Big Data For Engineers 2023 FS

Running RumbleDB on Azure HDInsight Clusters

We provide instructions for the **ungraded** Azure exploration you have time until the exam (we will try to push for that with Azure now), feedback on the exploration is welcomed but not compulsory. Please use the [feedback box on Moodle](#).

The overall goal of the ****ungraded**** Azure assignment is that you will have enough time to explore and experiment with RumbleDB in clusters with a huge dataset to push the limit, but no strings attached.

We are working with the Azure team to speed up the process for the subscriptions that are still not activated.

Important: Remember to **delete** the cluster once you are done. If you want to stop doing the assignments at any point, delete it and recreate it using the same container name as you used the first time so that the resources are still there.

Please do not hesitate to contact us anytime to clarify details.

Happy rumbling!

Your Big Data For Engineers TA Team

● Enrol in the Azure Lab

You can enrol in the Azure Lab using the following links (accessible from Moodle). And you need to wait until we approve your request. Once your request is approved, you need to accept the subscription provided to you. Note that the new code is 4URTUV.

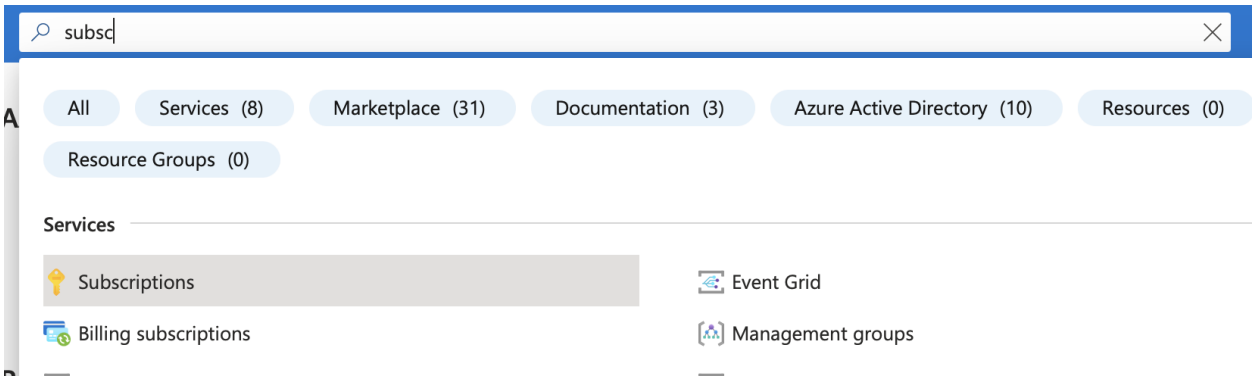
Online class room @ Azure

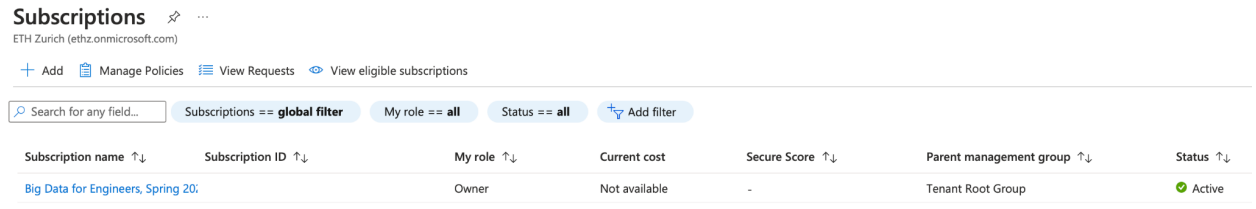
Here is a code 4URTUV that you can use **with your address <login>@ethz.ch (very important) and ETH password** to enroll in the course Lab in the Azure Education Hub under https://aka.ms/JoinEduLab.

If you wish to use an anonymous email address (anonymous to Azure) instead, this is possible too, please send an email to the TA team before informing them of this address so we can grant you access.

● Accept the subscription

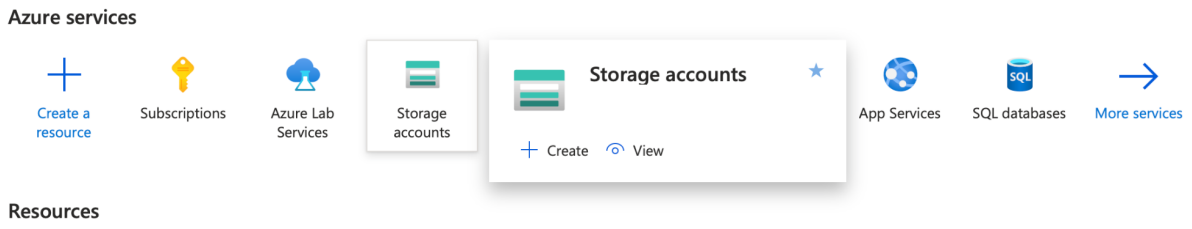
Once you have accepted your subscription, you could login to the Azure portal with your account, search for “subscriptions" in the search tab and click on the option with the key logo. Make sure that you can see the big-data subscription. You will be able to create resources with this subscription account, e.g., VMs and HDInsight clusters.



