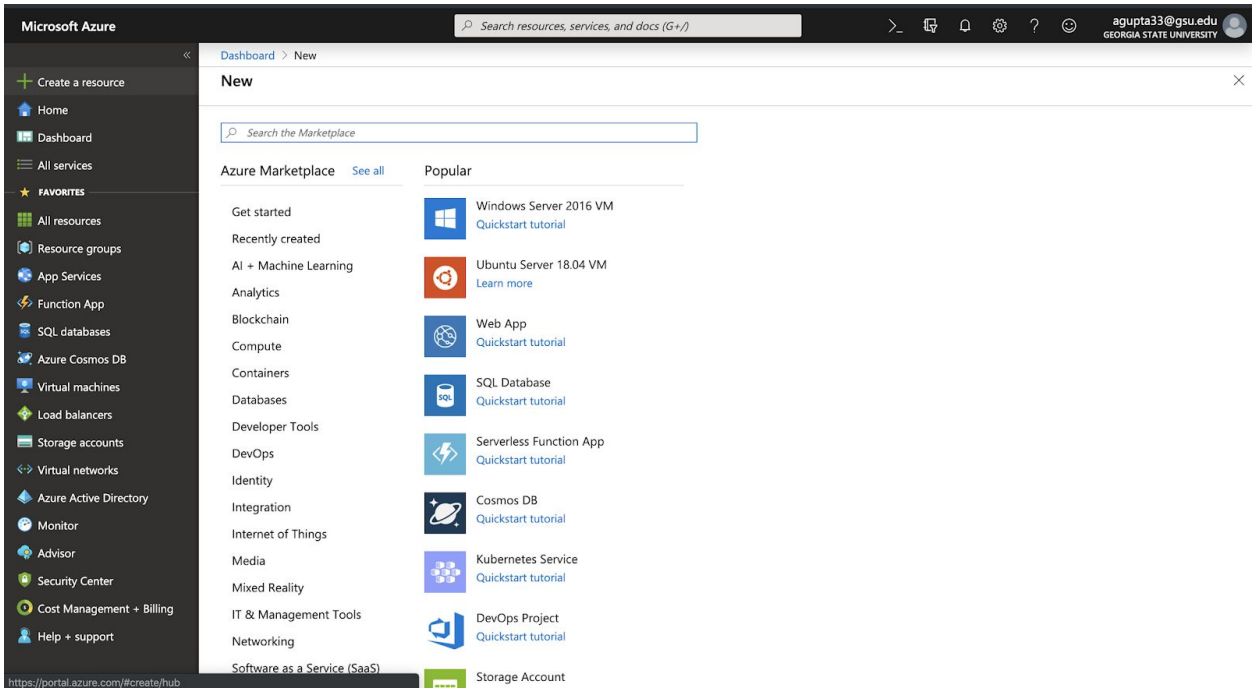
READ ME INSTRUCTIONS

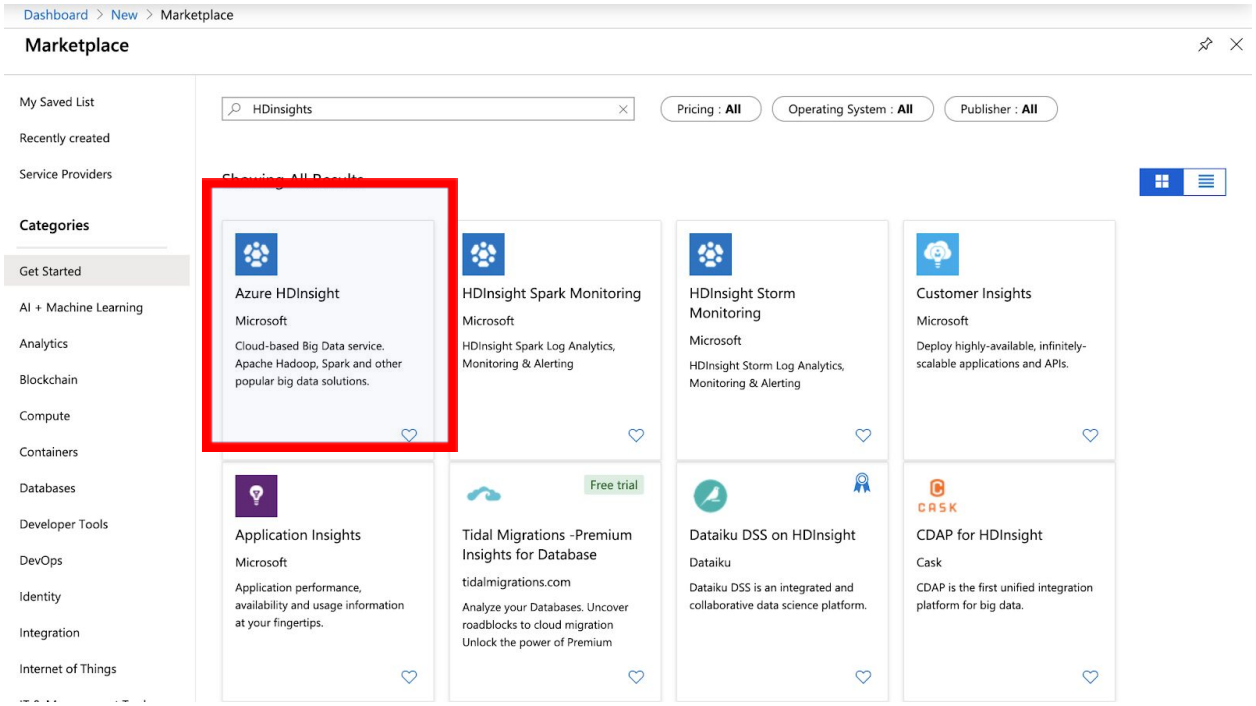
The file is divided into 2 parts.
First part explains how we **can create cluster using azure and run the preprocessing code**.
The second part deals **with installations of the important libraries** which will be required to run the questions.

