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: https://moodle-app2.let.ethz.ch/mod/journal/view.php?id=836348 .

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.

For the subscriptions that are still not activated, we are working with the Azure team to speed up the process.

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 TA Team

● Enroll to the Azure Lab

You can enroll 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 ZED9XP.

Online class room @ Azure

You can enroll in our online classroom in the Azure Education Hub under https://aka.ms/JoinEduLab using the code **ZED9XP**. Once enrolled, you can access it under https://aka.ms/startedu.
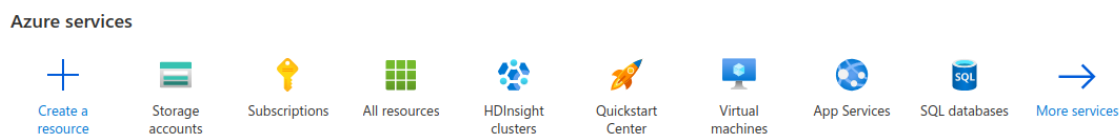
● Accept the subscription

Once you have accepted your subscription, you could login on 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.
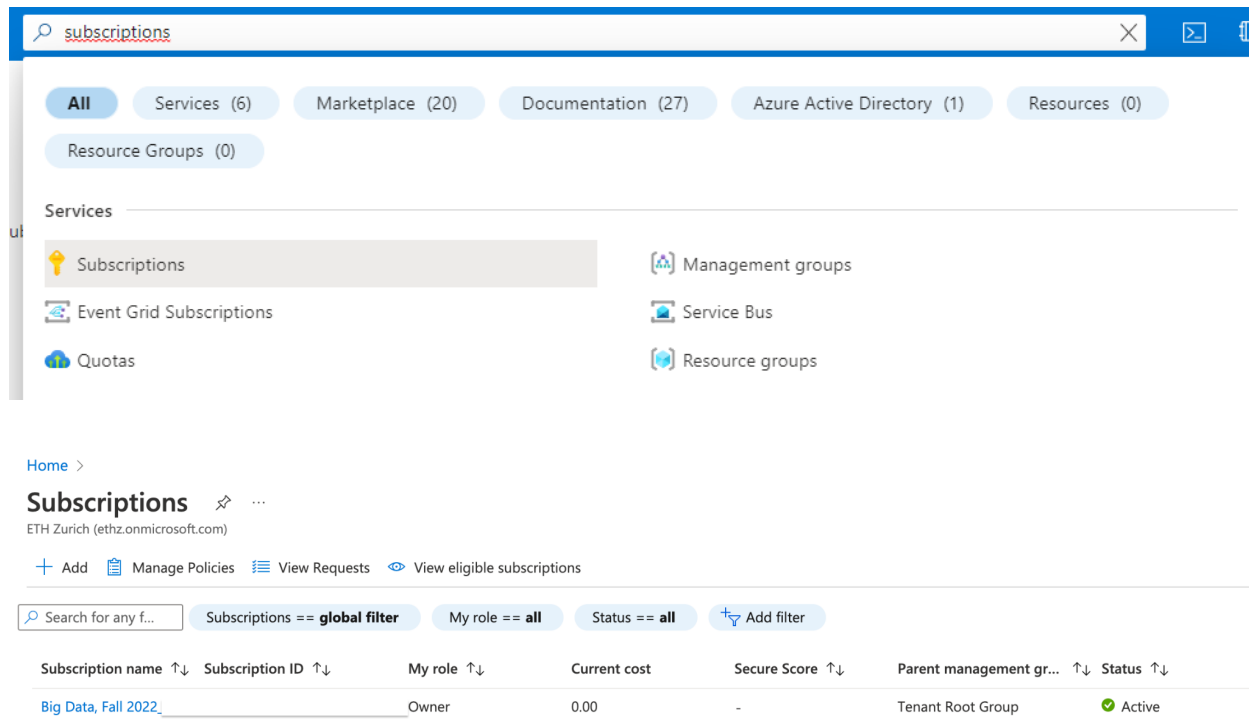
● Create a storage account

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

Please make sure you choose the subscription that matches your Azure lab subscription. And please make sure that you choose West Europe for your storage region. For the

rest of the configurations, you can leave it to default settings.

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *   (Disabled) Big Data, Fall 2022_Zhenhao_He ˅

        Resource group *   big-data ˅

                            Create new

Instance details

If you need to create a legacy storage account type, please click here.