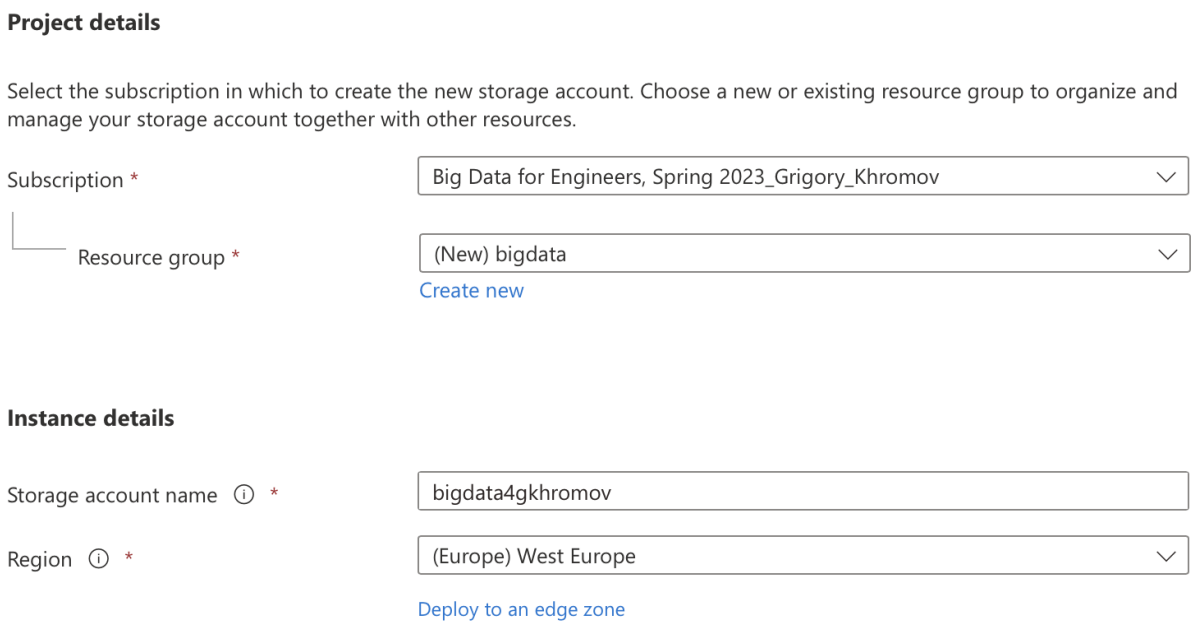


● Create a storage account

You need first to create an Azure storage account to accommodate the dataset that you use for the tasks. To create a storage account, go to the home page and click the icon of “Storage accounts” below.



Please make sure you choose the subscription that matches your Azure lab subscription. Do not forget to create a resource group with any name you like, and also make sure that you choose West Europe for your storage region. For the rest of the configurations, you can leave it to default settings. After that, click “Review” and then “Create”.



● Upload the dataset to the storage account

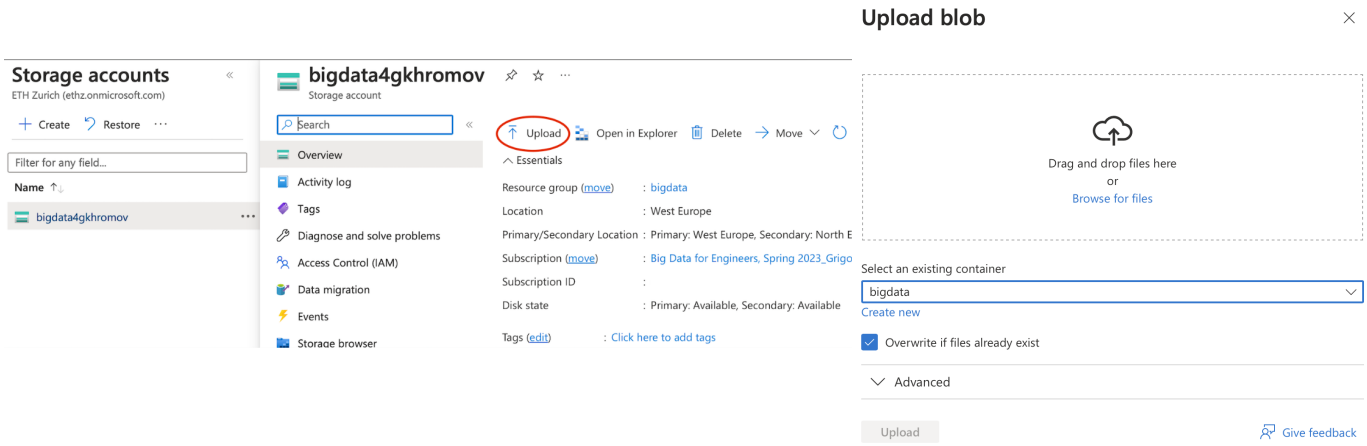
Once you have created your storage account, you could upload the following dataset in your account:

