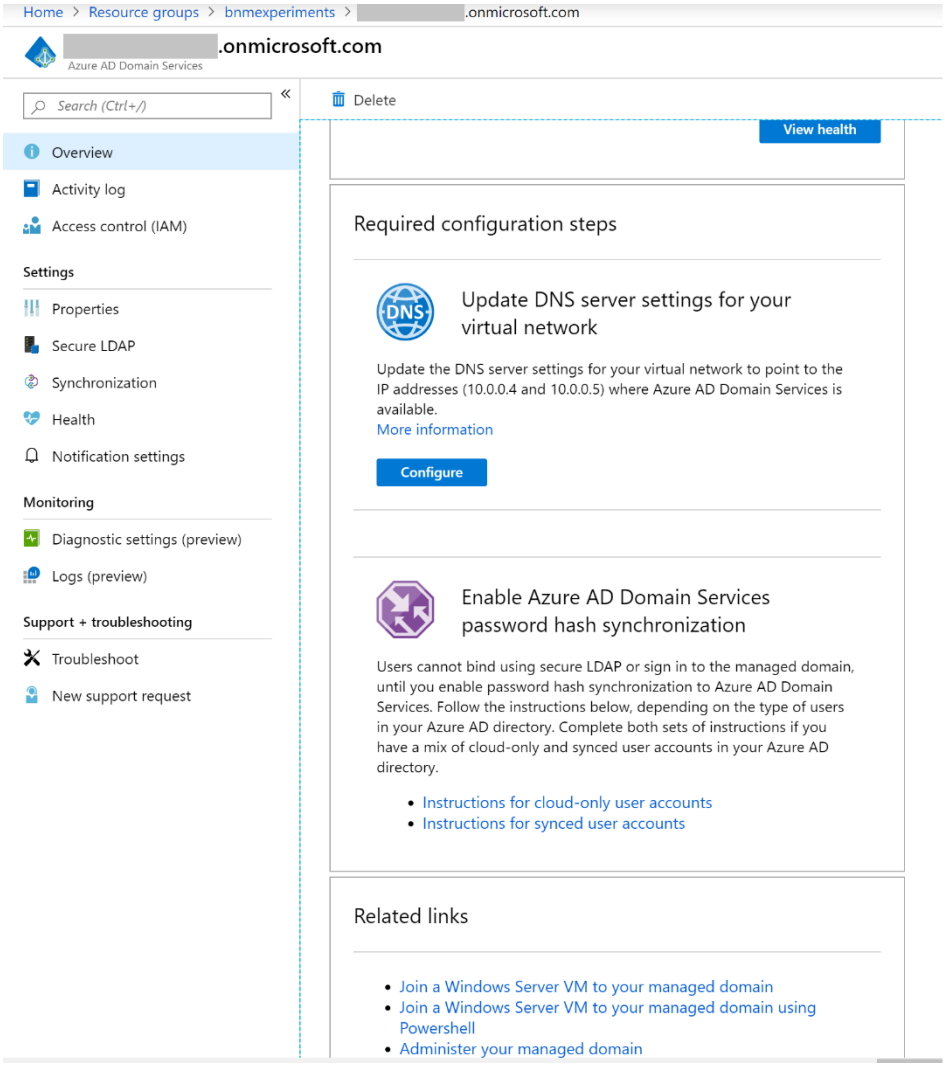


Figure 15: Azure AD Domain Services - Required Configuration Steps

3.4 Update DNS Server Settings for your Virtual Network

Getting back to the DNS configuration step, it is important to update the IP addresses as shown under the *Required configuration next steps* section (see the previous visual above). This configuration will allow dependent services like HDInsight Ranger to have a line of sight to your directory services. And such access is verified during the cluster setup process. Failure to do so will result in a terminal provisioning state failure error. Following is the sample of such error message:

```
{"code":"DeploymentFailed","message":"At least one resource deployment operation failed. Please list deployment operations for details. Please see https://aka.ms/arm-debug for usage details.","details":[{"code":"Conflict","message":"{\r\n \"status\": \"Failed\",\r\n \"error\": {\r\n \"code\": \"ResourceDeploymentFailure\",\r\n \"message\": \"The resource operation completed with terminal provisioning state 'Failed'.\",\r\n \"details\": [\r\n {\r\n \"code\": \"DomainNotFoundInActiveDirectory\",\r\n \"message\": \"Domain youradddsdomainname.onmicrosoft.com not found in the Active Directory.\"\r\n }\r\n ]\r\n }\r\n}"}]}
```


Here's the visual reference that details the deployment status:

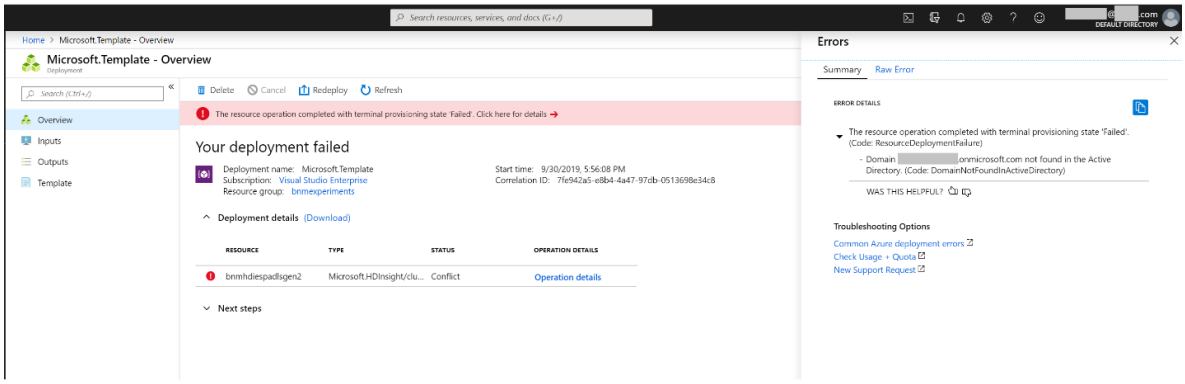


Figure 16: HDInsight Cluster Deployment Failure - DomainNotFoundInActiveDirectory

Following visual shows the inconsistent state of your cluster deployment process, where the necessary resources are deployed, but are not usable due to above failure condition:

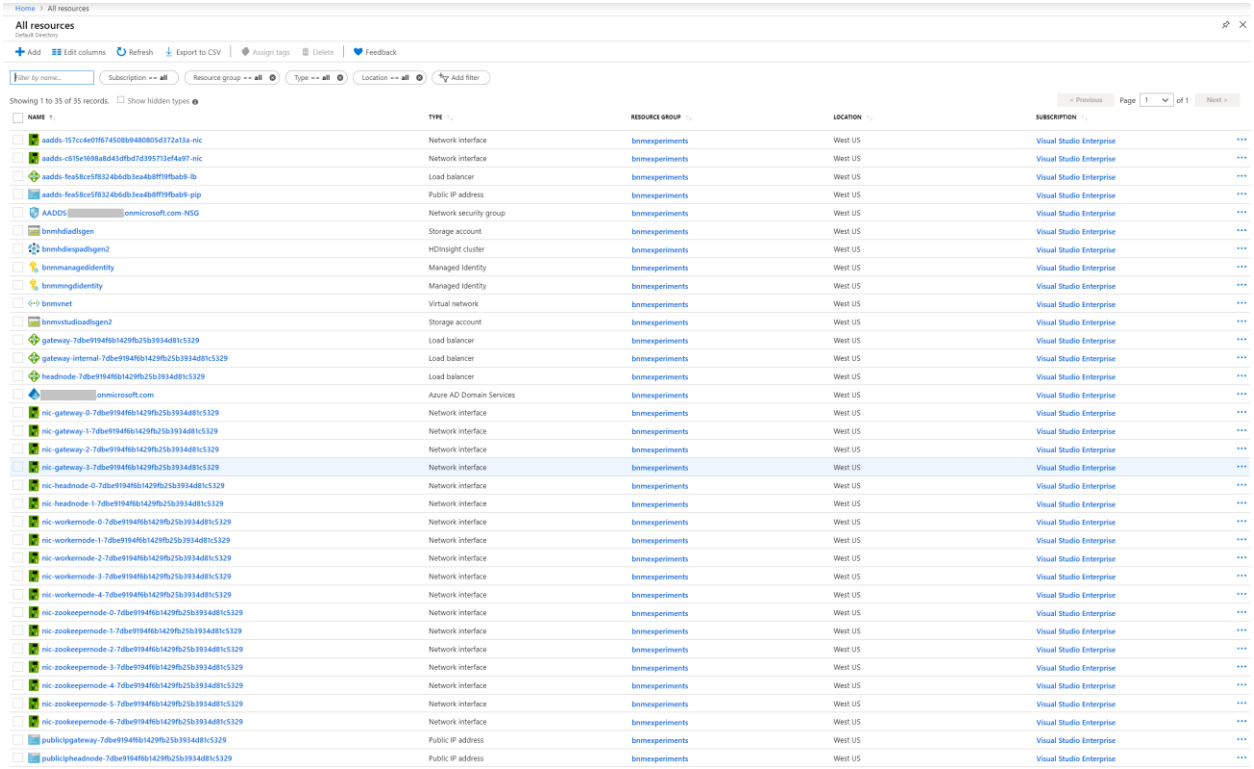


Figure 17: HDInsight Deployment - Set of Resources Deployed!

Remediation steps or the steps that must be performed as part of AAD-DS setup sequence are as follows:

- In the AAD-DS resource properties select the *Available in virtual network/subnet* link as shown below:

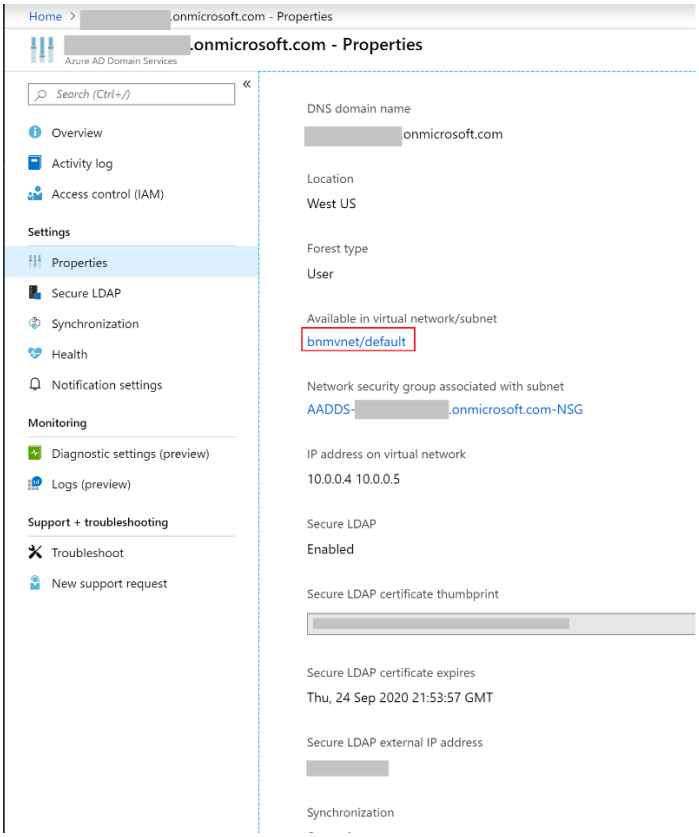


Figure 18: Azure AD Domain Services - Properties

- As shown in the following visual reference, instead of Azure-provided choice, select Custom and enter both the IP addresses from the previous step.

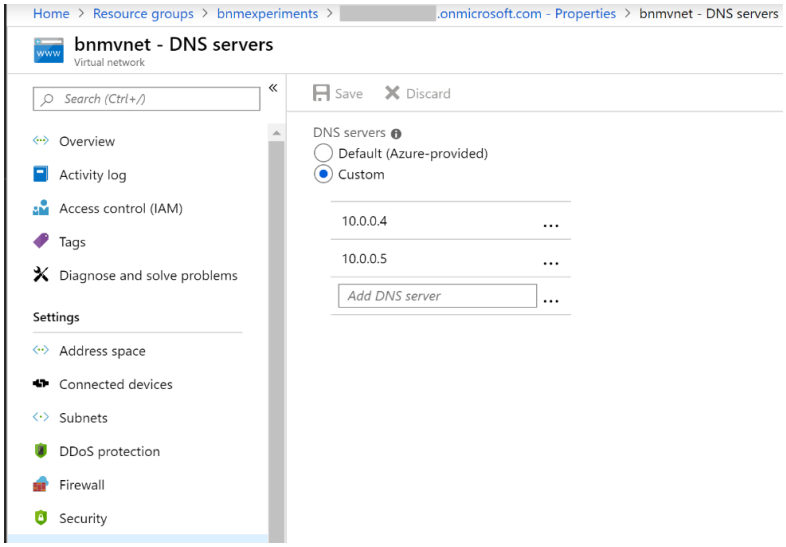


Figure 19: Azure AD Domain Services - VNET Properties - DNS IP address updates

3.5 Configure NSG Rules to support HDInsight Management & Health services

By default, access to HDInsight cluster components & services is disabled. Based on your requirements and requirements of the Azure management and health services certain IP and Ports must be enabled and configured using the NSG rules. In this section will go through the process of enabling essential NSG rules.

It is important to note that configurations applied here assume that you are going to leverage Azure-provided DNS service. Click here to access more on this topic from the official documentation. In this section will enable Azure Management and Health service IP address and ports on the NSG rules. *Skipping this step will result in cluster deployment failures*. Once the cluster setup is complete, we can come back and configure a rule to enable traffic on destination port 443 within the VNET IP address range (CIDR). This will enable secure HTTP access to cluster services. Click here to learn more about ports used by Apache Hadoop services on HDInsight).

As a next step follow the sequence below:

- In the Azure portal, navigate to the NSG resource. You can either use All Resources link in the left navigation pane of the portal or go to your resource group and select the NSG resource. Here's an example that uses the all resources view:

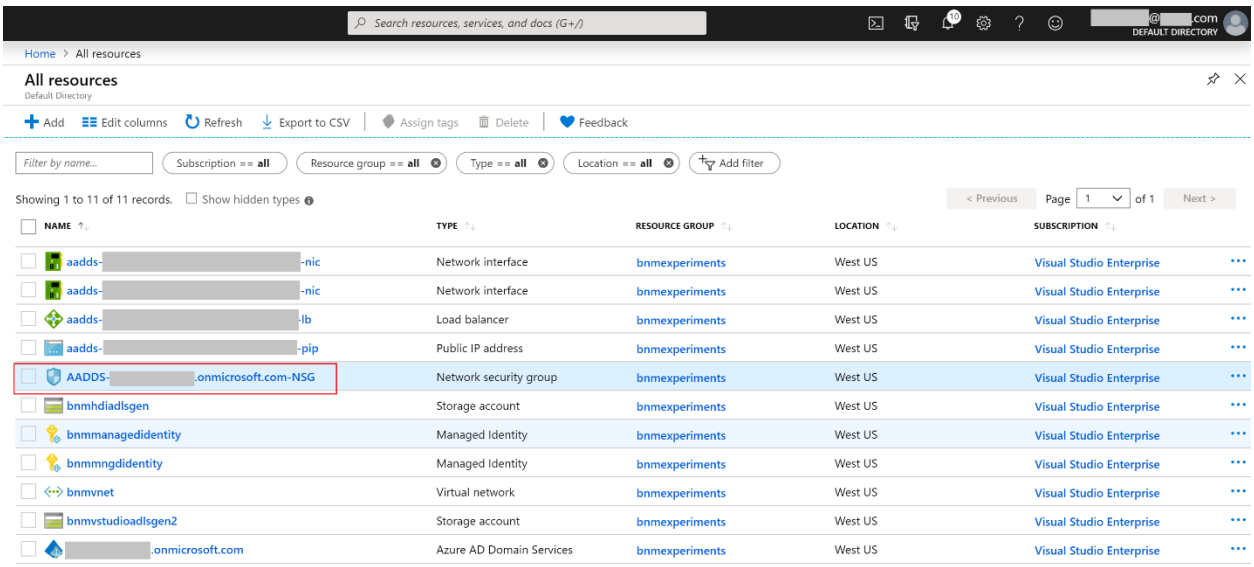


Figure 20: Subscription/All Resources View - AAD-DS NSG Resource

- You can reference the required HDInsight management and health service IP addresses and ports from the official documentation link here.

- You must configure IP address and ports that apply to all regions as well as those that are specific to a certain region where your HDInsight resource will be deployed.
- Following visual provides a reference example of all the necessary addresses and ports that are enabled in the NSG inbound rule records:

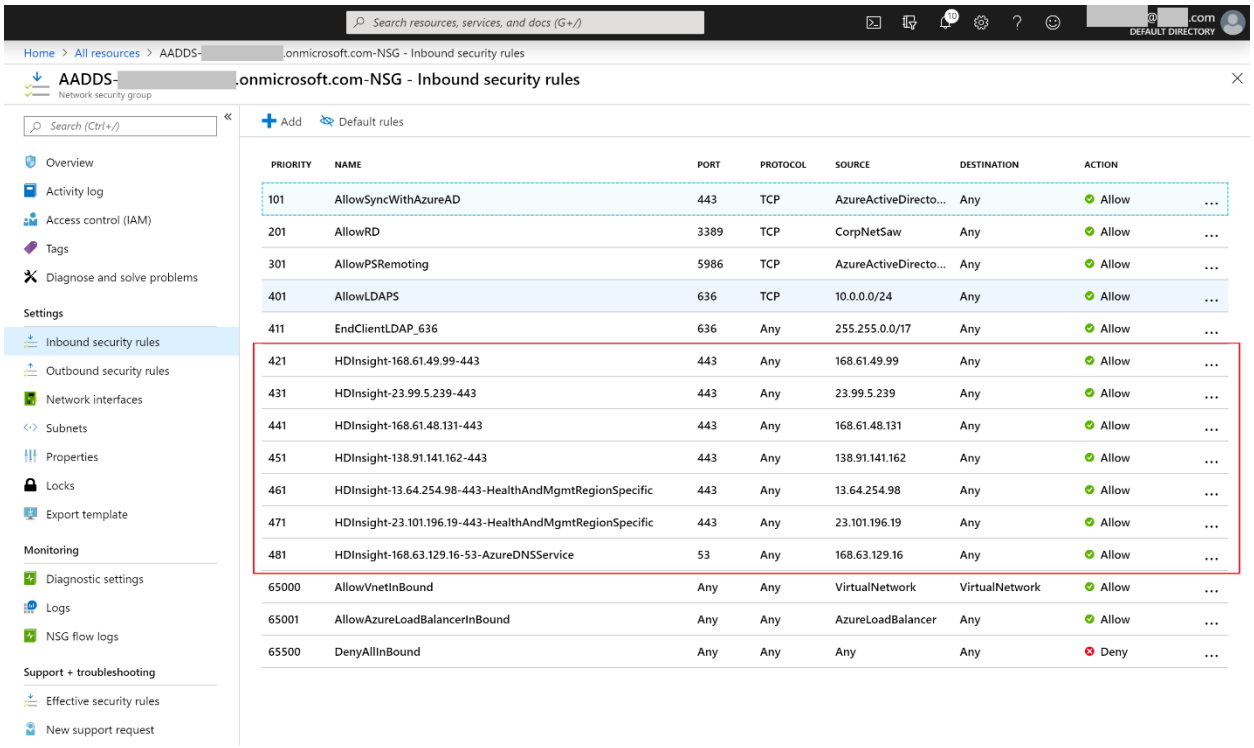


Figure 21: Configured NSG Rules for HDInsight Management & Health Service IP Addresses

4 Configure Access Controls

4.1 Setup User Assigned Managed Identity

User assigned managed identity that we create here will be configured as part of the HDI cluster setup. It will allow HDI services to authenticate and authorize with the ADLS Gen2 storage account, and offer HDFS service to its own components like Hive, etc., as well to the users who'd like to interact with HDI cluster's HDFS service. There are caveats to watch when interacting with HDFS, in comparison to how open source would work. Here, all storage interactions are delegated to the ADLS Gen2 service. Hence, configurations like authentication and authorization must be provisioned following ADLS Gen2 recommended methods. See more about ADLS Gen2 here.

Follow the below sequence to create a User Assigned Managed Identity resource:

- From the portal, click on + Create Resource link on the left navigation and search based on the string - User Assigned Managed Identity, as in the following visual:

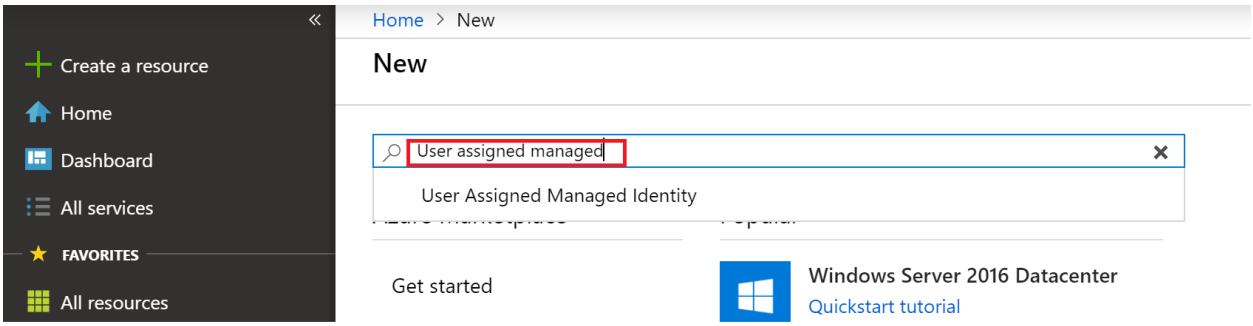


Figure 22: Search for User Assigned Managed Identity Resource type

- In the subsequent view, create the resource by clicking on the Create button.

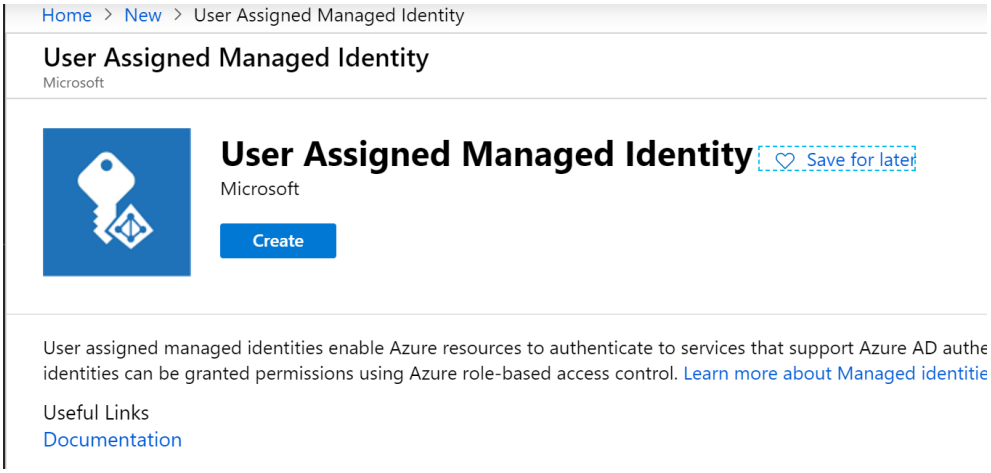


Figure 23: Create User Assigned Managed Identity

- Then, input all the necessary configuration detail as shown in the following visual:

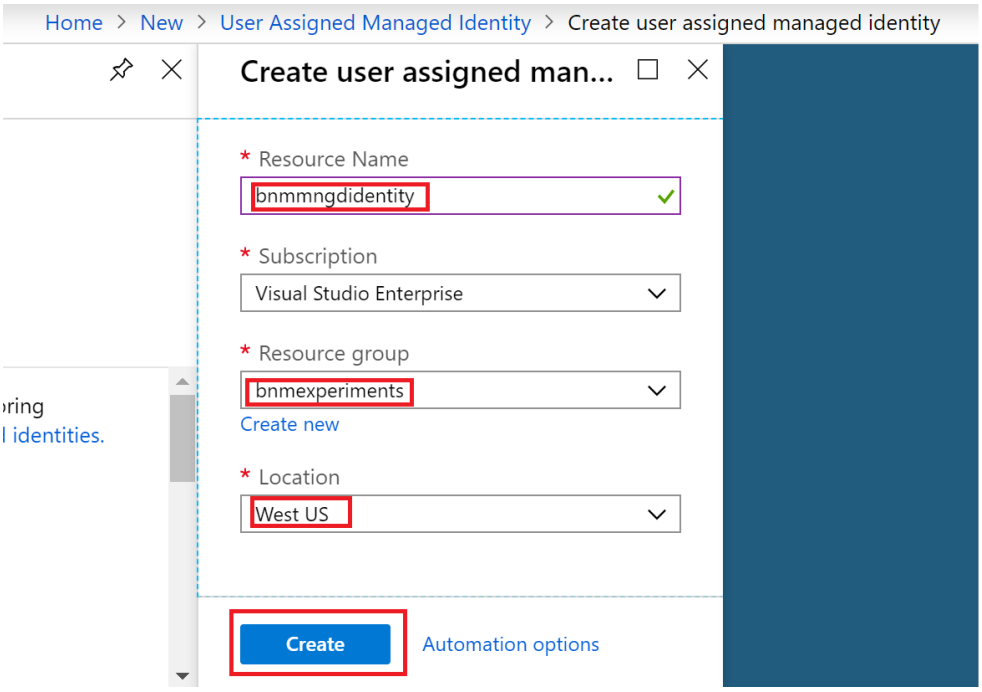


Figure 24: User Assigned Managed Identity - Create Form

- Note - for illustration, I have used a slightly different resource name. In the subsequent section about ADLS Gen2, the name of the resource that is used (and created for the purpose following same instructions) will be *bnmmanagedidentity*.
- You can verify the details of the managed identity from the all resources pane, as shown in the following visual:

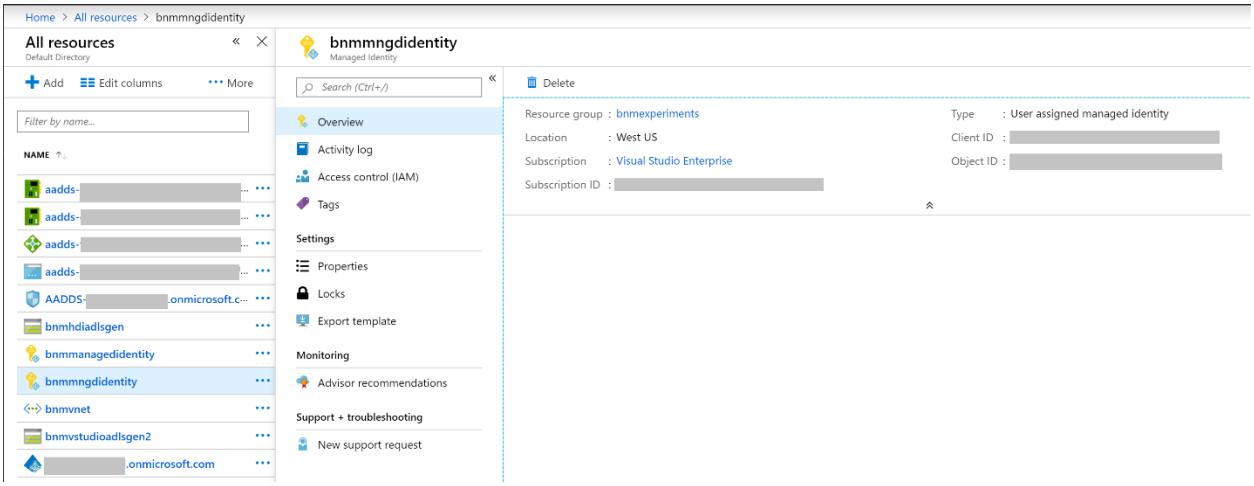


Figure 25: User Assigned Managed Identity - Blade View/Overview

4.2 Create & Configure *hadoopadmins* AD Group

While HDInsight depends on a managed identity to represent it against storage layer access, we need similar administration handle to administer user level access controls to the storage. For example, in order for Hadoop users to be able to successfully query data, besides granting them permission to the respective tables, they should also be granted access to the storage path where the table data is maintained and managed. Hence, create an Azure Directory group called *hadoopadmins* using the Azure Active Directory pane for Groups.

