

Create Apache Hadoop cluster in Azure HDInsight using Azure portal

In this document:

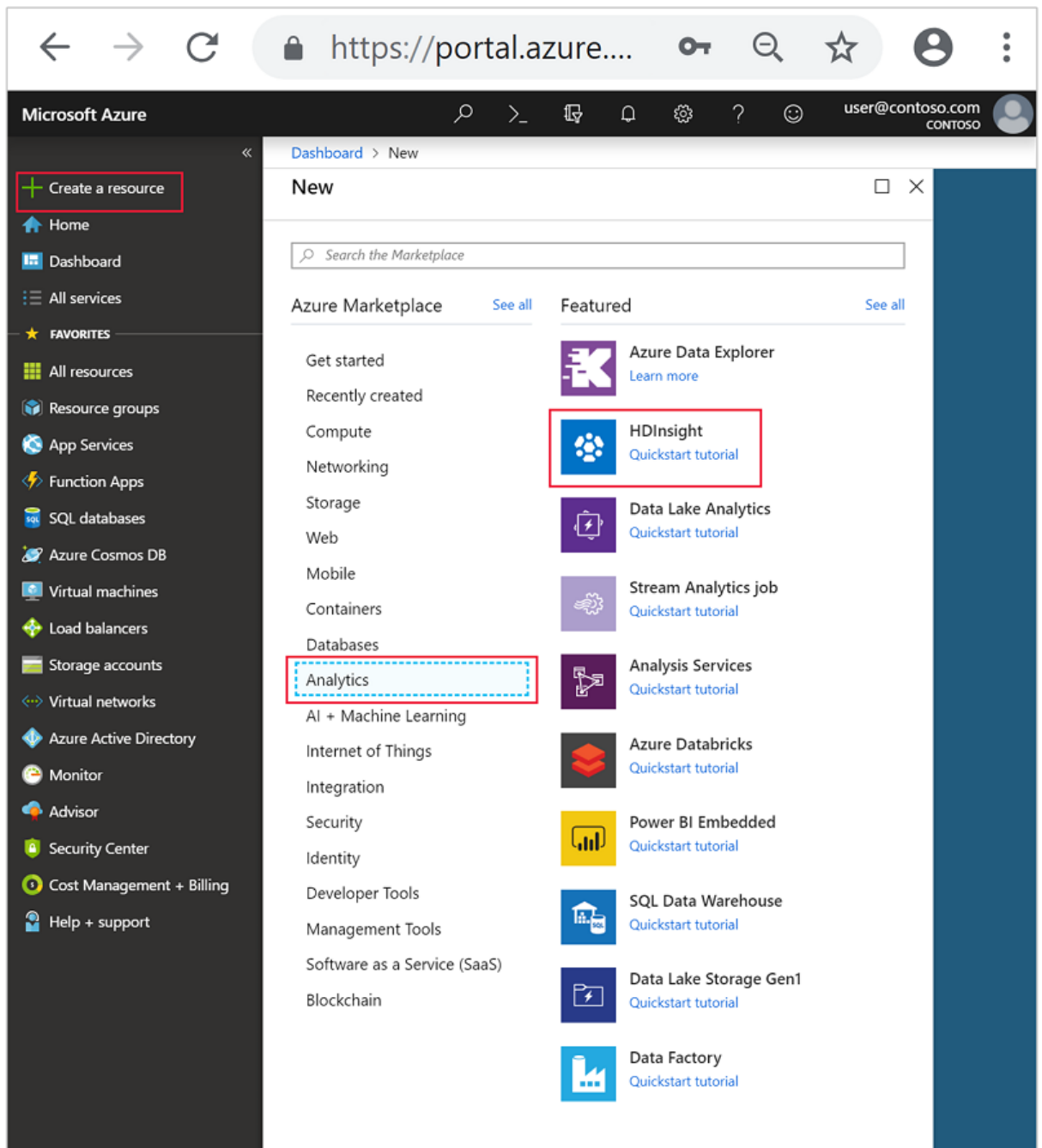
1. [Create an Apache Hadoop cluster](#)
2. [Run Apache Hive queries](#)
3. [Clean up resources](#)

In this document, you learn how to create [Apache Hadoop](#) clusters in HDInsight using Azure portal, and then run Apache Hive jobs in HDInsight. Most of Hadoop jobs are batch jobs. You create a cluster, run some jobs, and then delete the cluster. In this document, you perform all the three tasks.

Create an Apache Hadoop cluster

In this section, you create a Hadoop cluster in HDInsight using the Azure portal.