Small dataset: https://www.rumbledb.org/samples/git-archive.json

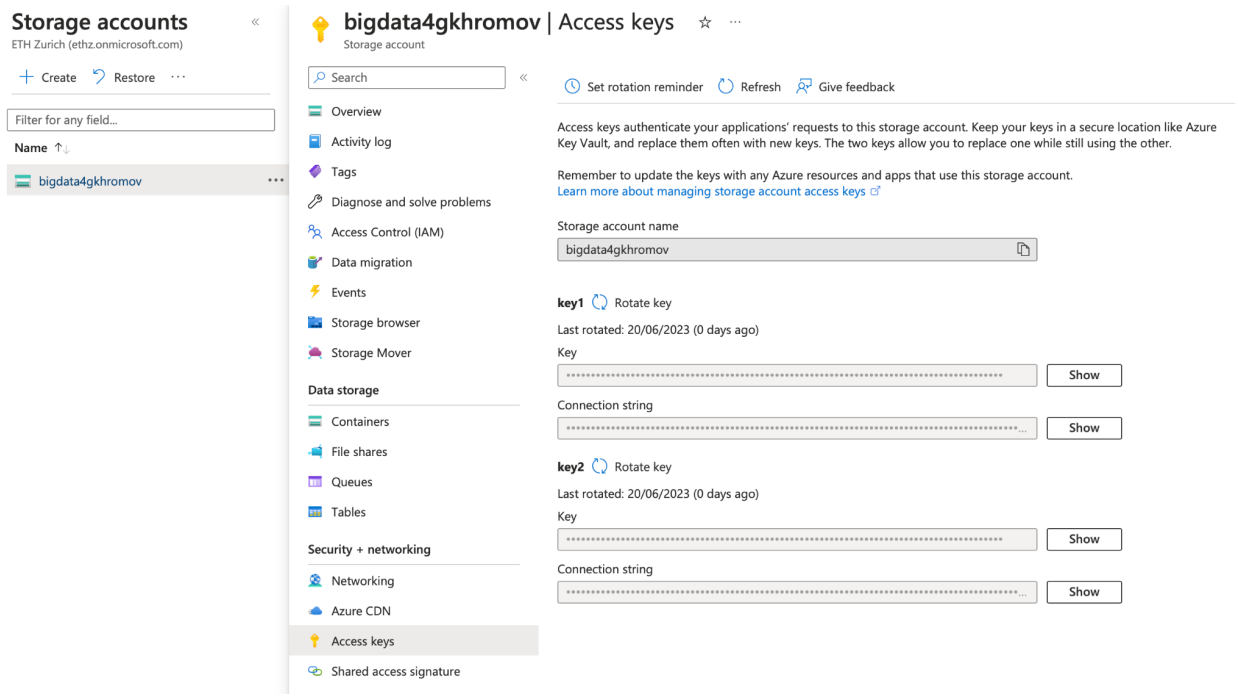
Larger dataset: https://www.rumbledb.org/samples/git-archive-big.json

Huge dataset: https://cloud.inf.ethz.ch/s/Ss5L7ASD2KKdrCx

You could update your datasets to the storage account using the “Upload” button in the storage account:



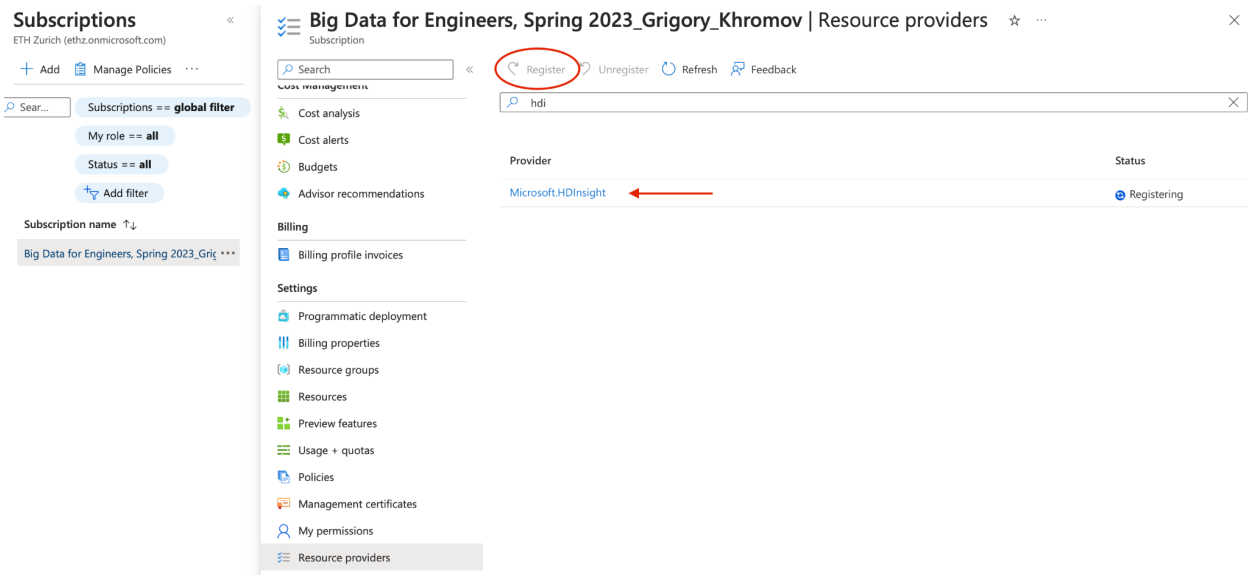
Note it could take a while to upload the huge dataset (or an even bigger dataset you want to test). For large datasets, it is recommended to use a script to do so. The access keys (account name, account key) can be acquired from this page. There might also be ways to extract compressed files on a blob, feel free to explore.



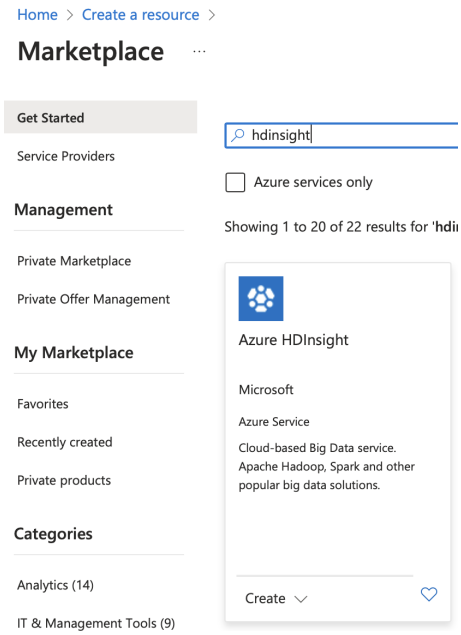
• Create HDInsight cluster

In this assignment, you will use the HDInsight cluster (https://learn.microsoft.com/en-us/azure/hdinsight/hdinsight-overview). To create an HDInsight cluster, you need to do the following steps:

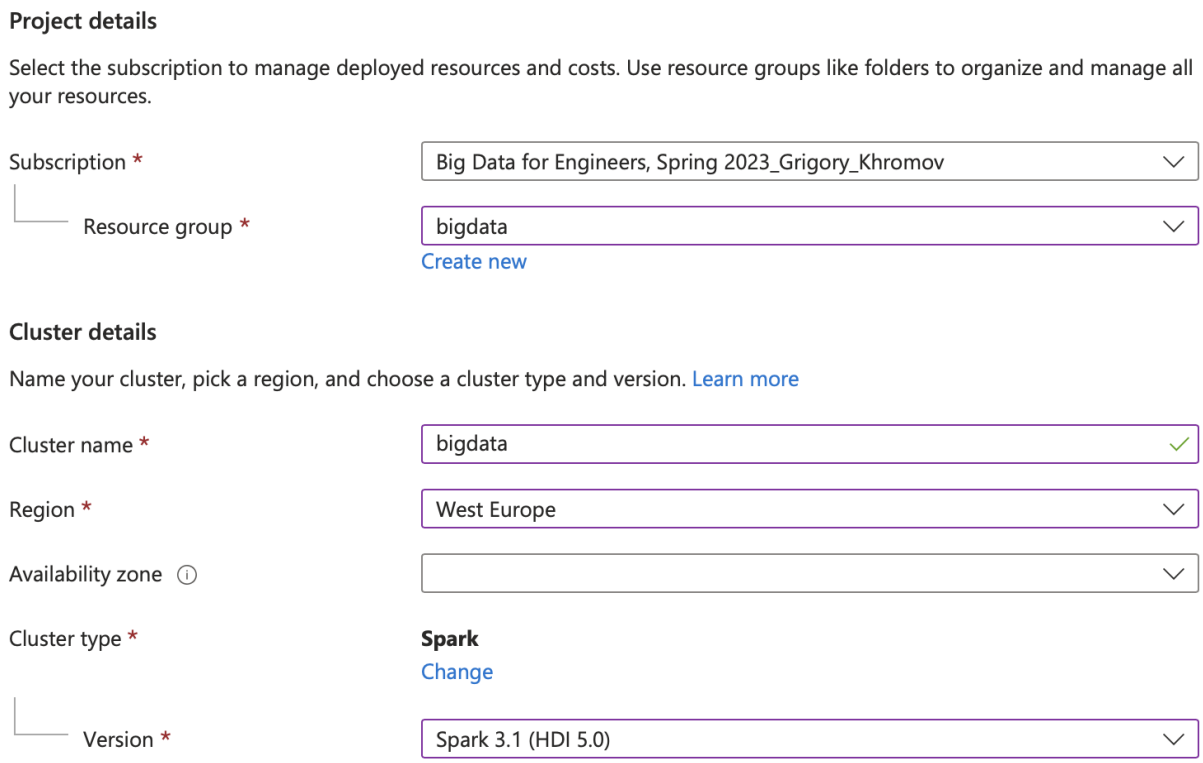
○ The first step is to register the service provider of HDInsight. To do so, you need to first go into “Subscriptions”. You could find the tab “Resource providers” on the left-hand side of the “Settingsa” panel and then search “HDInsight” for registration. After you click on the “Register” tab, it takes only a few seconds to register the resources. Then you should be able to see the status of “Microsoft.HDInsight” change to “Registered”.



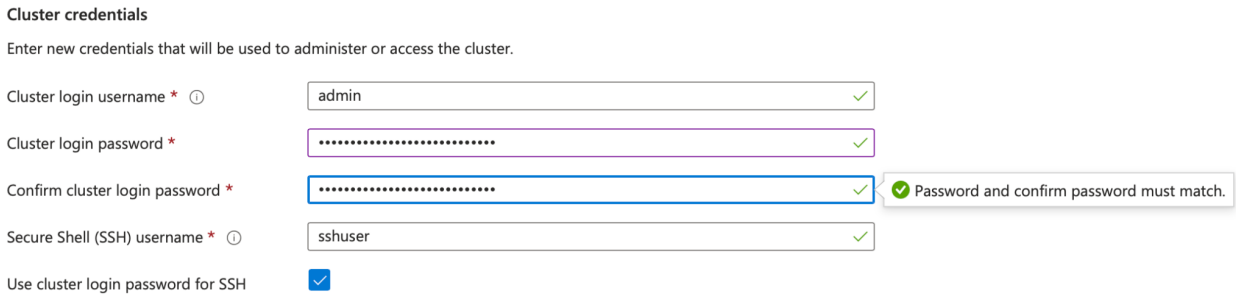
○ Then you could create the resource of HDInsight. You could search for the resources in the “Marketplace”, which you can find by going to the home page and clicking on the “Create a resource” button.



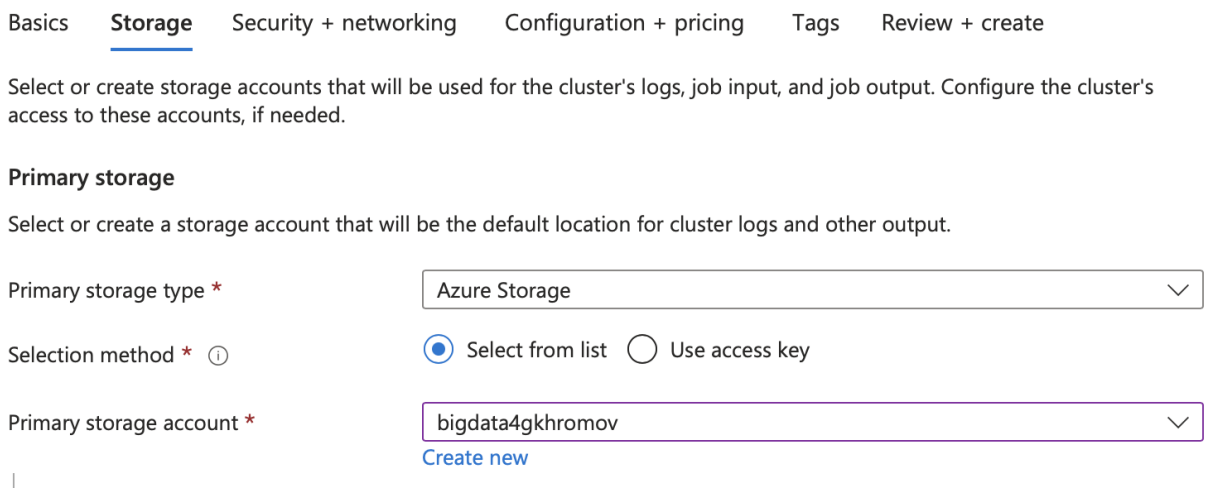
○ During the basic configuration of the HDInsight cluster, please make sure that you are using the proper big data Azure subscription and the region is **West Europe**. And please choose cluster type as **Spark with version 3.1**.



