Readme file for Spark Cluster

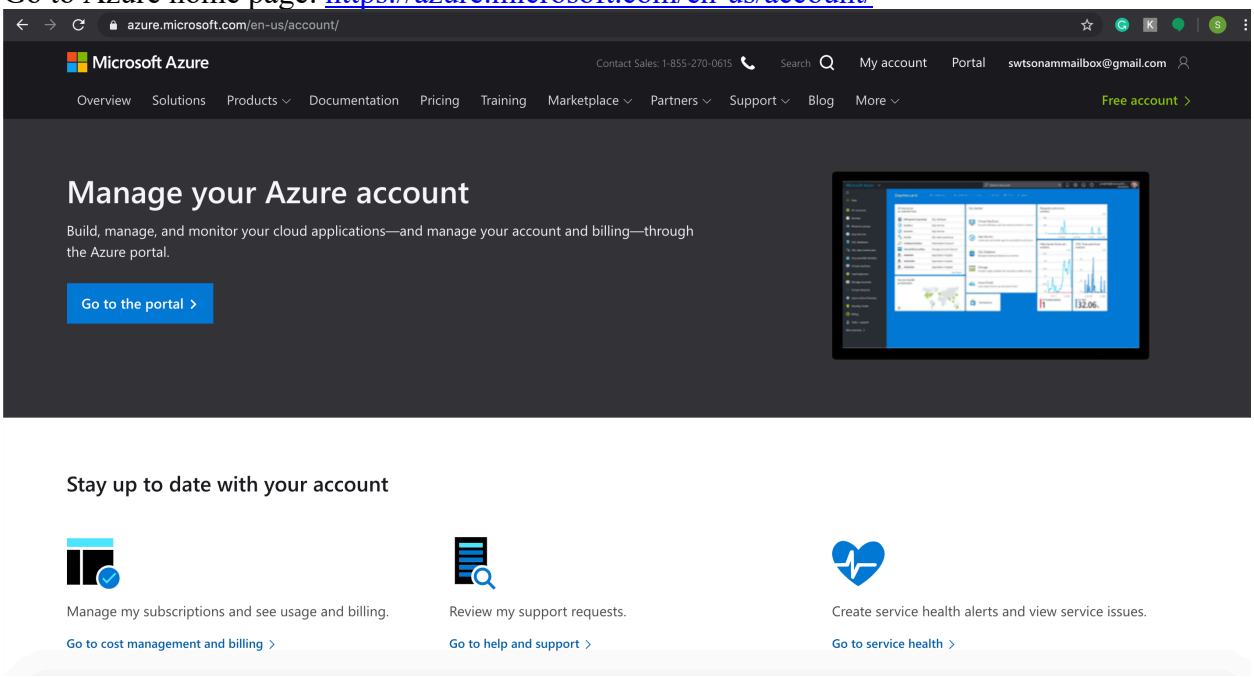
We have two Readme files for this project. One for the local environment and other is for spark cluster environment. Local environment readme file is named as 'Readme_local.md' where 'March2016_Twitter_Processing_Local.ipynb' notebook is to be executed. This readme file is for the spark cluster setup, where Jupyter notebook 'March2016_Tweet_Processing_Spark.ipynb' needs to be executed. Order of execution needs to be as readme for spark cluster and then readme for the local environment. We are using Twitter archive data of March 2016, which can be downloaded from below link: *https://archive.org/details/archiveteam-twitter-stream-2016-03*

If you are facing trouble at any point, please reach out to any of us
email-ID: sdasari3@student.gsu.edu; sdawani1@student.gsu.edu; hwoodyard1@student.gsu.edu; ngupta9@student.gsu.edu; asubramanian5@student.gsu.edu
or on phone - *+13134132140*

We tried to make things as straight-forward as possible for you.

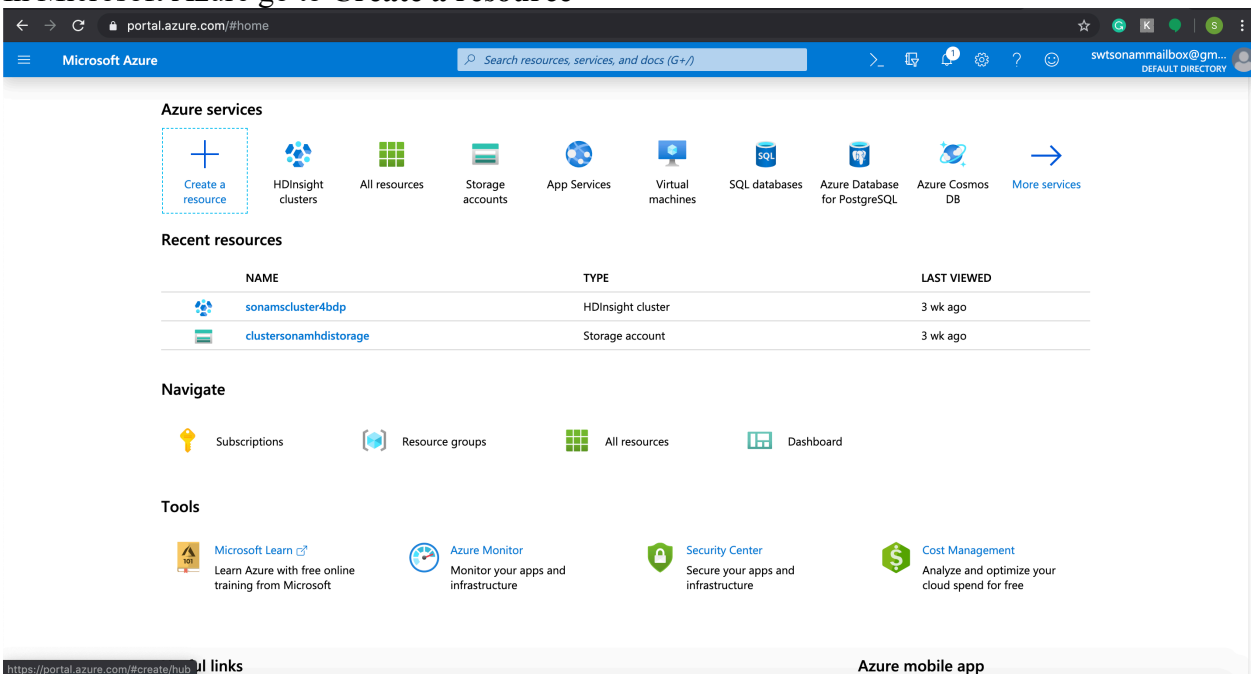
Instructions for setting up the cluster:

1. Go to Azure home page: https://azure.microsoft.com/en-us/account/

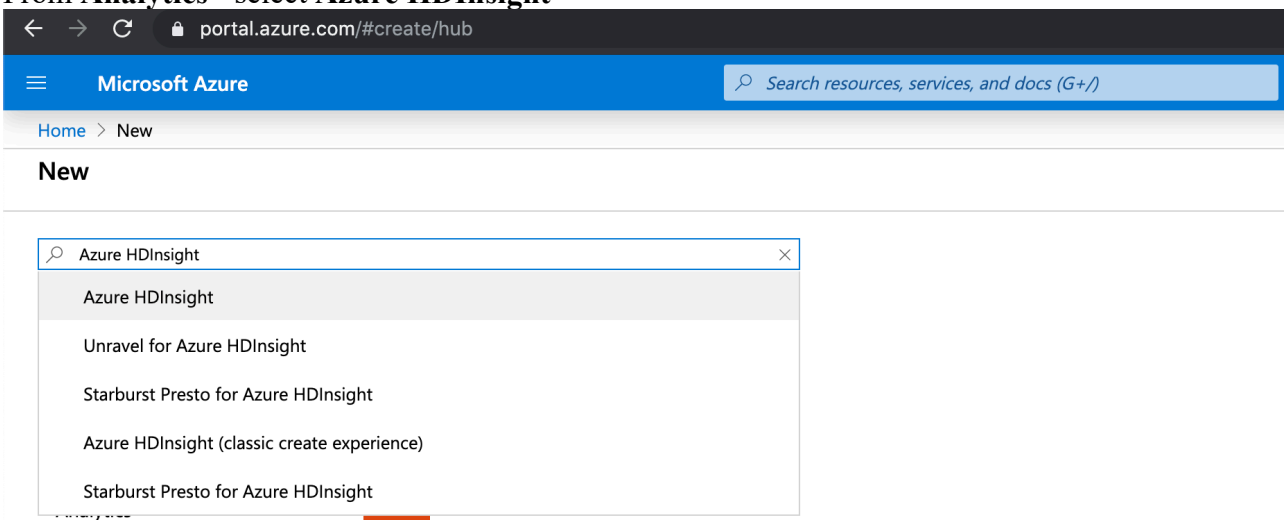


2. Either Sign in or create account
3. Go to the portal

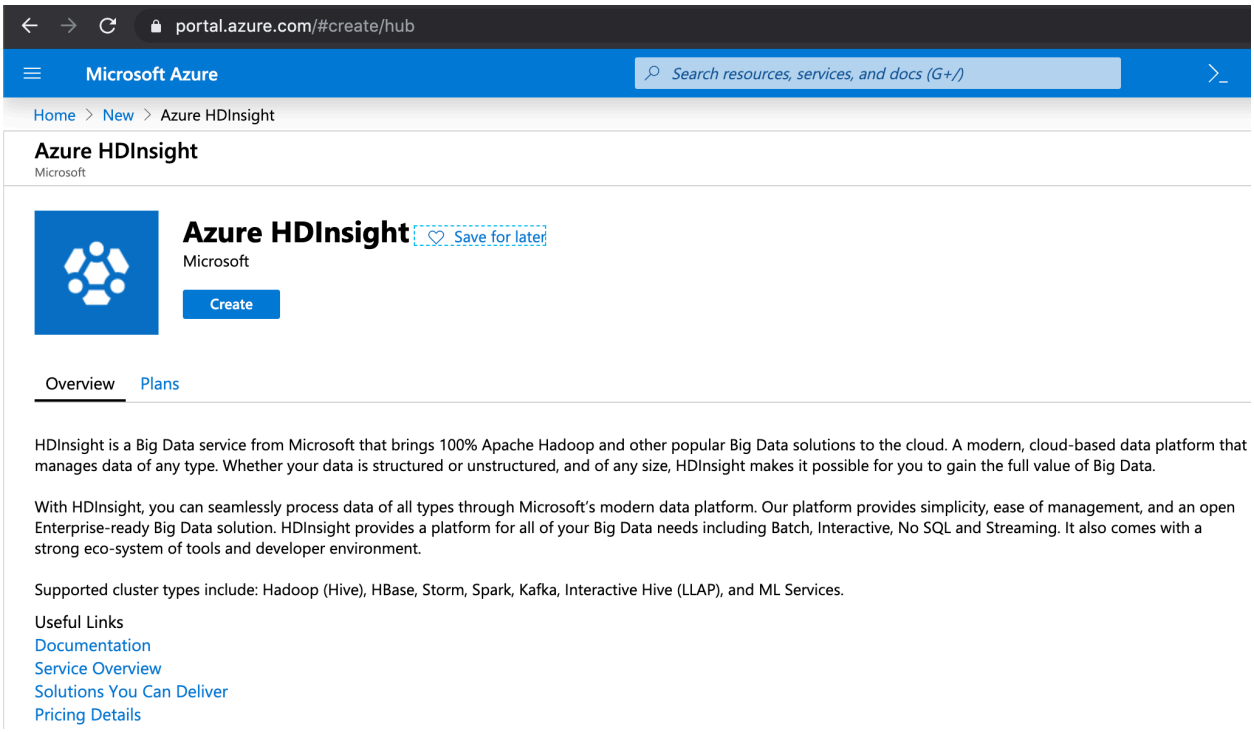
4. In Microsoft Azure go to **Create a resource**



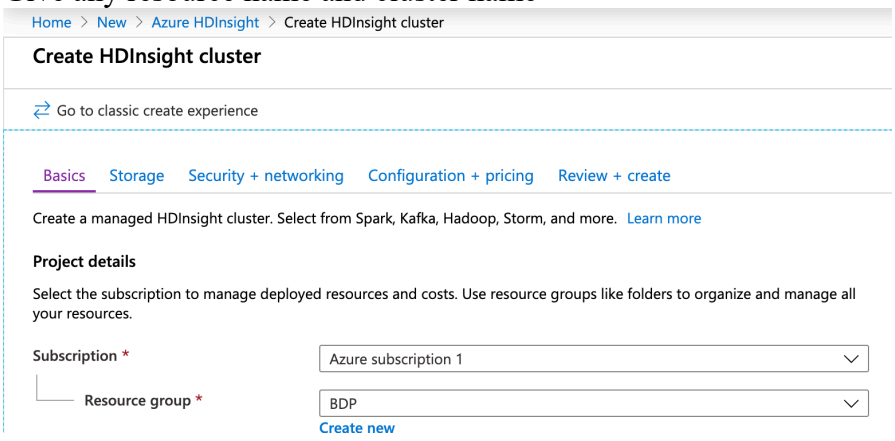
5. From **Analytics**> select **Azure HDInsight**



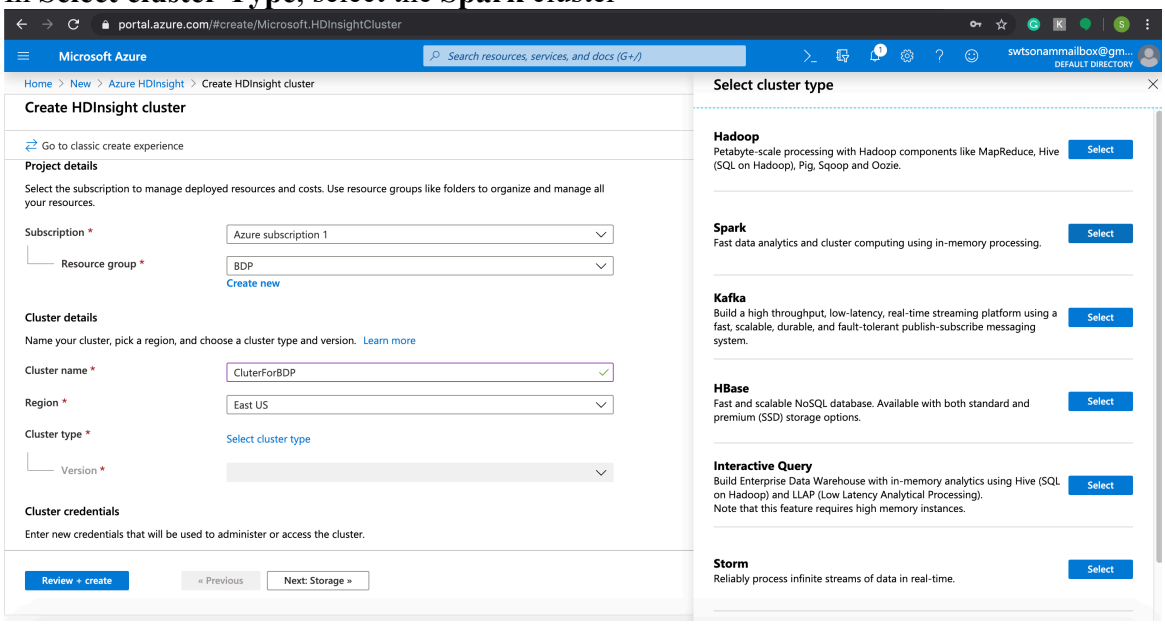
6. Click on Create



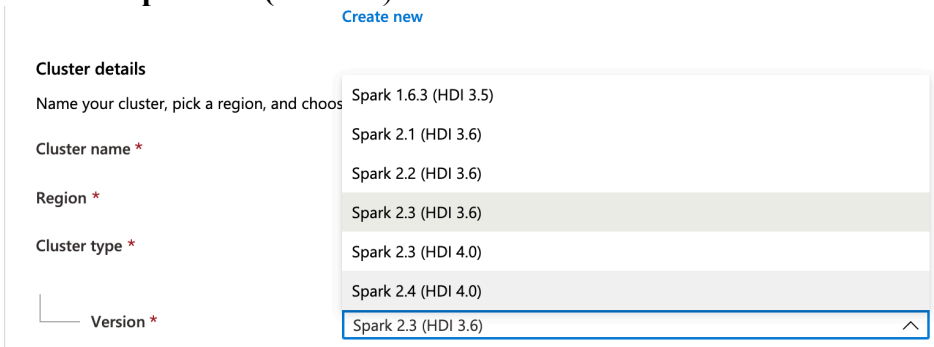
7. In **Basics**
   - Give any resource name and cluster name