Part I - CLUSTER CREATION AND PREPROCESSING

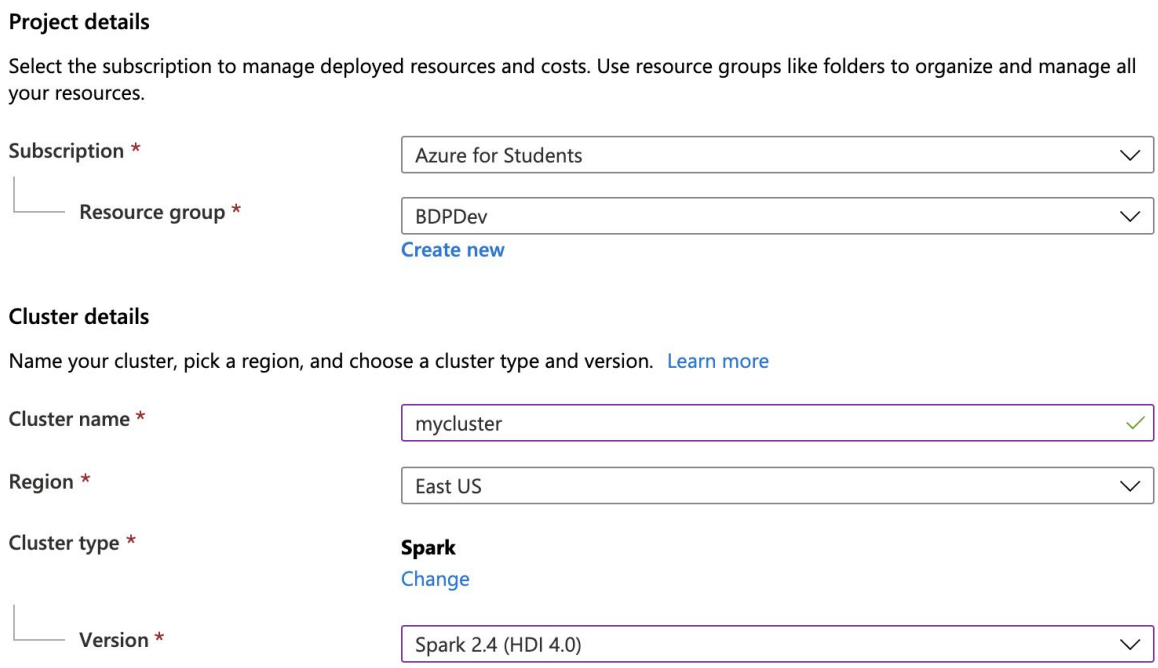
Step 1: Download Microsoft Azure Storage Explorer
Step2: login with your same credentials as you have in hdinstights.
Step3 : login to Microsoft azure account
Step 4: click on create resource button on top left and search for HD insights.



Step 5: Click on HDInsight and start the creation of the cluster



Step 6: Fill **project details** as below make sure you are choosing the correct version of spark and the name of your cluster should be unique



Step7: Fill in **cluster credentials**

Cluster credentials

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

Cluster login username * ⓘ	admin
Cluster login password *	
Confirm cluster login password *	
Secure Shell (SSH) username * ⓘ	sshuser
Use cluster login password for SSH	☑

Step8: This is important please fill the same details if you are using our (Anit's) storage account.

Or can click on create new and create a new storage account.

Also if you are using our storage account you would like to add your own storage account to run scripts which are explained below.

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 *	Azure Storage
Selection method *	Select from list / Use access key
Primary storage account *	bdpproj1 Create new
Container *	new

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

Click next, There is no need to change anything in Security + networking

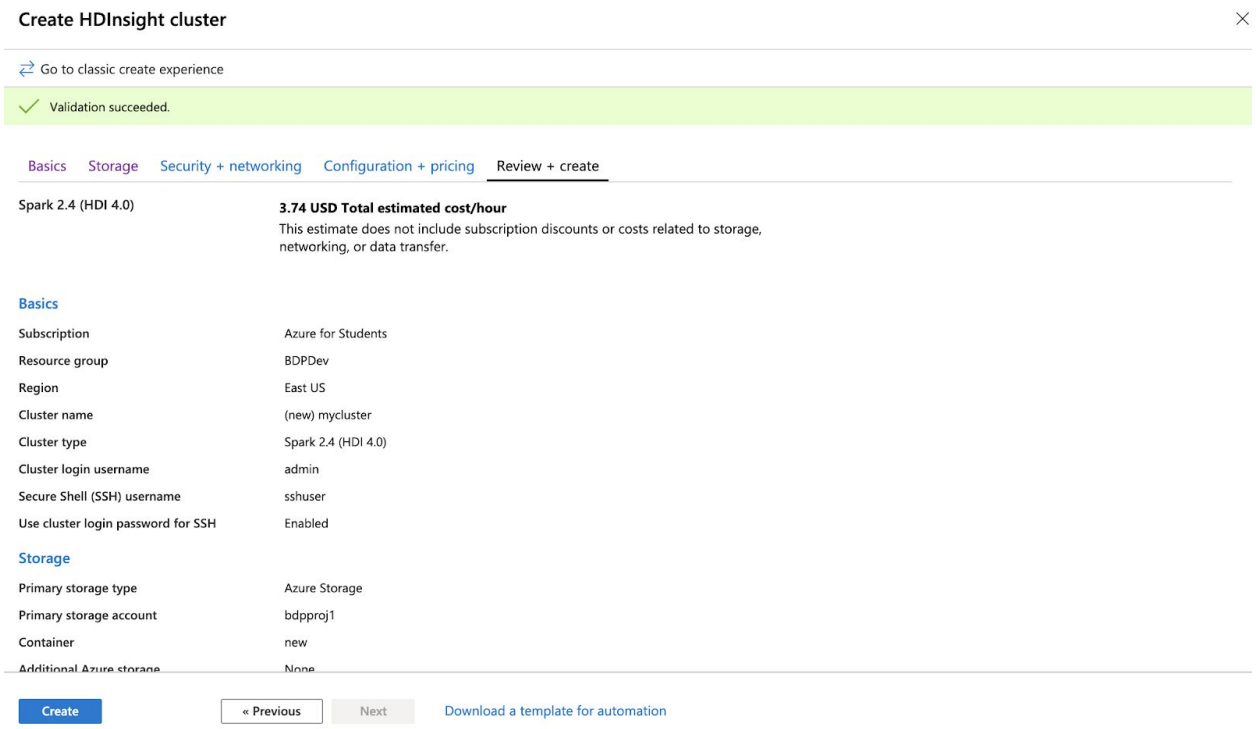
Step 9: Click next then on Configuration + pricing select the configuration of the head node and worker node. And then click next