○ Set a password that you will remember (best to write it down on a piece of paper). Leave login and SSH username as is.



○ Please also make sure that when you configure the storage, choose the Azure Storage as your primary storage source and link it to the storage account that you have just created.



○ We recommend you create the cluster with **the following options**. The number of nodes can be chosen by you, as it is within a reasonable amount of cost (say 2-3 USD per hour).

Basics | Storage | Security + networking | **Configuration + pricing** | Tags | Review + create

Configure cluster performance and pricing. Learn More

Node configuration

Configure your cluster's size and performance, and view estimated cost information.

The cost estimate represented in the table does not include subscription discounts or costs related to storage, networking, or data transfer.

This configuration will use 30 of 100 available cores in the West Europe region.
View cores usage
Open an HDInsight quota increase support case

+ Add application

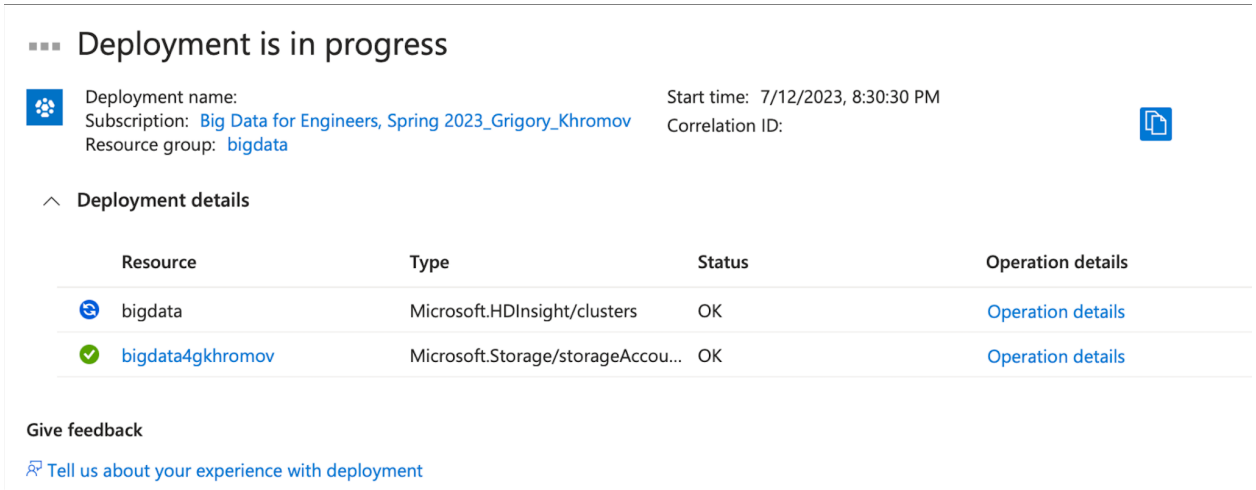
Node type	Node size	Number of ...	Estimated cost/h...
Head node	E4a v4 (4 Cores, 32 GB RAM), 0.38 USD/ho...	2	0.76 USD
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.12 USD/hour	3	0.00 (FREE)
Worker node	E4 V3 (4 Cores, 32 GB RAM), 0.38 USD/hour	4	1.52 USD

