# **Objective**

This training is intended for newly hired engineers aiming to learn Microsoft Azure HDInsight and support customers using HDInsight in various roles.

# **Pre-requisite**

1. General understanding of apache Hadoop
2. General familiarity with Windows and Linux platform
3. An Azure subscription
4. Access to Ramweb projects and Security groups to be able to use internal tools:
5. For using tools like Kusto, your email address needs to be added to hadoopsupport alias/SG
6. Get access to the following Ramweb project:

<https://ramweb>



# **Core Hadoop Rampup**

1. Familiarity with Hadoop, HDFS, MapReduce and Yarn
2. For Hadoop and HDFS fundamentals, please read <https://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520>
3. For Yarn fundamentals, please read <https://www.amazon.com/Apache-Hadoop-YARN-Processing-Addison-Wesley/dp/0321934504>
4. For Yarn on HDInsight, please review the triage by HDInsight PG

<https://learningcentral/Training/CoursePlayer.aspx?courseId=170072>

1. Hortonworks Hadoop training

# **What is HDInsight? (1 hour)**

1. Review the article and get a high level overview of Hadoop and HDInsight ecosystem:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/>

1. Review the blog to get high level understanding of HDInsight architecture, gateway and Azure storage (WASB) – the information may be a little outdated, but the high level architecture is still valid today. You can ignore the details of components, versions etc – just focus on the big picture.

<https://blogs.msdn.microsoft.com/bigdatasupport/2013/11/01/the-hdinsight-support-team-is-open-for-business/>

1. Review the article and understand at a high level what the offerings, versions and components are- again focus on the big picture, not details. **DONE**

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-component-versioning/>

# **How HDFS works in HDInsight – the role of Azure Storage (2 hours)**

1. Review the article:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-use-blob-storage/>

1. Review the PG presentation (video and slides) on this:

[\\dsdb\sqlskills\HDInsight\20140724\_HDInsight\_AzureStorage\_MostafaE](file:///\\dsdb\sqlskills\HDInsight\20140724_HDInsight_AzureStorage_MostafaE)

1. How to access your Azure storage account:

There are multiple tools (pick one you like)-

<http://www.cloudberrylab.com/free-microsoft-azure-explorer.aspx>

<http://azurestorageexplorer.codeplex.com/>

# **Get started with HDInsight cluster – provision a cluster (1 Day)**

1. Follow the article to create an HDInsight cluster

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-provision-linux-clusters/>

1. Get familiar with the portal (<https://portal.azure.com>)
2. Get familiar with the different options to create a cluster and different cluster types
3. Create at least one Linux Hadoop cluster (for Linux clusters, use a password for SSH for simplicity, instead of a key) and one windows Hadoop cluster – see the next section to learn how you can access a cluster.
4. If time permits, you can create other cluster types such HBase, storm and Spark/Linux clusters.
5. To reduce billing, if you don’t plan to actively use a cluster, delete a cluster. Also, create a cluster with 1 or 2 worker nodes.

Other relevant documentation:

You can pre-provision the azure SQL database and Azure storage accounts to be used in the HDInsight cluster.

Create a SQL database in minutes using the Azure portal.

<https://azure.microsoft.com/en-us/documentation/articles/sql-database-get-started/>

# **How to Access and manage your HDInsight cluster (4 hours)**

1. Review the article to learn how you can use SSH to access a Linux cluster:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-linux-use-ssh-windows/>

Then download Mobaxterm from here <http://mobaxterm.mobatek.net/> - this is much more user friendly than plain putty.

1. Review various management options on the cluster:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-administer-use-portal-linux/#pauseshut-down-clusters>

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-linux-information/>

1. Review the Qucikstart guide to learn how to navigate and perform certain actions on your Linux cluster

<https://microsoft.sharepoint.com/teams/bidpwiki/Documents/HDInsightQuickStart-rev2a.pdf>

1. On Windows clusters created, use Remote desktop to connect to HDInsight cluster and access various resources on the cluster.
2. Review the article below to learn about SSH tunneling on Linux HDInsight cluster:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-linux-ambari-ssh-tunnel/>

1. Learn how to manage HDInsight Linux cluster using Ambari UI:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-manage-ambari/>

1. Learn how you can customize your cluster using script action:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-customize-cluster-linux/>

# **How to run various workloads or jobs on HDInsight cluster (1 Day)**

1. Review and follow the article to learn how to submit Hive jobs:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/>

Try the methods such as Hive CLI, Ambari Hive view, Visual Studio and PowerShell. To install PowerShell and CLI, you can review the articles below:

<https://azure.microsoft.com/en-us/documentation/articles/xplat-cli-install/>

<https://azure.microsoft.com/en-us/documentation/articles/powershell-install-configure/>