Add application

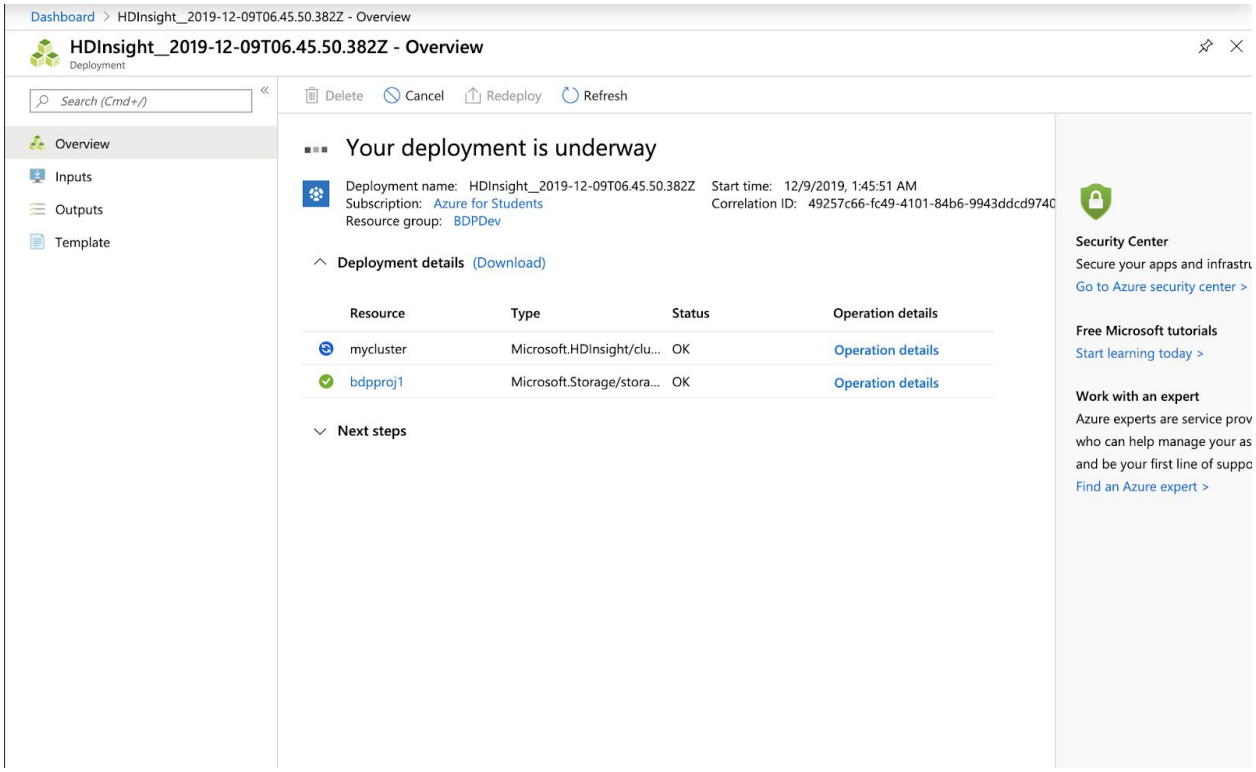
Node type	Node size	Number of ...	Estimated cost/hour
Head node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/...	2	0.75 USD
Worker node	D13 v2 (8 Cores, 56 GB RAM), 0.75 USD/...	4	2.99 USD