Storage account name ⓘ *

Region ⓘ *   (Europe) West Europe ˅

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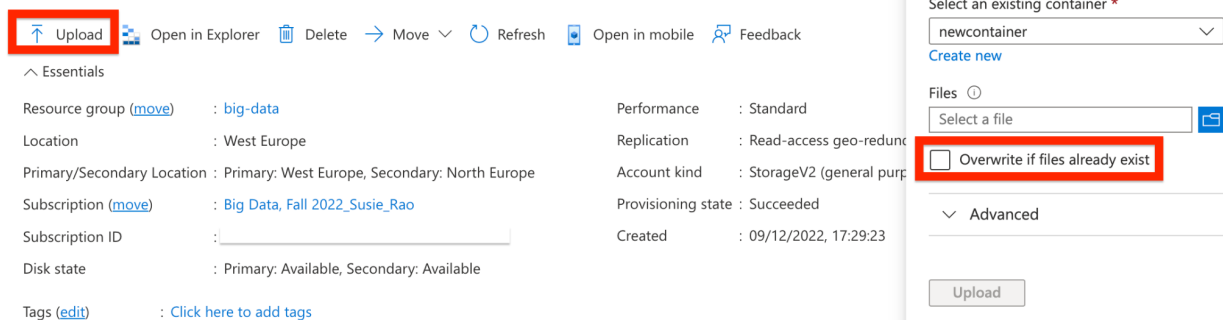
Note it could take a while to upload the huge dataset (or an even bigger dataset you want to test). For large datasets, it is recommended to use a script to do so. The access keys (account name, account key can be acquired from this page). There might also be ways to extract compressed files on blob,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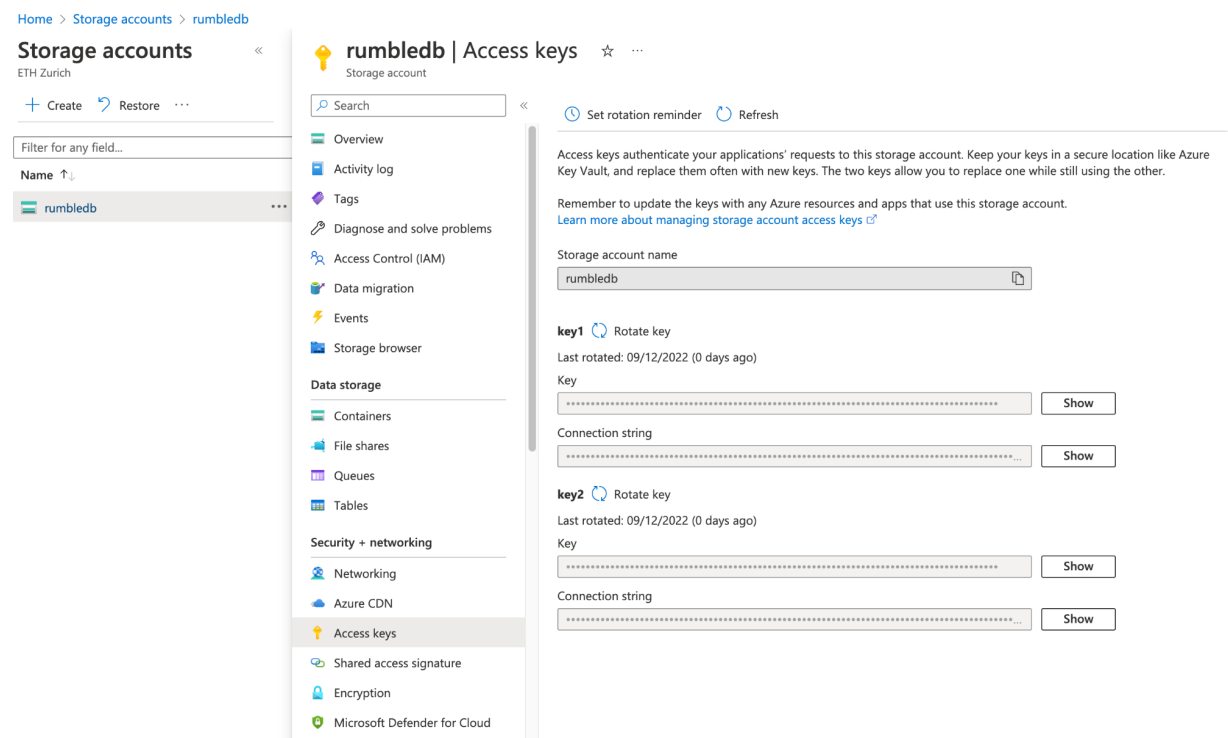
● Create HDInsight cluster

In this assignment, you will use the HDInsight cluster (https://learn.microsoft.com/en-us/azure/hdinsight/hdinsight-overview). To create a 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 of "Resource Provider" on the left hand side of the "Setting" 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" is changed to "Registered".