1. Review and follow the article to learn how to run MapReduce jobs:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-run-samples-linux/>

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-mapreduce/>

1. Review the article and learn how to interact and run HBase workloads

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hbase-tutorial-get-started-linux/>

1. Review the article and learn how to interact and run Storm workloads

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-storm-tutorial-get-started-linux/>

# **HDInsight Spark cluster (2 Days)**

1. Complete the training ‘HDInsight Spark 101’ from the wiki <https://microsoft.sharepoint.com/teams/bidpwiki/Pages1/Spark.aspx>
2. Complete the training ‘HDInsight Spark 201’ from the wiki <https://microsoft.sharepoint.com/teams/bidpwiki/Pages1/Spark.aspx>

Overview: Apache Spark on HDInsight Linux

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-overview/>

Spark Streaming

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-eventhub-streaming/>

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-jupyter-spark-sql/>

# **HDInsight HBase (Optional)**

Get started using Apache HBase

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hbase-tutorial-get-started-linux/>

Introduction to hbase Schema Design

<http://0b4af6cdc2f0c5998459-c0245c5c937c5dedcca3f1764ecc9b2f.r43.cf2.rackcdn.com/9353-login1210_khurana.pdf>

[Apache HBase reference guide](http://hbase.apache.org/book.html#quickstart).

<http://hbase.apache.org/book.html#quickstart>

# **How to Provision a cluster via other methods (4 hours)**

1. Review the article to get familiar on how to create a cluster via PowerShell:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-create-linux-clusters-azure-powershell/>

## Sample powershell script:

Add-AzureRmAccount

$MyClusterName = "AzimHdi32MetastorePs";

$MyClusterLocation = "South Central US";

$NumClusterNodes = 2;

$MyClusterVersion = "3.2";

$MyHDInsightUserName = ""

$MyHDInsightPwd = ""

$MySqlAzureUserName = ""

$MySqlAzurePwd = ""

$MySqlAzureServerName = "azimsqlsvrsouth.database.windows.net"

$MySqlAzureDbName = "AzimSqlDbSouth2"

$MyDefaultContainerName = "azimhdi32metastoreps"

$clusterResourceGroupName = "azimhdirg"

# Use the correct Azure Subscription!

$subid = ""

Select-AzureRmSubscription -SubscriptionId $subid

# Storage key

$primaryStorageAcctName = "azimstoragesouth"

$primaryStorageResourceGroupName = "azimstoragerg"

$storageAccountKey = Get-AzureRmStorageAccountKey -ResourceGroupName $primaryStorageResourceGroupName -Name $primaryStorageAcctName | %{ $\_.Key1 }

# credentials

$HdInsightPwd = ConvertTo-SecureString $MyHDInsightPwd -AsPlainText -Force

$HdInsightCreds = New-Object System.Management.Automation.PSCredential ($MyHDInsightUserName, $HdInsightPwd)

$SqlAzurePwd = ConvertTo-SecureString $MySqlAzurePwd -AsPlainText -Force

$SqlAzureCreds = New-Object System.Management.Automation.PSCredential ($MySqlAzureUserName, $SqlAzurePwd)

$config = New-AzureRmHDInsightClusterConfig -ClusterType Hadoop |

Add-AzureRmHDInsightMetastore -SqlAzureServerName $MySqlAzureServerName -DatabaseName $MySqlAzureDbName -Credential $SqlAzureCreds -MetastoreType HiveMetastore |

Add-AzureRmHDInsightMetastore -SqlAzureServerName $MySqlAzureServerName -DatabaseName $MySqlAzureDbName -Credential $SqlAzureCreds -MetastoreType OozieMetastore

$config.DefaultStorageAccountName="$primaryStorageAcctName.blob.core.windows.net"

$config.DefaultStorageAccountKey=$storageAccountKey

#create cluster

New-AzureRmHDInsightCluster -config $config -OSType Windows -clustername $MyClusterName -HttpCredential $HdInsightCreds -DefaultStorageContainer $MyDefaultContainerName -Location $MyClusterLocation -ResourceGroupName $clusterResourceGroupName -ClusterSizeInNodes $NumClusterNodes -Version $MyClusterVersion

1. Review the article to get familiar on how to create a cluster via ARM template:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-create-linux-clusters-arm-templates/>

1. Learn how to provision a cluster in a Virtual Network:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-extend-hadoop-virtual-network/>

1. Learn how you can customize a cluster via Script action:

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-customize-cluster-linux/>

# **Troubleshooting:**

## **Troubleshoot Provisioning Issue (Create, Delete, scale up/down):**

### Understand how HDinsight provisioning works:

1. Client tool like Azure portal or PowerShell collects the data from user and send a REST request to Azure Management Servers
2. Azure management servers perform initial validation and forwards the request to HDInsight provisioning servers.