Enable autoscale
Learn more

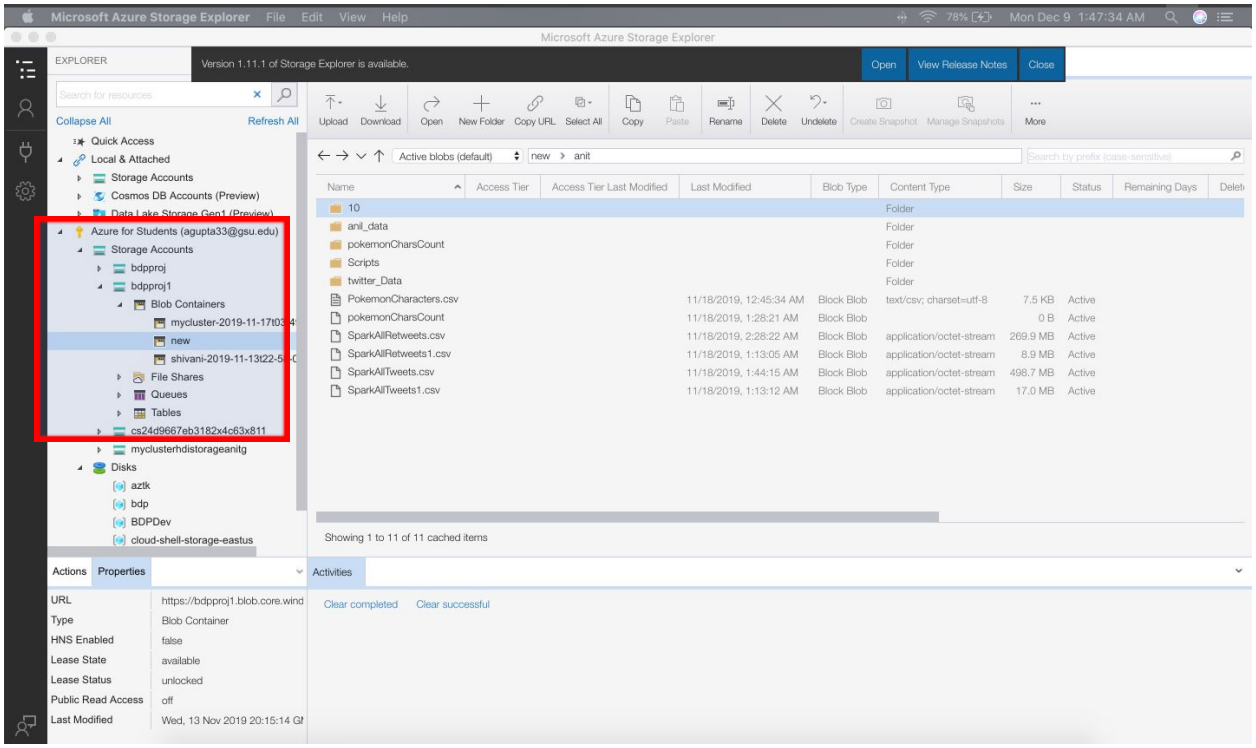
Total estimated cost/hour 3.74 USD



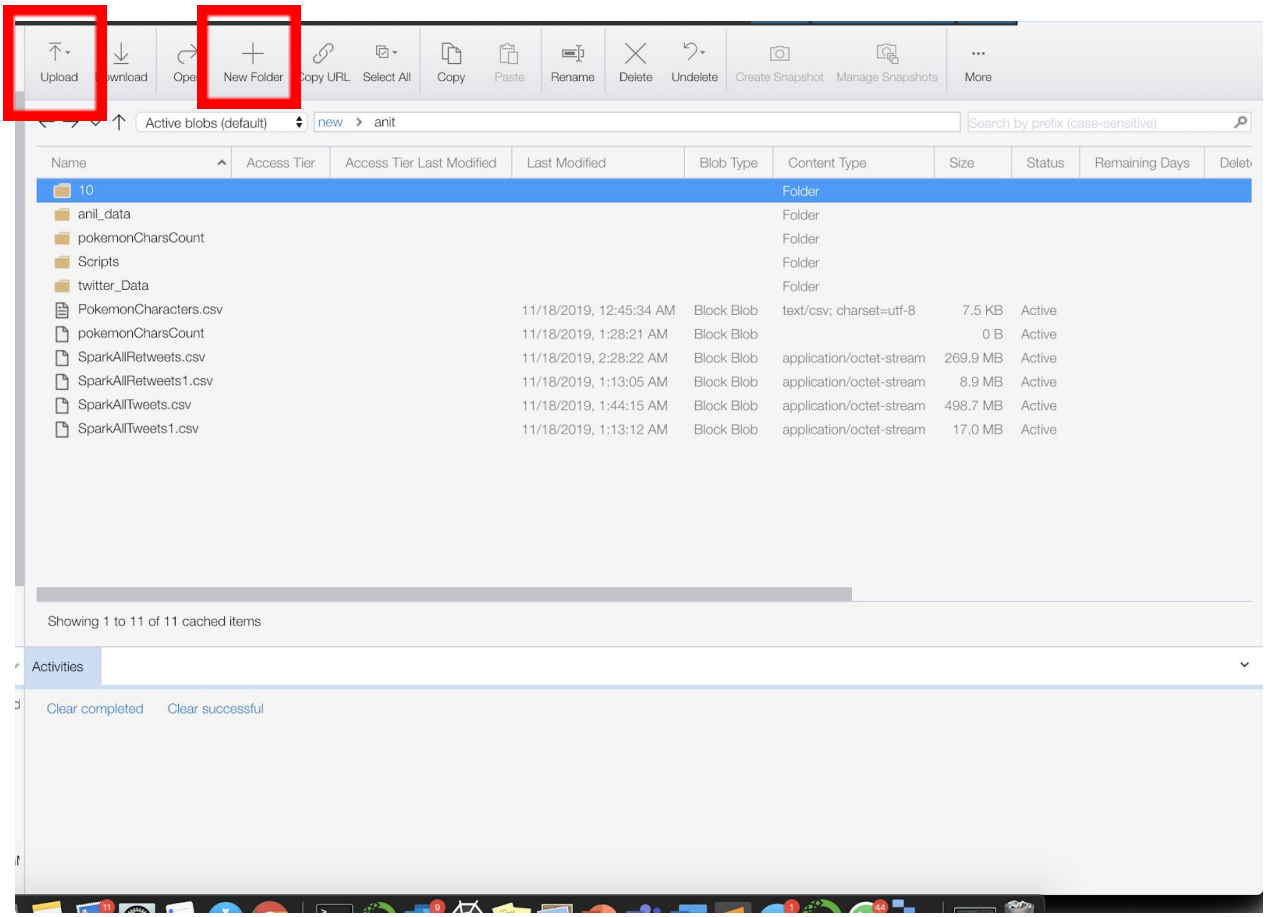
Once the validation is complete click on create button. It will take around 20-25 mins



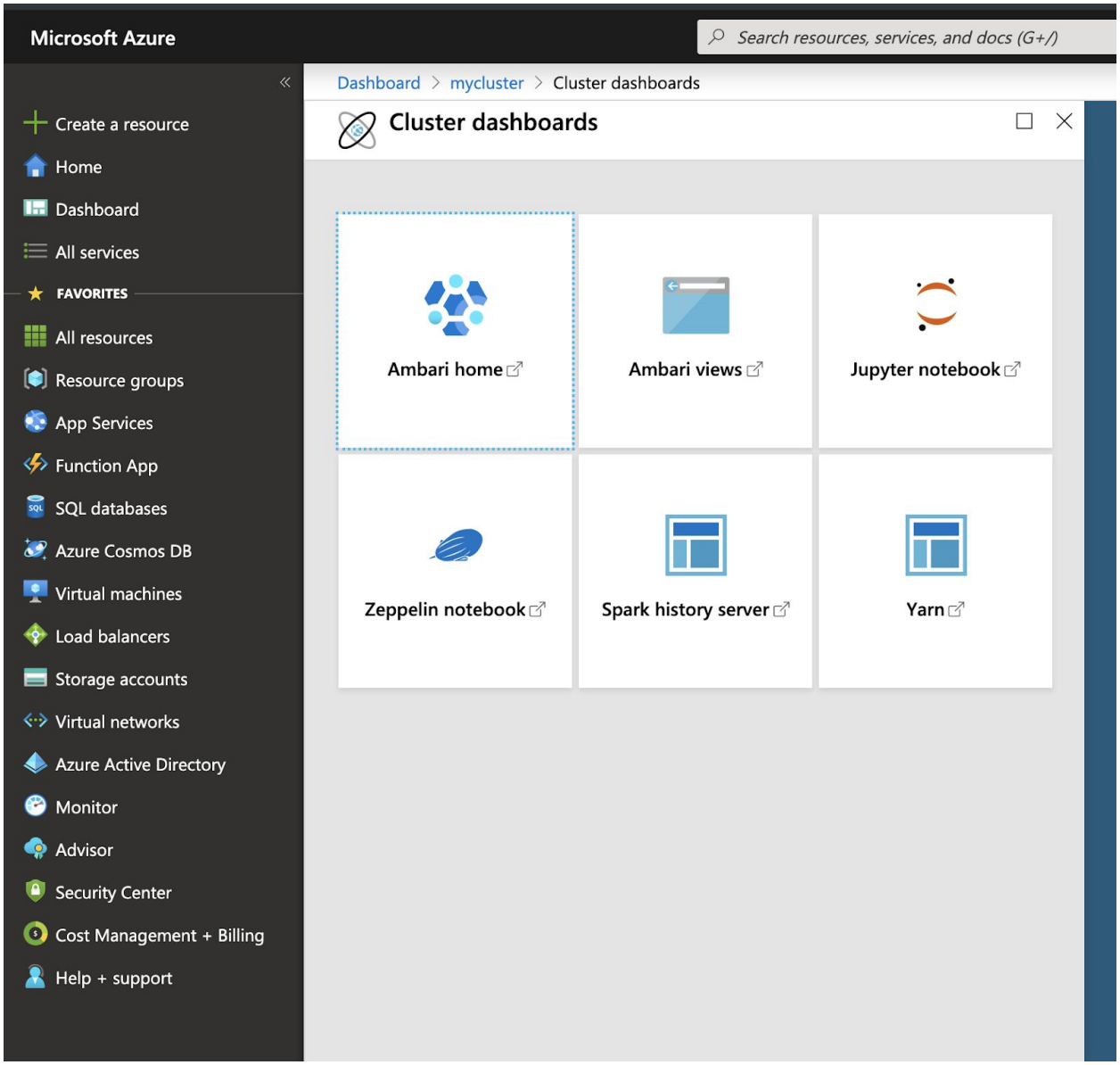
Step 10: After the cluster is created go to Azure Storage Explorer and refresh you will see your storage account or if you are using ours you will see below storage account