○ Then you could create the resource of HDInsight. You could search for the resources in the "Marketplace".
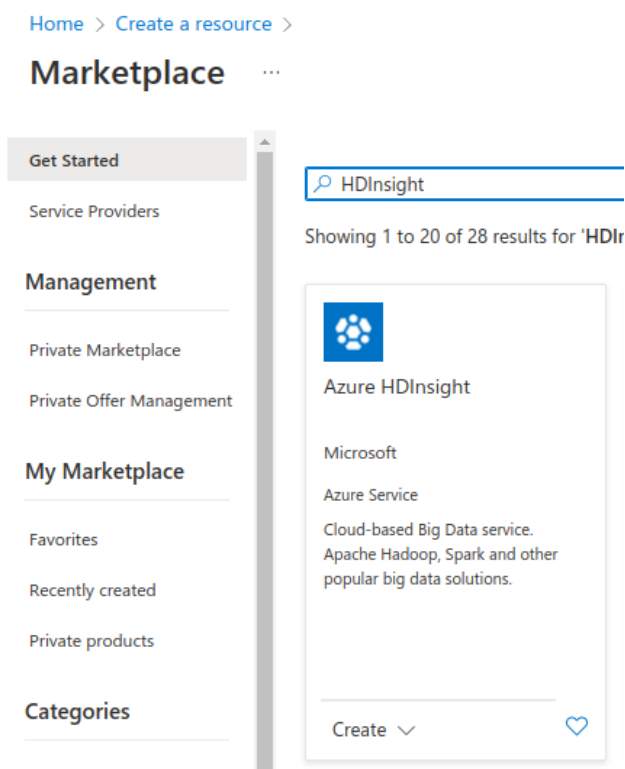
○ During the basic configuration of the HDInsight cluster, please make sure that you are using the proper big data Azure subscription and the region is **West Europe**. And please choose cluster type as **Spark with version 3.1**.

○ 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.

Create HDInsight cluster ...

Basics   **Storage**   Security + networking   Configuration + pricing   Tags   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 *	Azure Storage ⌄
Selection method * ⓘ	⦿ Select from list ◯ Use access key
Primary storage account *	Select an existing storage account ⌄
	Create new

○ We recommend you to 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).

Create HDInsight cluster ...

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 76 available cores in the West Europe region.
View cores usage
Open an HDInsight quota increase support case

+ Add application

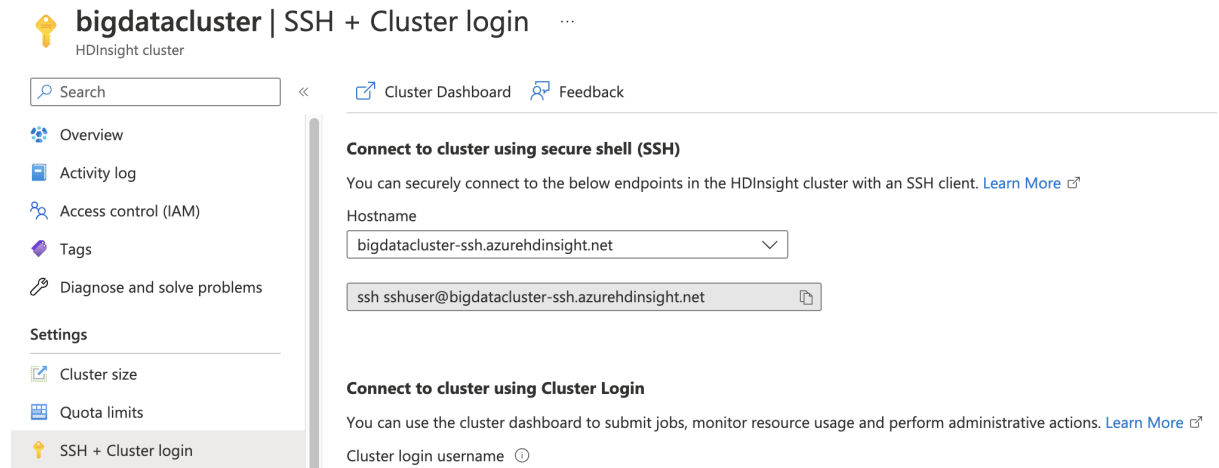| Node type	Node size	Number of ...	Estimated cost/h...
Head node	E4a v4 (4 Cores, 32 GB RAM), 0.38 USD/hour ⌄	2	0.76 USD
Zookeeper node	A2 v2 (2 Cores, 4 GB RAM), 0.12 USD/hour ⌄	3	0.00 (FREE)
Worker node	E4 V3 (4 Cores, 32 GB RAM), 0.38 USD/hour ⌄	4 ✓	1.52 USD