Following visual provides a reference to the AD Groups pane (note, the default directory here is the name of the Active Directory):

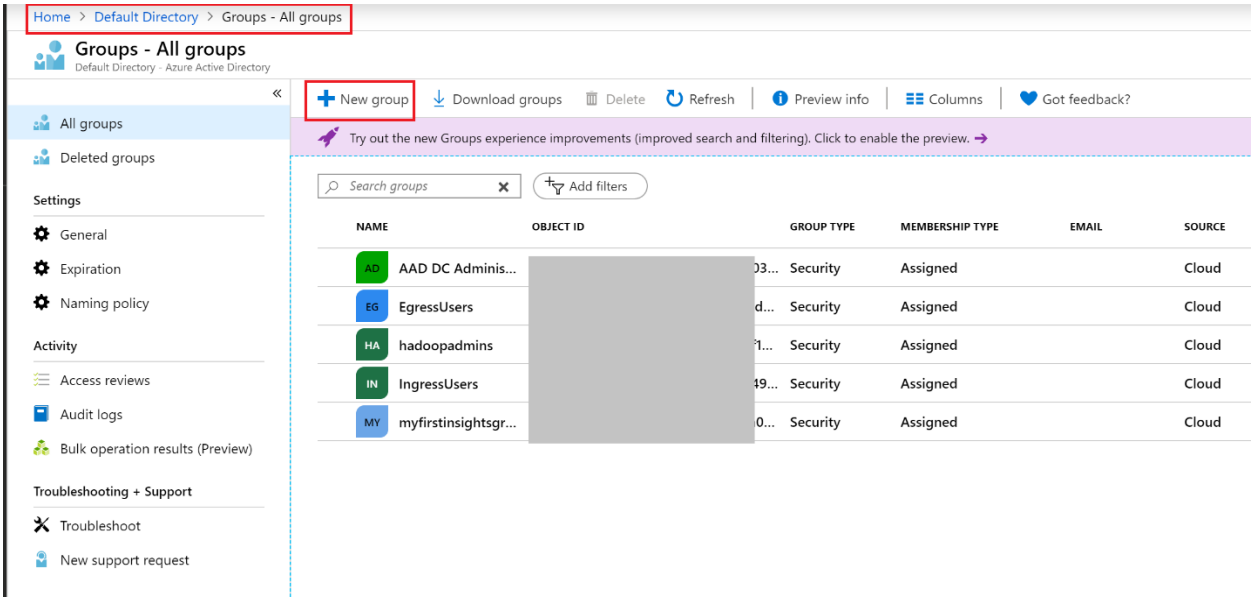


Figure 26: Azure AD - Groups - All Groups view

As shown in the sample reference above, use the *New group* link to add the group *hadoopadmins*. In subsequent sections will add this group to the storage account with a role that allows the members of the group to administer access control to other users.

Next, we need to ensure the group and its members from Azure Active Directory are also synchronized to the Azure Active Directory – Domain Services (AAD-DS). To satisfy this requirement it is required to add the group *hadoopadmins* to the AD Synchronization scope. Navigate to the AAD-DS Synchronization pane and add the group as shown below:

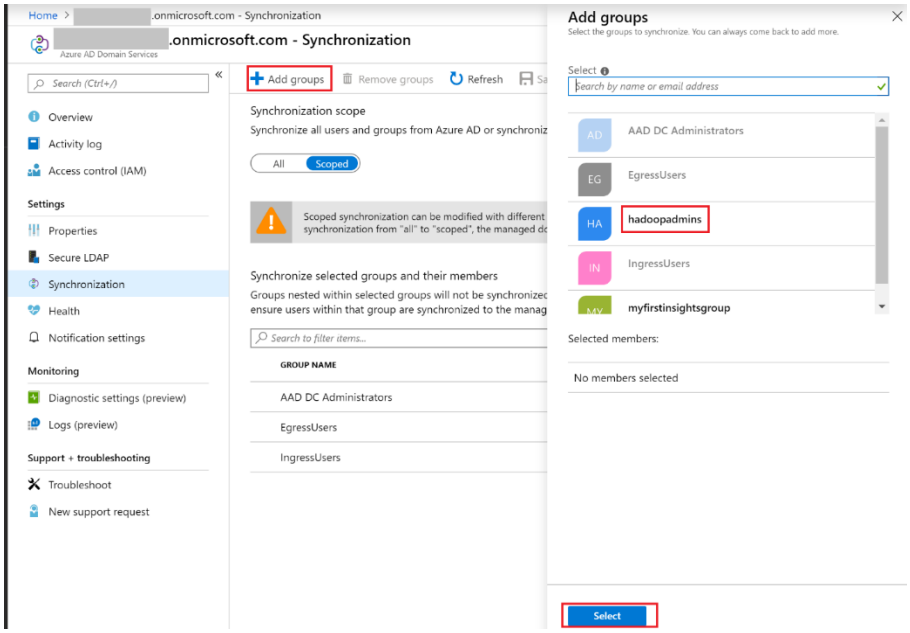


Figure 27: Azure AD Domain Services - Add Group to Synchronization scope

Once added, the group must be part of the list of groups that are enabled for AAD-DS synchronization scope:

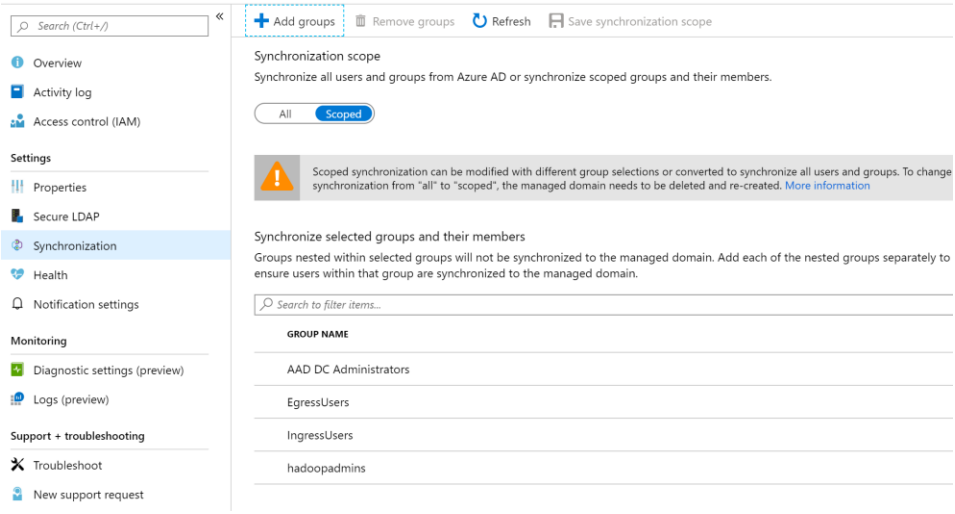


Figure 28: Updated Azure AD Domain Services - Synchronization Scope

4.3 Assign HDInsight Domain Services Contributor Role

HDInsight ESP cluster requires a managed identity to be configured, such that it can perform all necessary directory service interactions. HDInsight Domain Services Contributor role is a default role provided by AAD-DS integration to allow any assigned managed identity to perform HDInsight related operations with the AAD-DS. Click here to learn more on this topic from the official documentation. In this section we will assign the services contributor role to the managed identity that we have created in the previous section.

Here is the sequence of visuals that will guide you to configure the role to the managed identity:

- Navigate to the Access Control (IAM) pane of your AAD-DS resource, and configure the HDInsight Domain Services Contributor role, as shown in the following visual reference:

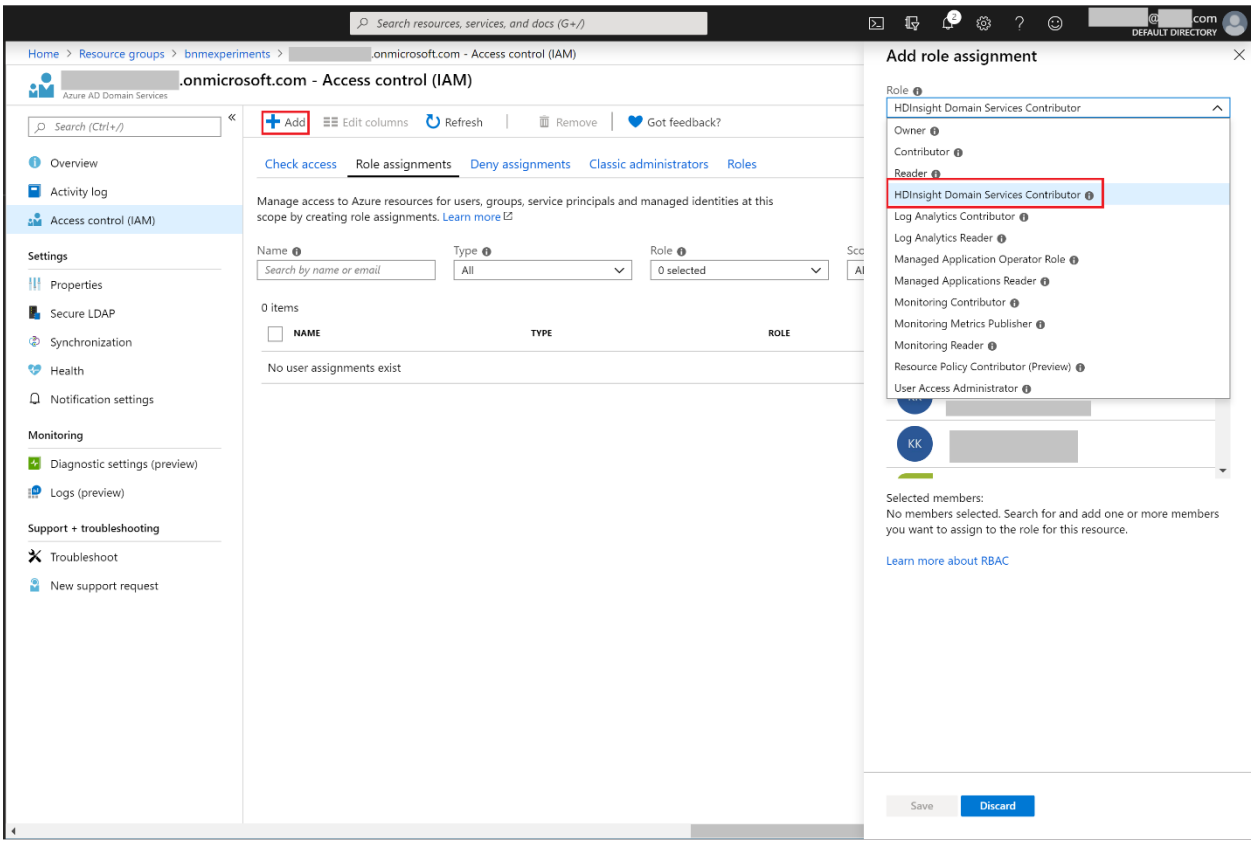


Figure 29: Assign HDInsight Domain Services Contributor role to User Assigned Managed Identity

- In the same view, once you have selected the role, next select the appropriate User assigned Managed Identity (example – *bnmmanagedidentity*) that you have defined in the prior sections. Following visual offers a reference to such selection:

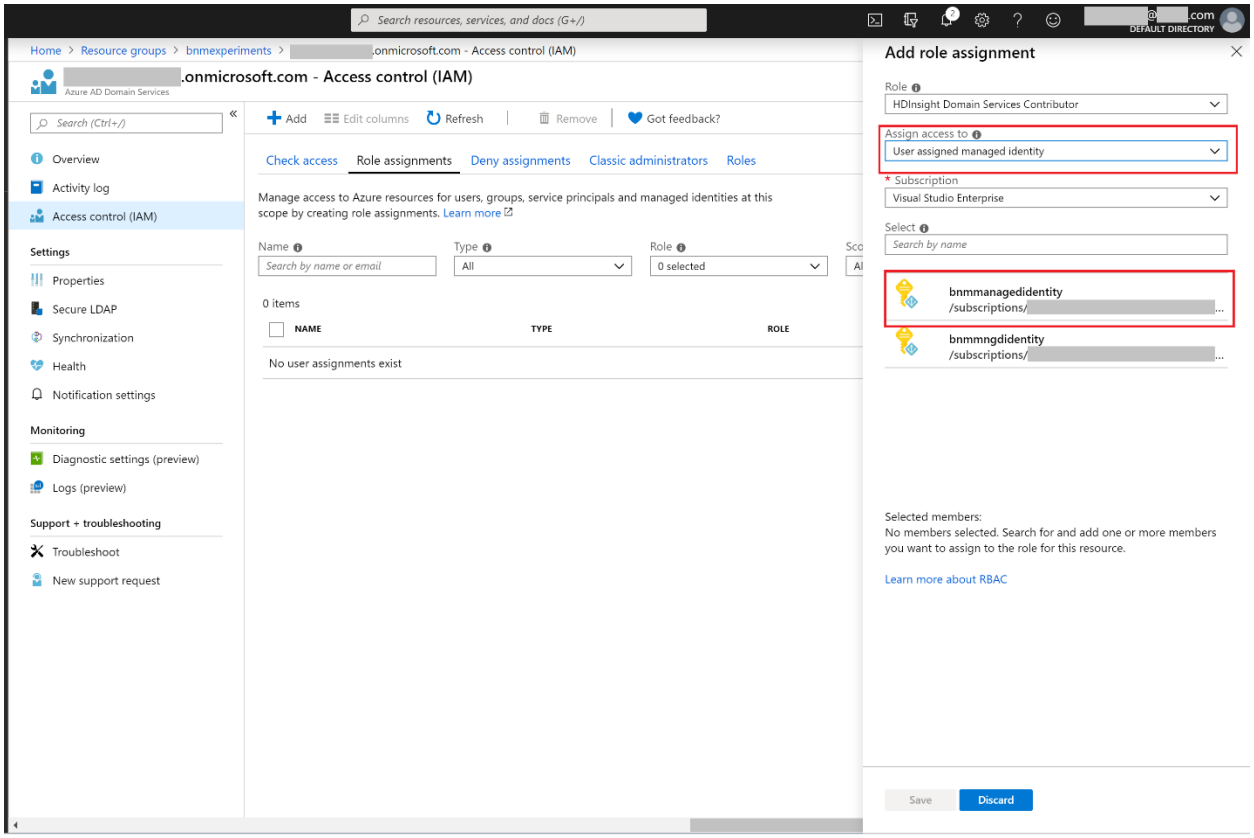


Figure 30: Select Managed Identity for the HDInsight Domain Services Contributor Role assignment

- Subsequently save the identity to role mapping configuration. The saved configuration will be shown as in the following visual:

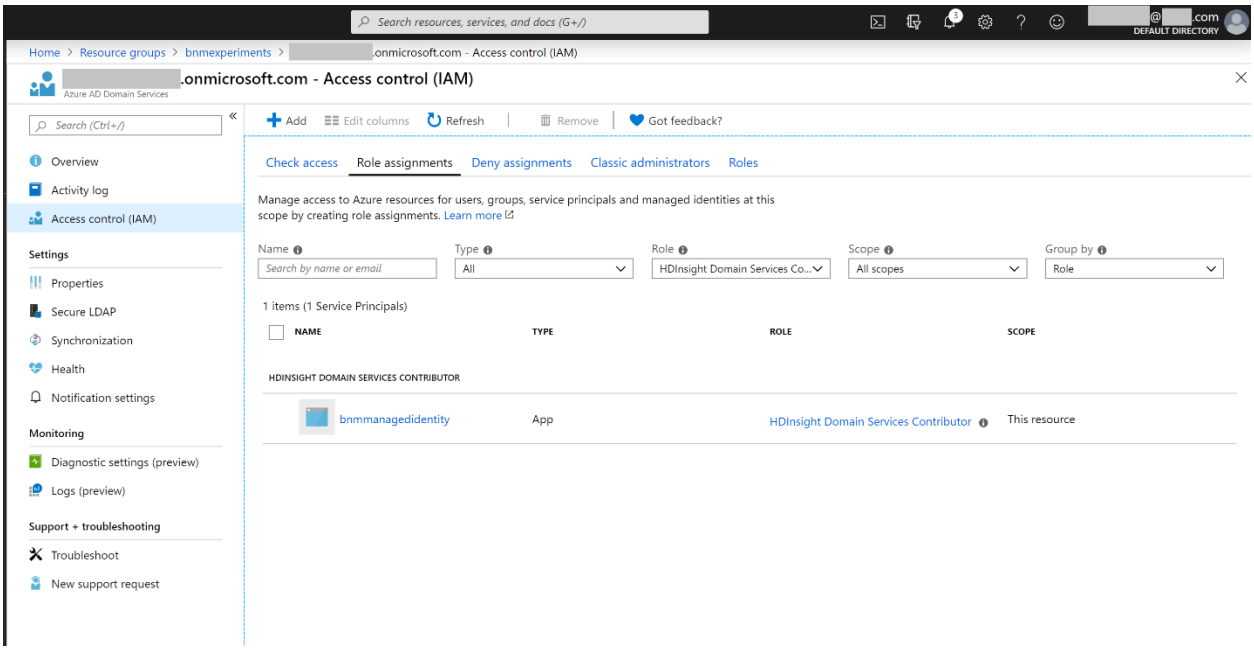


Figure 31: Updated HDInsight Domain Services Contributor Role Assignment

5 Create & Configure ADLS Gen2 Storage Account

ADLS Gen2 with support for Hierarchical Namespace (HNS) will be setup as a primary storage layer for the EPS enabled HDInsight cluster. In this section will go through the necessary steps to create an ADLS Gen2 resource and configure required administrator access controls.

5.1 Create ADLS Gen2 Storage Account

Using the Azure Portal interface, follow the steps detailed below to create an ADLS Gen2 Storage Account:

- Assuming you are logged in to Azure Portal, navigate to the Storage Accounts blade.

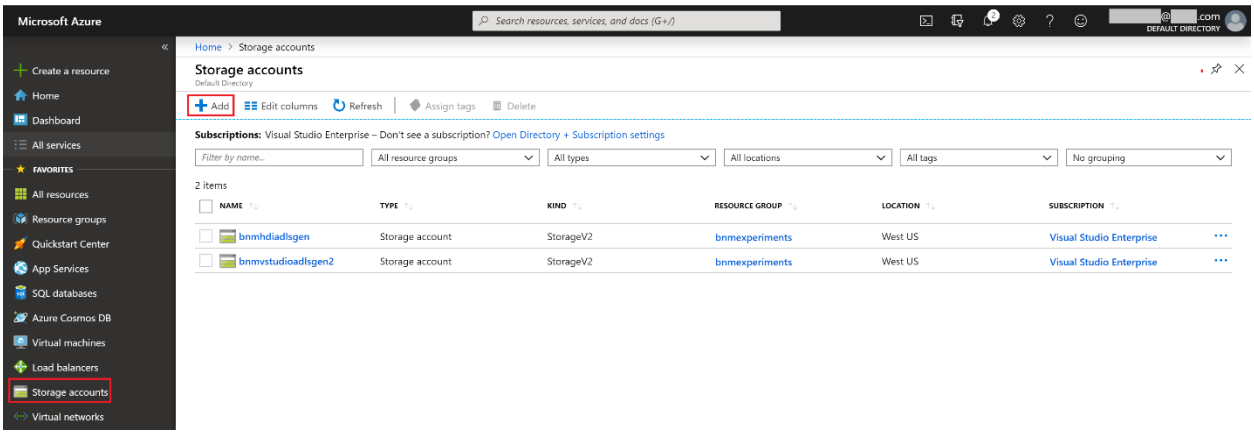


Figure 32: List of Storage Accounts

- Click on + Add to start defining the resource configuration. Configuration parameters to specify include:
    - Subscription – choose a subscription where you AAD-DS is enabled and the same where you would have the HDInsight cluster setup.
    - Resource Group – choose a resource group where HDInsight cluster will be setup. For example - `*bnmexperiments*`.
    - Storage account name – contextually relevant name of the ADLS Gen2 storage account. For example - `*bnmhdiespadlsgen2*`.
    - Location – same location where you will have HDInsight setup.
    - Performance – if your use case requires low-latent I/O needs, you can boost them by selecting a premium storage option. It provides SSD backed storage layer. If otherwise, you can use a standard choice.

- Account Kind – Select *StorageV2 (general purpose v2)*.
- Replication – you can choose a Locally-redundant Storage (LRS) that can sustain an available zone failure, but not a region-wide failure.
    - Infer the desired replication strategy based on your fault-tolerance requirements.
    - Choices other than LRS will have additional costs. Something to avoid if your business case does not need one.
- Access tier (default) – considering we will have frequent read/write operations on the storage layer, select hot access tier.

- Following visual provides a sample configuration setting:

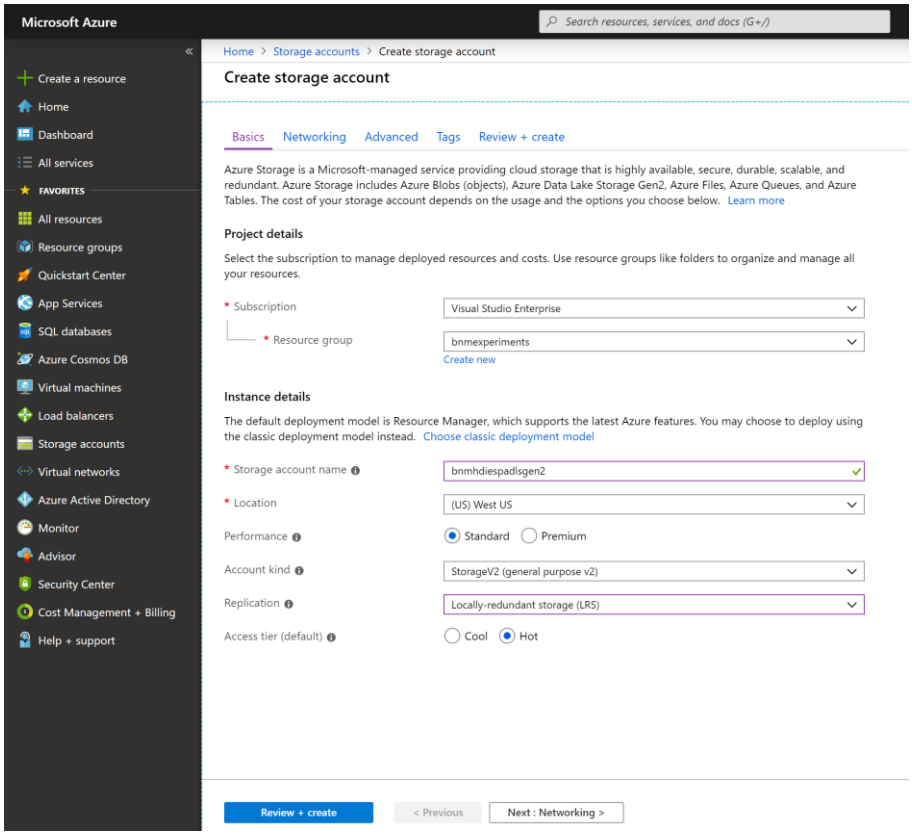


Figure 33: Create Storage Account - Basics Tab

- Click on *Next: Networking* to configure networking options. Access to the storage account can be restricted to select IP addresses or VNETs. For example, we can limit access from just the VNET where HDInsight will be setup.