### Provisioning issues from Portal or Azure Management Server or PowerShell:

1. Use IE or Chrome F12 developer tools -> Network trace
2. HDInsight Kusto or Jarvis may not have any logs for such issues.
3. From PowerShell, you can use the –debug switch to collect debug logs, for example:

New-AzureRmHDInsightClusterConfig |

Add-AzureRmHDInsightMetastore -SqlAzureServerName $MySqlAzureServerName -DatabaseName $MySqlAzureDbName -Credential $SqlAzureCreds -MetastoreType OozieMetastore |

Add-AzureRmHDInsightMetastore -SqlAzureServerName $MySqlAzureServerName -DatabaseName $MySqlAzureDbName -Credential $SqlAzureCreds -MetastoreType HiveMetastore |

New-AzureRmHDInsightCluster `

-ClusterType Hadoop `

-OSType Windows `

-ClusterSizeInNodes $NumClusterNodes `

-ResourceGroupName $clusterResourceGroupName `

-ClusterName $MyClusterName `

-HttpCredential $HdInsightCreds `

-Location $MyClusterLocation `

-DefaultStorageAccountName "$primaryStorageAcctName.blob.core.windows.net" `

-DefaultStorageAccountKey $key1 `

-DefaultStorageContainer $MyDefaultContainerName `

-Debug

### Provisioning issues from HDInsight Provisioning servers:

These are the most common provisioning errors, such as internal error or AzureResourceCreationFailedErrorCode. For these CRUD failures, we have a few options to review backend provisioning logs-

1. Review Watson logs:

For every failed cluster, there are Watson logs generated and are saved under that specific date, such as –

[\\sqlcl\team\HDInsight\Watson\mdslogsIaas\2016-1-22](file:///\\sqlcl\team\HDInsight\Watson\mdslogsIaas\2016-1-22)

The above folder will contain all failed cluster logs on January 22 – locate your cluster by ClusterDnsName and get the logs for the cluster.

1. Use a tool like Jarvis or Kusto to review backend provisioning logs

### Using Kusto for Troubleshooting HDInsight issues:

Review the HDInsight Kusto Wiki page:

<https://microsoft.sharepoint.com/teams/bidpwiki/Pages1/HDInsight%20-%20Querying%20Kusto.aspx>

Some simple Kusto queries for provisioning:

General:

IaasClusterCRUDEvent | where UserSubscriptionId == "301f3f35-7867-4afe-a7b7-4fbe407b8193"

| order by PreciseTimeStamp asc

| where PreciseTimeStamp >= datetime("2016-05-18")

| project PreciseTimeStamp, Role, EventName, ClusterDnsName, Location, State, ClusterNodesAsJson

Error only:

IaasClusterCRUDEvent | where UserSubscriptionId == "0bb13fe6-5eba-476c-ab22-1524c55dc69e"

| order by PreciseTimeStamp desc

| where State == "Error"

| where PreciseTimeStamp >= datetime("2016-05-05")

IaasClusterCRUDEvent | where UserSubscriptionId == "059ed1ab-6824-4344-9a65-a0504248340f"

| order by PreciseTimeStamp asc

| where PreciseTimeStamp >= datetime("2016-05-09")

| where ClusterDnsName =="scdwetlcmg"

| where State == "Error"

Windows cluster:

ClusterCRUDEvent | where UserSubscriptionId == "206878b8-5f28-476e-acc8-1cff1b03efe3"

| order by PreciseTimeStamp asc

| where PreciseTimeStamp >= datetime("2016-05-24")

| where DataCenterLocation == "Southeast Asia"

| project PreciseTimeStamp, Role, EventName, ClusterDnsName, DataCenterLocation, ContainerState, DeploymentState

Verbose logs (both Linux and Windows):

LogEntry| where ClusterDnsName  == "brysmihb"

| order by PreciseTimeStamp asc

| where PreciseTimeStamp >= datetime("2016-06-16 14:15:00")

| where PreciseTimeStamp <=  datetime("2016-06-16 15:30:00")

| where TraceLevel == "Error"

Scale down Linux:

IaasClusterScaleEvent

| where UserSubscriptionId == "c5105a87-bb09-4f6c-9481-d7041638077a"

| where ClusterDnsName == "dc305cl01"

| order by PreciseTimeStamp asc

| where PreciseTimeStamp >= datetime("2016-06-13")

| project PreciseTimeStamp, ClusterDnsName, Location, ScaleOperation, ScaleFrom, RequestedScaleTo , ScaledTo, State, ErrorInfoAsJson