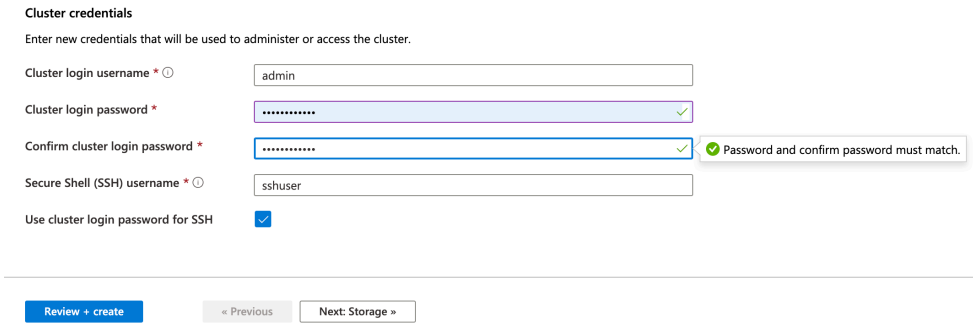
- In **Select cluster Type**, select the **Spark** cluster



- Version**: Spark 2.4 (HDI 4.0)**



- Fill in your usernames and passwords as necessary



- Press ‘Next Storage’

8. In **Storage,**

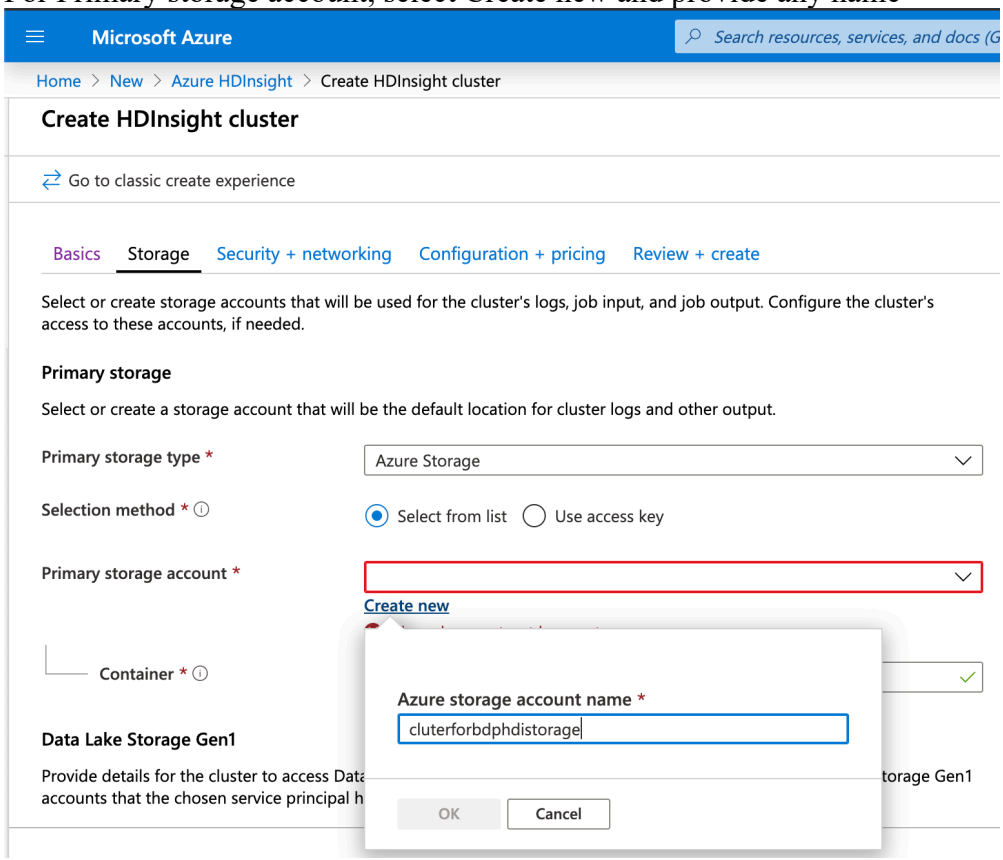
We have two options:

1. To create new storage and upload the data files on the storage (Primary storage)
2. Linking with already created storage where we have already uploaded data (Secondary storage)

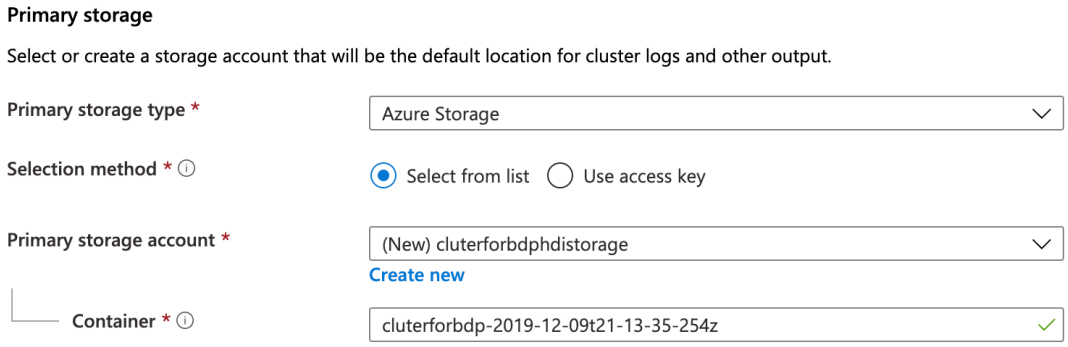
We are mentioning instruction for both the methods.

For the first method, which is to create primary storage:

- Select the primary storage type as: Azure Storage
- Use Selection method as: Select from list
- For Primary storage account, select Create new and provide any name



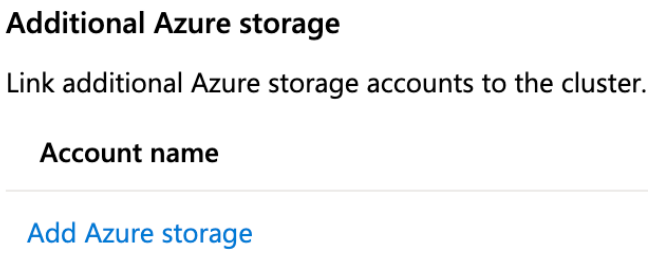
- Let the container as it is



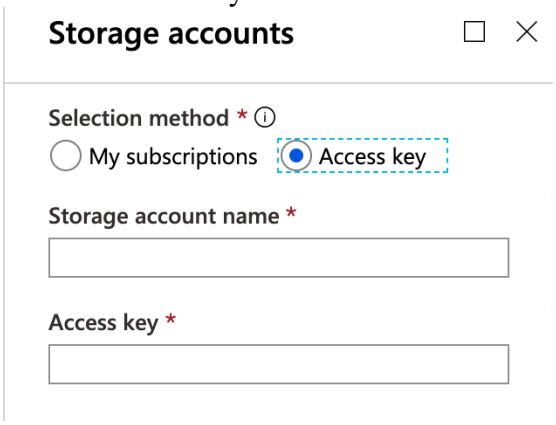
To add the data in Primary storage, follow **Instructions to upload the data into Azure blob storage account** given at the end of this section.

For the linking the existing storage where we have uploaded untouched March Twitter data for your convenience as secondary storage:

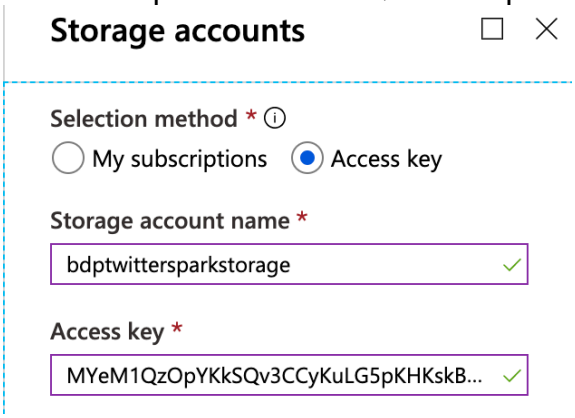
- Follow the above instructions to create a primary storage for your cluster in which the code will reside.
- Select the link 'Add Azure storage', it will open panel on right side



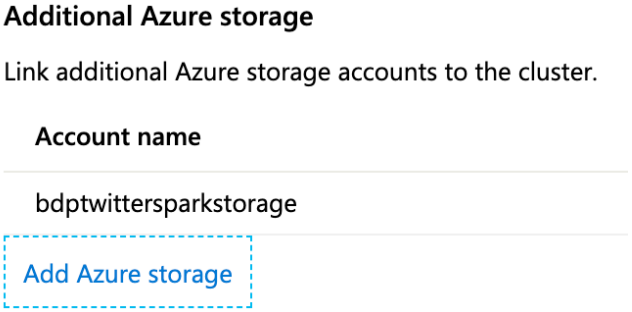
- Select 'Access key' as selection method:



- Use the below details:
  Storage account name: bdptwittersparkstorage
  Access Key:
  MYeM1QzOpYKkSQv3CCyKuLG5pKHKskBA5D2p0wlPmD1yQHtdfnRv0j
  X+mEtAzspsv7DnGc6OaQw15IY9qLZBOw==

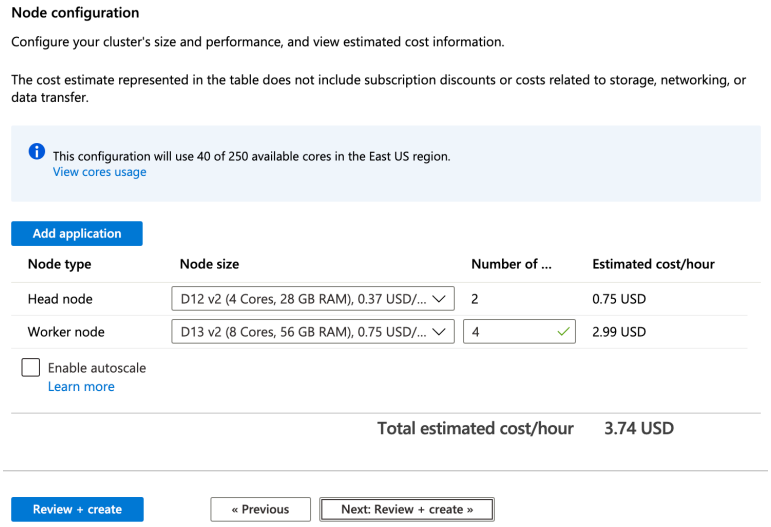


- After clicking Select, you will find the storage added



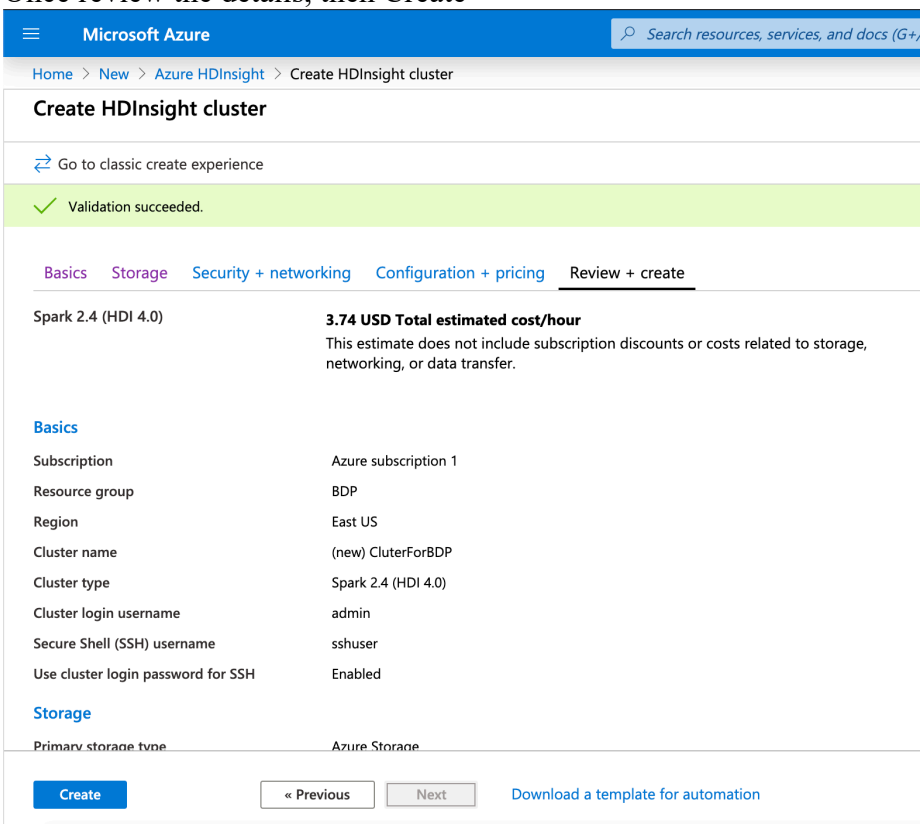
- Click Next Security + Networking

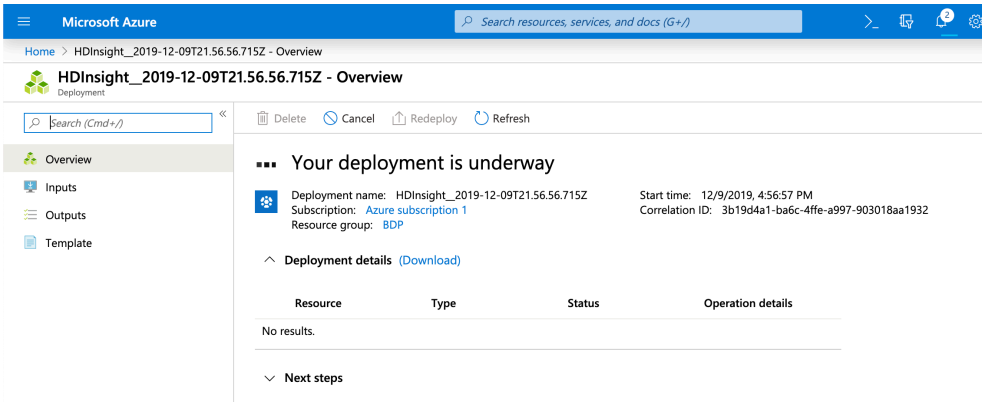
9. Leave the **Security + Networking** as is.
10. In **Configuration + Pricing** set up cores as necessary for the process you want to run. Recommended:



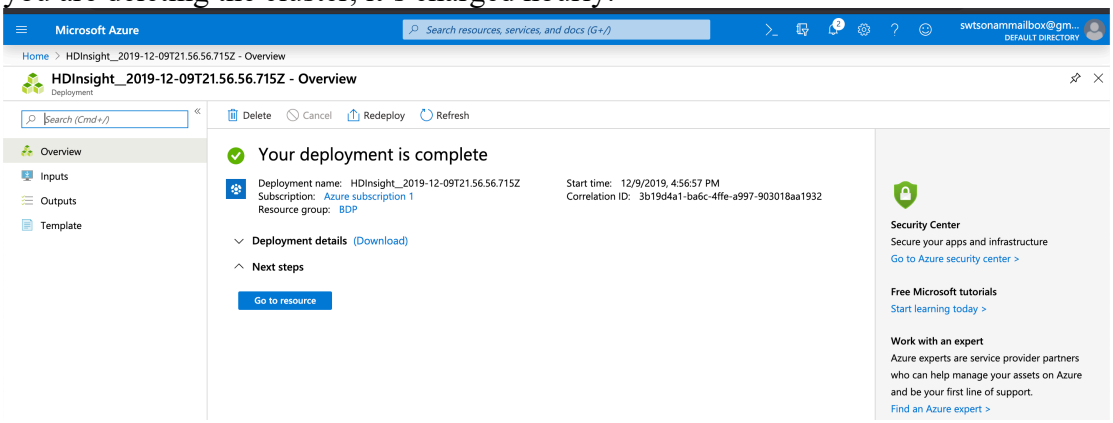
11. Click ***Review+Create***

12. Once review the details, then Create

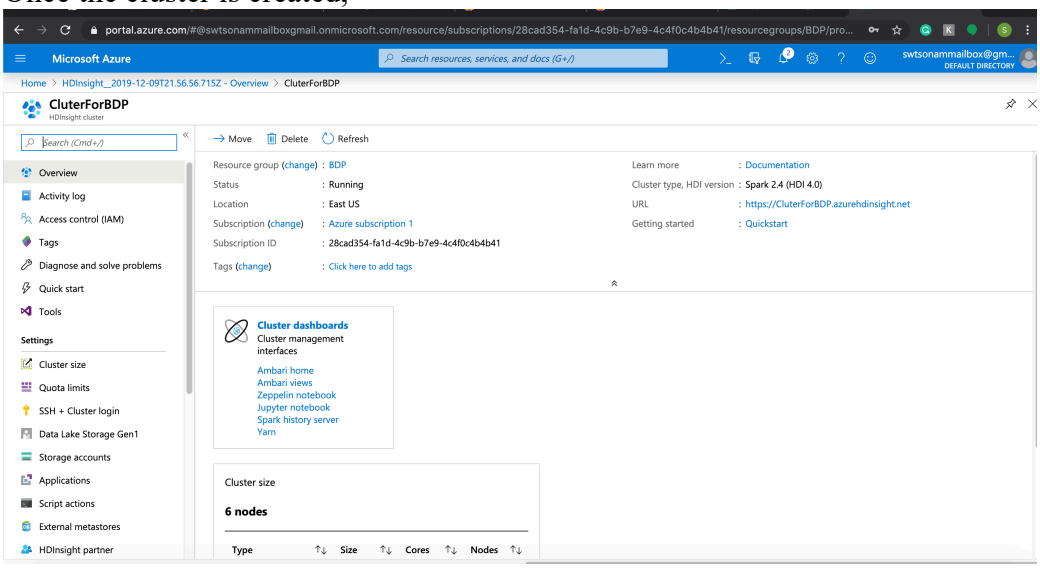




13. It takes around 10-15 mins for the cluster to start. Once you are done using it make sure you are deleting the cluster, it's charged hourly.