☐ Enable managed disk

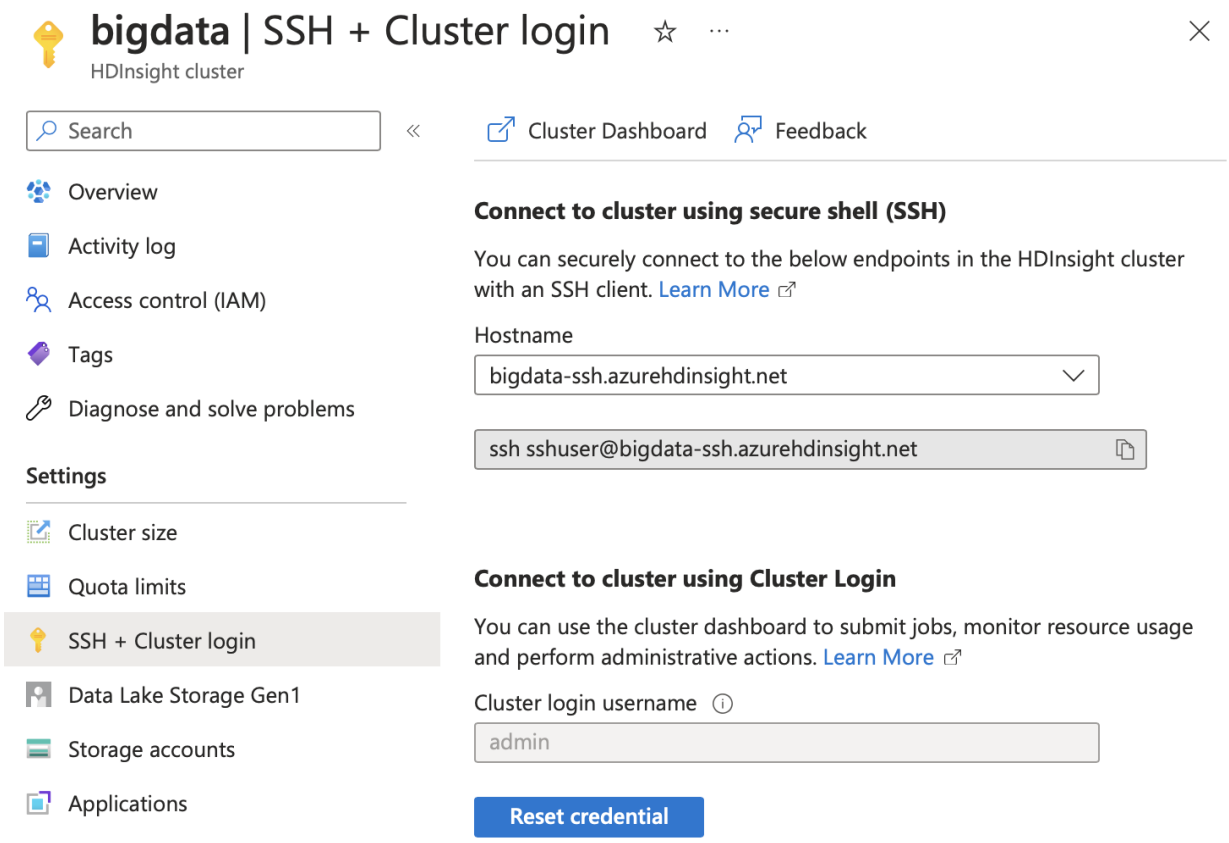
☐ Enable autoscale
Learn More

Total estimated cost/hour 2.28 USD

○ It takes about 10-15 minutes to create the cluster. Once you have created your HDInsight cluster, you can access it using SSH.



○ Access your cluster.
Make sure you can access your cluster (the NameNode) via SSH: *$ ssh <ssh_user_name>@<cluster_name>-ssh.azurehdinsight.net* . You can find this string in “Home” -> name of your HDInsight cluster in the “Resources” section -> SSH + Cluster login left tab. E.g., *ssh sshuser@bigdata-ssh.azurehdinsight.net.* The password is the one you specified when you created the cluster.



● Quick test of the cluster

- On the cluster (via ssh): Download RumbleDB (to the local disk of the remote machine on the head node e.g., sshuser@hn0-bigdat):
  wget https://github.com/RumbleDB/rumble/releases/download/v1.20.0/rumbledb-1.20.0-for-spark-3.1.jar
- Run the shell:
  spark-submit rumbledb-1.20.0-for-spark-3.1.jar repl
- In the shell, you could run the following command to read the json file that is in your storage account.
  This is how you access the azure blob storage (https://learn.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-use-blob-storage):
  wasbs://<containername>@<accountname>.blob.core.windows.net/<file.path>/

  You can find the container name in “Home” -> “Resources” -> name of the “Storage account” -> “Containers” left tab. Account name is identical to the name of the “Storage account”.