### Using Jarvis for Troubleshooting HDInsight issues:

Please review the wiki page <https://microsoft.sharepoint.com/teams/bidpwiki/Pages1/HDInsight%20Troubleshooting%20Dashboards.aspx>

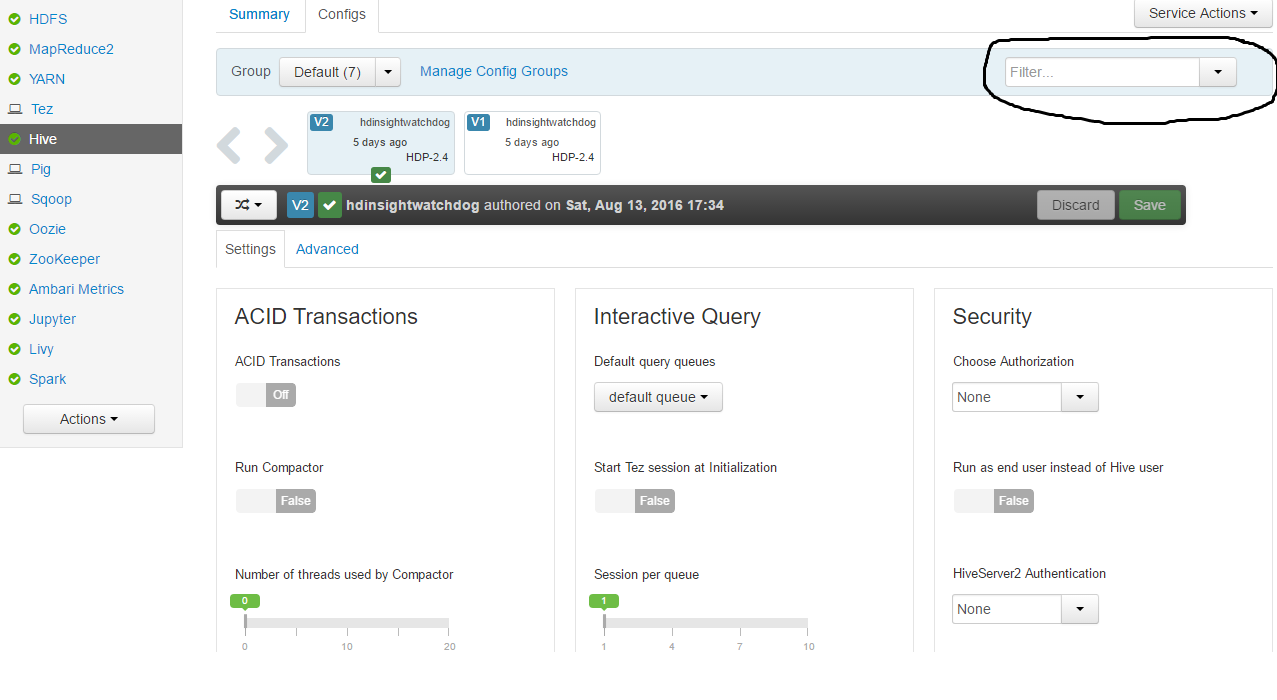
### Using MDS (deprecated) for troubleshooting HDInsight issues:

<https://microsoft.sharepoint.com/teams/bidpwiki/Pages1/HDInsight%20Troubleshooting%20Dashboards.aspx>

## **Troubleshoot HDInsight Linux cluster issues**

**Finding component configuration:**

1. You can use Ambari dashboard, and select a specific component on the left, like Hive. Search for the specific configuration in the ‘Filter’ combobbox, like below –



1. The component specific configuration files (such as hive-site.xml, oozie-site.xml etc) are located under the /etc/<componentName>/conf file location, for example -

/etc/oozie/conf

/etc/hive/conf

**Finding Component log:**

1. Server side component logs are located under /var/log, such as

/var/log/oozie,

/var/log/hive etc

1. User log: /tmp/<sshUserName> <---- hive.log will be found here
2. **TIP:** When looking for a log, it’s important to understand if the log you are looking for will be located on the headnode or one of the worker nodes where the specific application like Hive CLI was launched – it varies depending on the way the job was submitted.

**Troubleshooting docs in the wiki:**

Review the wiki page under the section ‘Troubleshooting’ on the wiki <https://microsoft.sharepoint.com/teams/bidpwiki/Pages1/Humboldt.aspx>

## **Troubleshoot Job submission (errors or performance) Issues**

1. How to collect Yarn application logs on windows cluster <https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-access-yarn-app-logs/>
2. How to collect Yarn application logs on Linux cluster <https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-access-yarn-app-logs-linux/>
3. Review the wiki to understand what other logs are available at various locations – may be a little outdated, but still a good reference

<http://hadoopwiki/w/Cluster_Troubleshooting_Guide/Supportability_Data>