☐ Enable managed disk

Total estimated cost/hour   2.28 USD

Script actions

Use script actions to run custom PowerShell or Bash scripts on cluster nodes during cluster provisioning. Learn about script actions

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

🔑 **bigdatacluster** | SSH + Cluster login ···
HDInsight cluster

🔍 Search	≪	☑ Cluster Dashboard 🗫 Feedback

Connect to cluster using secure shell (SSH)

You can securely connect to the below endpoints in the HDInsight cluster with an SSH client. Learn More ↗

Hostname

bigdatacluster-ssh.azurehdinsight.net ⌄

ssh sshuser@bigdatacluster-ssh.azurehdinsight.net 📋

Connect to cluster using Cluster Login

You can use the cluster dashboard to submit jobs, monitor resource usage and perform administrative actions. Learn More ↗

Cluster login username ⓘ

Left sidebar navigation:
- 👥 Overview
- 🗎 Activity log
- 🗝 Access control (IAM)
- 🏷 Tags
- 🔗 Diagnose and solve problems

Settings
- ✏ Cluster size
- ▦ Quota limits
- 🔑 SSH + Cluster login

○ Access your cluster.
Make sure you can access your cluster (the NameNode) via SSH:
$ ssh <ssh_user_name>@<cluster_name>-ssh.azurehdinsight.net

E.g., *ssh sshuser@bigdatacluster-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>/*

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

> *json-file("wasbs://big-data-2022-11-07t16-17-03-407z@bigdatablobstorage.blob.*
> *core.windows.net/*.json")*
>
> *json-file("wasbs://big-data-2022-11-07t16-17-03-407z@bigdatablobstorage.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 9090

- ○ SSH forwarding
  After running RumbleDB as a server, we can user a juypter notebook to interact
  with it. We recommend to use SSH forwarding. For that, make sure you have run:
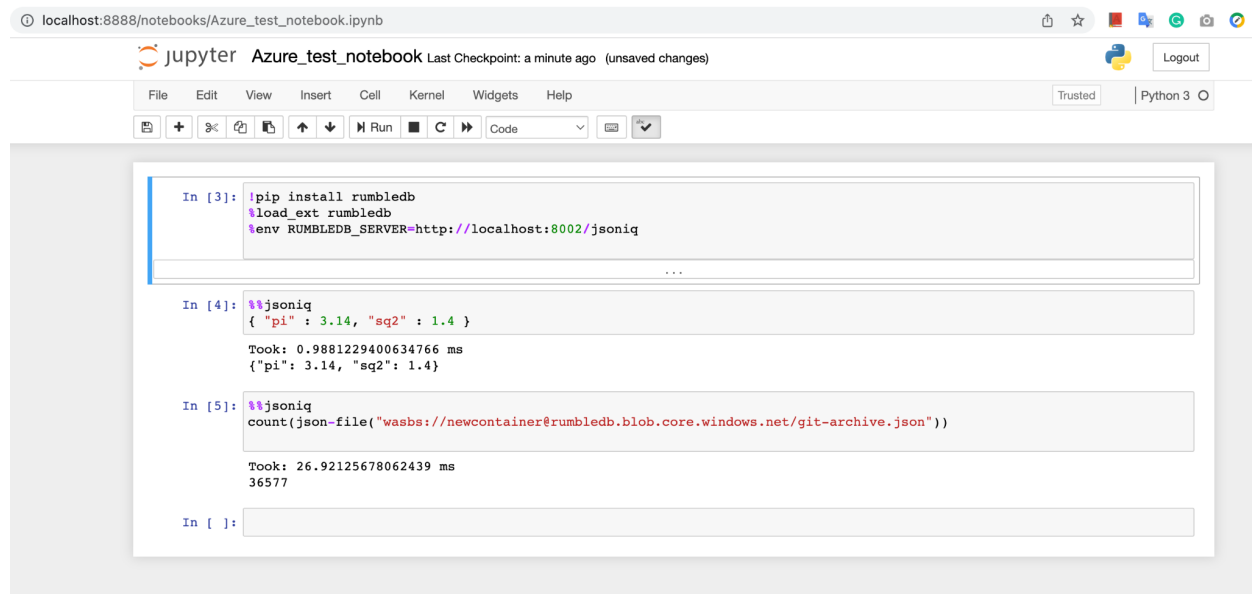 *spark-submit rumbledb-1.20.0-for-spark-3.1.jar --server yes --port 8002*

  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@bigdatacluster-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 [this week'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.