1. Sign in to the [Azure portal](#).
2. From the Azure portal, go to **Create a resource > Analytics > HDInsight**.



3. Under **Basics**, enter or select the following values:

Property	Description
Subscription	Select your Azure subscription.
Resource group	Create a resource group or select an existing resource group. A resource group is a container of Azure components. In this case,

Property	Description
	the resource group contains the HDInsight cluster and the dependent Azure Storage account.
Cluster name	Enter a name for the Hadoop cluster. Because all clusters in HDInsight share the same DNS namespace this name needs to be unique. The name can consist of up to 59 characters including letters, numbers, and hyphens. The first and last characters of the name cannot be hyphens.
Location	Select an Azure location where you want to create your cluster. Choose a location closer to you for better performance.
Cluster type	Select Select cluster type . Then select Hadoop as the cluster type.
Version	The default version for the cluster type will be specified. Select from the drop-down list if you wish to specify a different version.
Cluster login username and password	The default login name is admin . The password must be at least 10 characters in length and must contain at least one digit, one uppercase, and one lower case letter, one non-alphanumeric character (except characters ' " `). Make sure you do not provide common passwords such as "Pass@word1".
Secure Shell (SSH) username	The default username is sshuser . You can provide another name for the SSH username.
Use cluster login password for SSH	Select this check box to use the same password for SSH user as the one you provided for the cluster login user.

Create HDInsight cluster

[Go to classic create experience](#)

[Basics](#)
[Storage](#)
[Security + networking](#)
[Configuration + pricing](#)
[Review + create](#)

Create a managed HDInsight cluster. Select from Spark, Kafka, Hadoop, Storm, and more. [Learn more](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

* Subscription

* Resource group [Create new](#)

Cluster details

Name your cluster, pick a location, and choose a cluster type and version. [Learn more](#)

* Cluster name

* Location

* Cluster type [Select cluster type](#)

* Version

Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

* Cluster login username

* Cluster login password

* Confirm cluster login password

* Secure Shell (SSH) username

Use cluster login password for SSH ☒

[Review + create](#)
[« Previous](#)
[Next: Storage »](#)

Select the **Next: Storage >>** to advance to the storage settings.

4. From the **Storage** tab, provide the following values:

Property	Description
Primary storage type	Use the default value Azure Storage .
Selection method	Use the default value Select from list .
Primary storage account	Use the drop-down list to select an existing storage account, or select Create new . If you create a new account, the name must be between 3 and 24 characters in length, and can include numbers and lowercase letters only
Container	Use the autopopulated value.

Create HDInsight cluster

[Go to classic create experience](#)

[Basics](#)
[Storage](#)
[Security + networking](#)
[Configuration + pricing](#)
[Review + create](#)

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

* Primary storage type

* Selection method ☒ Select from list ☐ Use access key

* Primary storage account
[Create new](#)

* Container

Data Lake Storage Gen1

Provide details for the cluster to access Data Lake Storage Gen1. The cluster will be able to access any Data Lake Storage Gen1 accounts that the chosen service principal has access to.

Data Lake Storage Gen1 access [Configure access settings](#)

Additional Azure storage

Link additional Azure storage accounts to the cluster.

[Add Azure storage](#)

Metastore settings

To preserve your Hive and/or Oozie metadata outside of this cluster, select a SQL database for this cluster.

SQL database for Hive

SQL database for Oozie

[Review + create](#)
[« Previous](#)
[Next: Security + networking »](#)

Select the **Review + create** tab.

- From the **Review + create** tab, verify the values you selected in the earlier steps.

Create HDInsight cluster

[Go to classic create experience](#)

✓ Validation succeeded.

[Basics](#) [Storage](#) [Security + networking](#) [Configuration + pricing](#) [Review + create](#)

Hadoop 2.7.3 (HDI 3.6) **<price> USD Total estimated cost/hour**
This estimate does not include subscription discounts or costs related to storage, networking, or data transfer.

Basics

Subscription	Azure
Resource group	(new) myResourceGroup
Location	East US
Cluster name	(new) myHadoop
Cluster type	Hadoop 2.7.3 (HDI 3.6)
Cluster login username	admin
Secure Shell (SSH) username	sshuser
Use cluster login password for SSH	Enabled

Storage

Primary storage type	Azure Storage
Primary storage account	(new) myhadoopdistorage
Container	myhadoop-2019-09-27t22-00-07-427z
Additional Azure storage	None
Data Lake Storage Gen1 access	Disabled

Cluster configuration

Head	2 nodes, D12 v2 (4 Cores, 28 GB RAM)
Worker	4 nodes, D4 v2 (8 Cores, 28 GB RAM)

[Create](#) [« Previous](#) [Next »](#) [Download a template for automation](#)