Once the cluster is created,



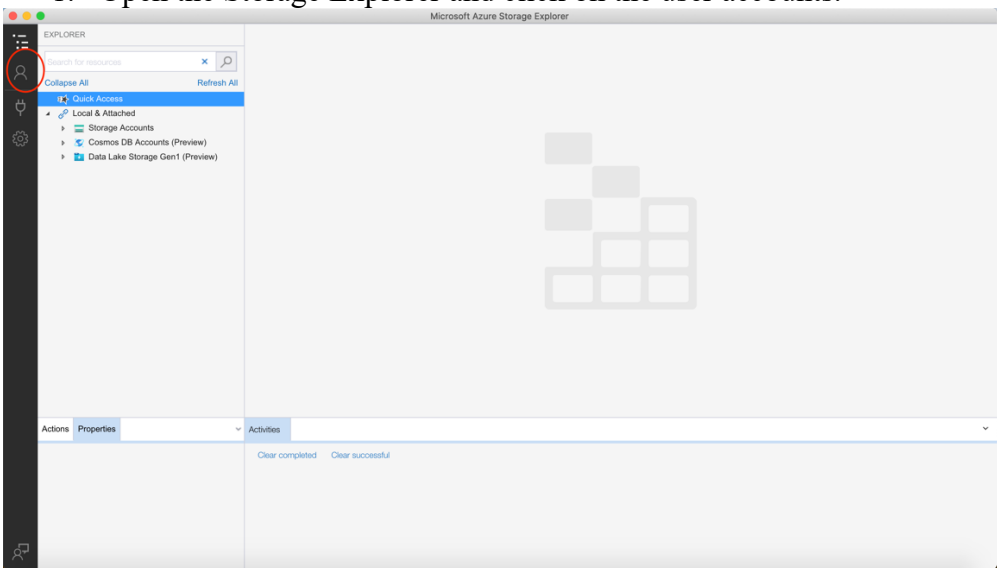
we have to update some settings on cluster according to the requirement of the project. The instructions for the same are listed under **Instructions to update cluster setting and installing external libraries**

Instructions to upload the data into Azure blob storage account:

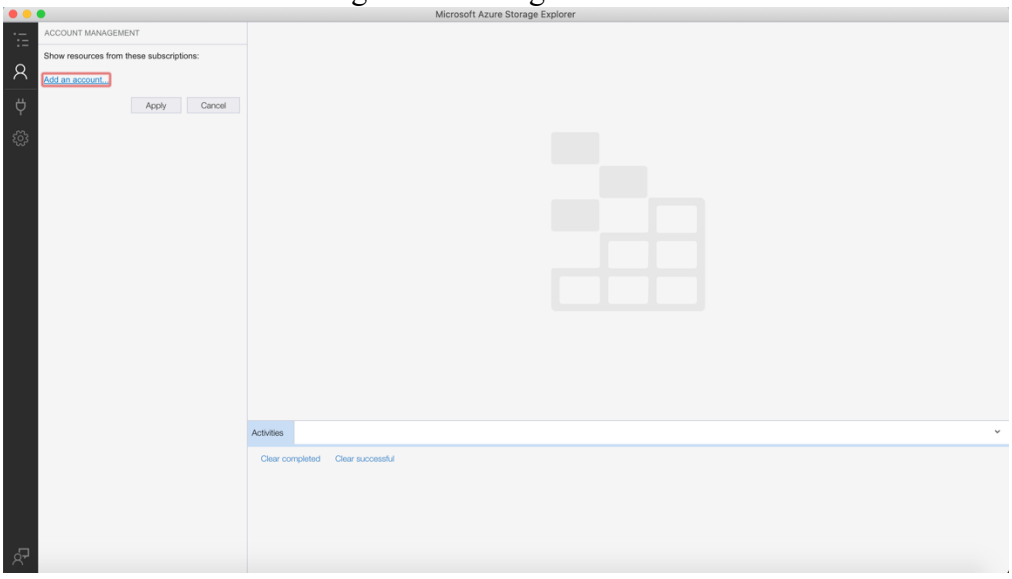
Please find below the link to download Microsoft Azure Storage Explorer:
https://azure.microsoft.com/en-us/features/storage-explorer/

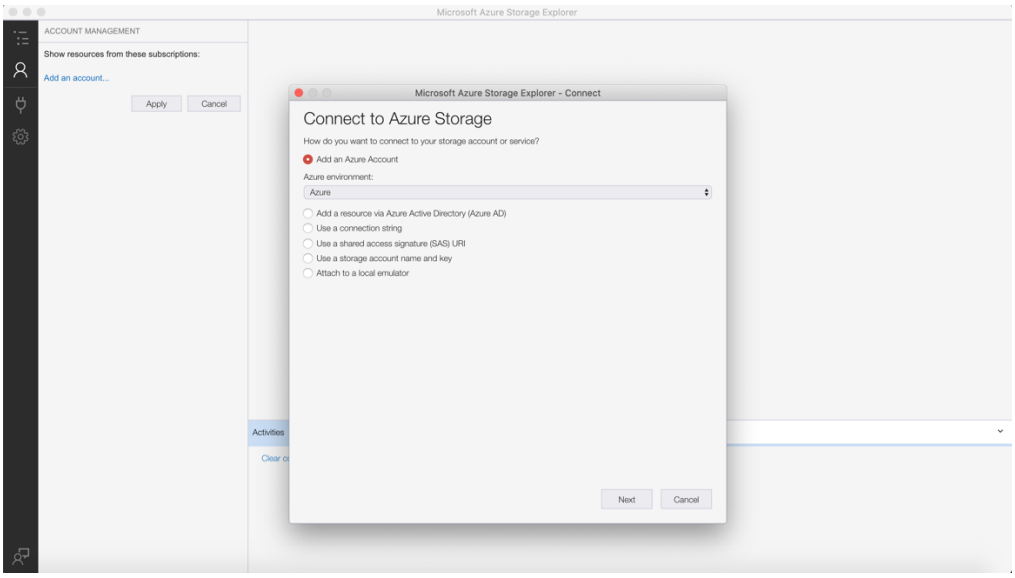
Once the Storage explorer is downloaded and installed, follow the below instructions:
The necessary parts are highlighted in red in below screenshots.

1. Open the Storage Explorer and click on the user accounts.

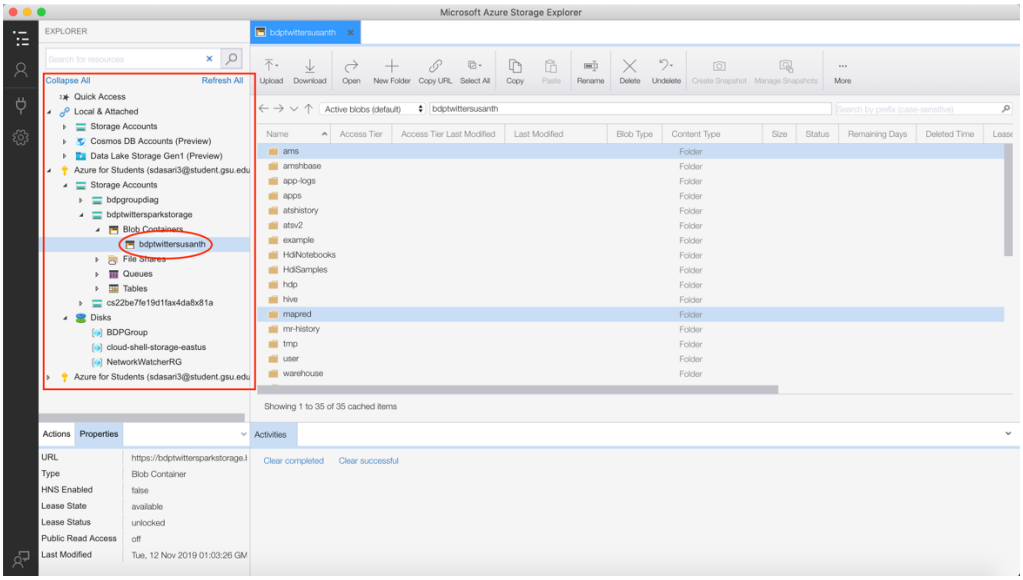


2. Click on the **Add an account** option and follow instructions to add your account on which the blobstorage for HDInsights cluster is created.