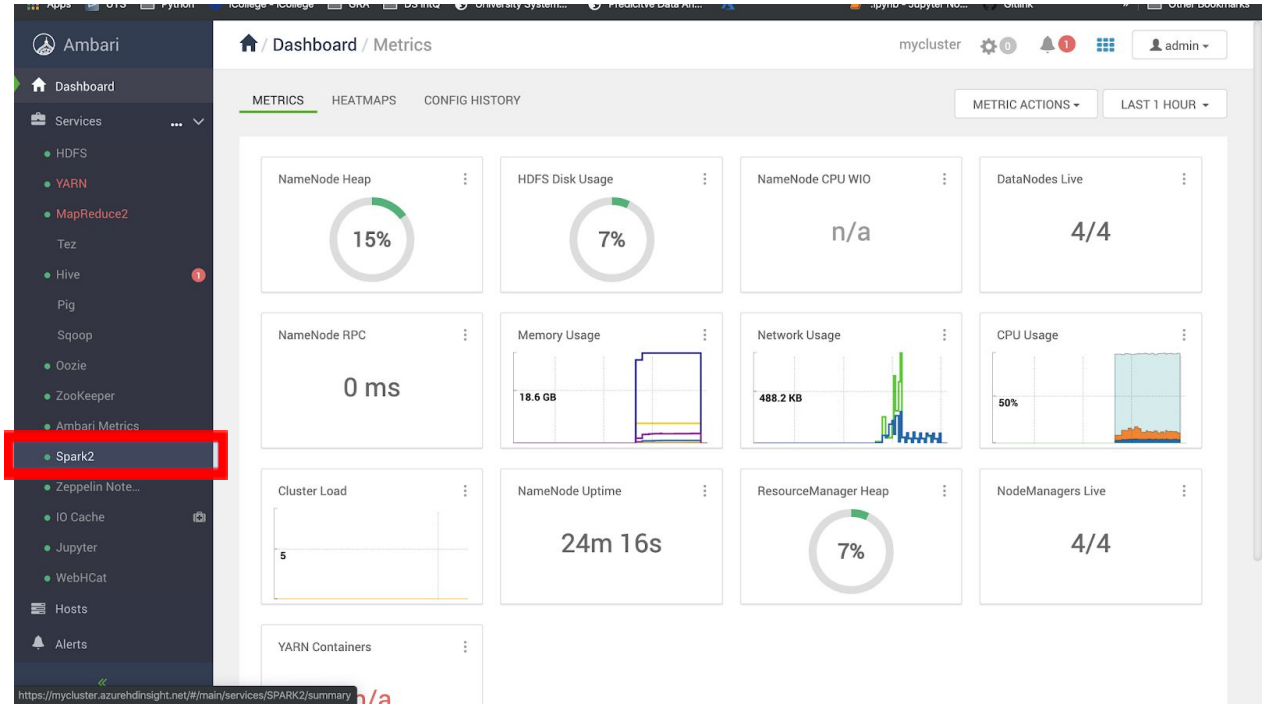
Step 11: Here you can see that you October data with the as folder 10 if you are using our storage account or you will need to upload your own data folder from your local machine. That can be simply done by creating a new folder and then by clicking on the upload button.