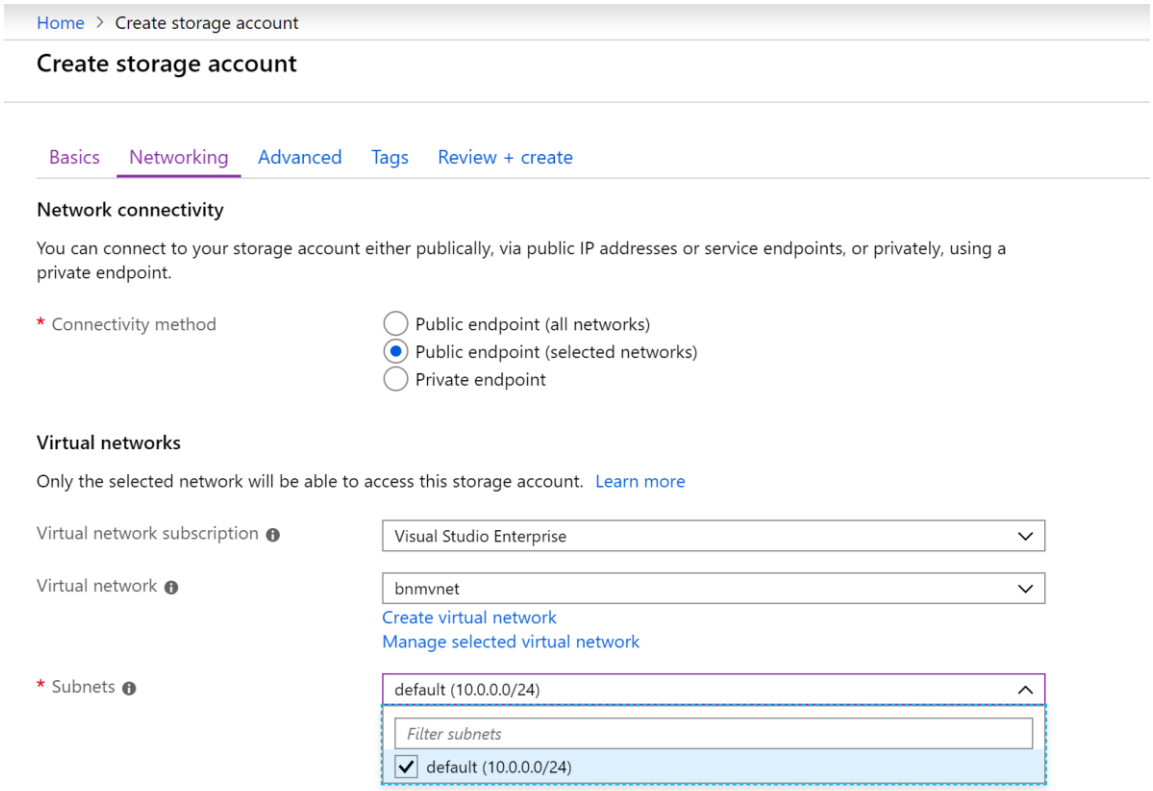


Figure 34: Create Storage Account - Networking Tab

- Then, will configure Advanced configuration settings that include following:
  - Enable *Security – Secure transfer required* to ensure storage account will respect only secured connections. Example, RESTful services that depend on HTTPS rather HTTP, and encrypted SMB connection.
    - Non-secure connection attempts will be rejected.
  - Enable *Hierarchical namespace* to take advantage of the file system semantics and POSIX style support offered by the ADLS Gen2.

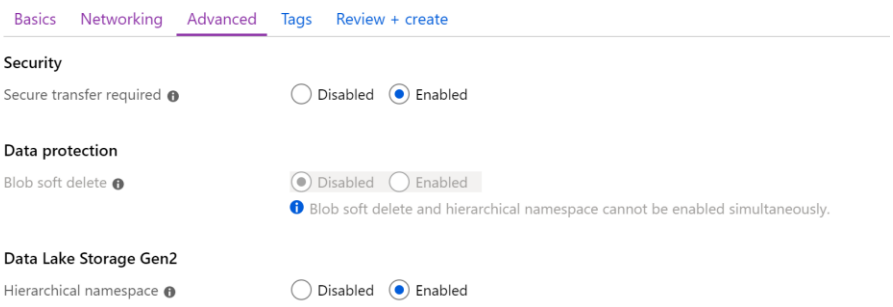


Figure 35: Create Storage Account - Advanced Tab

- Resources can be assigned common tags for manageability reasons. The next step will allow you to add tags. Further, you can review the configuration choices, as shown in the following visual reference:

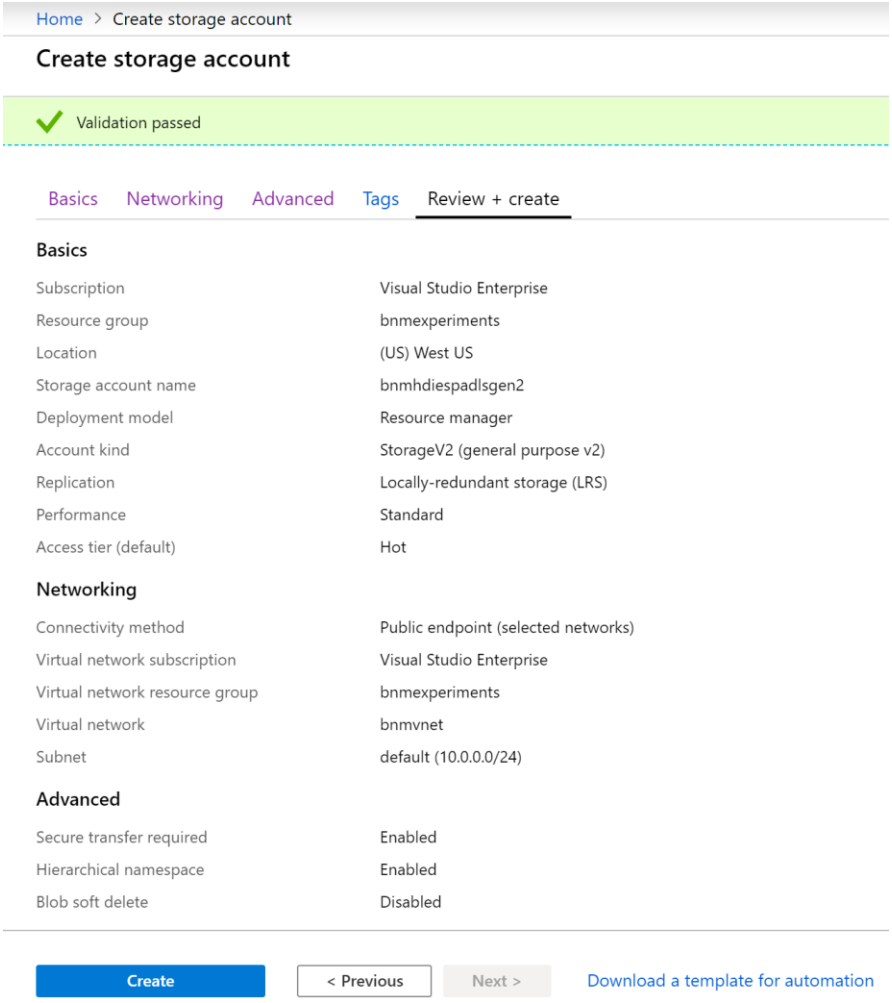


Figure 36: Create Storage Account - Review + Create - Validation Passed

- Click on create to submit the request to deploy and create the storage account resource. Following is the set of visual references:

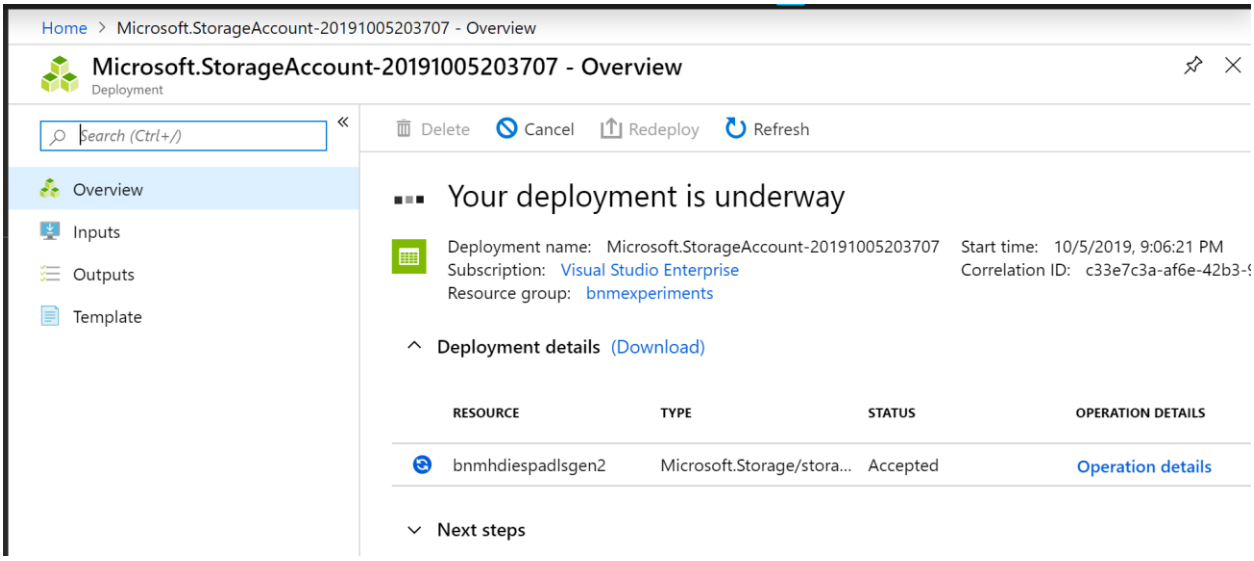


Figure 37: Storage Account Deployment (underway)

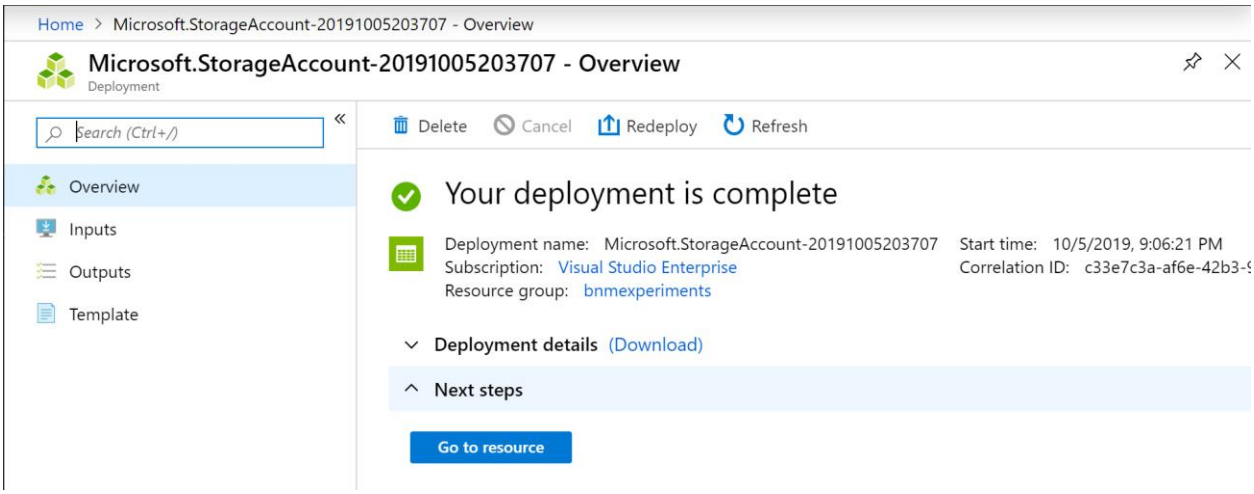


Figure 38: Storage Account Deployment Complete

- Click on the *Go to resource* button to navigate to the just created ADLS Gen2 storage account.

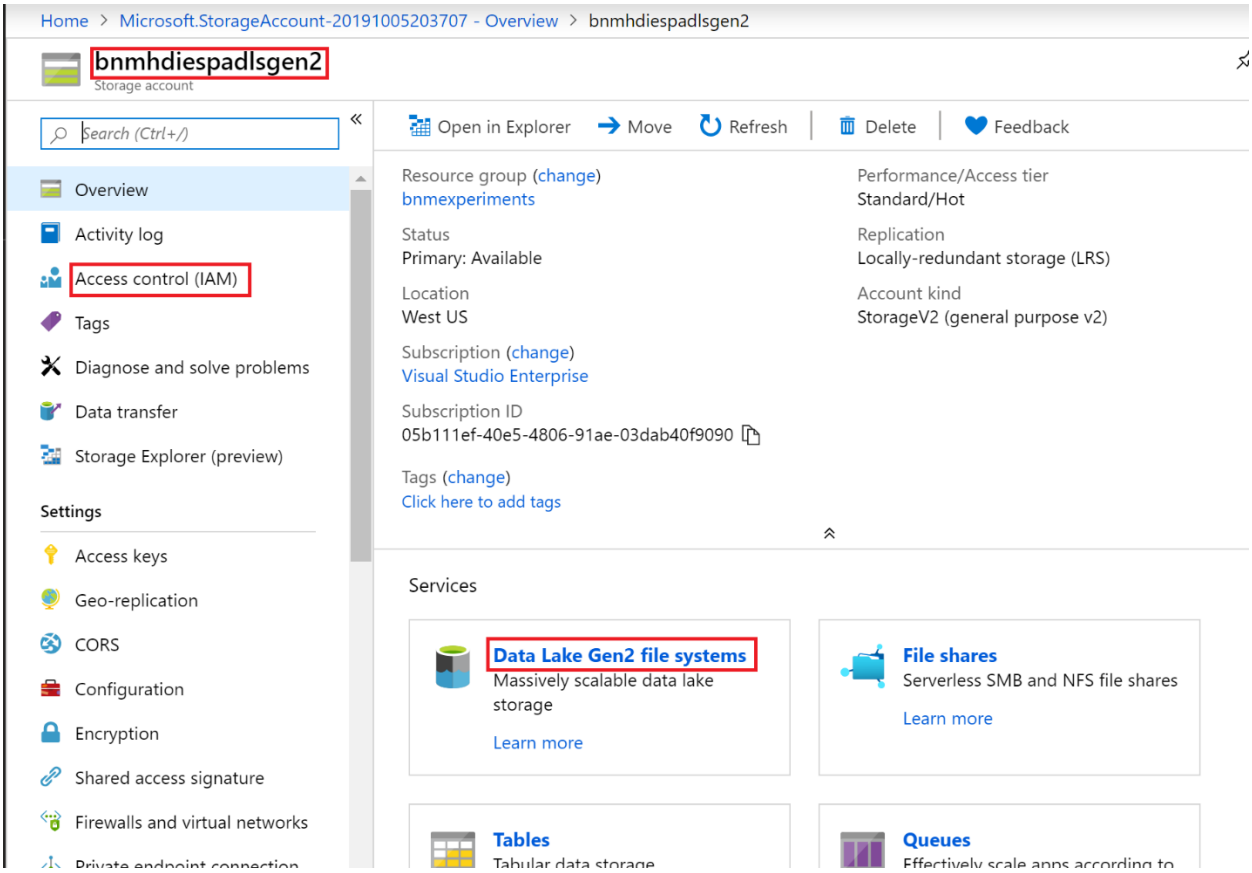


Figure 39: ADLS Gen2 - Overview

- Note the highlighted links in the above visual. For example, Access control (IAM) allows you to configure security access at the storage account level, while the link to Data Lake Gen2 file systems will help you navigate to the individual storage containers and configure appropriate settings at each container level.

5.2 Validate network restrictions

Since we have restricted access to the ADLS Gen2 Storage Account for a select VNET and Subnet, we can quickly verify its impact. This can be quickly validated by logging into your Azure Subscription using Storage Explorer and attempt to access the storage account container resources. Following visual provides a reference to a failure condition that must be expected when you have restricted access to the storage account:

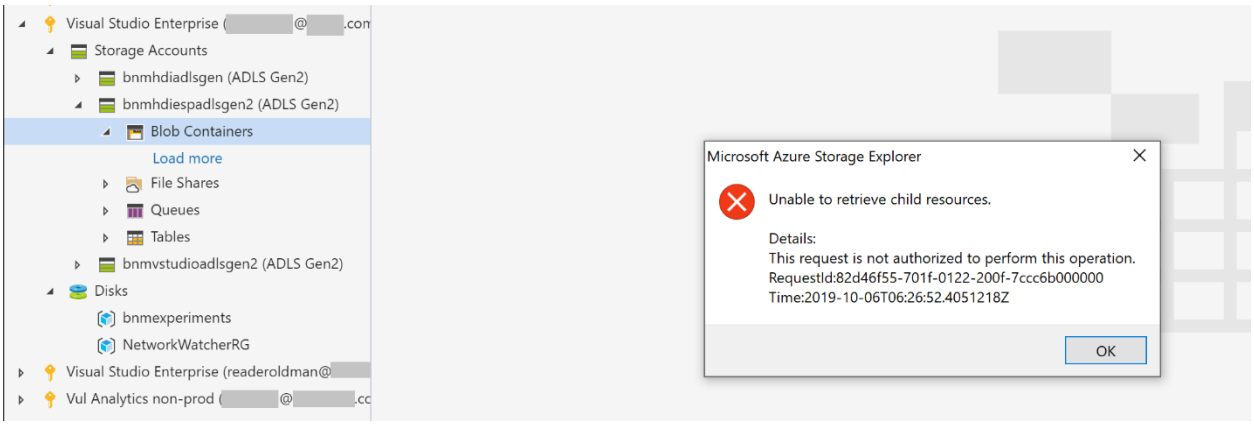


Figure 40: ADLS Gen2 - Validate Network Restrictions - Sample Error Message

Also, the same effect can be observed when attempting to access the storage explorer from within the Azure portal.

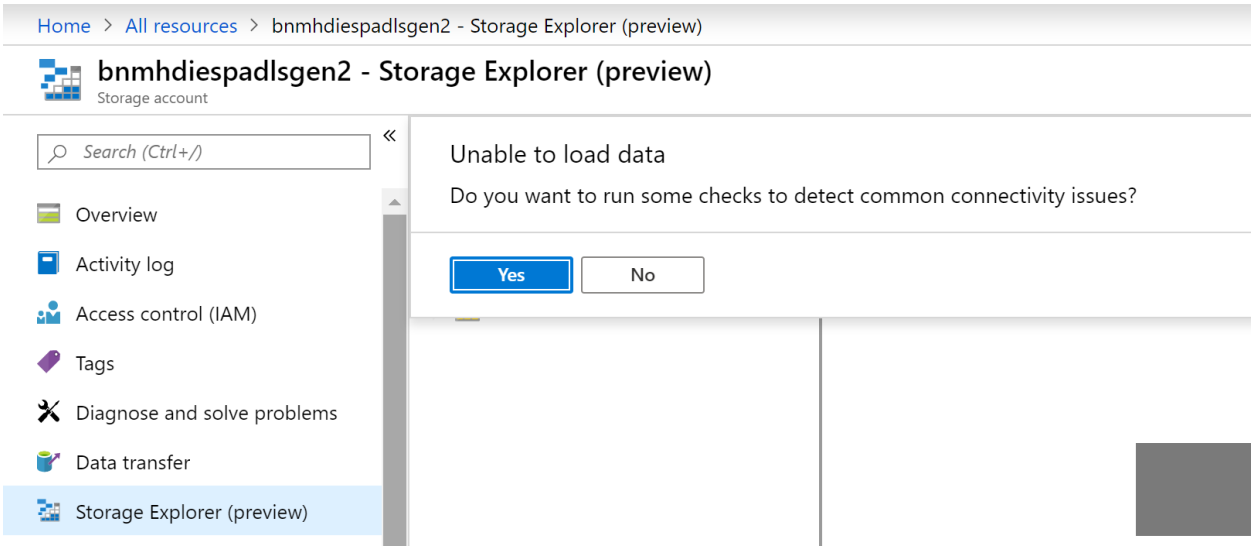


Figure 41: ADLS Gen2 - Validate Network Restrictions - Unable to Load Data (Sample)

5.3 Configure Storage Blob Data Owner Role

We need two kinds of owners assigned to the storage account, such that HDInsight services can successful read and write to the underlying storage containers, and that members of the group *hadoopadmins* can administer access to HDFS layer for users intending to query data from data stores like Hive. And, also be able to write to shared folders that are part of the HDInsight storage container.

5.3.1 Add role to User Assigned Managed Identity

Following are the necessary steps to assign the Storage Blob Data Owner role to the managed identity – *bnmmanagedidentity*:

- Click on the *+Add* link to configure the desired role.

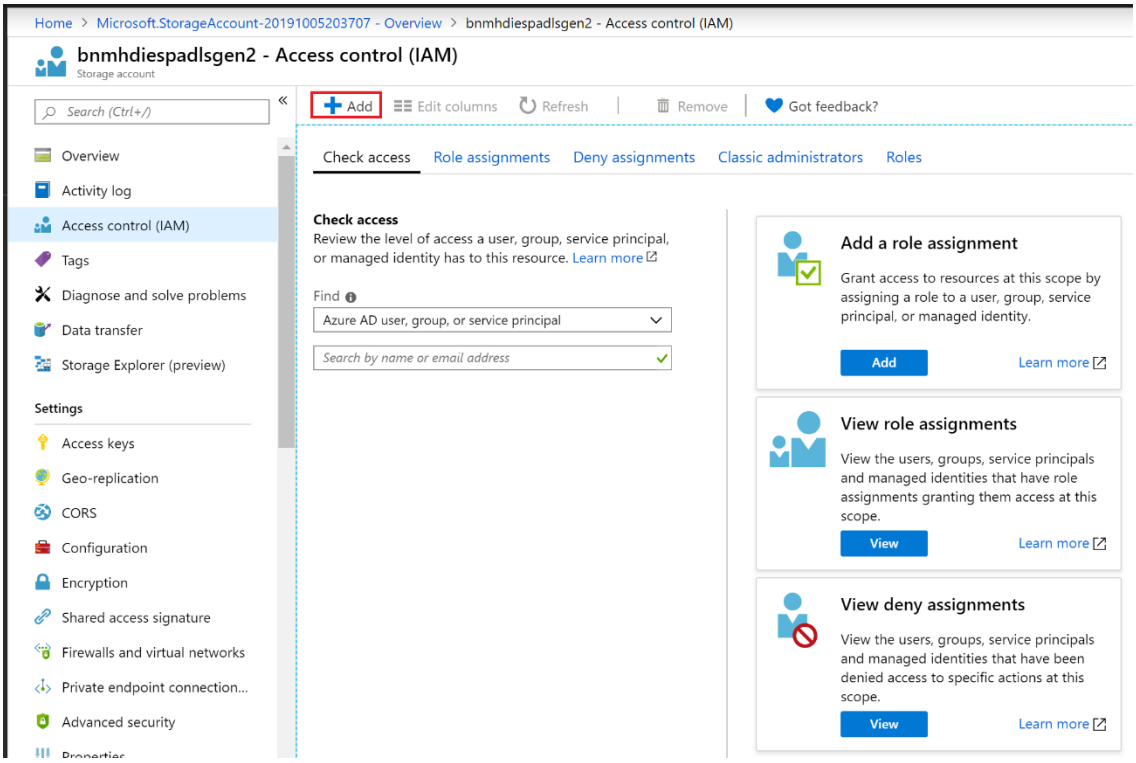


Figure 42: ADLS Gen2 - Add Assign User Assigned Managed Identity

- Here's the sample reference visual:

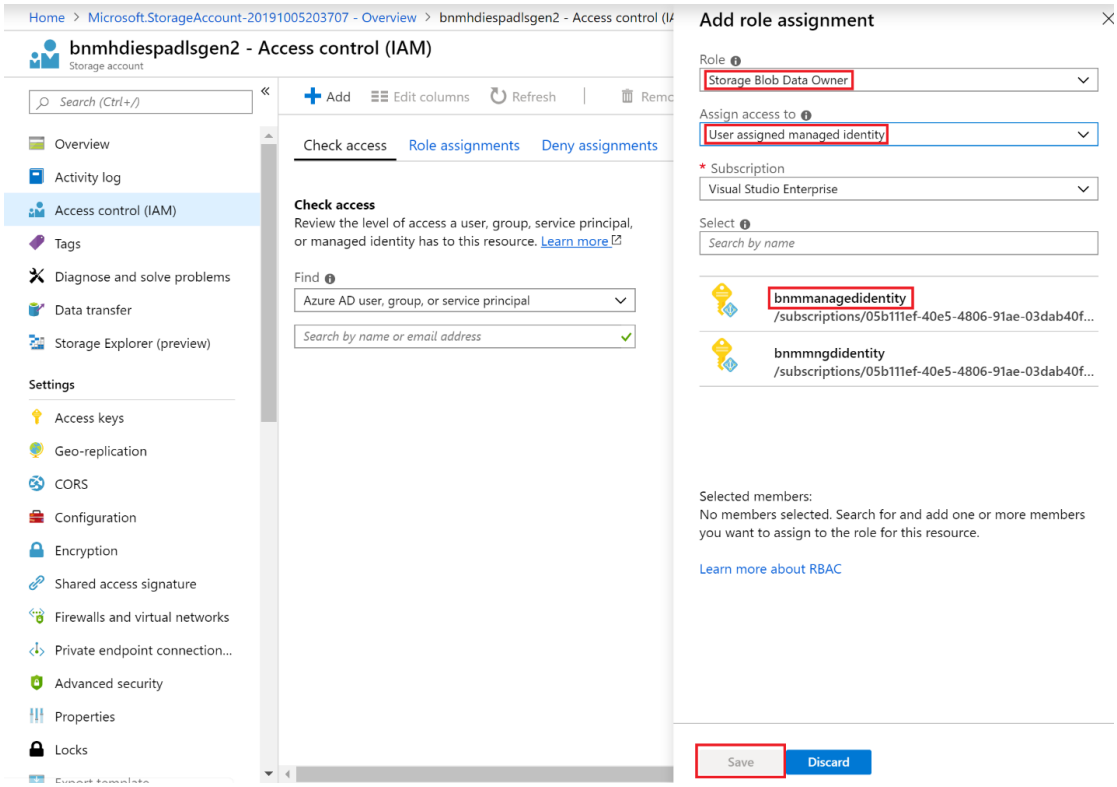


Figure 43: ADLS Gen2 - Managed Identity selection

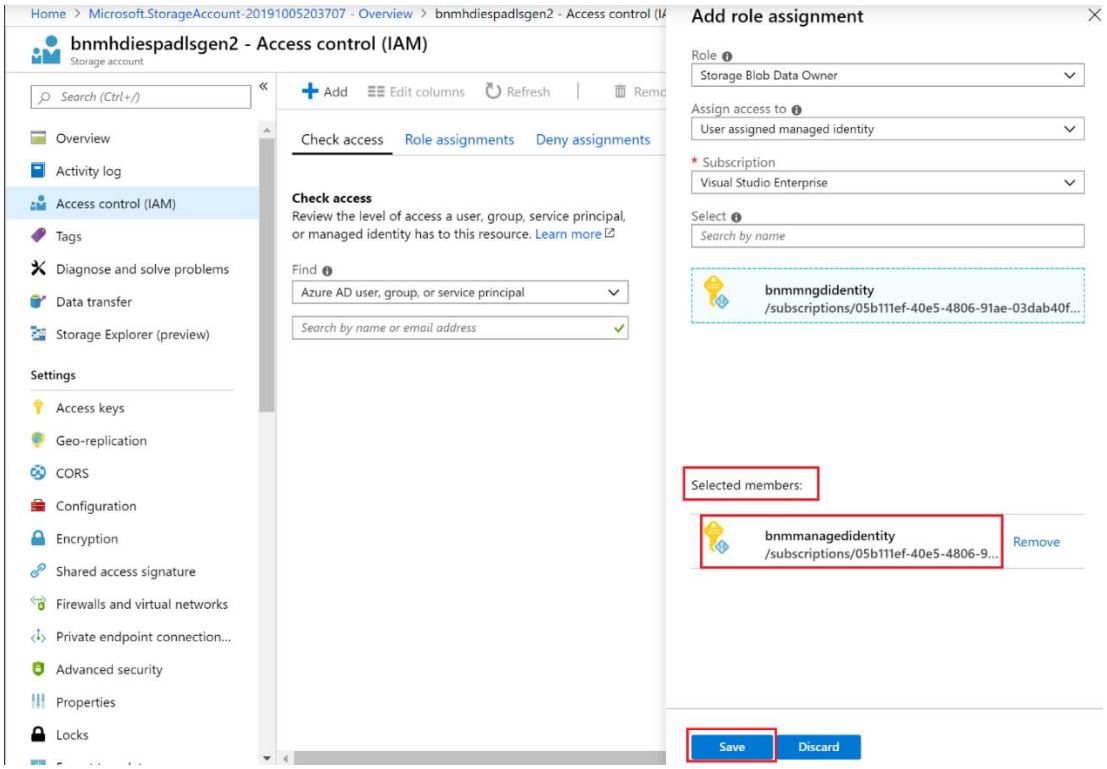


Figure 44: ADLS Gen2 - Save selected Managed Identity

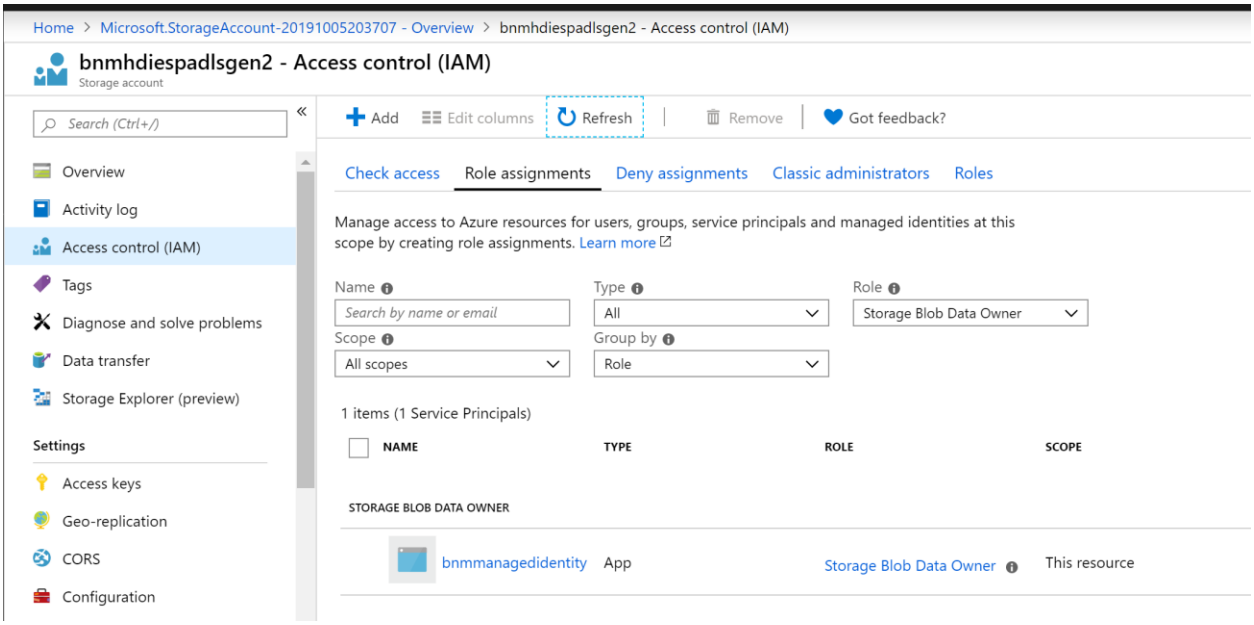


Figure 45: ADLS Gen2 - Updated Storage Blob Data Owner Role view

- Configuring the role will give the managed identity the ability to define a storage container for the HDInsight cluster, at the time of cluster deployment process.

5.3.2 Add role to *hadoopadmins* group

Like the managed identity, add the group *hadoopadmins* the role of a Storage Blob Data Owner. For reference, use the following sample visual:

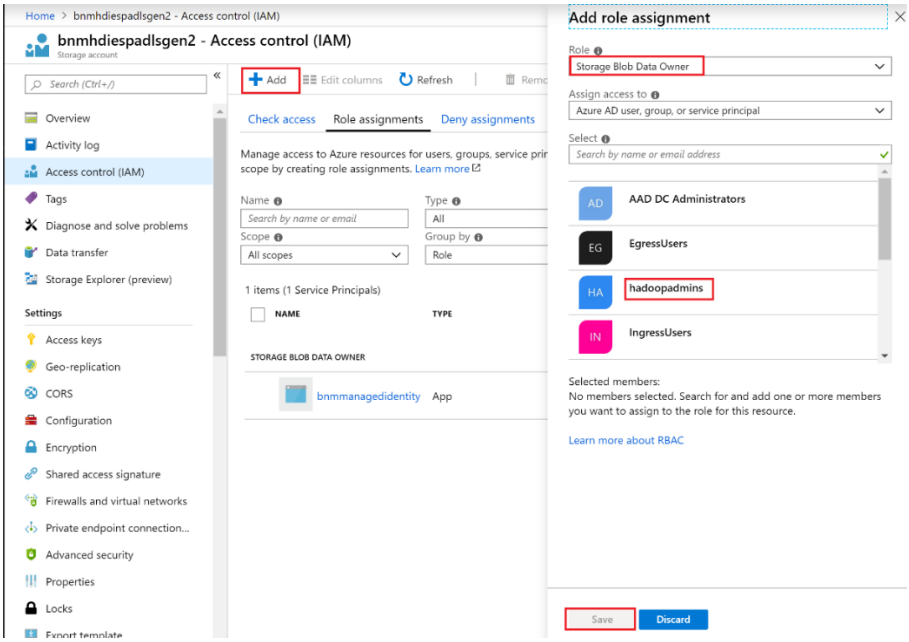


Figure 46: ADLS Gen2 - Add role to *hadoopadmins* group

Select the group *hadoopadmins*, then save the selection. Here's the updated list of role assignments:

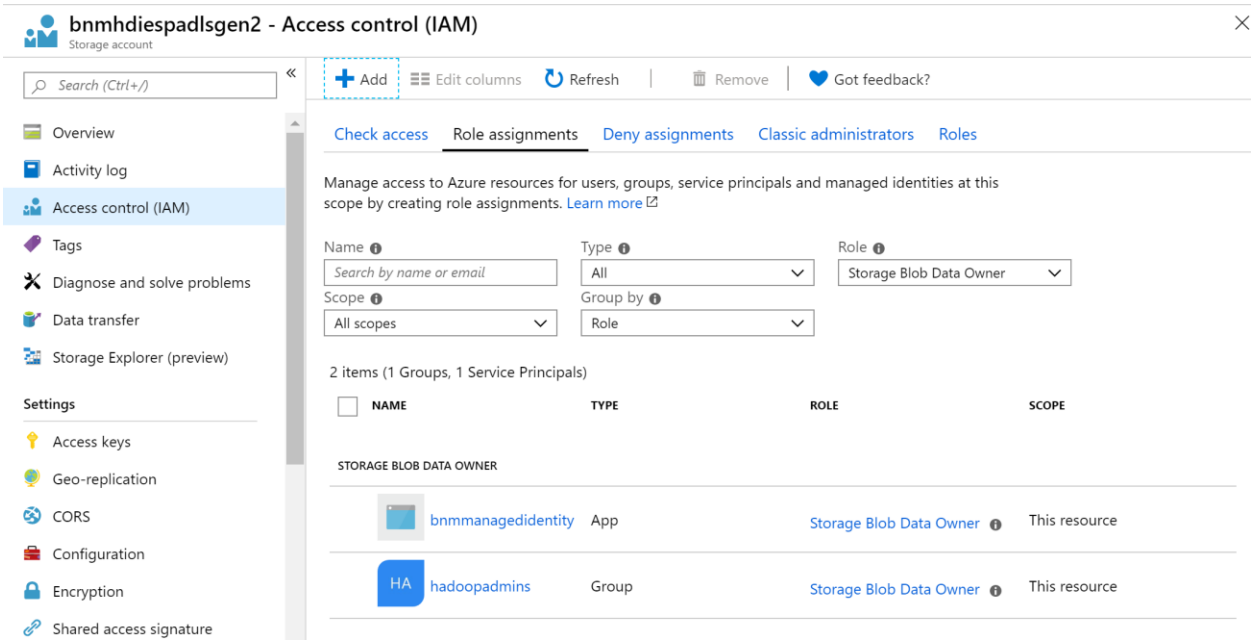


Figure 47: ADLS Gen2 - Updated Role Assignments View

5.4 Add HDInsight Cluster Admin User

In this section will add and configure a dedicated user to administer all Hadoop service activities. The user will also be part of the group *hadoopadmins*. Following sequence describes the steps involved in adding this new user:

- Navigate to the Azure Active Directory in the portal, as shown in the following visual reference, and select the group *hadoopadmins* that will be assigned as part of user creation.

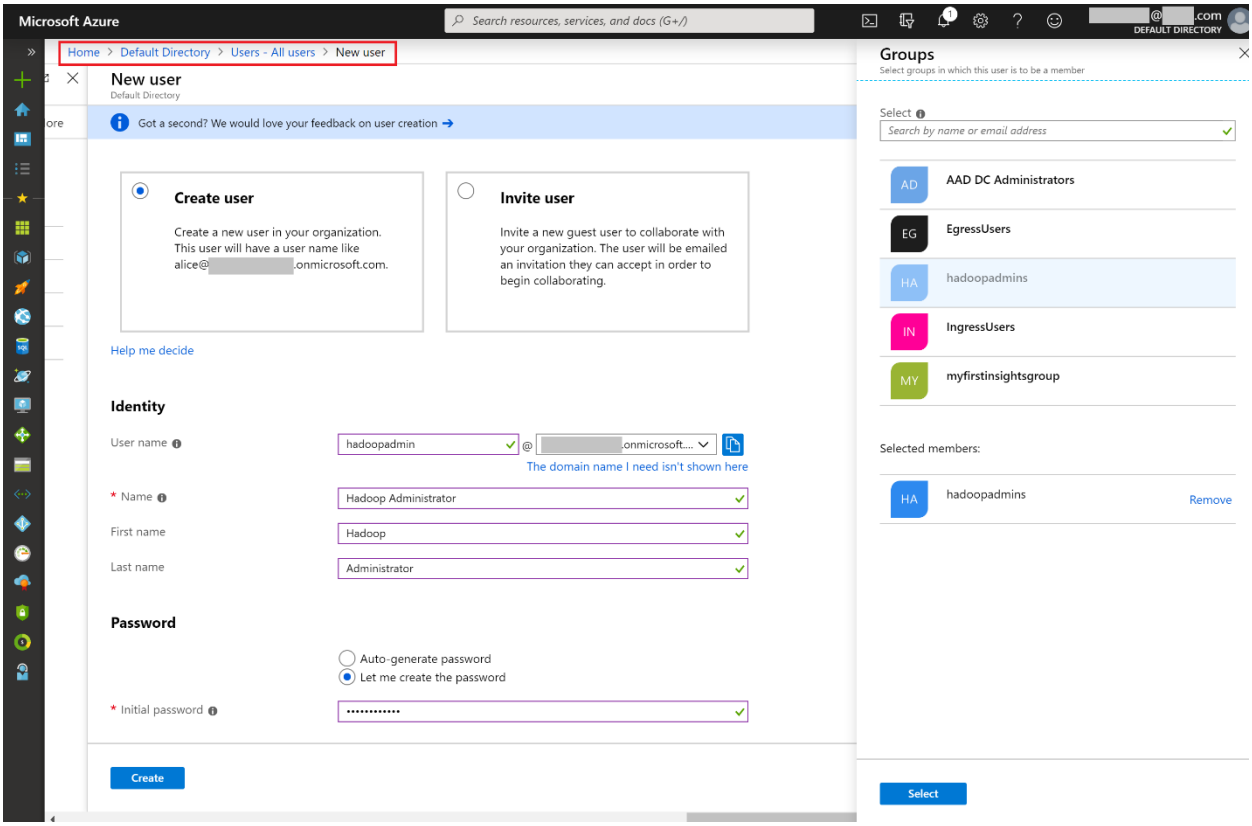


Figure 48: Add *hadoopadmin* user to *hadoopadmins* group

- You can restrict the sign-in of this user by limiting the geo-location from where the user can login into Azure. Ensure to leave the Block Sign-in to No. Else, the user will not be able to sign-in i.e., login to the Azure.

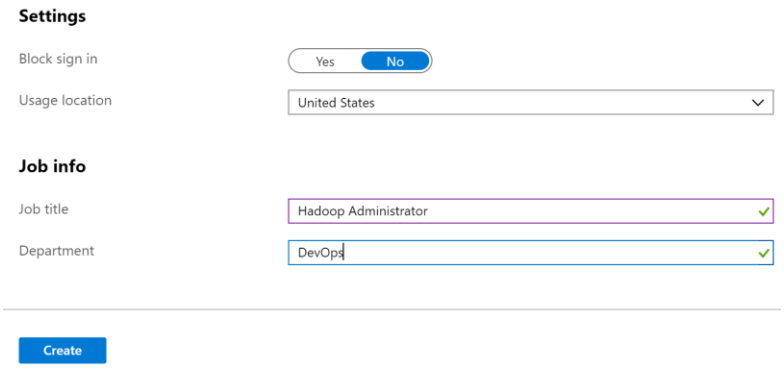


Figure 49: Add hadoopadmin user to hadoopadmins group (Contd...)

- Once successfully created, the user would be listed in the Azure Active Directory set of users.

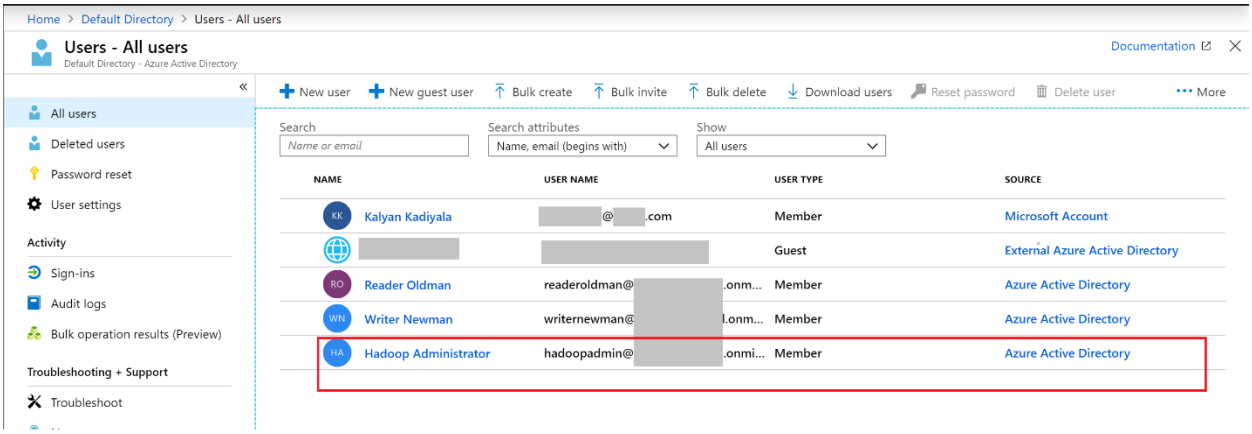


Figure 50: Azure Active Directory - All Users view

- And, here is a quick look at the Hadoop Administrator's profile:

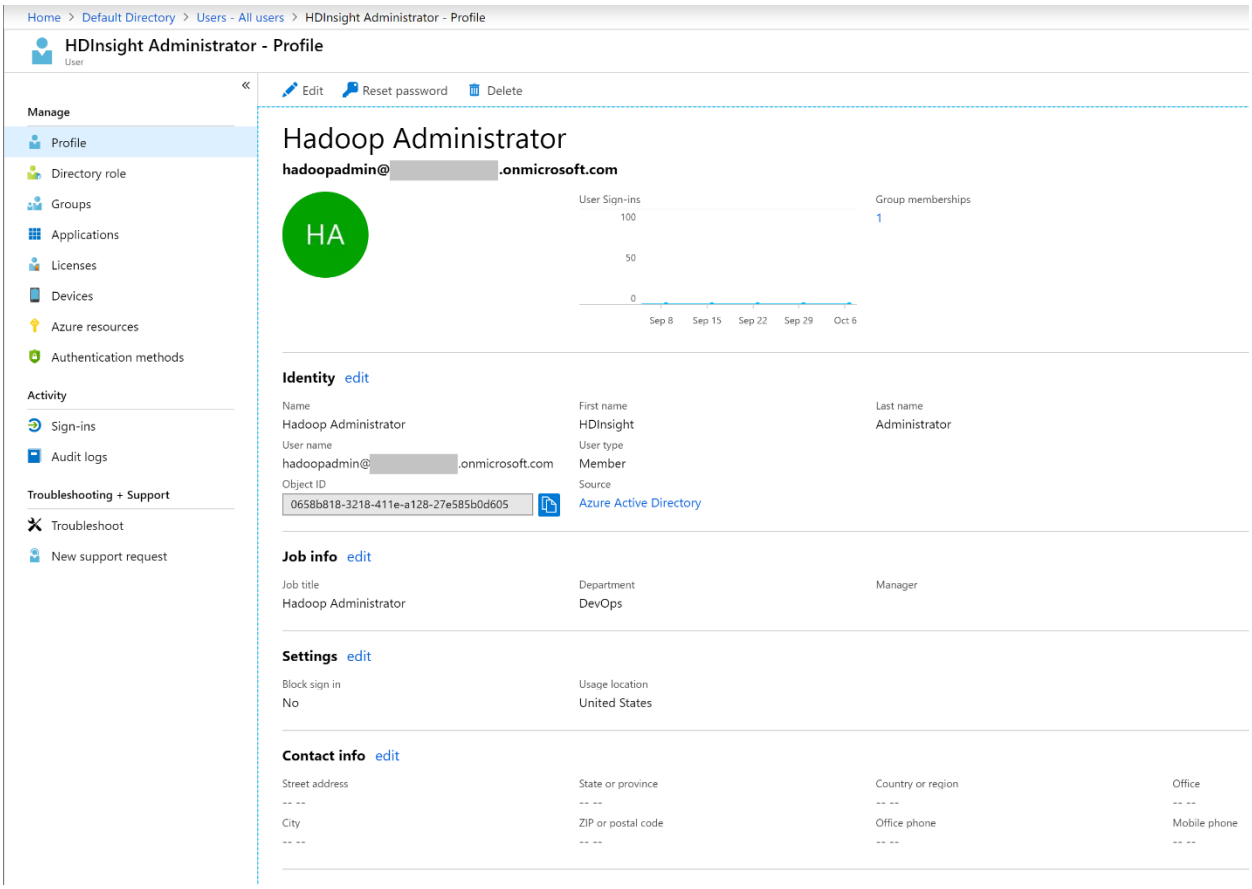


Figure 51: Hadoop Administrator User Profile (Azure Active Directory)

- The next step is to verify login to Azure subscription. To do so, click on the profile icon on the top right corner of your Azure portal and choose *Sign in with a different account link*.

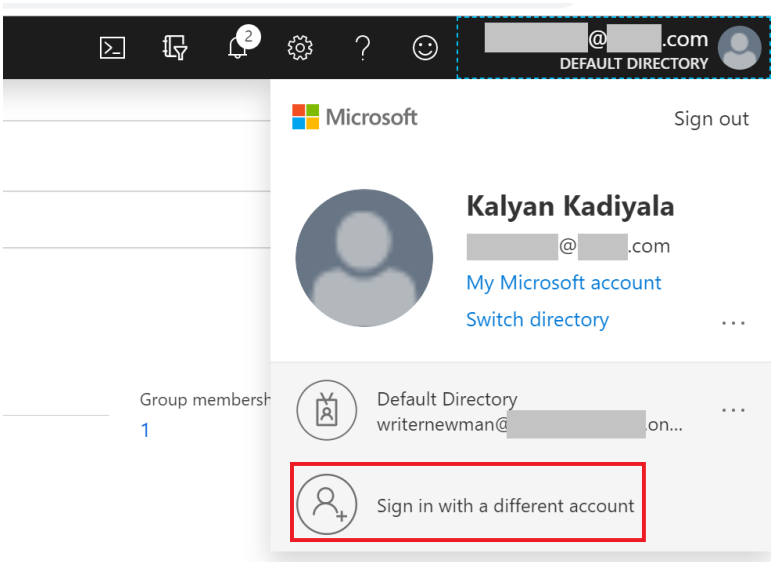
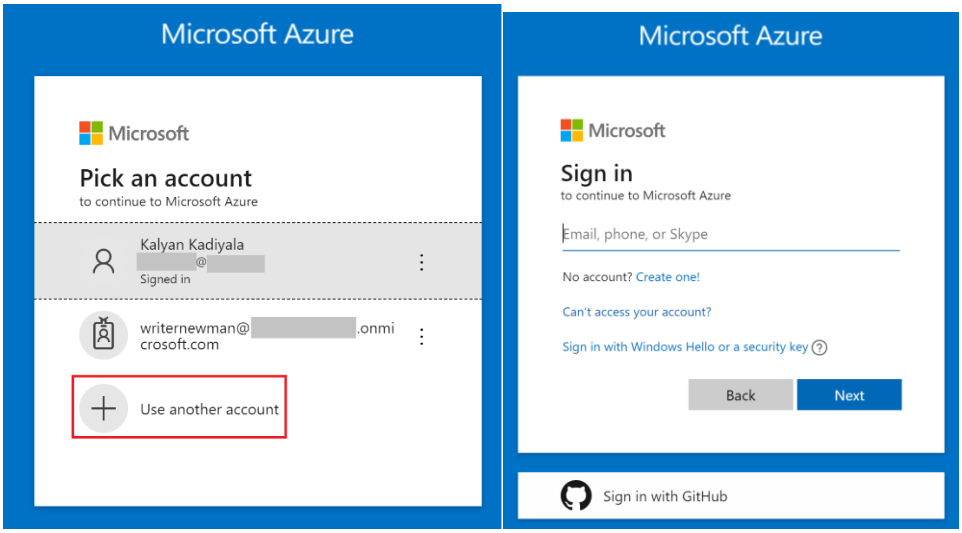


Figure 52: Azure Portal - Sign-in with a different account



- Select + Use another account.
- In the login prompt enter the credentials for our new user – hadoopadmin@youradddsdomainname.onmicrosoft.com

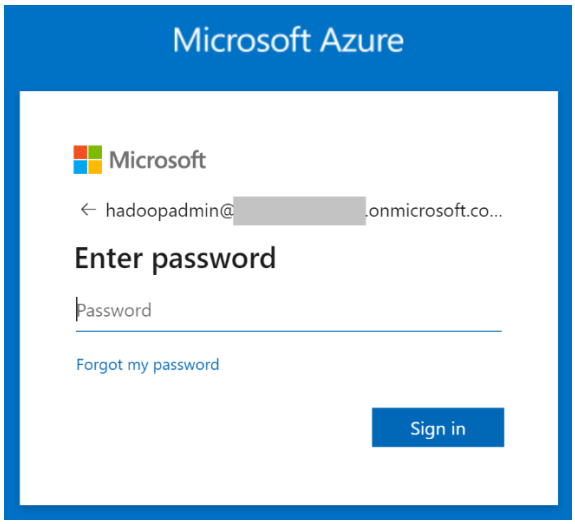


Figure 53: Azure Portal Login

- On first successful login you will be prompted to change your password

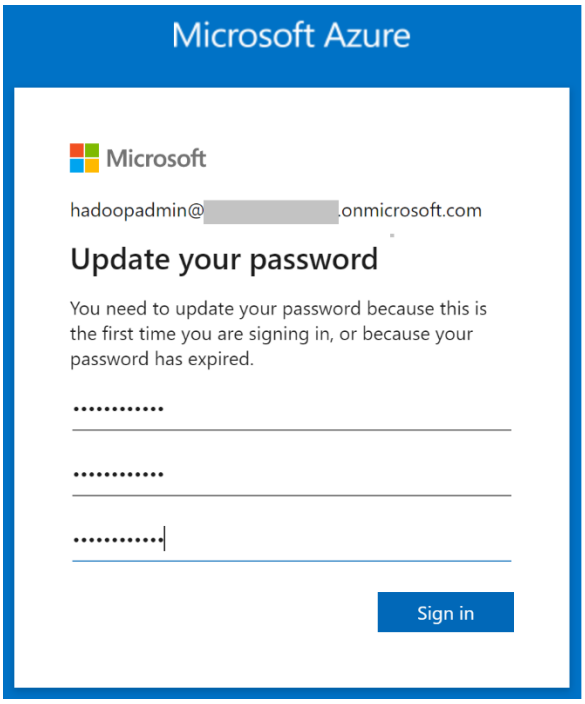


Figure 54: Azure Portal First time login - Update Password

- Visual reference that shows a landing page post successful login (with password change).