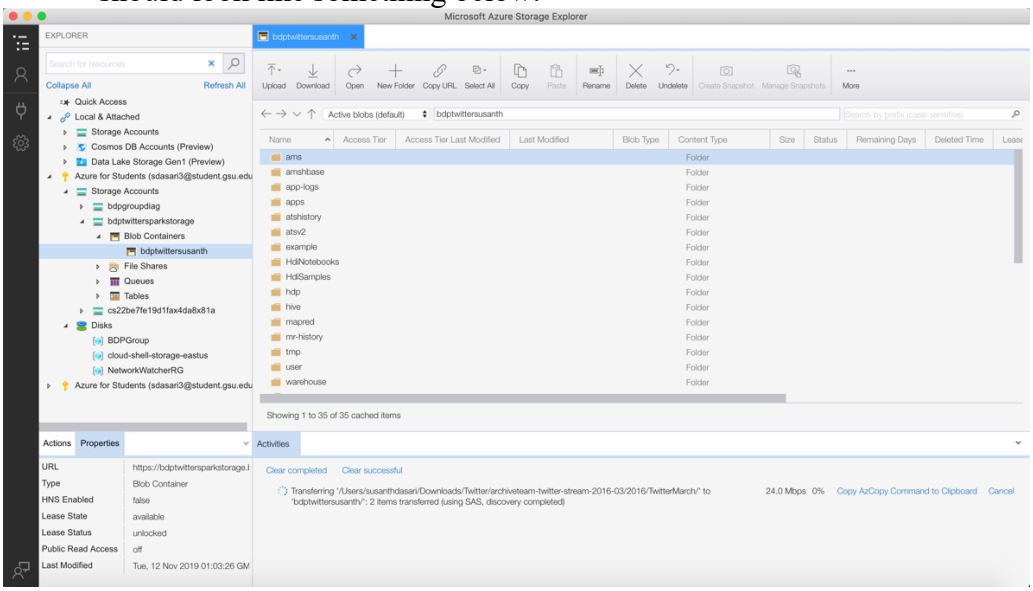


3. Once the account is added, please navigate to the blobstorage created for HDInsights cluster as shown below.



4. Use the **upload** option in the menu bar and select **Upload Folder**.
5. Navigate to the folder which is extracted from the downloaded **Twitter March TAR file (**https://archive.org/download/archiveteam-twitter-stream-2016-03)
6. Open the folder **archiveteam-twitter-stream-2016-03 > 2016 > TwitterMarch** and select **TwitterMarch.**

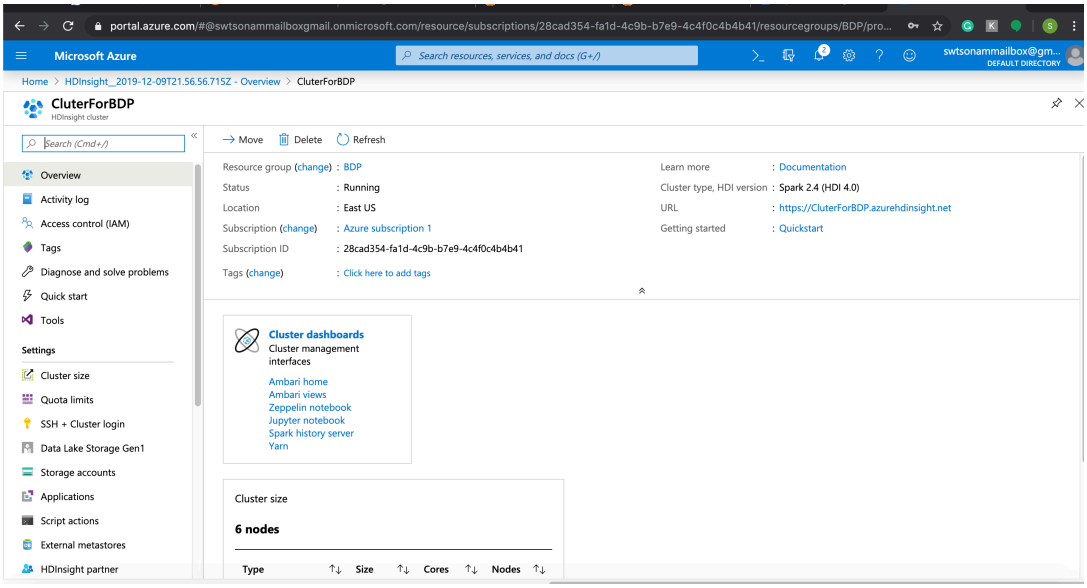
7. The upload of all the subdirectories of **TwitterMarch** should begin and the progress should look like something below.



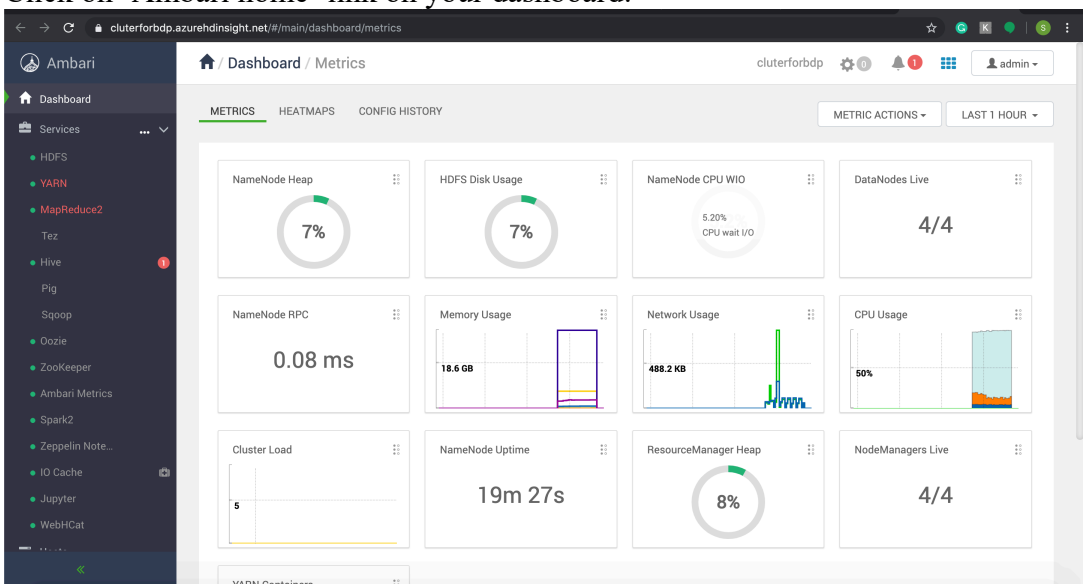
Instructions to update cluster setting and installing external libraries:

So we will be starting with dashboard to add and update some configuration parameters

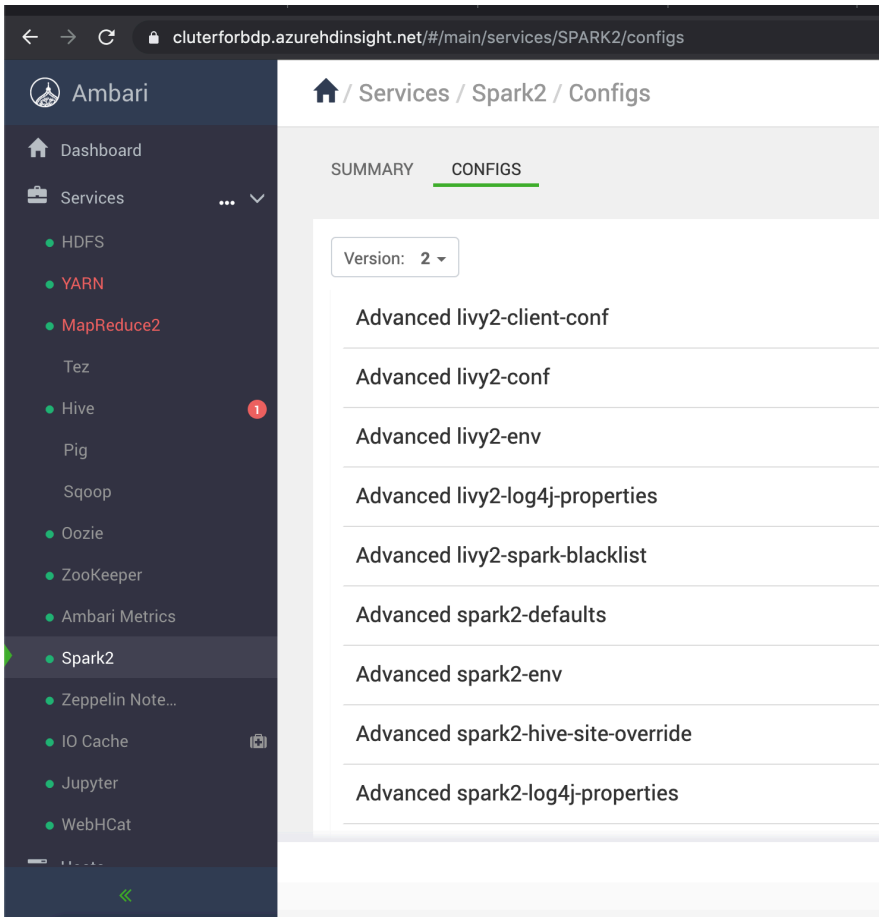
Dashboard will look like:



- Click on ‘Ambari home’ link on your dashboard:

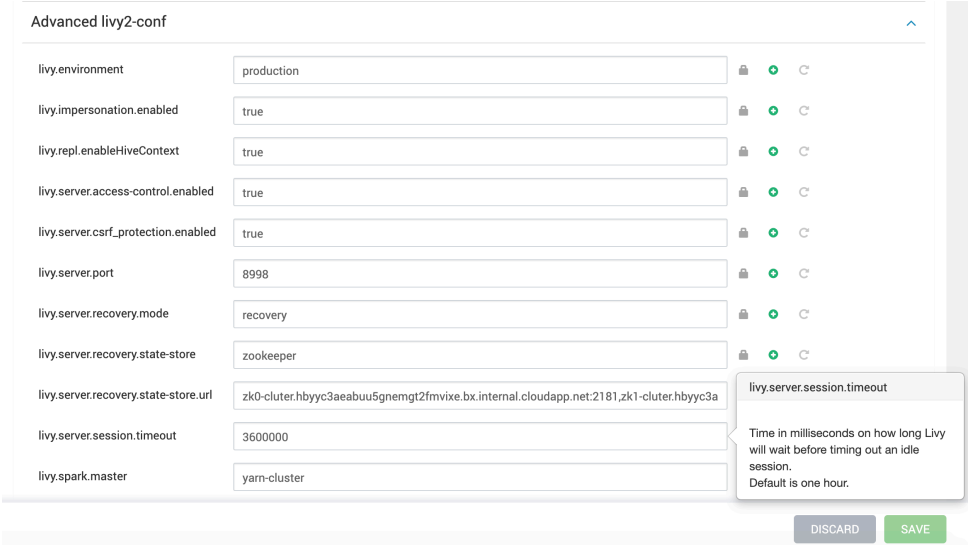


- Click on Spark2 from the left panel
- Click on CONFIGS

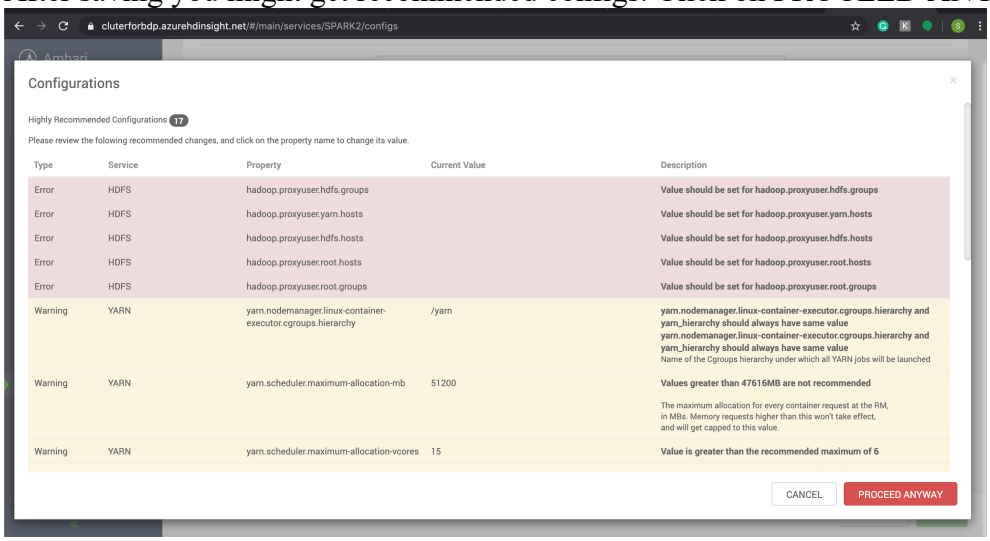


Parameters to update:

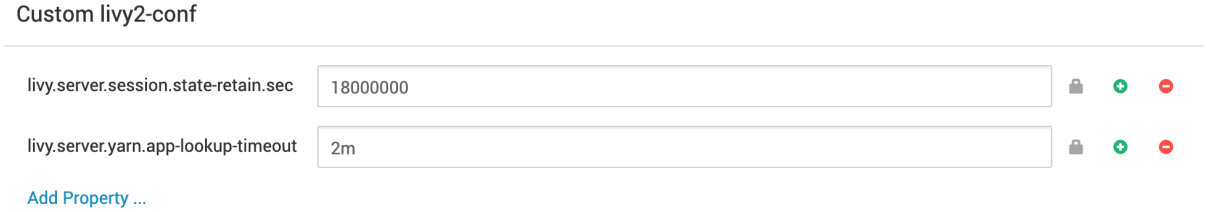
1. ‘Advanced livy2-conf’ Under ‘Spark2’
   - Select the ‘Advanced livy2-conf’
   - Update ‘livy.server.session.timeout’ from 36000000 to 180000000



   - After saving you might get recommended configs. Click on PROCEED ANYWAY

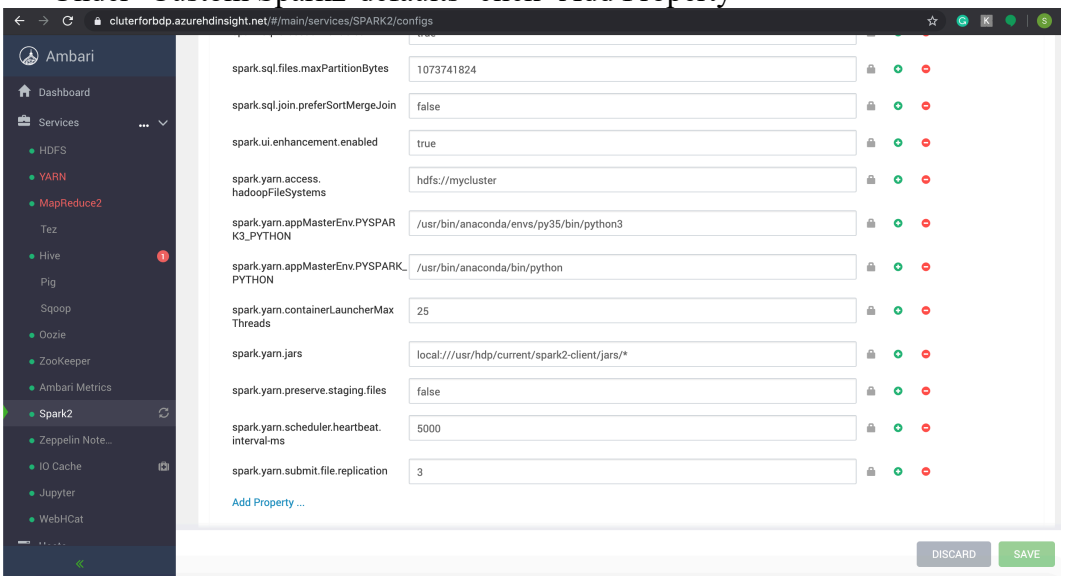


2. ‘Custom livy2-conf’ Under ‘Spark2’
   - Select the ‘livy.server.session.state-retain.sec’
   - Update ‘livy.server.session.state-retain.sec’ from 36000000 to 180000000