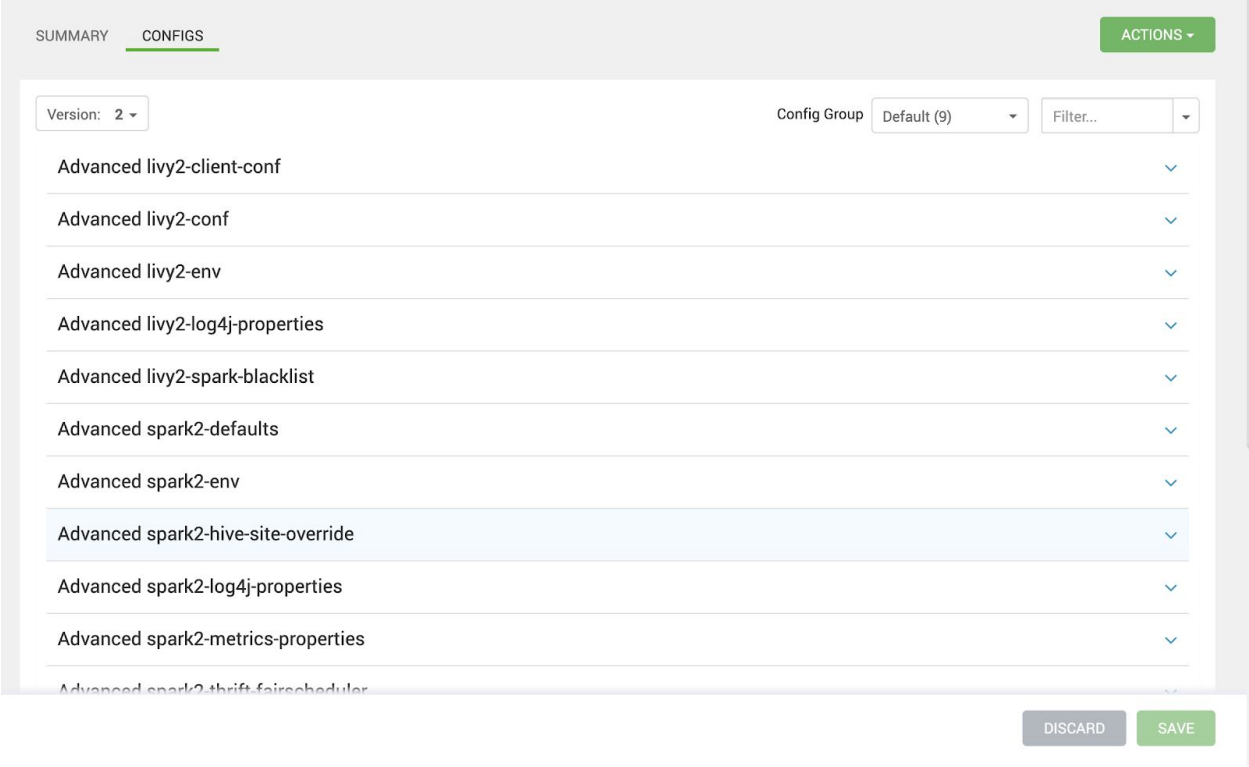
Step 12: Once the cluster is setup go-to resource and click on Ambari home



Step 13: Then go to Spark 2



STEP 14: Click on config and change the below-mentioned parameters



Step 15: You need to update 4 parameters so your kernel doesn't timeout while the code is running

Step 15. i) Update livy.server.session.timeout from 36000000 to 180000000: make sure you put right number of zeros. (we did this mistake every single time we created the clusters)

Advanced livy2-conf

livy.environment	production
livy.impersonation.enabled	true
livy.repl.enableHiveContext	true
livy.server.access-control.enabled	true
livy.server.csrf_protection.enabled	true
livy.server.port	8998
livy.server.recovery.mode	recovery
livy.server.recovery.state-store	zookeeper
livy.server.recovery.state-store.url	zk1-bdppro.2h5ffrof4nveneybdi1ajpa2qh.bx.internal.cloudapp.net:2181,zk2-bdppro.2h5ffrof
livy.server.session.timeout	18000000
livy.spark.master	yarn-cluster

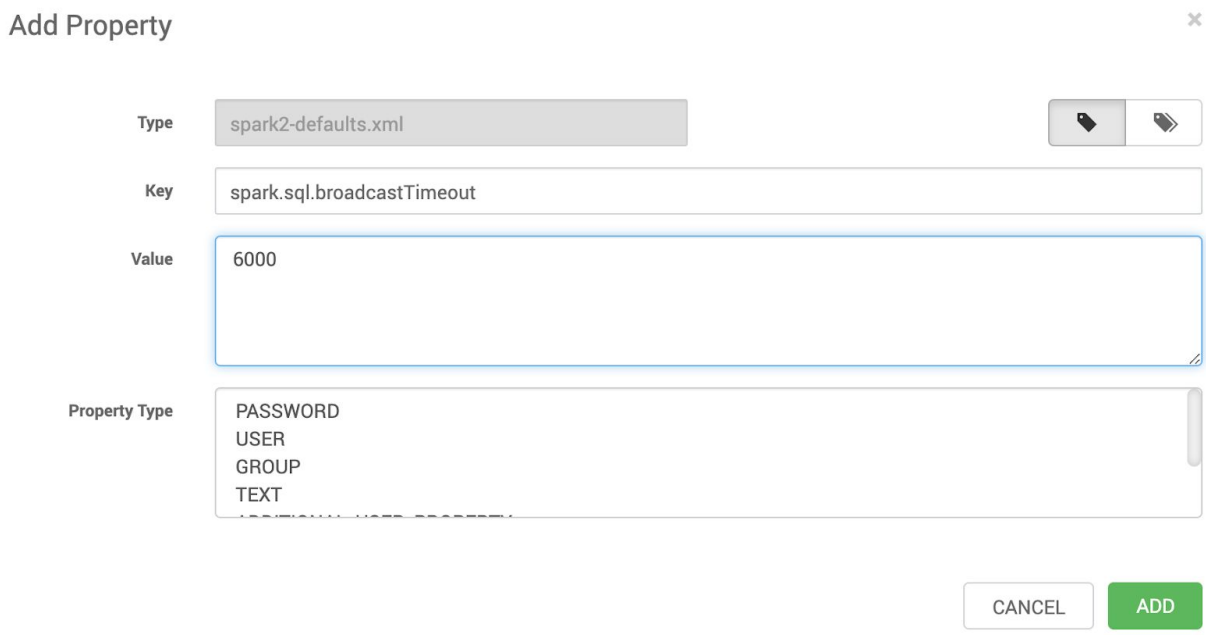
Step 15.ii) Change livy.server.yarn.app-loolup-timeout from 2m to 10m as shown below