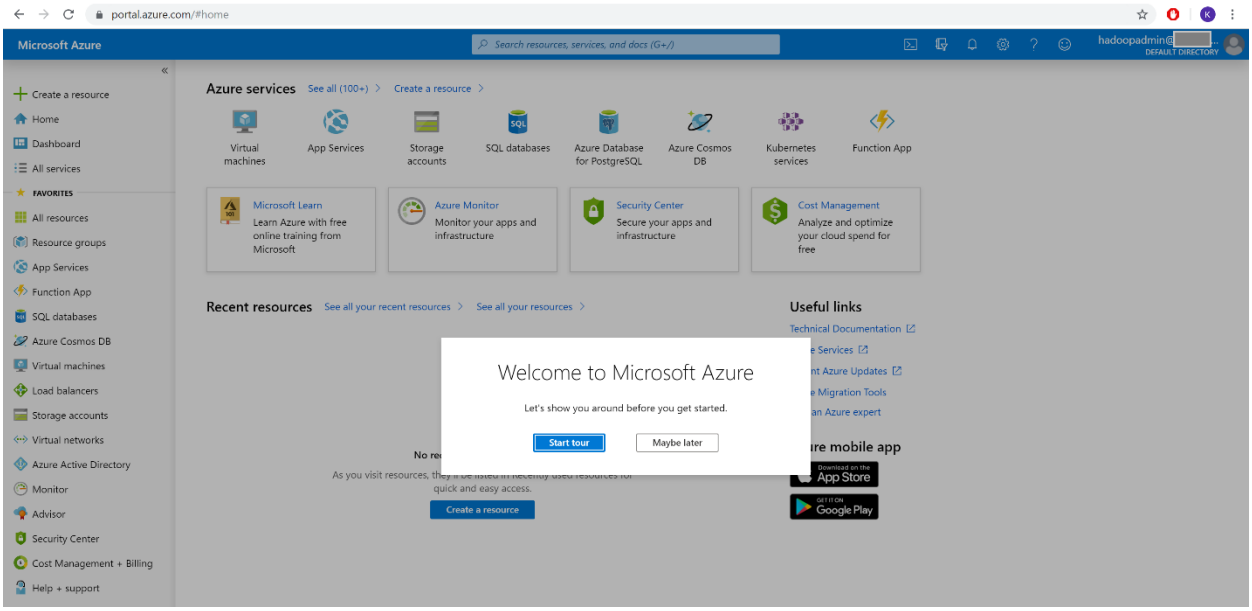


Figure 55: Azure Portal New User Login - Landing Page

Note – able to successfully login and being part of the *hadoopadmins* user group does not grant automatic access to the resources like Hadoop Cluster and the underlying ADLS Gen2. This is accomplished by enabling roles. More on this under the section – *Enforce Storage Access-Controls*.

5.5 Additional role configurations for group *hadoopadmins*

To be able to administer user access controls, the members of the group *hadoopadmins* must be enabled with additional roles at parent resource levels in the hierarchy. This includes the resource, resource-group and at the subscription. Following is a visual excerpt from the official documentation that demonstrates how the RBAC roles are enforced at various scope levels of the resource hierarchy:

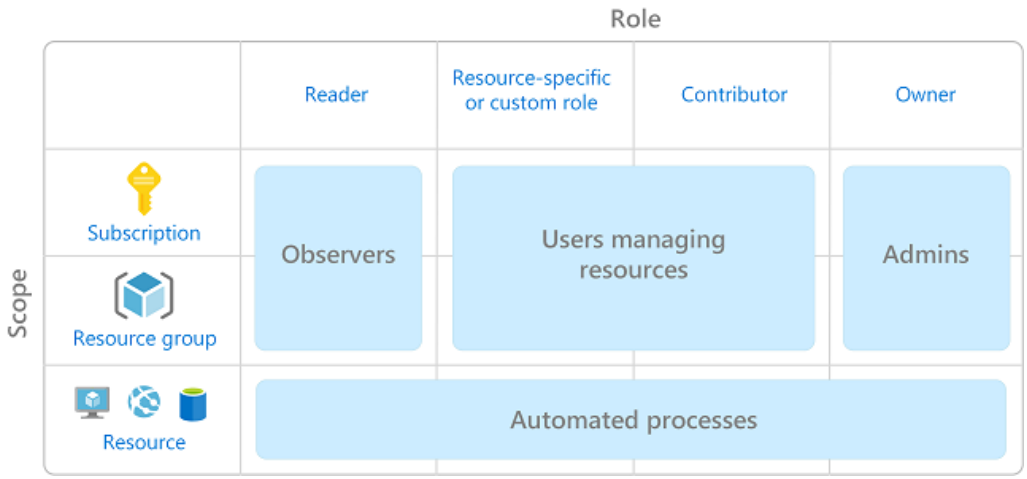


Figure 56: RBAC Roles Application Overview

Based on the above guidance following role assignments must be made for the group *hadoopadmins*:

Role	Scope (resource type)	Additional Description
Reader	Storage Account, Resource Group	Use Access Control (IAM) from respective resources.
Reader and Data Access	Storage Account	Storage Account's Access Control (IAM); allows the respective users to be able to view folder contents (say from storage explorer)
Storage Blob Data Owner	Storage Account	Storage Account's Access Control (IAM)

Table 1: hadoopadmins group RBAC roles Overview

Since we have already assigned the Storage Blob Data Owner access to the *hadoopadmins* group, we also need to enable the Reader and the Reader & Data Access roles at respective scopes. The process is similar to that described in the section *Add role to hadoopadmins group*, where you would be using the resource specific Access Control (IAM) pane.

Following visual provides a view of the expected role assignment at the storage account level for the group *hadoopadmins*:

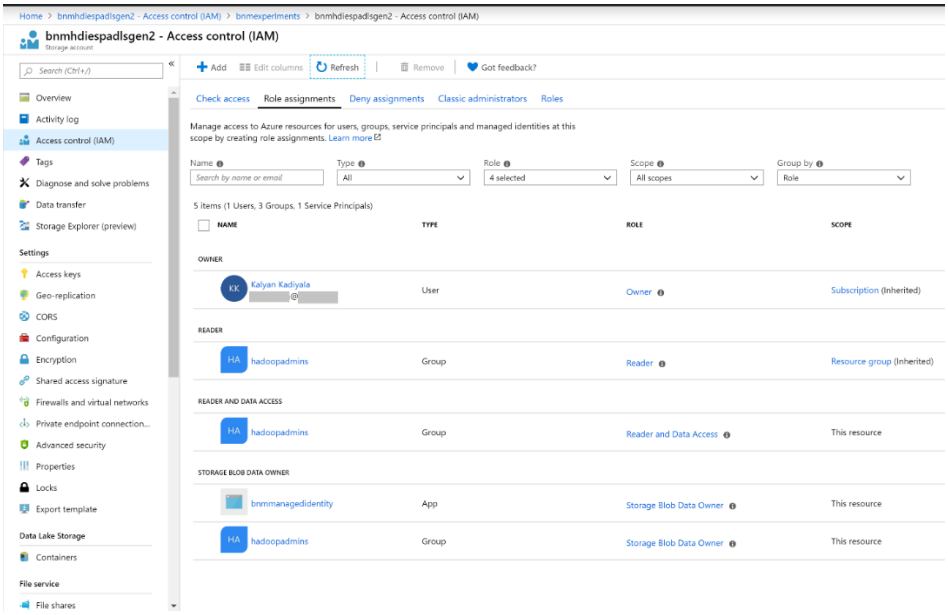


Figure 57: *hadoopadmins* group assigned roles - Access Control (IAM) overview

And, here is the expected storage explorer view that a user of the group *hadoopadmins* should see (in this case it is the admin user that we have added in preceding sections):

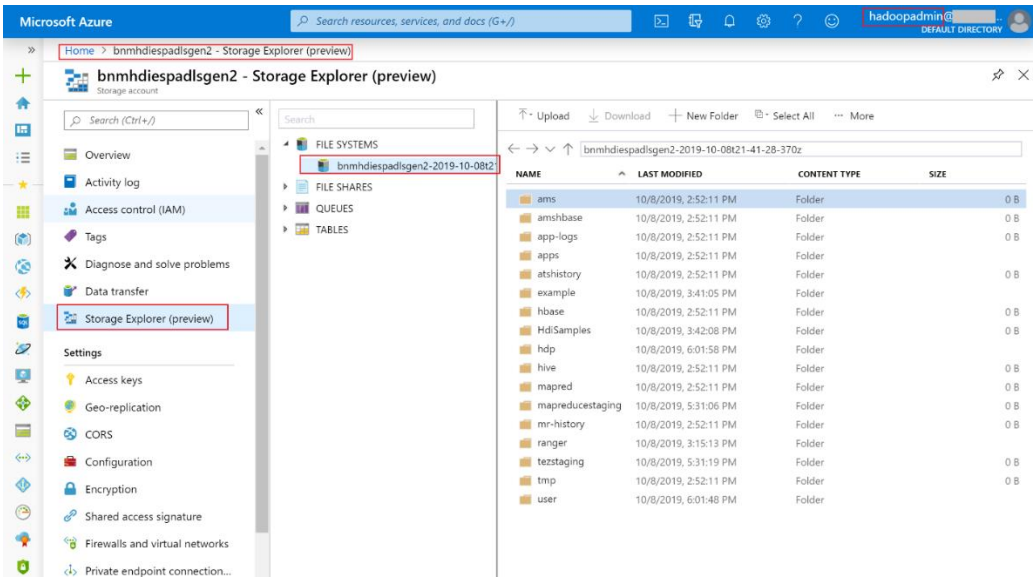


Figure 58: Hadoop Admin User - Storage Explorer View - Folder View Sample

6 Setup HDInsight ESP enabled cluster

Azure HDInsight supports several cluster types depending on your needs. For example, you could have a cluster of type Hadoop that focuses on Hive, HDFS and YARN/Map-reduce. If you intend to bring in Apache Spark workload then there is a cluster type called Spark. And so forth, for HBase if you intend to setup a low-latent, consistent NoSQL backend data store.

By enabling the Enterprise Security Package (ESP) support, HDInsight services can leverage the Azure Active Directory – Domain Services (AAD-DS) to address user & service principal identity management, and handle secure interactions with the cluster services that require Kerberos authentication. It is important to note that storage layer access controls must be managed by the interfaces supported by ADLS Gen2 itself. These include Role-based Access Controls (RBAC) and Access Control Lists (ACLs). RBAC's fine grain support is limited to assigning and managing roles at-most at the container level. While, ACLs allow you to enforce more fine-grained access control at folder paths and files within containers as well.

The sub-sections below will guide through the necessary steps to create and setup HDInsight cluster along with enabling ESP support.

6.1 Create HDInsight + ESP Resource

- Assuming you have are already logged in to the Azure Portal, click the + *Create Resource* link from the left navigation pane.

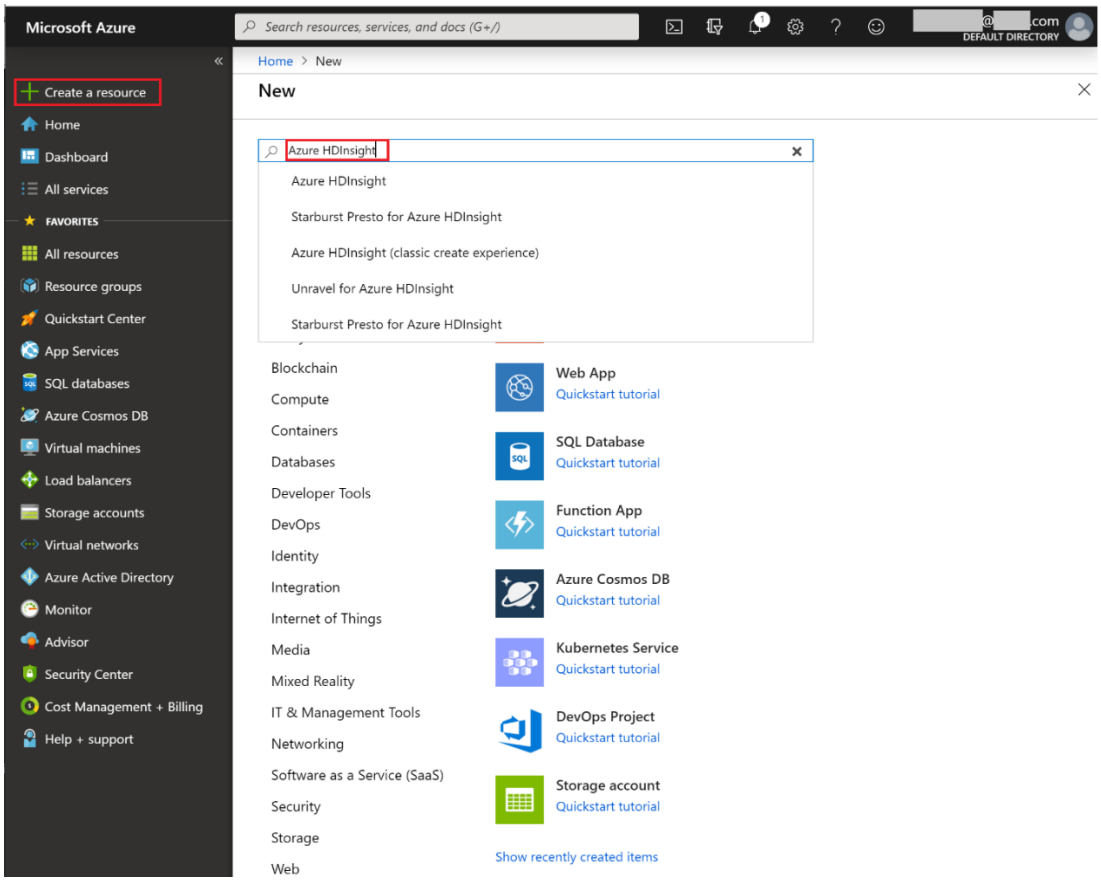


Figure 59: Create Azure HDInsight - Search

- Once you have selected the resource type, the next step is to begin the creation process.

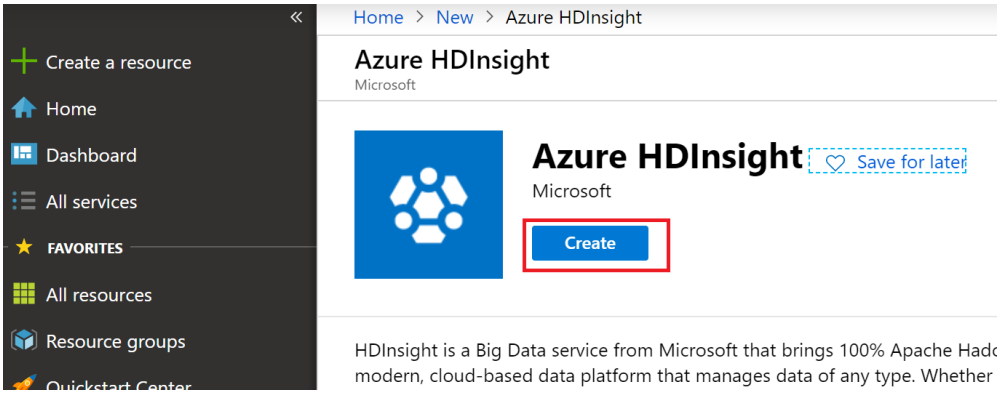


Figure 60: Create Azure HDInsight Resource

- In the *Basics* tab select the Subscription, Resource Group, name for the cluster (of your choice), location (Azure region), the type of the cluster and the cluster credentials. Following visual references provide some contextual examples:

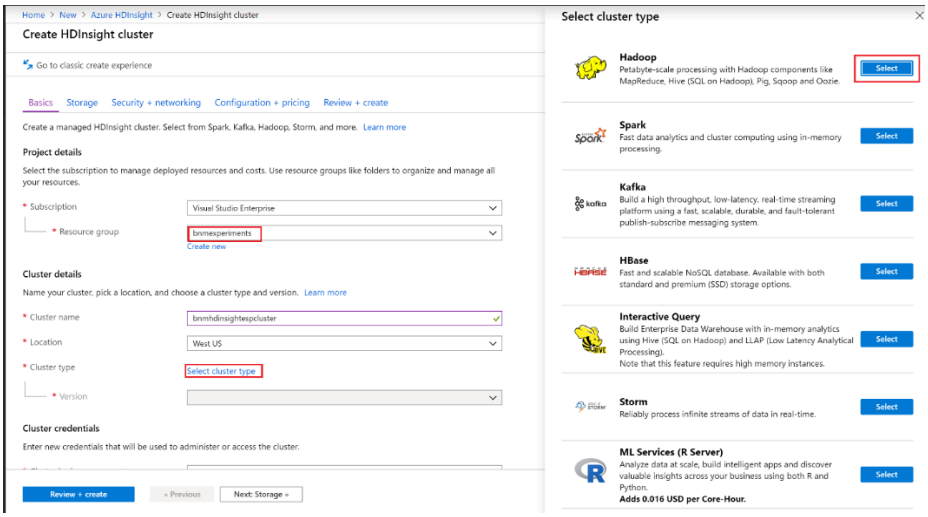


Figure 61: Create HDInsight + ESP - Choose cluster type

- In this experiment we are going to select HDI 3.6, which is observed to be a common compatible version in the Azure migration case studies. For example, it is compatible with Hortonworks Distribution version 2.6. Please note, the way connected components are integrated in HDInsight distribution will still vary with that of open source or other vendor frameworks. For example, HDInsight relies heavily on ADLS choices to satisfy HDFS at-rest needs.

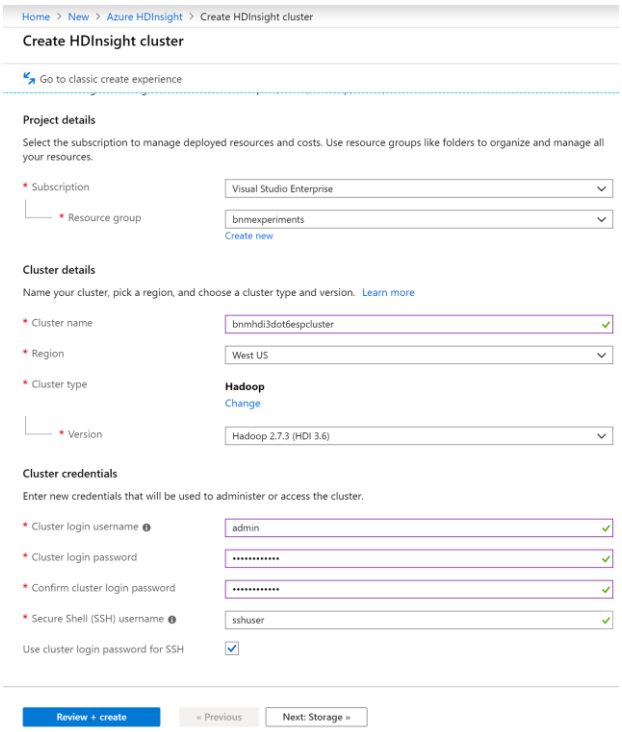


Figure 62: Create HDInsight + ESP cluster - Details

- Quick note on the cluster login name parameter above. It maps to the HTTP-User who can login to Ambari with administrator privileges. Note – as part of the Security + Network tab configuration we will enable ESP and add a cluster admin user. The identity of this user will be part of the AAD and synchronized by AAD-DS. The configuration parameter here is – *Cluster admin user*. We will need to select the AAD admin user (*hadoopadmin@youraaddsdomainname.onmicrosoft.com*) that we have added in prior sections above.
- Click on `*Next: Storage* >>` to configure the storage.
    - Select ADLS Gen2 Storage Account that we have created earlier - *bnmhdiespadlsgen2*.
    - Then specify the User Assigned Managed Identity *bnmmanagedidentity*, that the cluster will use to represent on its behalf when interacting with the storage containers.

- You can choose to give the file system a name or leave it at its default.

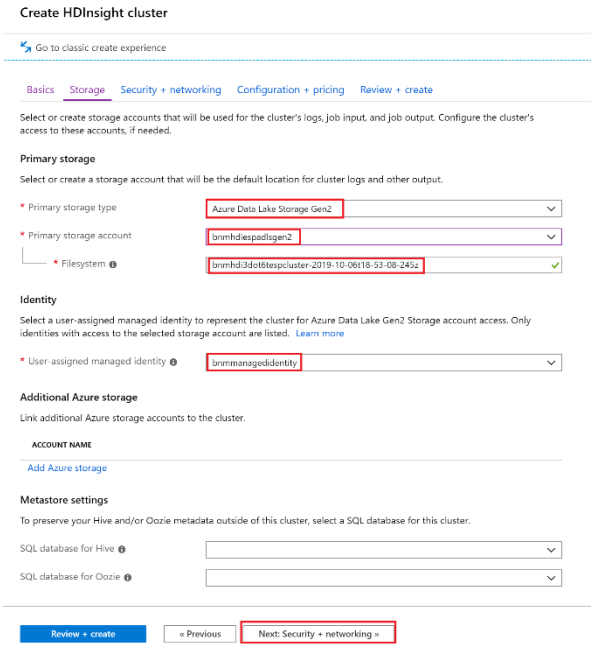


Figure 63: Create HDInsight + ESP cluster - Storage Tab

- It is recommended to consider using an external datastore of choice for services like Hive and Oozie, when setting up a production cluster. For example, SQL DB offers in-built replication strategies to satisfy DR strategies.

- Next step is to configure the Security and networking, where we also would enable the Enterprise Security Package.

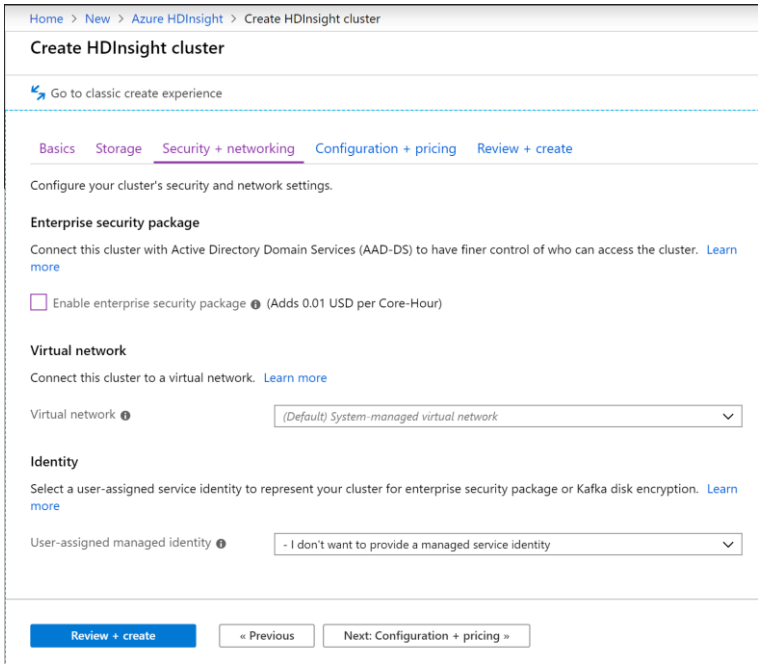


Figure 64: Create HDInsight + ESP cluster - Security + Networking Tab

- Enable the Enterprise Security Package.