6. Select **Create**. It takes about 20 minutes to create a cluster.

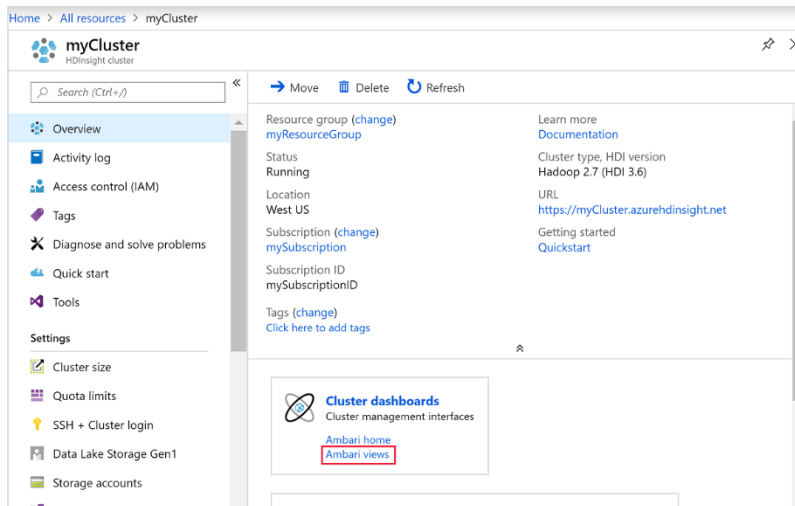
Once the cluster is created, you see the cluster overview page in the Azure portal.

Each cluster has an [Azure Storage account](#) or an [Azure Data Lake account](#) dependency. It is referred as the default storage account. HDInsight cluster and its default storage account must be colocated in the same Azure region. Deleting clusters does not delete the storage account.

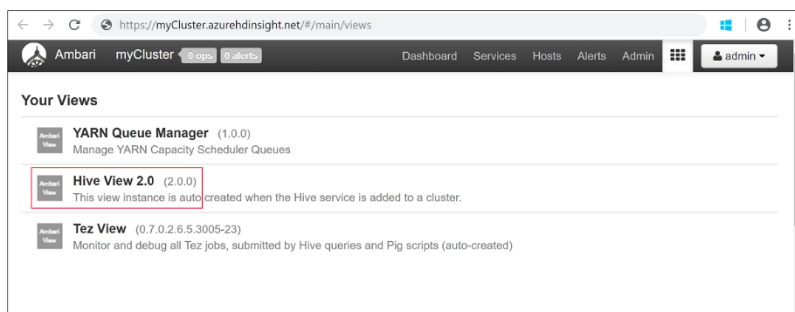
Run Apache Hive queries

[Apache Hive](#) is the most popular component used in HDInsight. There are many ways to run Hive jobs in HDInsight. In this guide, you use the Ambari Hive view from the portal. For other methods for submitting Hive jobs, see [Use Hive in HDInsight](#).

1. To open Ambari, from the previous screenshot, select **Cluster Dashboard**. You can also browse to <https://ClusterName.azurehdinsight.net>, where ClusterName is the cluster you created in the previous section.



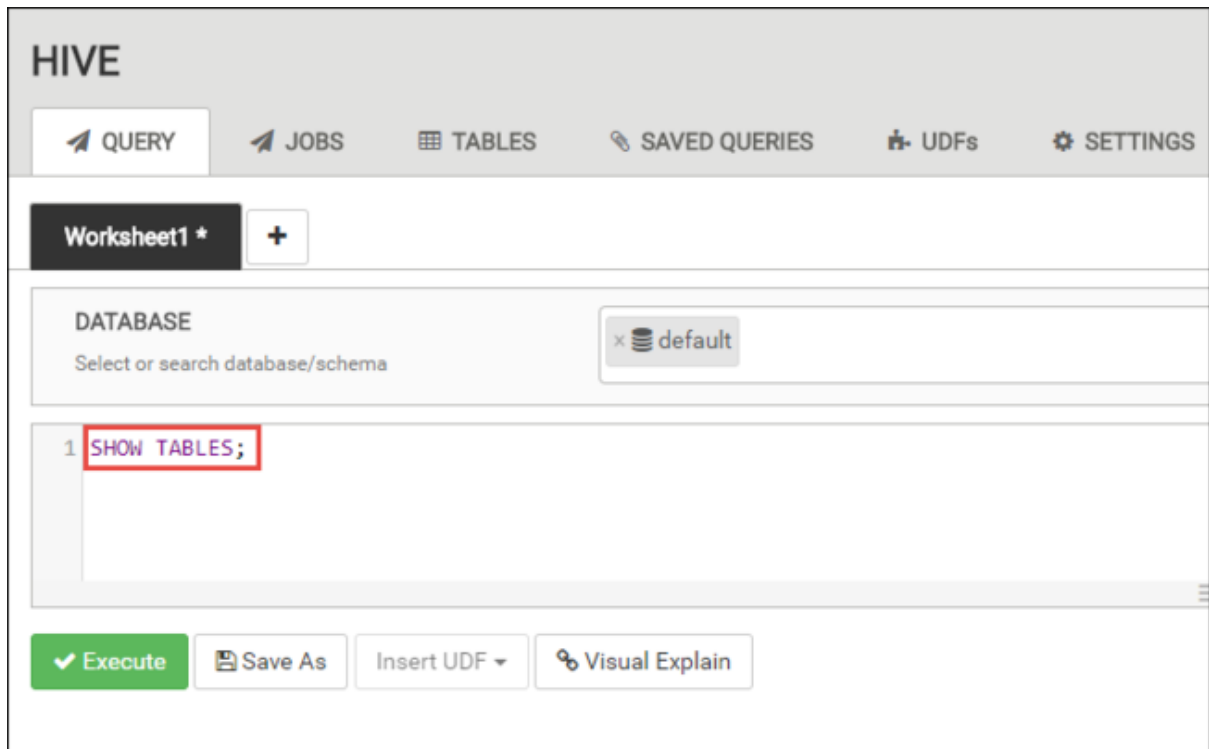
2. Enter the Hadoop username and password that you specified while creating the cluster. The default username is **admin**.
3. Open **Hive View** as shown in the following screenshot:



4. In the **QUERY** tab, paste the following HiveQL statements into the worksheet:

SQL Copy

```
SHOW TABLES;
```



5. Select **Execute**. A **RESULTS** tab appears beneath the **QUERY** tab and displays information about the job.

Once the query has finished, the **QUERY** tab displays the results of the operation. You shall see one table called **hivesampletable**. This sample Hive table comes with all the HDInsight clusters.

HIVE

QUERY JOBS TABLES SAVED QUERIES UDFs SETTINGS

Worksheet1 * +

DATABASE
Select or search database/schema default

1 SHOW TABLES;

Execute Save As Insert UDF Visual Explain

RESULTS LOG VISUAL EXPLAIN TEZ UI

Filter columns x

tab_name
hivesampletale

- Repeat step 4 and step 5 to run the following query:

SQL Copy

```
SELECT * FROM hivesampletale;
```

- You can also save the results of the query. Select the menu button on the right, and specify whether you want to download the results as a CSV file or store it to the storage account associated with the cluster.

RESULTS LOG VISUAL EXPLAIN TEZ UI

Filter columns x

hivesampletale.clientid	hivesampletale.querytime	hivesampletale.market	hivesampletale.de	state	hivesampletal
8	18:54:20	en-US	SCH-i500	California	United States
23	19:19:44	en-US	Incredible	Pennsylvania	United States
23	19:19:46	en-US	Incredible	Pennsylvania	United States
23	19:19:47	en-US	Incredible	Pennsylvania	United States
28	01:37:50	en-US	Droid X	Colorado	United States

SAVE TO HDFS
DOWNLOAD AS CSV

After you have completed a Hive job, you can [export the results to Azure SQL database or SQL Server database](#), you can also [visualize the results using Excel](#).

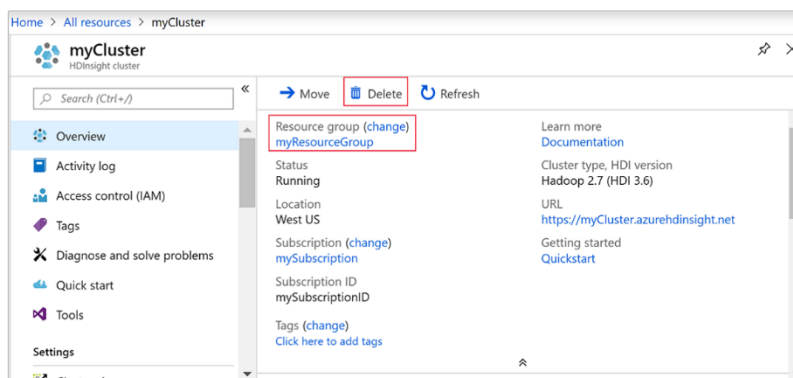
Clean up resources

After you complete the guide, you may want to delete the cluster. With HDInsight, your data is stored in Azure Storage, so you can safely delete a cluster when it is not in use. You are also charged for an HDInsight cluster, even when it is not in use. Since the charges for the cluster are many times more than the charges for storage, it makes economic sense to delete clusters when they are not in use.

Note

To delete the cluster and/or the default storage account

1. Go back to the browser tab where you have the Azure portal. You shall be on the cluster overview page. If you only want to delete the cluster but retain the default storage account, select **Delete**.



2. If you want to delete the cluster as well as the default storage account, select the resource group name (highlighted in the previous screenshot) to open the resource group page.
3. Select **Delete resource group** to delete the resource group, which contains the cluster and the default storage account. Note deleting the resource group deletes the storage account. If you want to keep the storage account, choose to delete the cluster only.