Custom livy2-conf

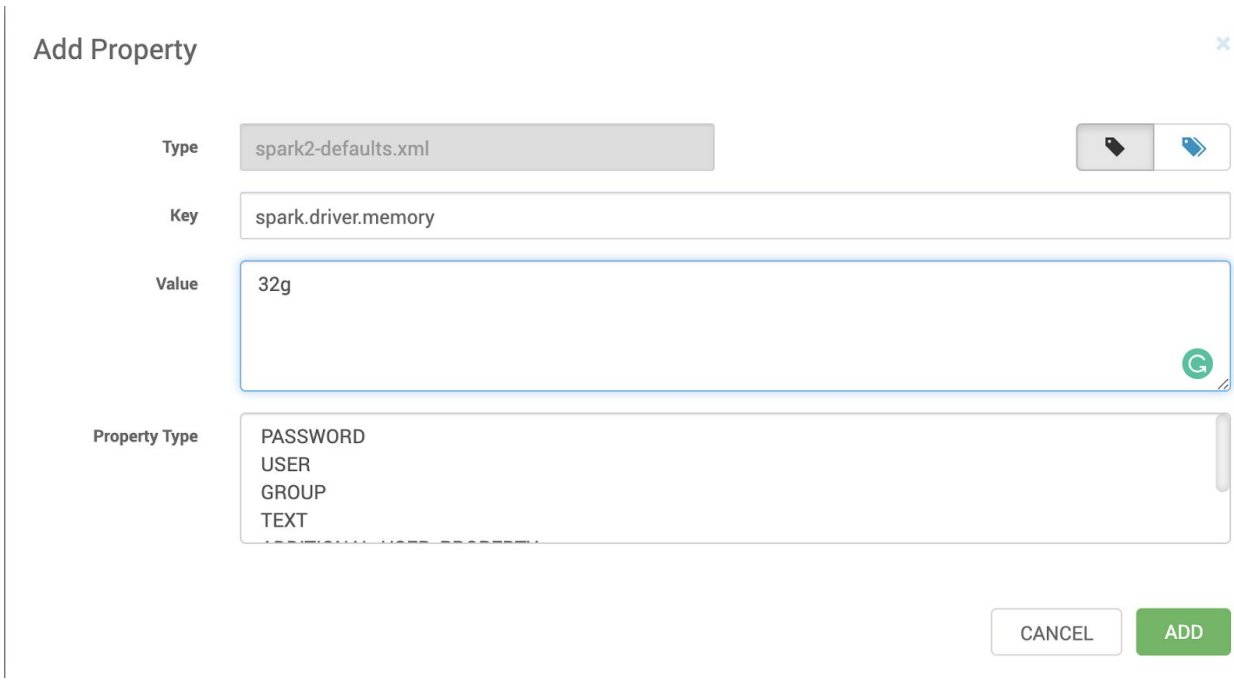
livy.server.session.state-retain.sec	3600000
livy.server.yarn.app-lookup-timeout	10m

Add Property ...

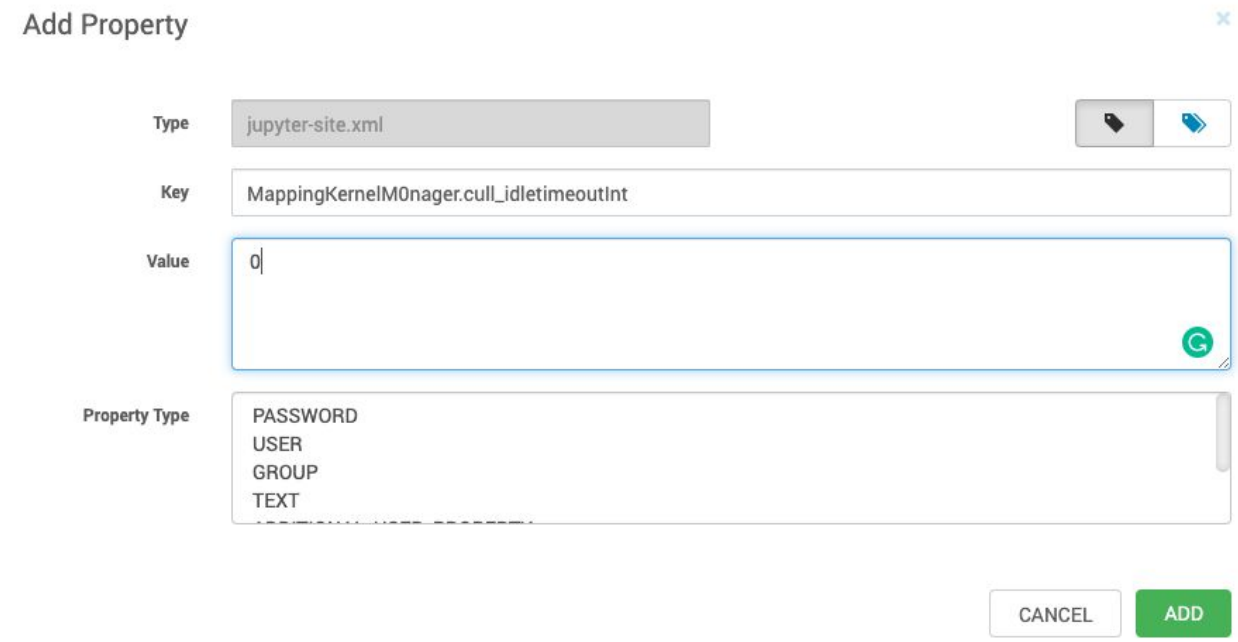
Step 15.iii) then in Custome spark2-daefaults you need to add a property