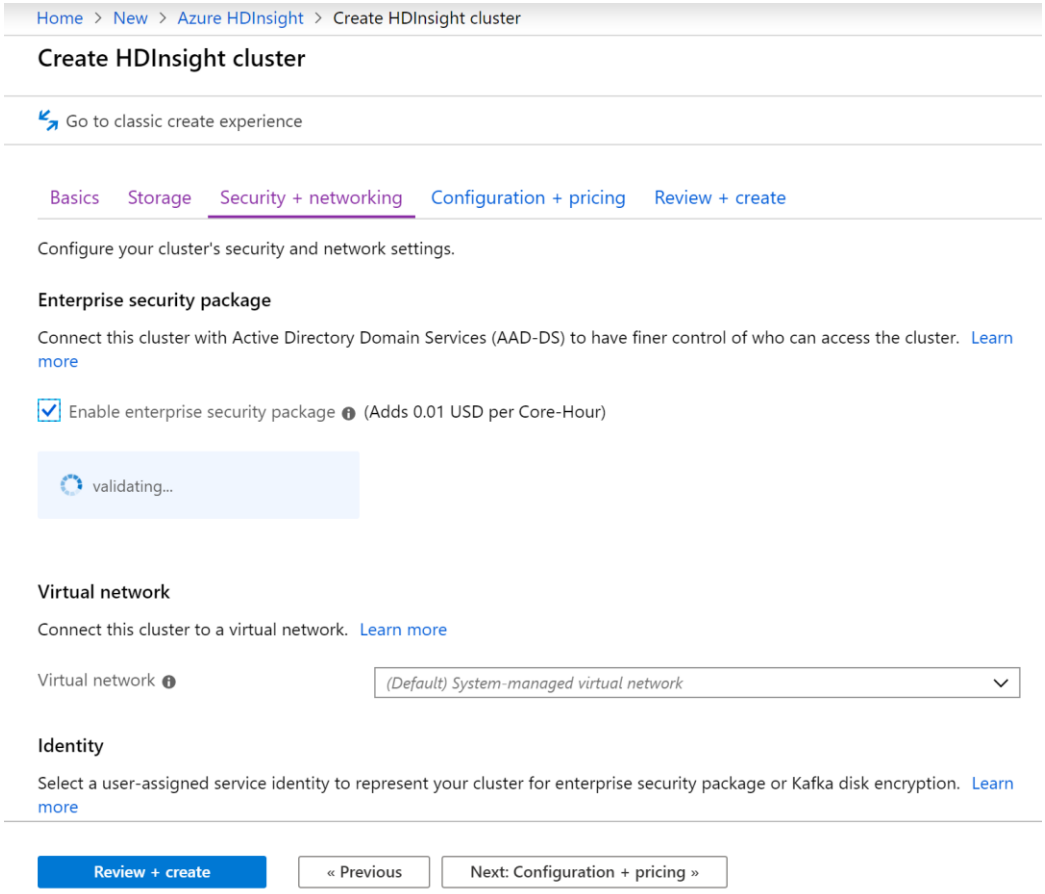


Figure 65: Create HDInsight + ESP cluster - Enable ESP

- Once enabled, the configuration process will validate existing AAD-DS setup, that is available in the subscription's tenant. It also derives necessary domain name parameters as shown below:

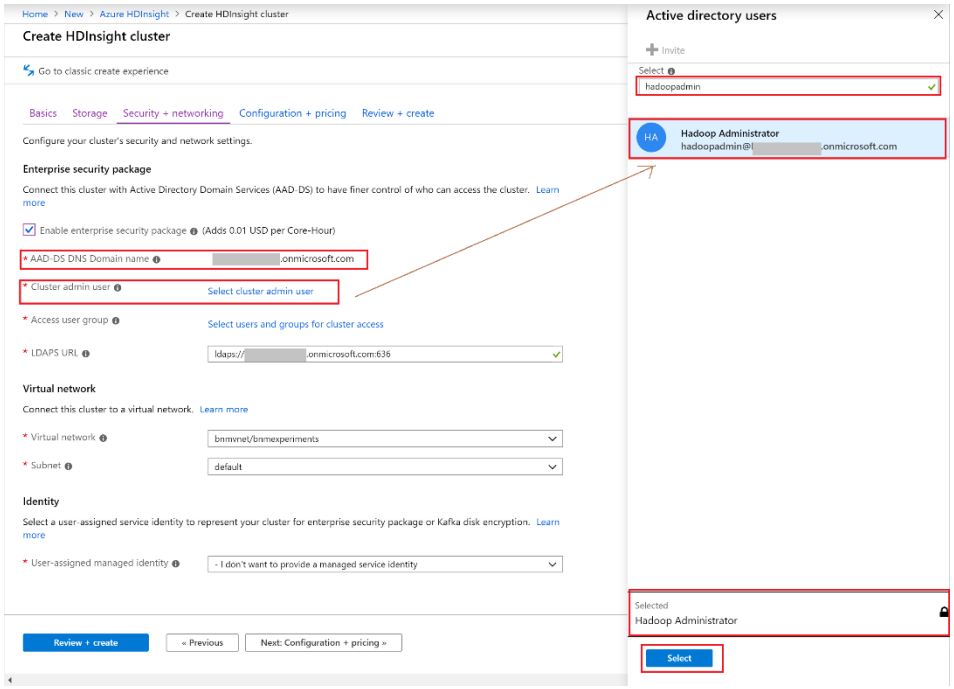
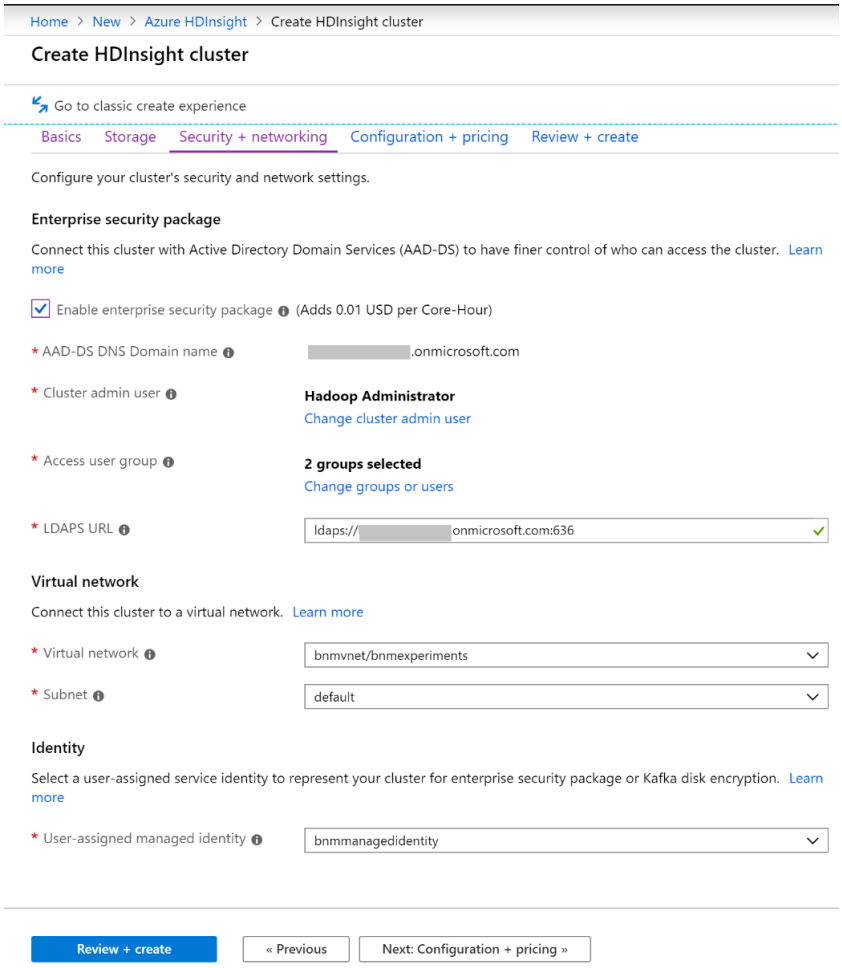


Figure 66: Create HDInsight + ESP cluster - Select Cluster Admin User

- Pay attention to the AAD-DS DNS Domain name and the LDAPS URL. These should map to your AAD-DS DNS name. For example, let's call it *youradddsdomainname.onmicrosoft.com*.
- Next will choose the *cluster admin user*, who will have permissions to manage cluster services as well grant permissions to other users for cluster access. Example, able to use Ranger and grant permission to a Hive table.
    - This user must be defined in the Azure Active Directory. For example, it should be the *hadoopadmin@youraaddsdomainname.onmicrosoft.com* user that we have created in the previous steps.
- Will next choose the additional AD user groups that should have user level access to the cluster. For example – *IngressUsers* and *EgressUsers*.
- Your configuration pane should look like, very similar to the following visual below:



- Also note the choice of VNET/Subnet choices above. The VNET choice here must be the same as that configured on Storage Account's firewall rules. Otherwise, it can be any other VNET with the exception your storage account does not block access to traffic from such virtual network.
- It is important to ensure the User-assigned Managed Identity is set to the identity that is created under the section *Configure Access Controls* section above. The identity is used by HDInsight cluster service represent on their behalf when interacting with the storage account, as well to handle on-desk encryption for Kafka.
- Next choose the configuration and pricing options, as show in the sample below:

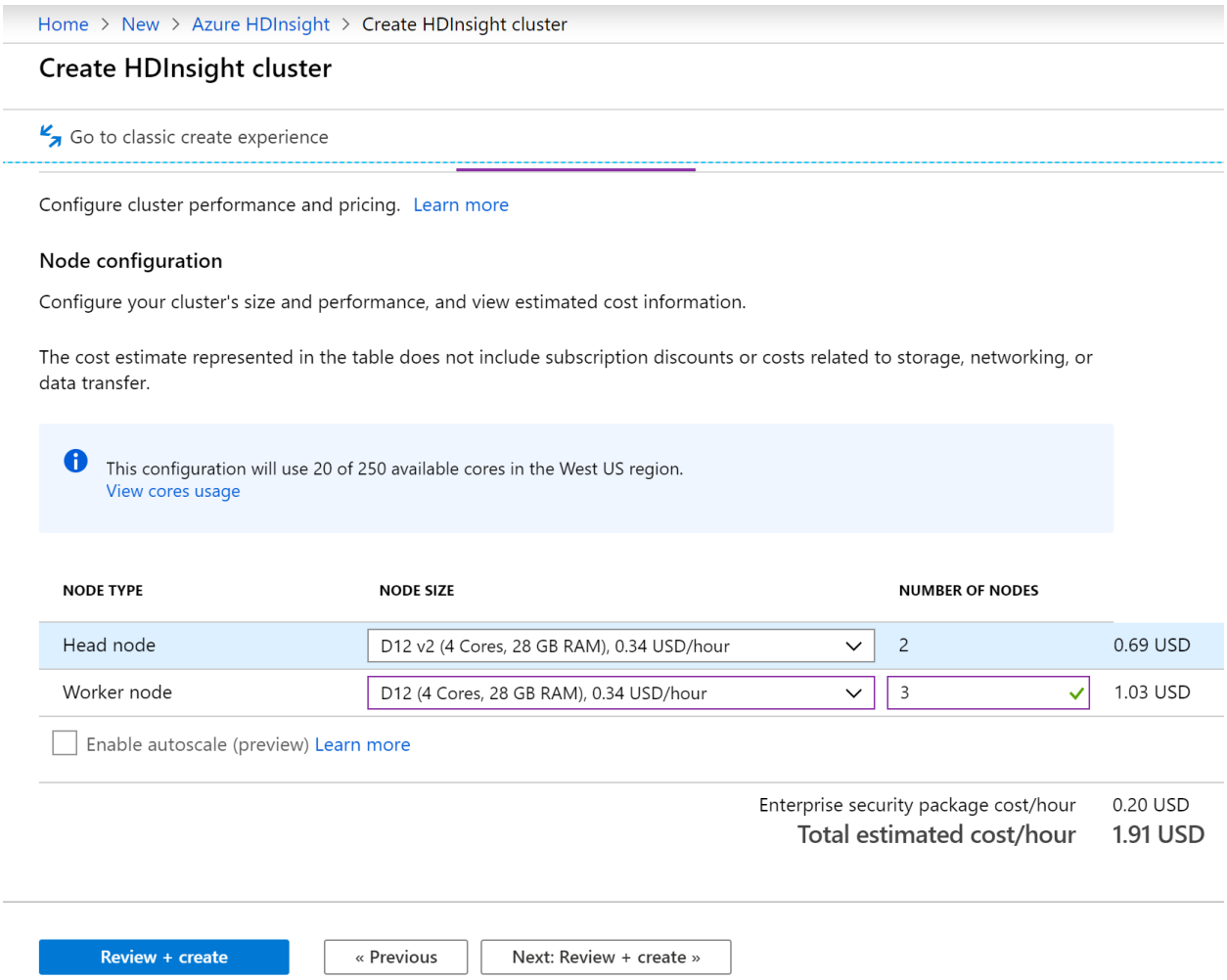


Figure 67: Create HDInsight + ESP cluster - Price Estimate

- Choice of nodes here must be derived based on the workloads and required capacity to support them. Head nodes host one or more coordinator services like YARN Resource Manager, HDFS Namenode, and Hive Master.

- Next step is to review the configurations and create the cluster.

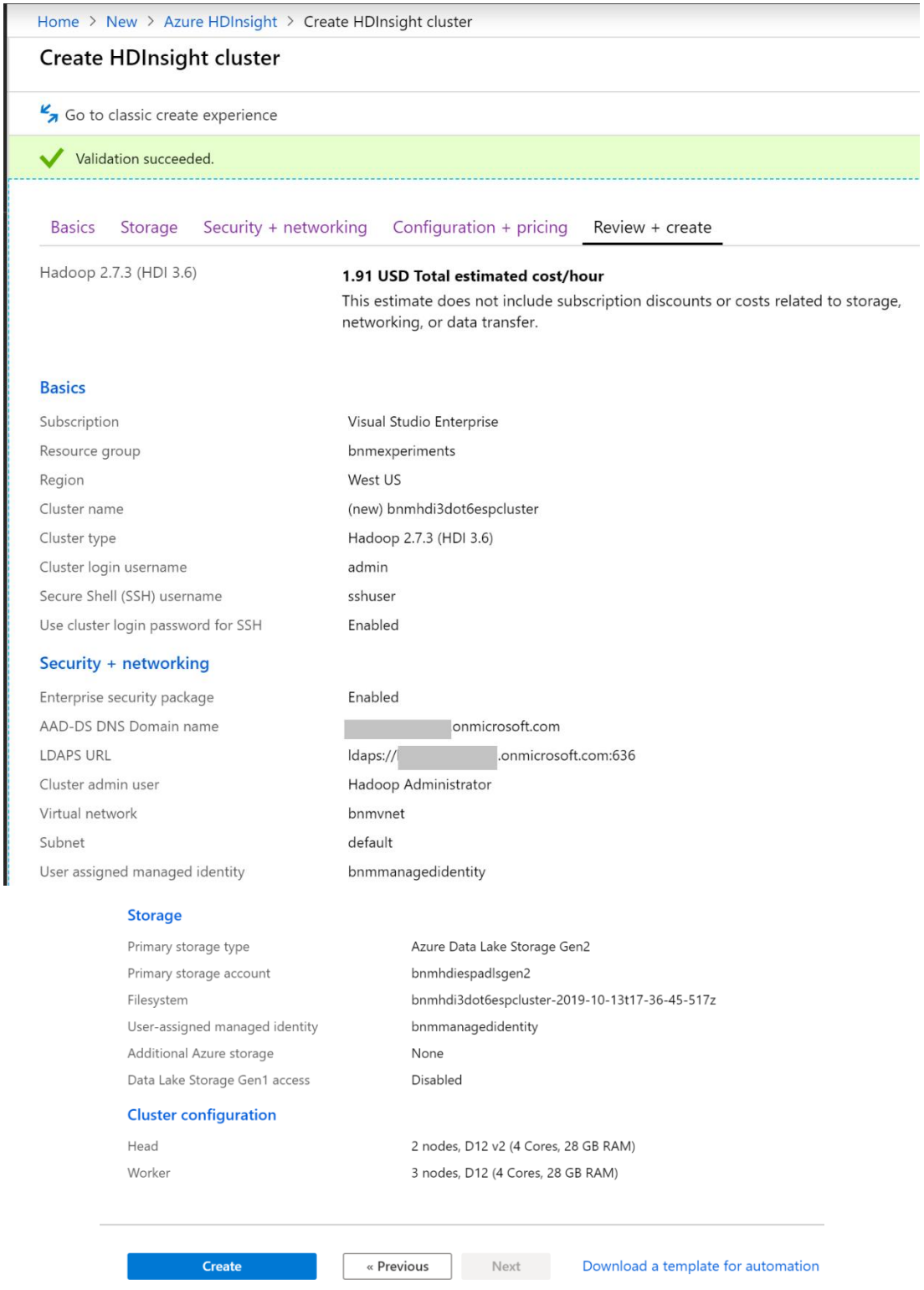


Figure 68: Create HDInsight + ESP cluster - Configuration Summary

- Submit configuration for resource creation and deployment.

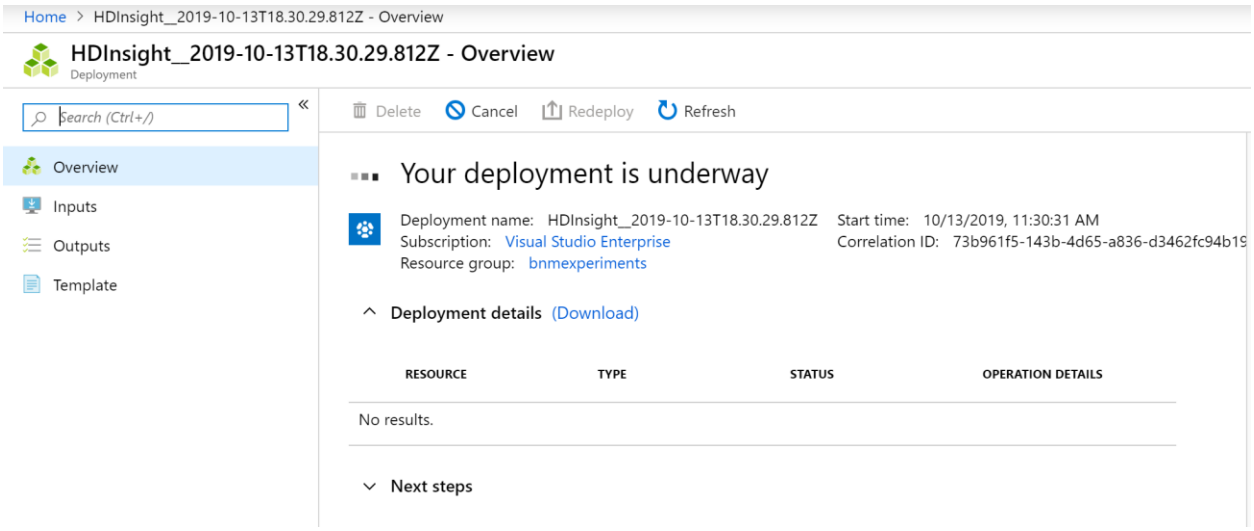
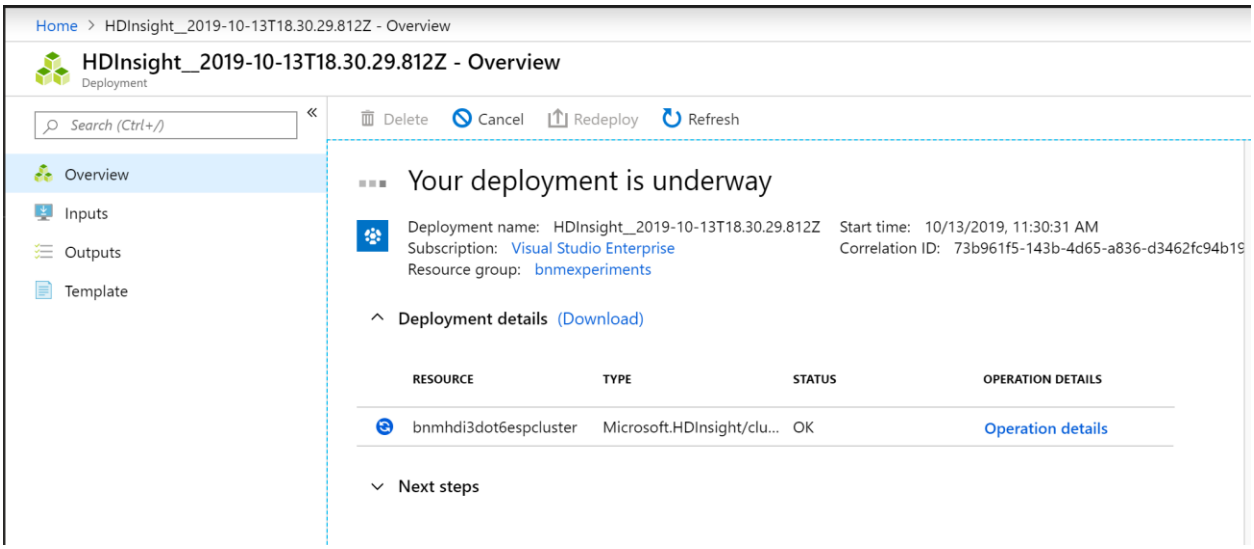


Figure 69: Create HDInsight + ESP cluster - Deployment Underway



- The deployment will take about 25 to 30 minutes. It would add whole bunch of resources including:
    - Network Interface Cards (NIC) for each of the nodes – gateway, head, zookeeper and worker nodes.
    - Load balancers for the gateway and head nodes.
    - HDInsight cluster
    - Etc. as required for a functional HDInsight cluster with ESP enabled.
- Here is the sample visual that shows the detailed provided by the portal on successful deployment of the cluster.

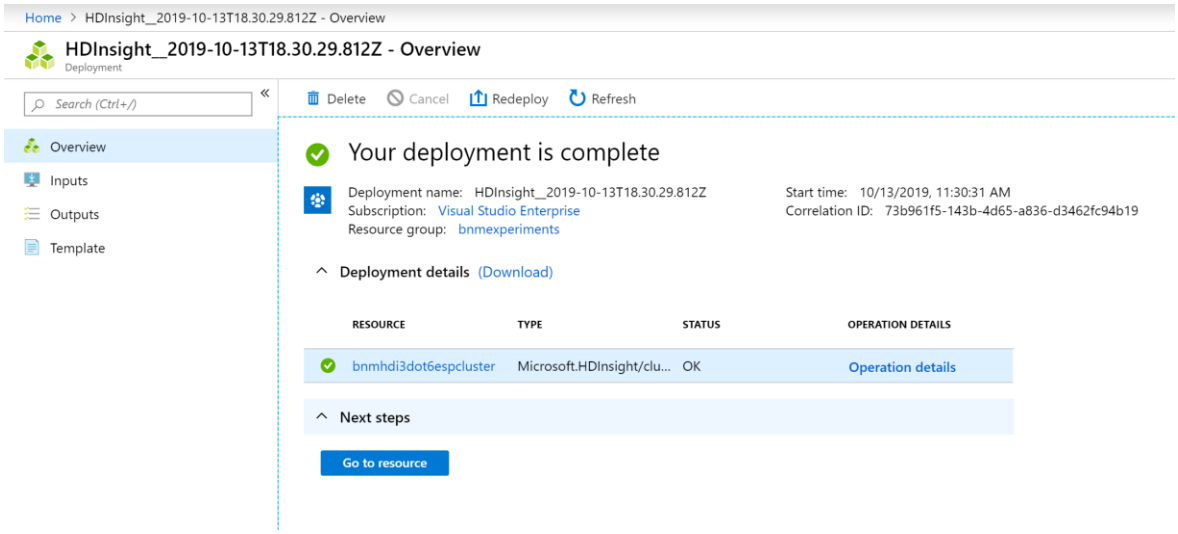


Figure 70: Create HDInsight + ESP cluster - Deployment Complete

- And, when you click on the *Operation details* link (see under the deployment details above), you should see detail that is very similar to following:

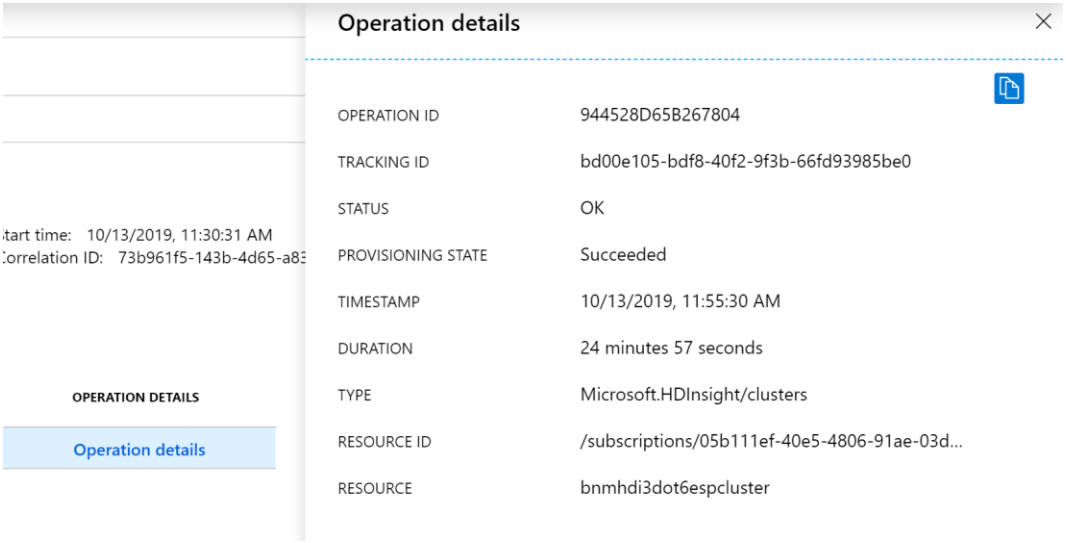


Figure 71: Create HDInsight + ESP cluster - Deployment Operation Details

- Here is the sample visual reference that shows the overview page of the newly created secure HDInsight cluster:

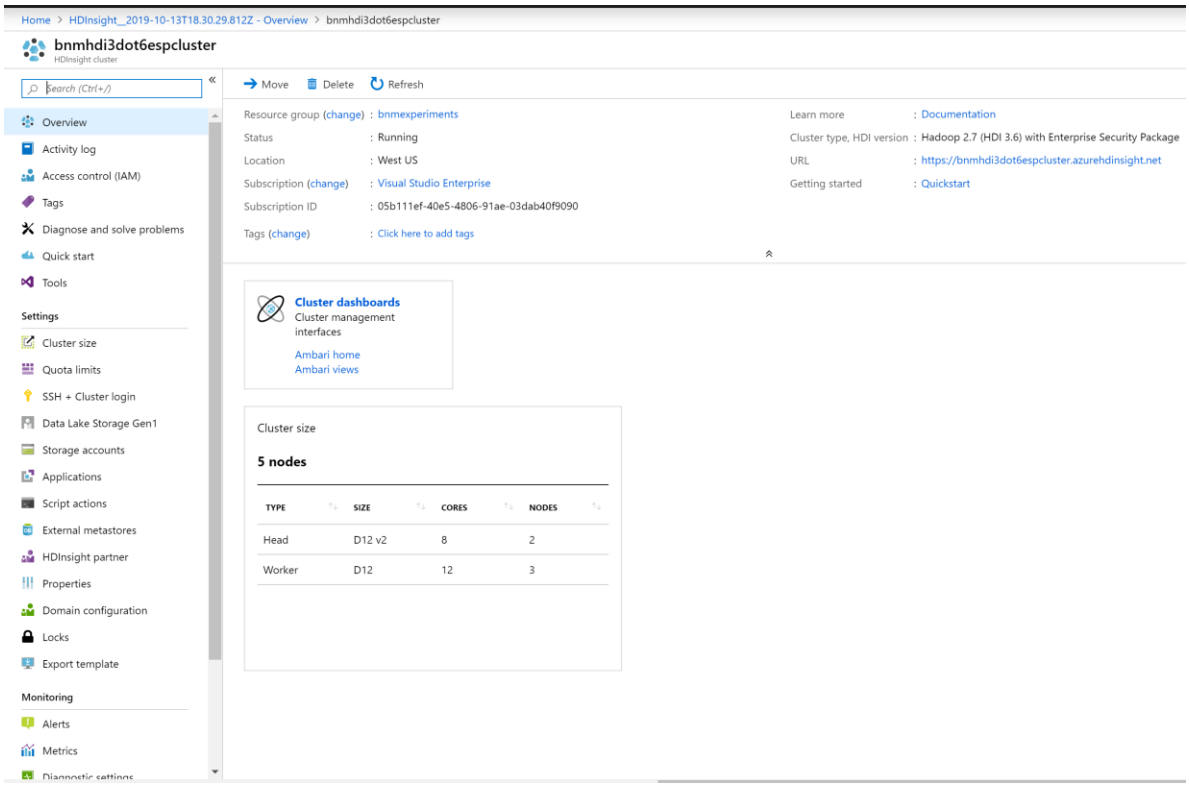


Figure 72: HDInsight + ESP cluster overview

- The URL shown to the top right of the pane above is the Ambari Web UI. Ambari listens on port 443. NSG rules must be configured to allow inbound traffic on port 443. Without which, access to the Ambari Web UI cannot be satisfied by your browser. Here's the sample error for additional details of specific browser error:

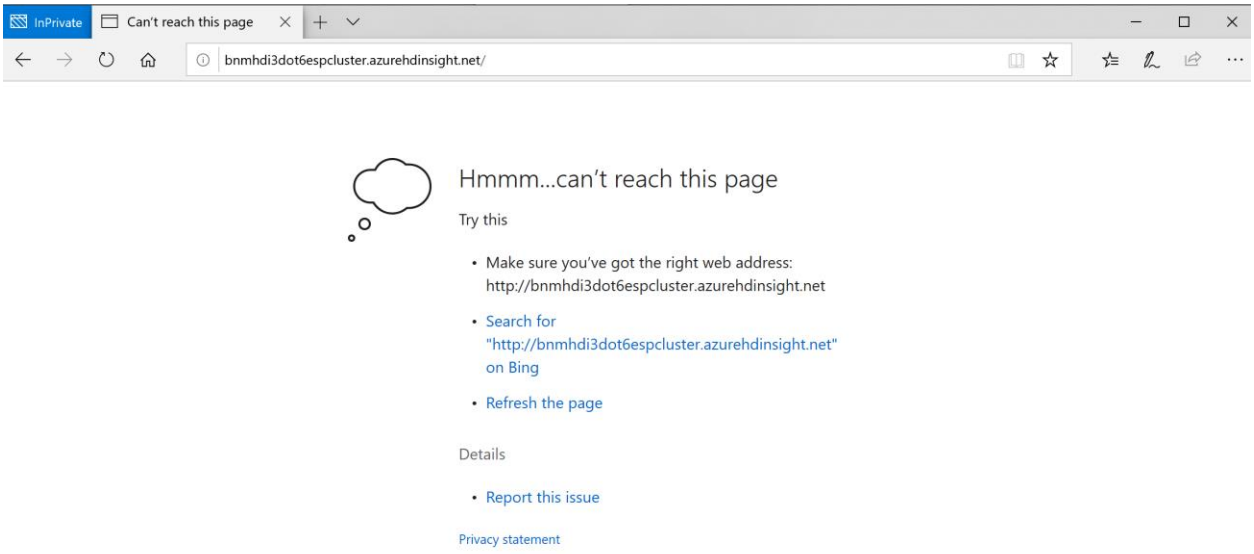
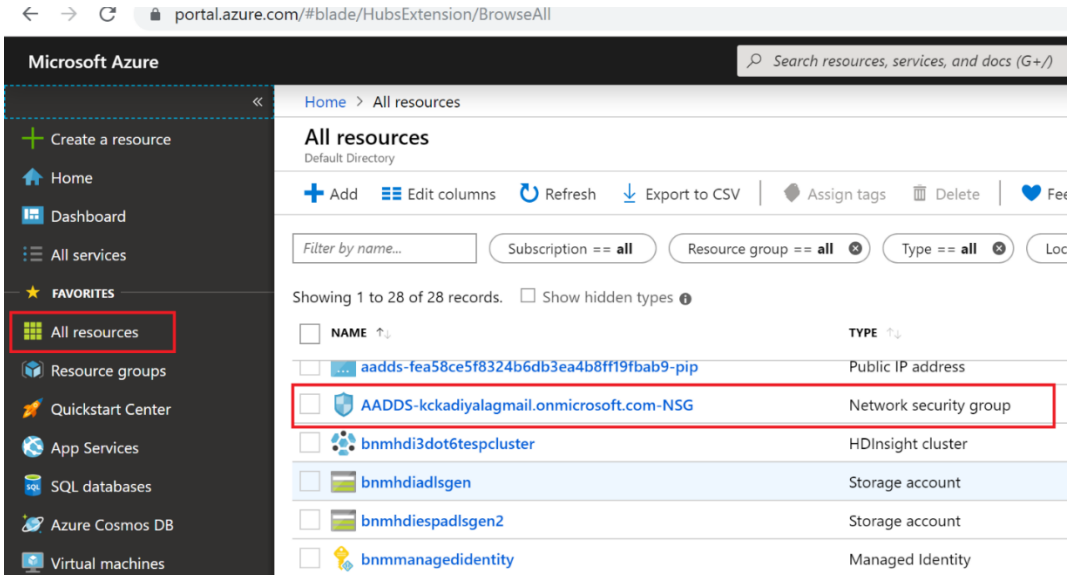


Figure 73: Ambari UI - Cannot reach the page

6.2 Enable NSG rules for Ambari UI (port 443)

As observed from the previous step, without enabling NSG rule for inbound traffic on secure HTTP port, users will not be able to access the Ambari Web UI. In this section will add NSG rule to enable such access.

- Navigate to the NSG resource using the *All Resources* link available on the left navigation pane of the portal.



- Click + Add link to add an entry for the port 443 inbound TCP traffic.

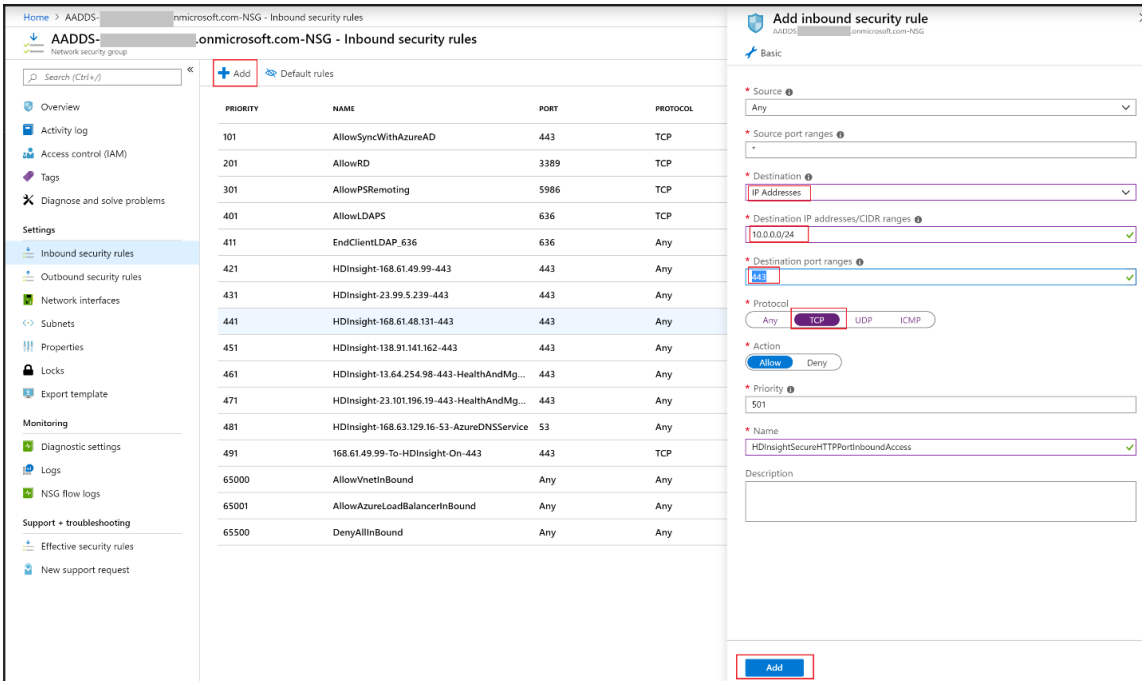


Figure 74: Enable Port 443 on NSG for Ambari UI

6.3 Verify Ambari UI HTTPs Port access

- Once the rule is added it will be part of the NSG Inbound Security rules list as shown below:

421	HDInsight-168.61.49.99-443	443	Any	168.61.49.99	Any	Allow	...
431	HDInsight-23.99.5.239-443	443	Any	23.99.5.239	Any	Allow	...
441	HDInsight-168.61.48.131-443	443	Any	168.61.48.131	Any	Allow	...
451	HDInsight-138.91.141.162-443	443	Any	138.91.141.162	Any	Allow	...
461	HDInsight-13.64.254.98-443-HealthAndMg...	443	Any	13.64.254.98	Any	Allow	...
471	HDInsight-23.101.196.19-443-HealthAndMg...	443	Any	23.101.196.19	Any	Allow	...
481	HDInsight-168.63.129.16-53-AzureDNSService	53	Any	168.63.129.16	Any	Allow	...
491	168.61.49.99-To-HDInsight-On-443	443	TCP	168.61.49.99	VirtualNetwork	Allow	...
501	HDInsightSecureHTTPPortInboundAccess	443	TCP	Any	10.0.0.0/24	Allow	...

Figure 75: NSG Inbound Rule Updated - Ambari UI - HTTPS port enabled

- Give it few seconds and refresh your browser. You should see following login prompt, where you can use the user name as the *hadoopadmin@youraaddsdomainname.onmicrosoft.com* and the password as configured from earlier AD user setup process.

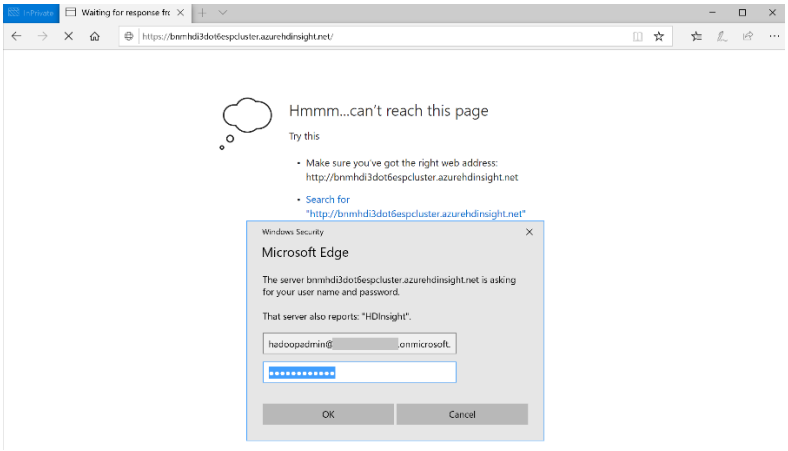


Figure 76: Retry accessing Ambari Web UI

- A successful login would land you at the Ambari Web UI – Dashboard view, as shown below:

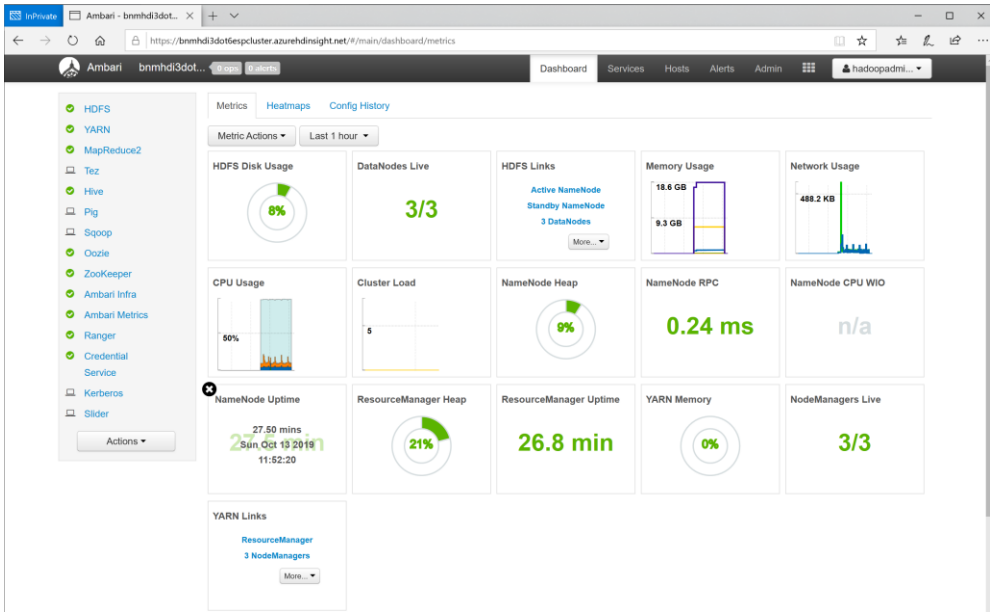


Figure 77: Ambari Dashboard View

- And, here is the visual reference to the HDFS service details:

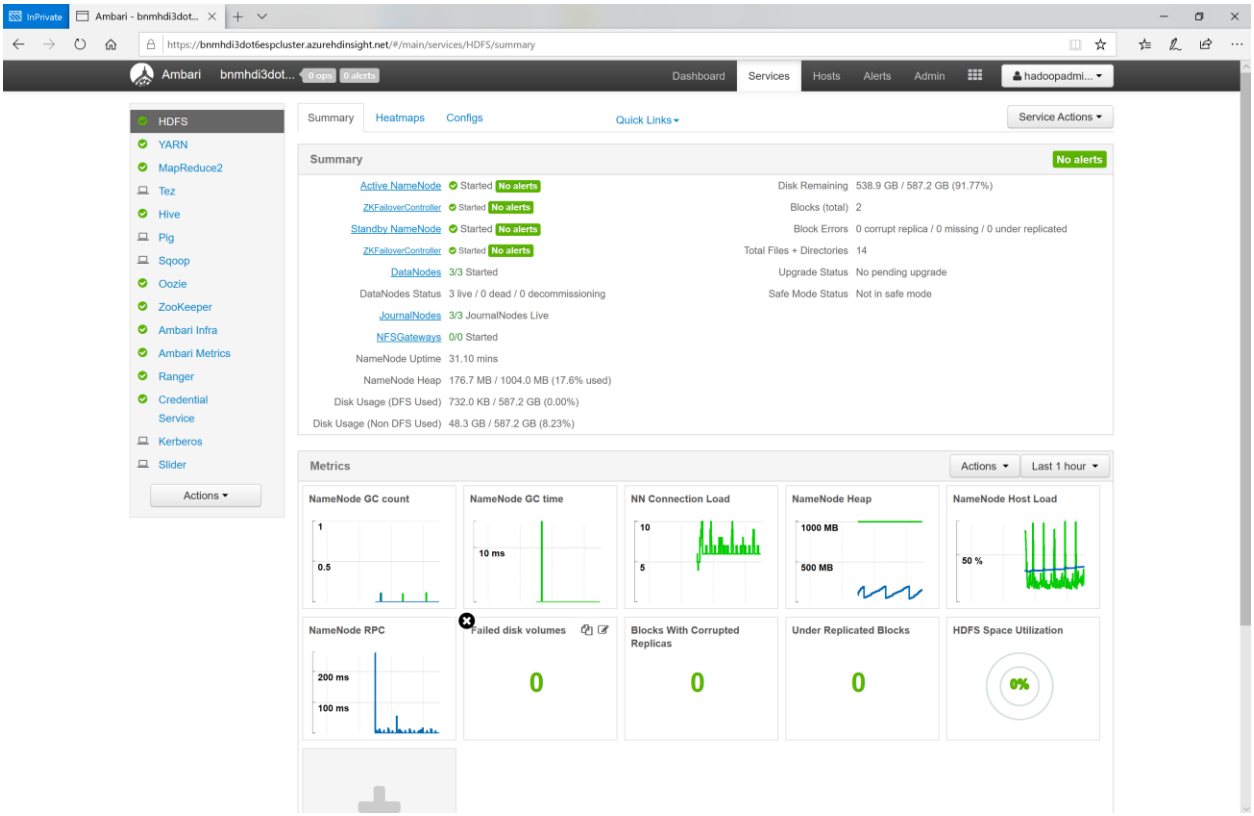


Figure 78: Ambari Web UI - HDFS Summary

6.4 Setup SSH-Tunneling to access Ambari URLs

To access resource URL's such as Namenode, Job History, etc., we must be configure SSH-tunneling. Before we attempt SSH-tunneling, the SSH port 22 must be enabled for in-bound traffic using the NSG rules. Such configuration must be limited to destination IP address range of 10.0.0.0/24 and protocol support limited to TCP. See below an entry for quick reference:

Figure 79: Enable SSH access on NSG

Click here to know more about methods to enable SSH-tunnel. For this experiment here, will use SSH-tunneling approach, where in a tunnel or port forwarding will be established on localhost @ 9876. All calls will be routed to destination FQDN – bnmhdi3dot6espcluster-ssh.azurehdinsight.net, which is the SSH hostname for the head node.

Following is the reference CLI output:

```
kalyan@myhost-name:/mnt/c/Users/userid$ ssh -C2qTnNv -D 9876 sshuser@bnmhdi3dot6espcluster-
ssh.azurehdinsight.net
OpenSSH_7.4p1 Debian-10+deb9u6, OpenSSL 1.0.2r  26 Feb 2019
debug1: Reading configuration data /home/kalyan/.ssh/config
debug1: Reading configuration data /etc/ssh/ssh_config
debug1: /etc/ssh/ssh_config line 19: Applying options for *
debug1: Connecting to bnmhdi3dot6espcluster-ssh.azurehdinsight.net [137.135.47.123] port 22.
debug1: Connection established.
debug1: identity file /home/kalyan/.ssh/id_rsa type 1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_rsa-cert type -1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_dsa type -1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_dsa-cert type -1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_ecdsa type -1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_ecdsa-cert type -1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_ed25519 type -1
debug1: key_load_public: No such file or directory
debug1: identity file /home/kalyan/.ssh/id_ed25519-cert type -1
debug1: Enabling compatibility mode for protocol 2.0
debug1: Local version string SSH-2.0-OpenSSH_7.4p1 Debian-10+deb9u6
debug1: Remote protocol version 2.0, remote software version OpenSSH_7.2p2 Ubuntu-4ubuntu2.8
debug1: match: OpenSSH_7.2p2 Ubuntu-4ubuntu2.8 pat OpenSSH* compat 0x04000000
debug1: Authenticating to bnmhdi3dot6espcluster-ssh.azurehdinsight.net:22 as 'sshuser'
debug1: SSH2_MSG_KEXINIT sent
debug1: SSH2_MSG_KEXINIT received
debug1: kex: algorithm: curve25519-sha256@libssh.org
debug1: kex: host key algorithm: ecdsa-sha2-nistp256
debug1: kex: server->client cipher: aes128-ctr MAC: umac-64-etm@openssh.com compression:
zlib@openssh.com
debug1: kex: client->server cipher: aes128-ctr MAC: umac-64-etm@openssh.com compression:
zlib@openssh.com
debug1: expecting SSH2_MSG_KEX_ECDH_REPLY
debug1: Server host key: ecdsa-sha2-nistp256 SHA256:9xy78s6sdMRCCPLcPGSuypodYstkkuD2dPYgijtyyto
The authenticity of host 'bnmhdi3dot6espcluster-ssh.azurehdinsight.net (137.135.47.123)' can't be
established.
ECDSA key fingerprint is SHA256:9xy78s6sdMRCCPLcPGSuypodYstkkuD2dPYgijtyyto.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'bnmhdi3dot6espcluster-ssh.azurehdinsight.net' (ECDSA) to the list of
known hosts.
debug1: rekey after 4294967296 blocks
debug1: SSH2_MSG_NEWKEYS sent
debug1: expecting SSH2_MSG_NEWKEYS
debug1: SSH2_MSG_NEWKEYS received
debug1: rekey after 4294967296 blocks
debug1: SSH2_MSG_EXT_INFO received
debug1: kex_input_ext_info: server-sig-algs=<rsa-sha2-256,rsa-sha2-512>
debug1: SSH2_MSG_SERVICE_ACCEPT received
Authorized uses only. All activity may be monitored and reported.
debug1: Authentications that can continue: publickey,password
debug1: Next authentication method: publickey
debug1: Offering RSA public key: /home/kalyan/.ssh/id_rsa
debug1: Authentications that can continue: publickey,password
```

```
debug1: Trying private key: /home/kalyan/.ssh/id_dsa
debug1: Trying private key: /home/kalyan/.ssh/id_ecdsa
debug1: Trying private key: /home/kalyan/.ssh/id_ed25519
debug1: Next authentication method: password
sshuser@bnmhdi3dot6espcluster-ssh.azurehdinsight.net's password:
debug1: Enabling compression at level 6.
debug1: Authentication succeeded (password).
Authenticated to bnmhdi3dot6espcluster-ssh.azurehdinsight.net ([137.135.47.123]:22).
debug1: Local connections to LOCALHOST:9876 forwarded to remote address socks:0
debug1: Local forwarding listening on ::1 port 9876.
debug1: channel 0: new [port listener]
debug1: Local forwarding listening on 127.0.0.1 port 9876.
debug1: channel 1: new [port listener]
debug1: Requesting no-more-sessions@openssh.com
debug1: Entering interactive session.
debug1: pledge: network
debug1: client_input_global_request: rtype hostkeys-00@openssh.com want_reply 0
```

The next step is the configure browser's proxy settings. For the context will use Firefox Browser, where in the proxy settings can be managed using the Browser properties itself. Unlike, Edge or Chrome where the proxy settings are delegated to the user's laptop network configurations. Following visual provides a sample view of how to set the SOCKS5 proxy configuration:

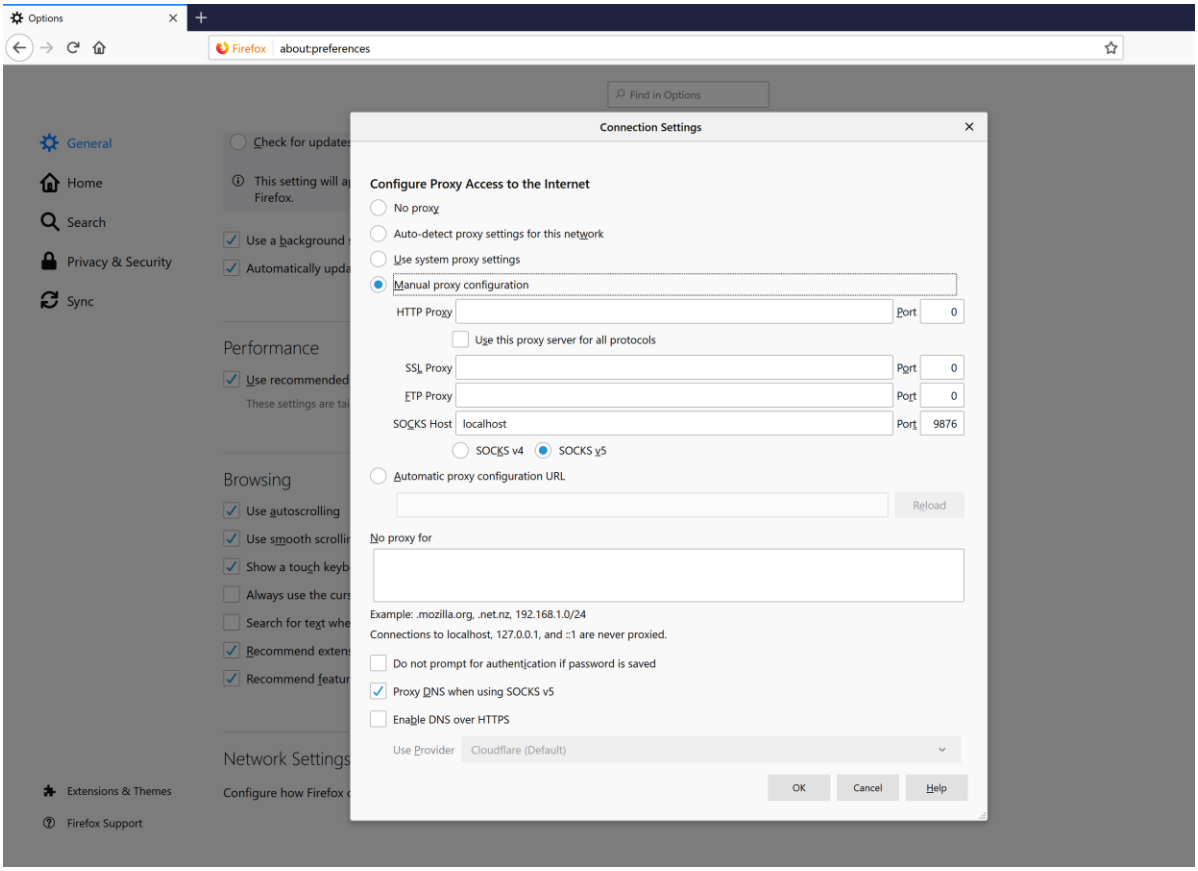


Figure 80: Firefox - SOCKS5 Configuration

6.5 Validate SSH-Tunnel

Now that you have updated the proxy settings, check if the tunneling works by accessing the site - https://www.whatismyip.com/. The web page should return the IP address of one of the Microsoft data center addresses, as shown in the following visual reference:

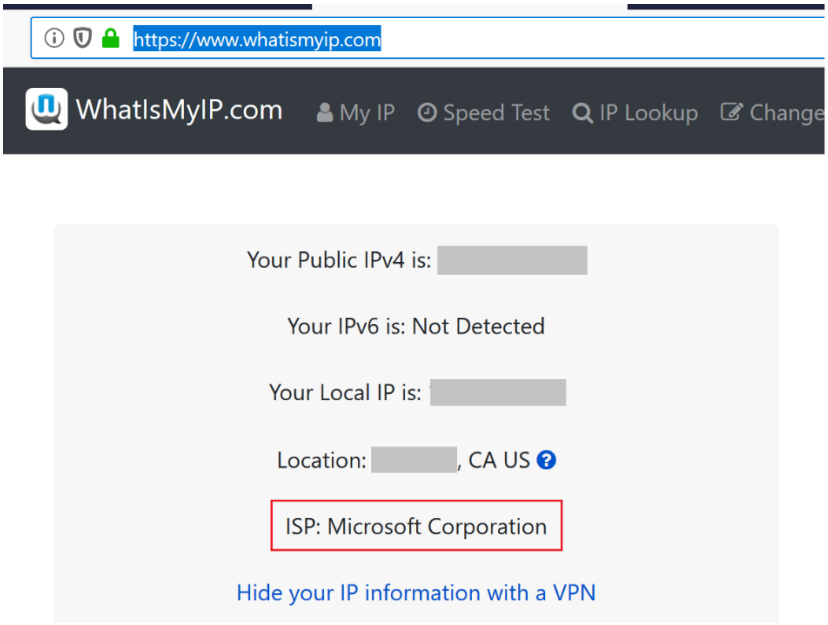


Figure 81: SOCKS5 Proxy Enabled - Whatismyip.com verification

Since we have started the SSH-tunnel (port forwarding) in debug mode, you should see console output that relates to browser activity. Here are the URLs that you can try to connect to the Ambari UI on the head node:

- Access via SSH-tunnel (HTTP) – secure over SSH:
    - http://hn0-bnmhdi.youraaddnsdomainname.onmicrosoft.com:8080
- Access via public internet (HTTPS) – secure HTTP:
    - https://bnmhdi3dot6espcluster.azurehdinsight.net

Let's try with the SSH-tunnel bound HTTP URL. Note all traffic is encrypted and routed through the secure tunnel (port forwarding mechanism). When accessed the browser will prompt you to login to the Ambari UI, as shown in the following visual reference:

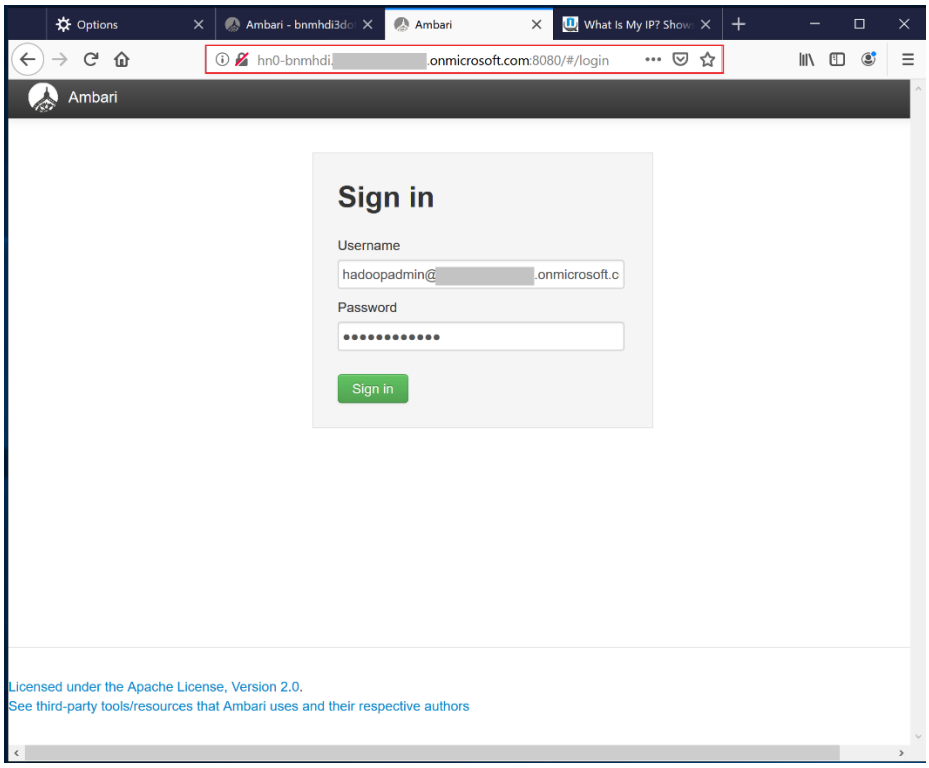


Figure 82: Ambari Web UI access using SSH-tunneling (port-forwarding)

6.6 Known Issues – Web UI

At the time, this document is being developed, there are few noted issues associated with ability to access Namenode UI. These are detailed below:

- The links in the Namenode UI does not work when logged in to Ambari UI using the HTTPS route (public internet). The user name provided in the login prompt is in the format – *hadoopadmin@youraaddsdomainname.onmicrosoft.com*. Here's the visual reference of the landing page view:

Figure 83: HDInsight - Ambari - Namenode UI (Known Issue)

- SSH tunneling provides an encrypted connection to the target head-node in the HDInsight cluster. Hence, the URL to the Ambari portal will be different as described in the validate

section above. However, the user cannot login to the portal using the complete identity, i.e., hadoopadmin@youraaddsdomainname.onmicrosoft.com. You can only use the name portion of it i.e., *hadoopadmin*.