  Example code to read the json file (you need to change the container name and account name) :

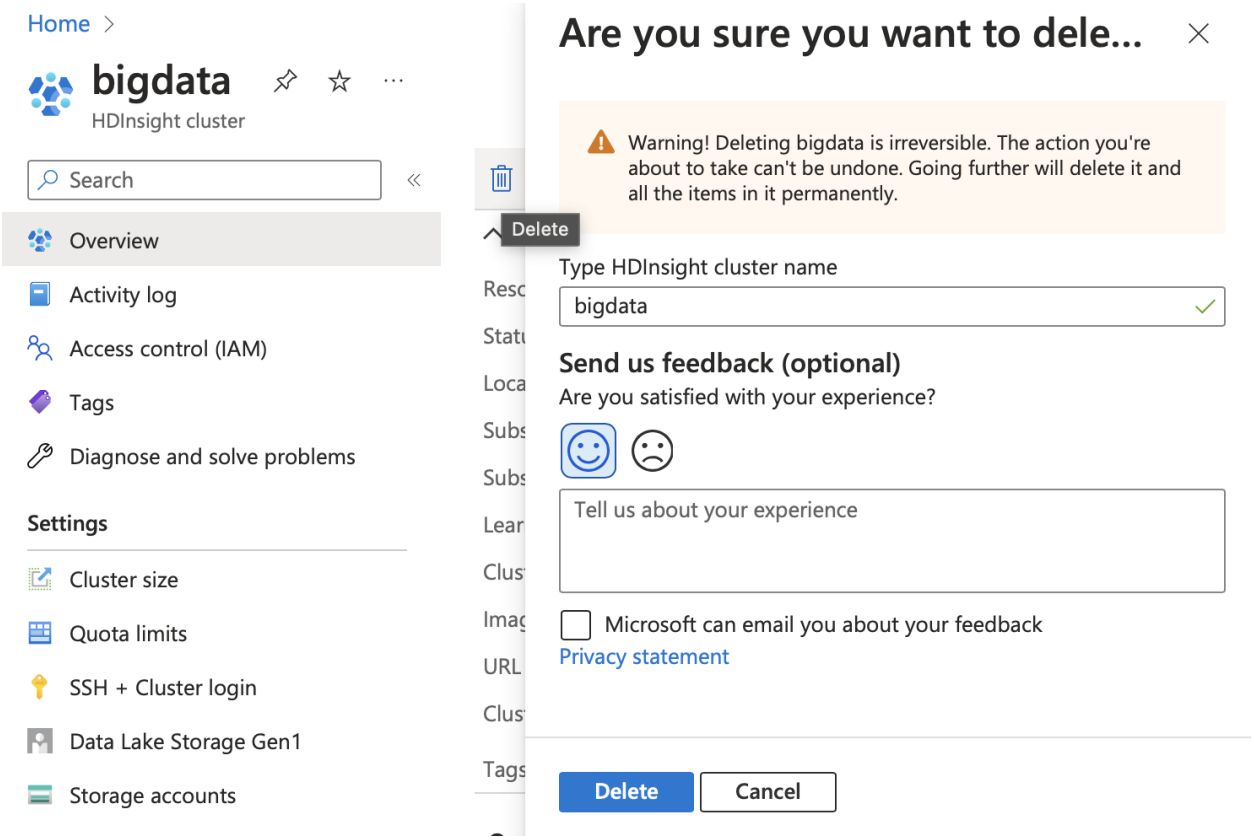
json-file("wasbs://bigdata@bigdata4gkhromov.blob.core.windows.net/.json")*

json-file("wasbs://bigdata@bigdata4gkhromov.blob.core.windows.net/git-archive.json").type=>distinct-values()

- You can also run the RumbleDB as a server:
  spark-submit rumbledb-1.20.0-for-spark-3.1.jar --server yes --port 8002
- SSH forwarding
  After running RumbleDB as a server, we can use a juypter notebook to interact with it. We recommend to use SSH forwarding. For that, make sure you have run: *spark-submit rumbledb-1.20.0-for-spark-3.1.jar --server yes --port 8002*
  and then on your local machine forward 8002 => localhost:8002
  ssh -N -L 8002*:localhost:*8002 *sshuser@[servername]-ssh.azurehdinsight.net E.g.,*
  ssh -N -L 8002*:localhost:*8002 *sshuser@bigdata-ssh.azurehdinsight.net*
  See an example of running a local notebook interacting with RumbleDB hosted on Azure
  You can now try out various queries on different datasets we have on RumbleDB's exercise sheet from (1) the RumbleDB shell on Azure and (2) your notebook on the cluster and evaluate the speed difference.

● Delete / down size the cluster

Important: Remember to **delete** the cluster once you are done. If you want to stop doing the assignments at any point, delete it and recreate it using the same container name as you used the first time, so that the resources are still there. It is very important that you remember to delete the cluster if you don't plan to use it as this is costly and soon you will use up all your credits.



If you don't want to delete your cluster, note that cluster cannot be shut down, but it's possible to scale down the worker nodes to minimize cost when you do not use it (https://learn.microsoft.com/en-us/azure/hdinsight/hdinsight-scaling-best-practices) . The storage is relatively cheap compared to the cluster so you can keep your storage account for a longer period of time. But please do remember to delete your storage once you don't need it anymore.