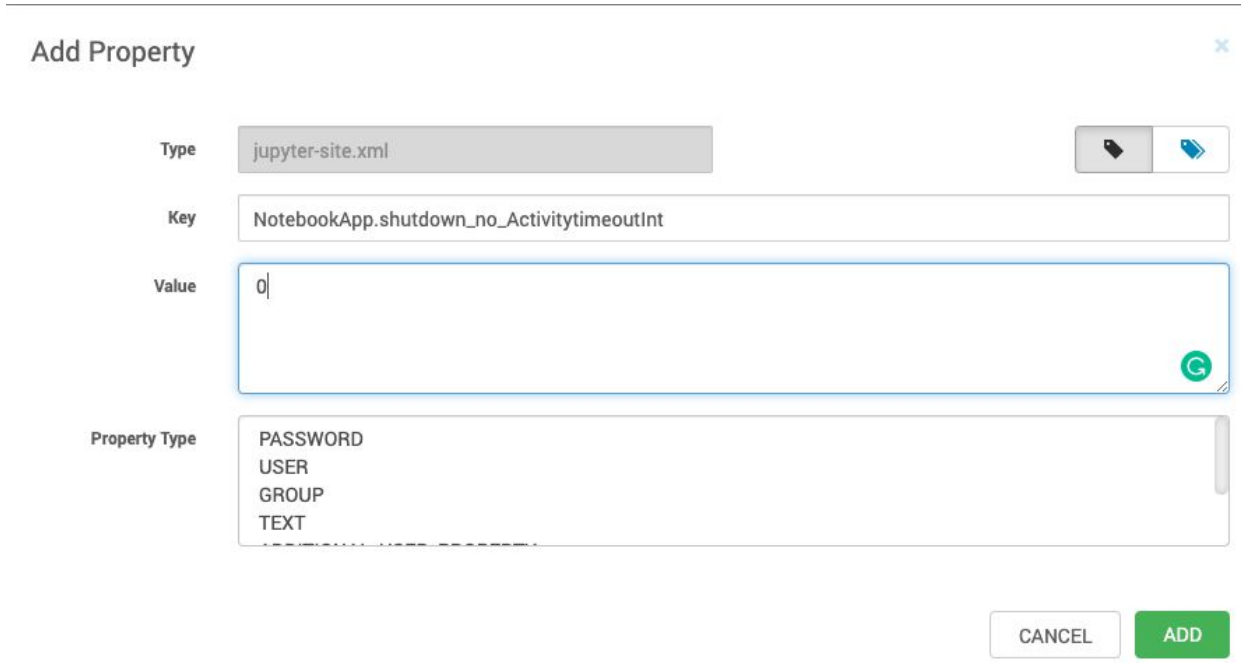


Step 15 iv) Add property spark.driver.memory as 32g



Step 16 i) Go to jupyter configuration and change below 2 parameters

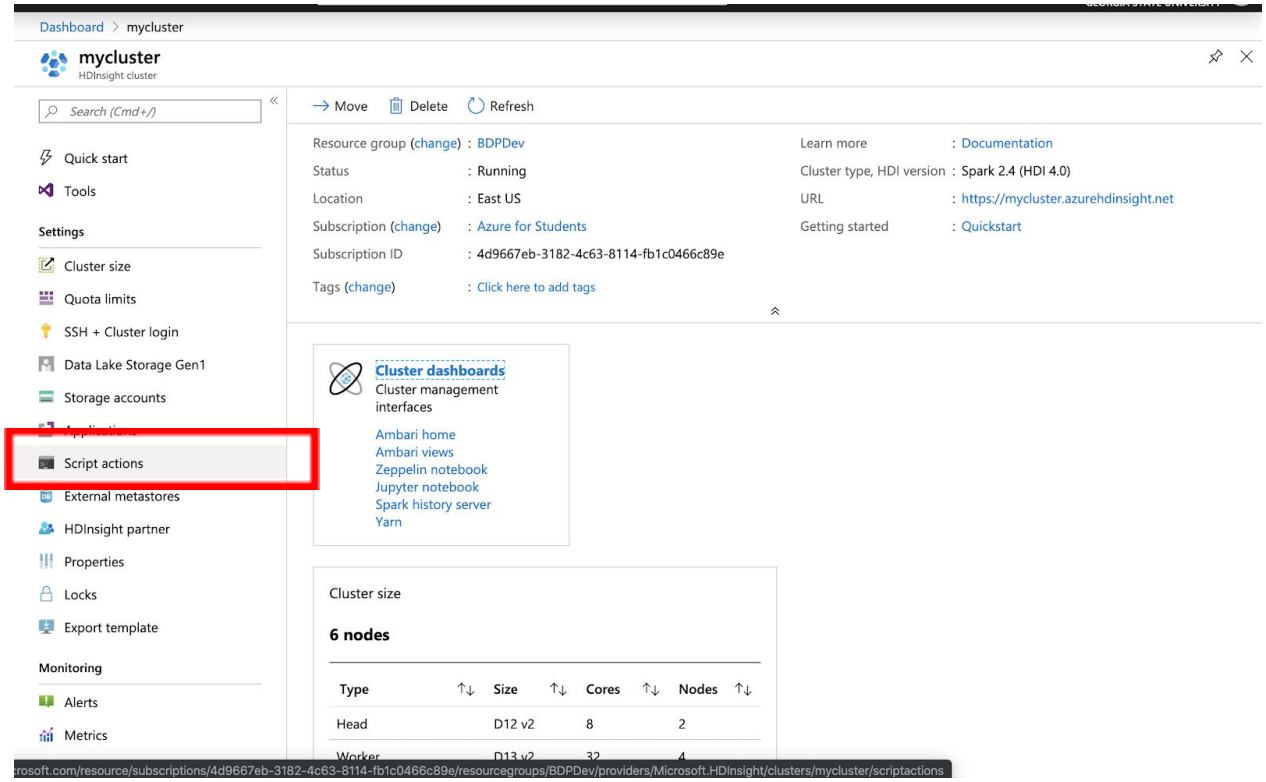




MappingKernelManager.cull_idle_timeoutInt - 0
NotebookApp.shutdown_no_activity_timeoutInt - 0

After this save and restart Spark and Jupyter both from the actions on the top right corner of the window.

Step 17) having done that add goto script actions in order to run your imports



Step 18) Then got to Submit new script option. Select script type as custom. Add the bash.sh raw file URL in `Bash script URI` or any public URL of bash file containing the commands to install the libraries.

In our case

```
/usr/bin/anaconda/envs/py35/bin/pip install azure
/usr/bin/anaconda/envs/py35/bin/pip install pandas==0.19.2
```

Then select the node types required. (head and worker in our case).
Optional: Select 'Persist this script action when new nodes are added to the cluster' option if you want to install these libraries when new worker nodes are added.

Upon clicking create. Our packages will be installed on all the nodes and we are ready to run our preprocessing code.

Step 19) Then go to Jupyter notebook under `Cluster management interfaces` in the overview section in our cluster home page. Then go ahead and run `SparkAllWords.ipynb`. This will generate the tweets and retweets files inside the Azure storage account.

PART II - Question Specific installations

Initial spark setup:

1. pip install pyspark

(we downloaded the spark-2.4.4-bin-hadoop2.7 version)

2. Install Java 8
3. In MAC OS, edit .bash_profile with following OR In Windows environment variables add:

```
export JAVA_HOME=$(/usr/libexec/java_home)
export SPARK_HOME=~/spark-2.4.4-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$PATH
export PYSPARK_PYTHON=python3
```

4. Start pyspark by running below command in terminal -
   pyspark

Question 1
---
pip install pygal

Question 2
---
pip install geopy
pip install plotly

Question 3
---
pip install nltk
nltk.download(‘stopwords’)
nltk.download('wordnet')

Question 4
---
pip install pyldavis

Question 5 - Introduction
---
pip install pygal

Question 6 - Time Series Analysis

```
pip install statsmodels
pip install datetime
pip install ipython
pip install ipywidgets
pip install strings
pip install seaborn
```