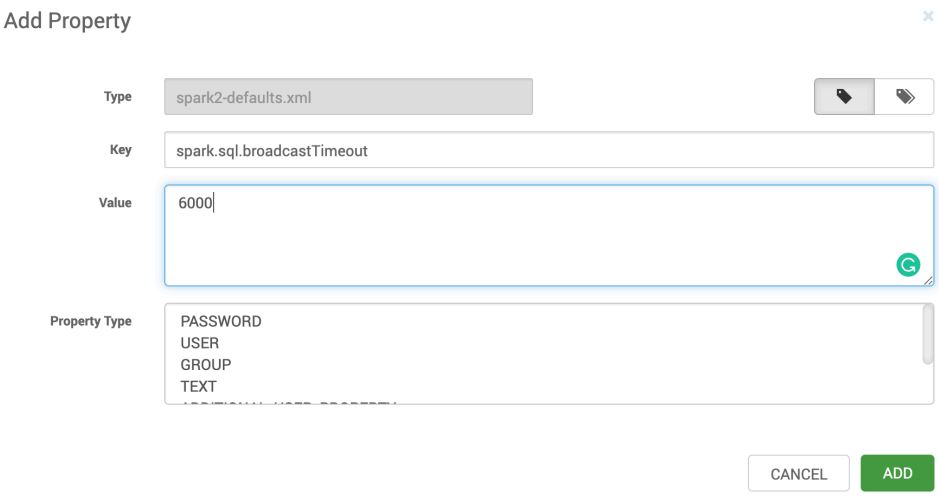


Parameters to add:

1. Under ‘Custom Spark2-defaults’
   - Under ‘Custom Spark2-defaults’ click ‘Add Property’



   - Fill the details as:
     Key: spark.sql.broadcastTimeout
     Value: 6000



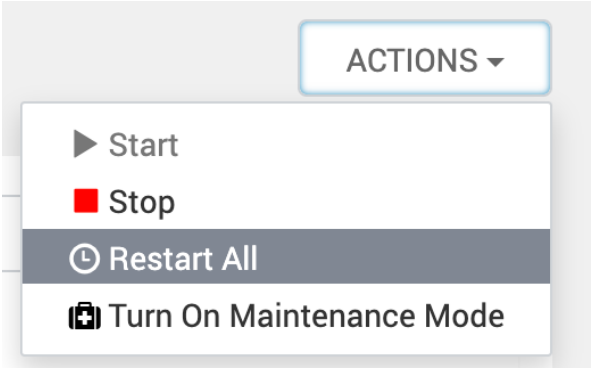
2. Under ‘Custom Spark2-defaults’
   Key: spark.driver.memory
   Value: 32g

3. From the left panel go to ‘Jupyter’
   - Under ‘Custom jupyter-site’
   - Add property
     Key: MappingKernelManager.cull_idle_timeoutInt
     Value: 0

4. Being in Jupyter config
   - Under ‘Custom jupyter-site’
   - Add property
     Key: NotebookApp.shutdown_no_activity_timeoutInt
     Value: 0

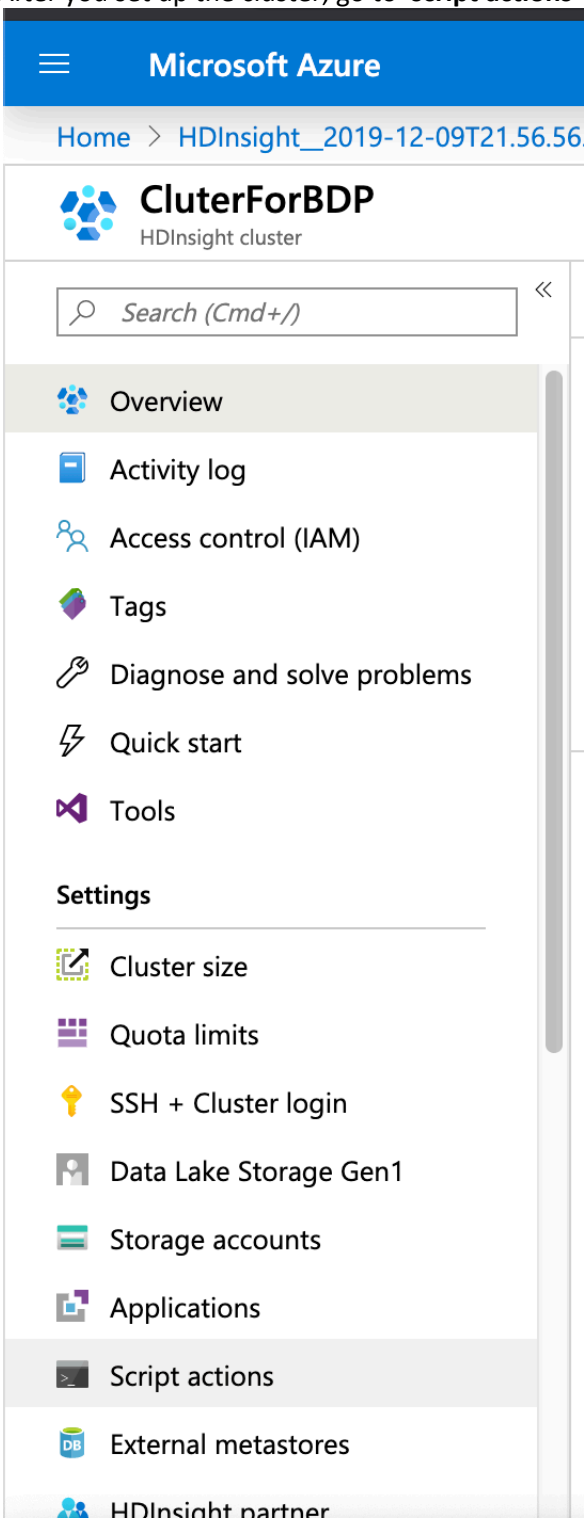
After updating and adding configs, for ‘Spark2’ and ‘Jupyter’ :

- Go to the ‘Actions’ on top right corner for each of them separately and select **restart all**

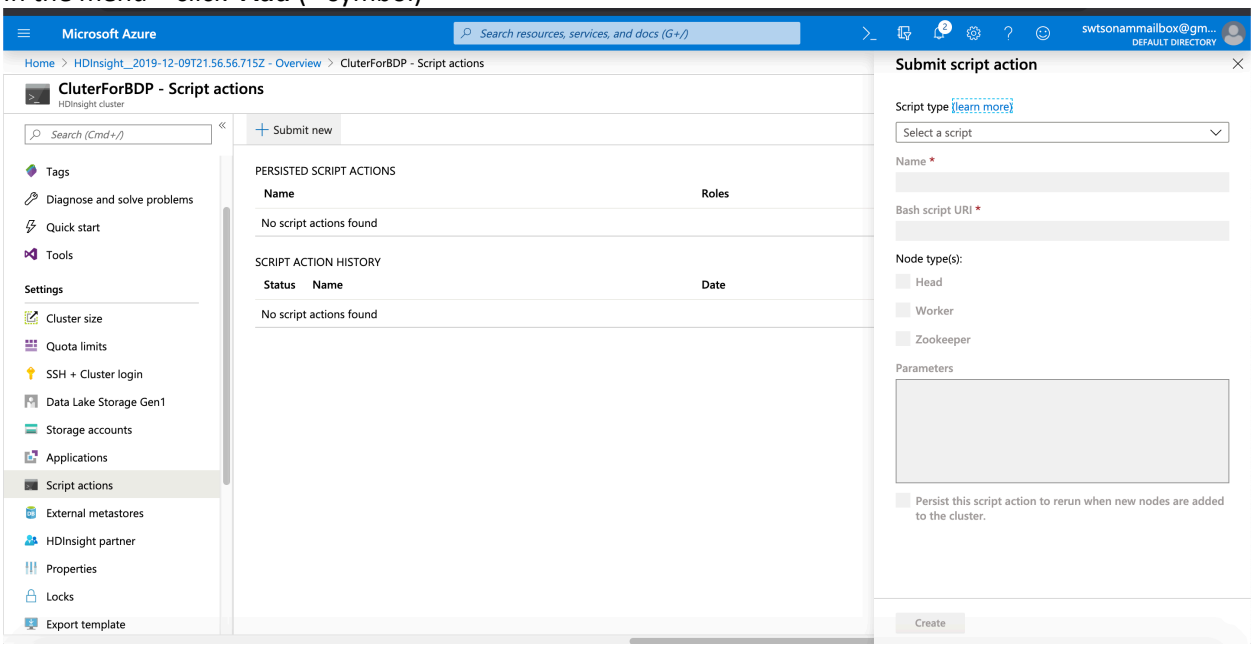


Adding external libraries:

- After you set up the cluster, go to '**script actions**'



- In the menu > click '**Add'**(+ symbol)



- Select Script as 'Custom'
- Provide any name
- For Bash script URI, we need to provide URI for the script (pip3-install-packages-bash.sh) which is given in the project zip.

  URI: *https://bdptwittersparkstorage.blob.core.windows.net/bdptwittersusanth/pip3-install-packages-bash.sh*
  This script will install required libraries for the project.