- When logged into Ambari using the SSH-tunneling approach, the login works successfully as described in the previous point. However, you cannot access the HDFS Namenode UI page. The UI throws an error that is like the following visual:

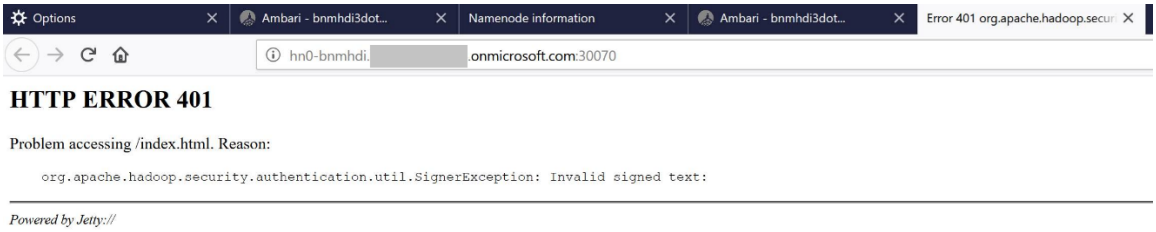
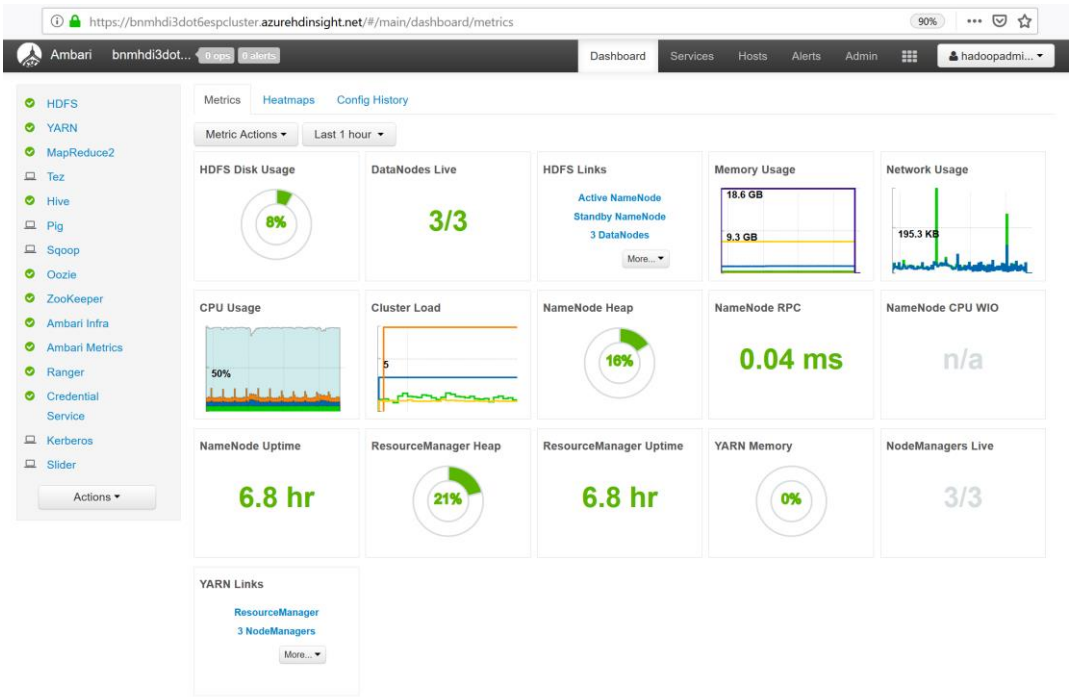


Figure 84: Ambari Web UI - Using SSH-tunnel - HTTP Error 401

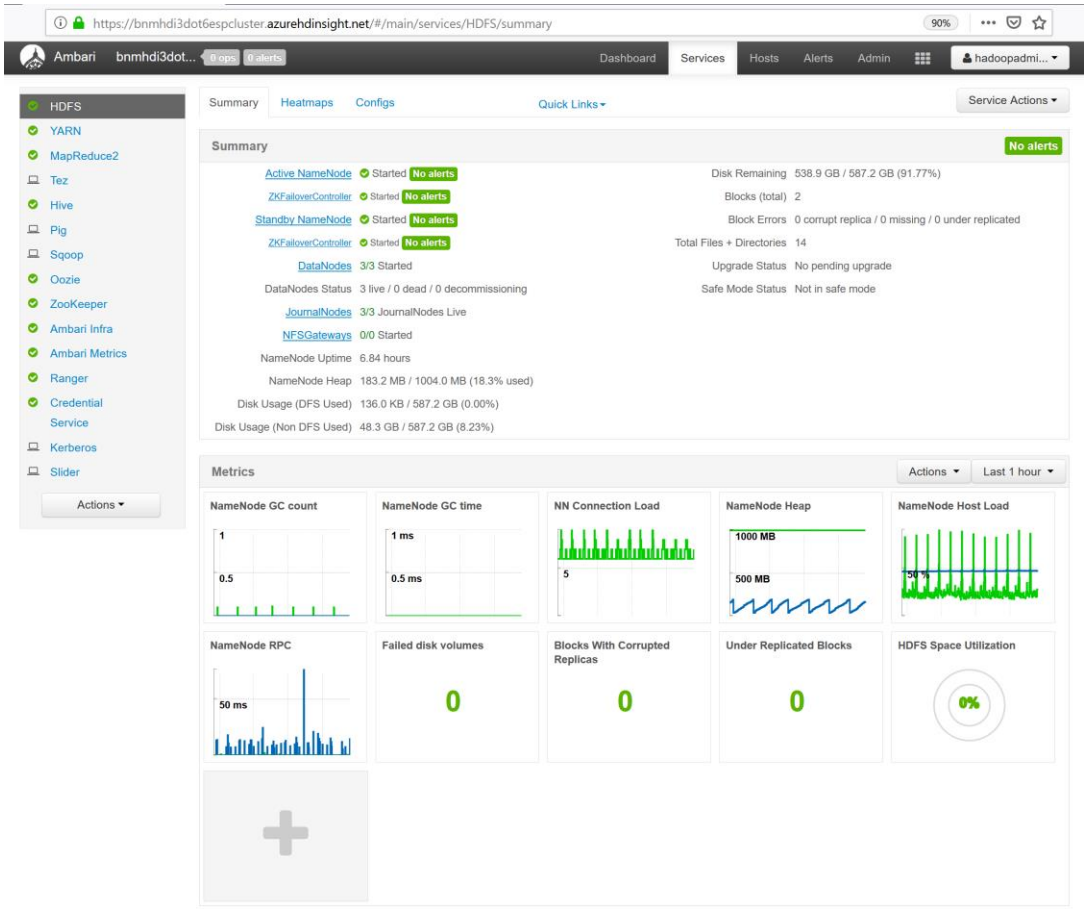
7 Post Installation Review

We have observed few issues accessing Namenode UI or the ability to access YARN Resource Manager UI (when using the SSH-tunnel/port-forwarding route), specifically using the Quick Links section. However, one can validate each of the services from the service specific link that is available on the landing page. The previously noted issues do not cause hindrance to access actions that you as an administrator can perform.

7.1 Ambari Landing Page



7.2 HDFS Service Summary



7.3 YARN Summary

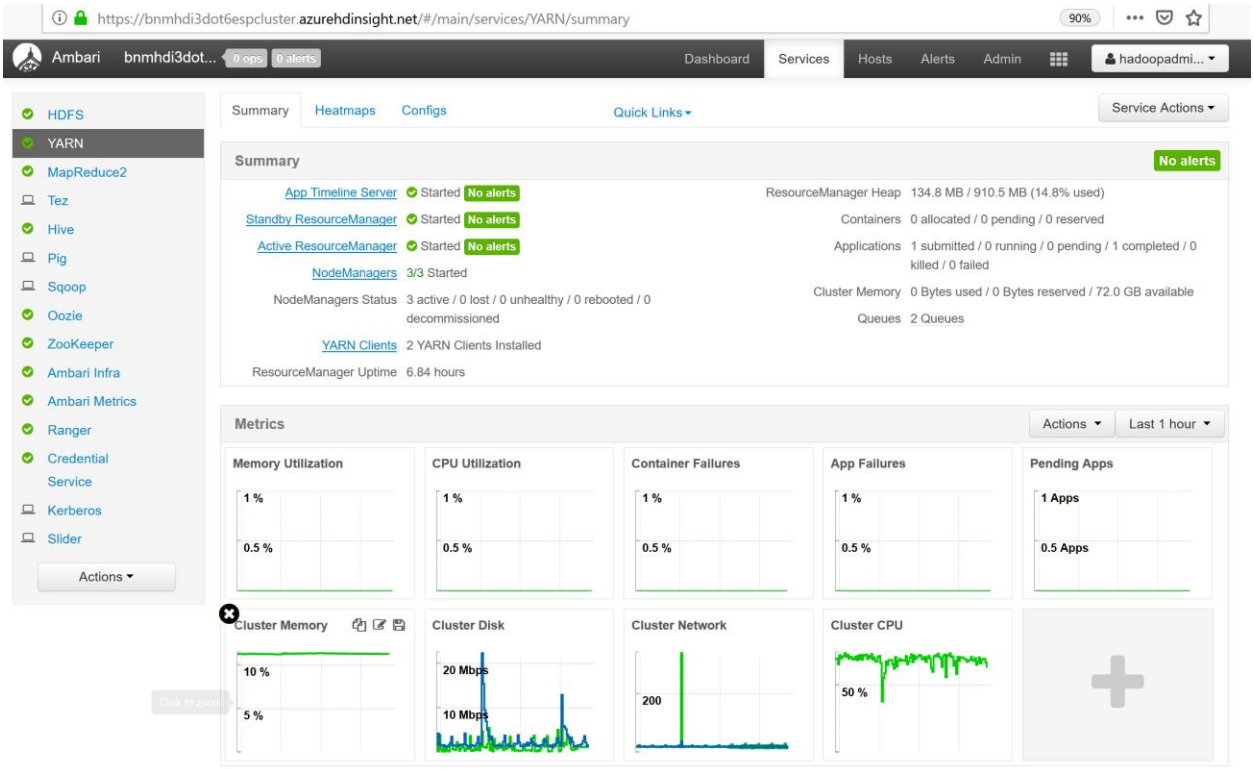


Figure 85: HDInsight YARN Summary

7.4 Ranger

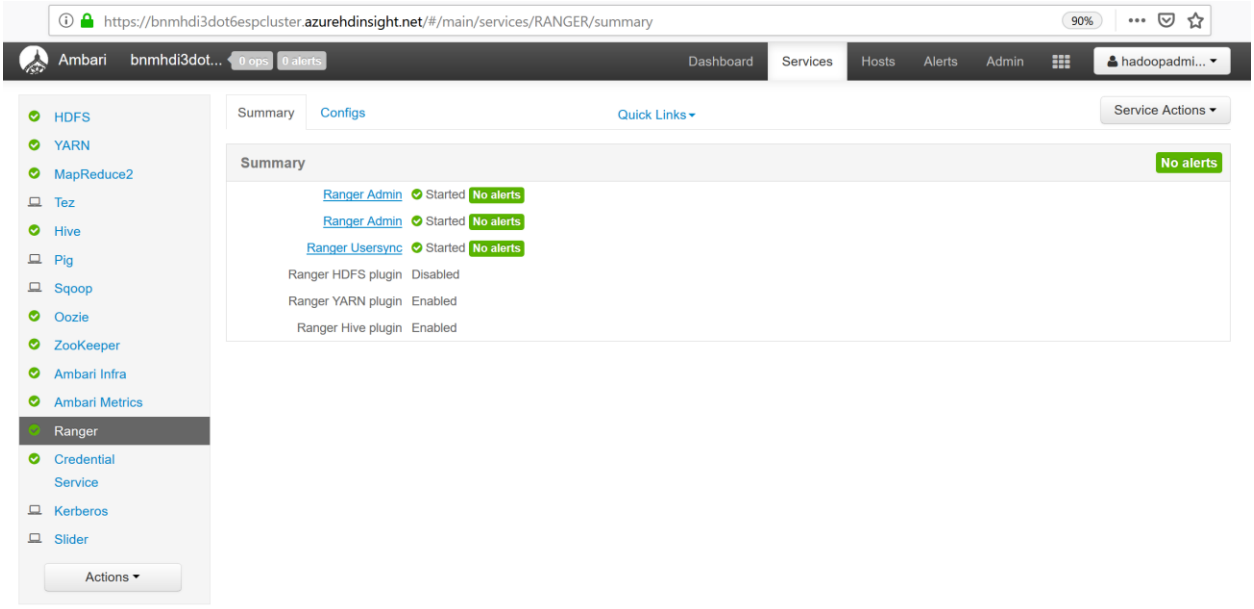


Figure 86: HDInsight - Ranger Overview

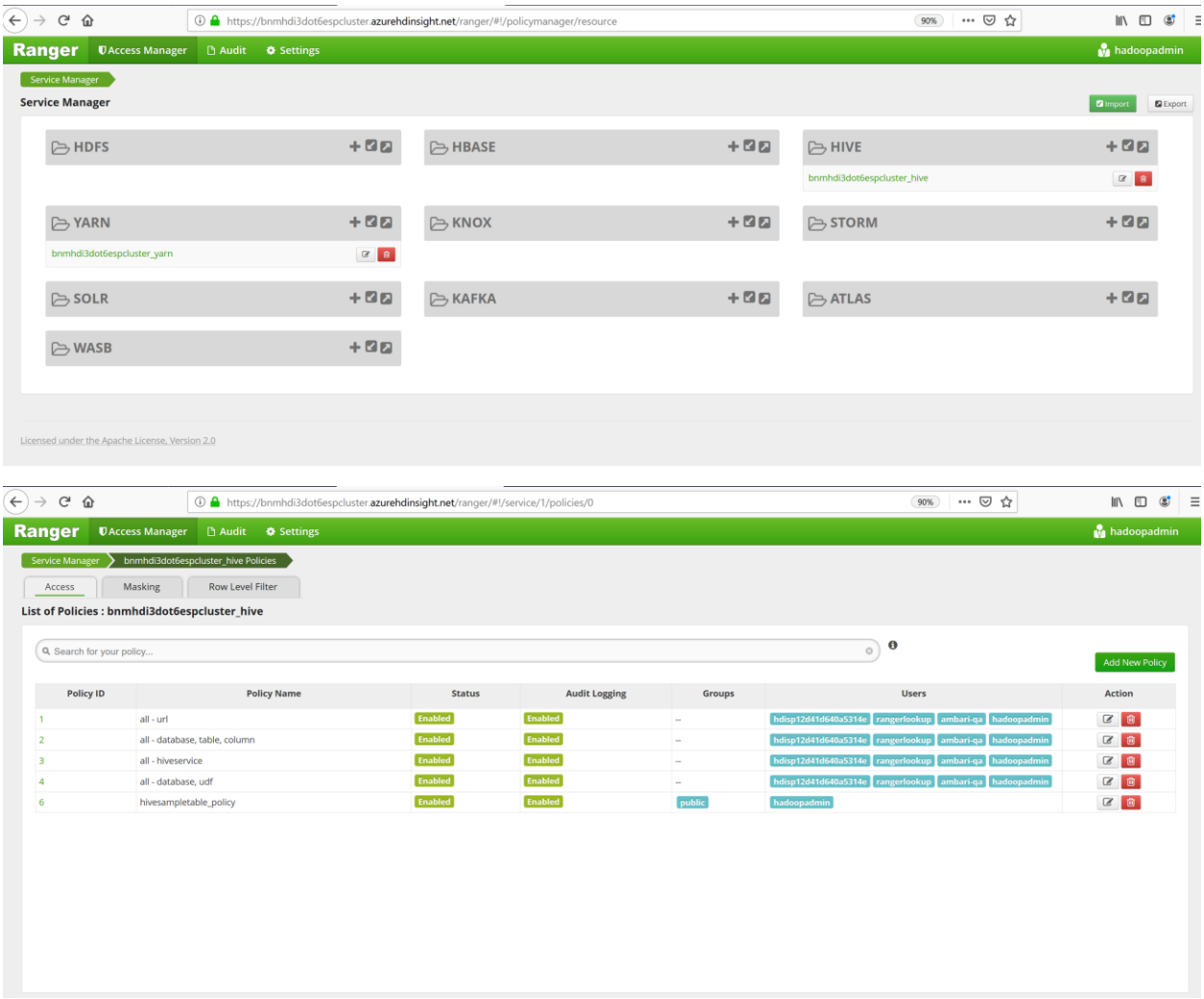


Figure 87: HDInsight - Ranger - Hive Overview

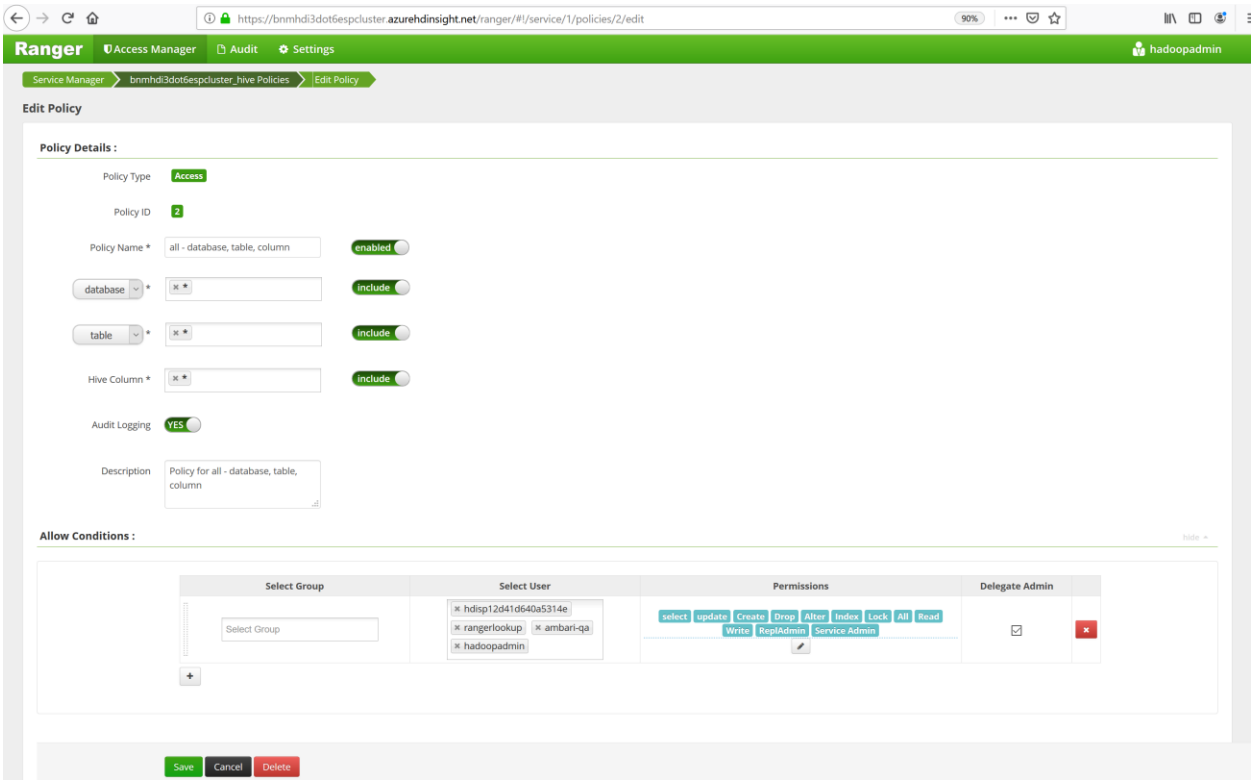


Figure 88: HDInsight - Ranger - Hive Sample Policy

7.5 Storage explorer – HDInsight Container view

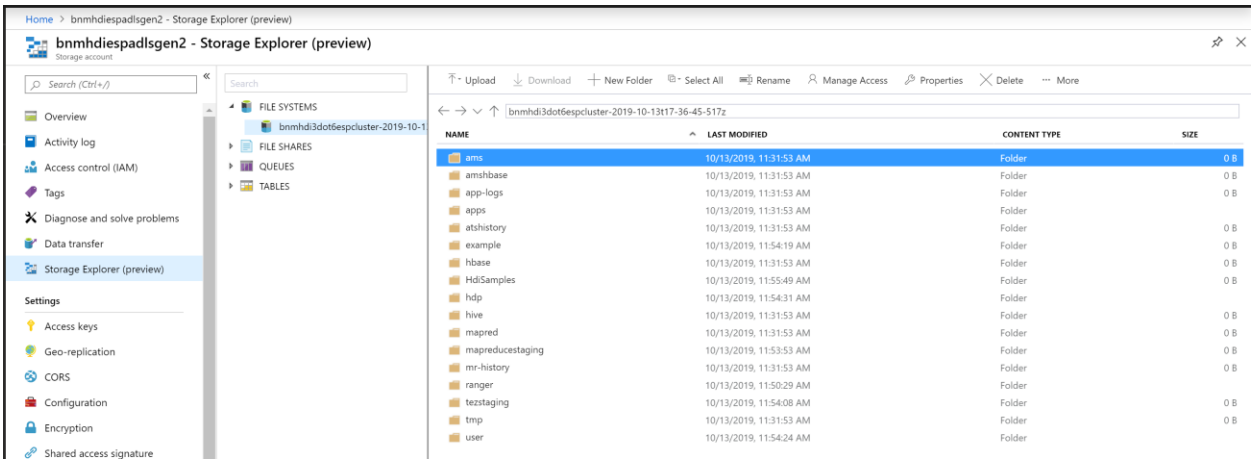


Figure 89: HDInsight Storage Container View using Storage Explorer for *hadoopadmin*

8 Appendix – Reference URLs

- Plan a virtual network for Azure HDInsight
- Directly connect to Apache Hadoop Services
- Configure DNS Zone for external access
- Remote Server Administration Tools
- Tutorial: Configure secure LDAP for an Azure Active Directory Domain Services managed domain - Test Queries to the managed domain
- Tutorial: Enable password synchronization in Azure Active Directory Domain Services for hybrid environments
- HDInsight management IP addresses
- Ports used by Apache Hadoop services on HDInsight
- Introduction to Azure Data Lake Storage Gen2
- Create and Authorize a managed identity
- What is Azure Private Endpoint?
- Use Azure Data Lake Storage Gen2 with Apache Hadoop in Azure HDInsight
- Azure Active Directory – Domain Services (AAD-DS)
- Troubleshoot cluster creation failures with Azure HDInsight
- Manage access to Azure resources using RBAC and the Azure Portal
- Best practice for using RBAC
- Users of HDInsight clusters with ESP
- Synchronize AAD users with Ambari REST API (immediate synchronization)
- Use SSH tunneling

9 Feedback and suggestions

If you have feedback or suggestions for improving this data migration asset, please contact the Data SQL Ninja Team (datasqlninja@microsoft.com). Thanks for your support!

Note: For additional information about migrating various source databases to Azure, see the Azure Database Migration Guide.