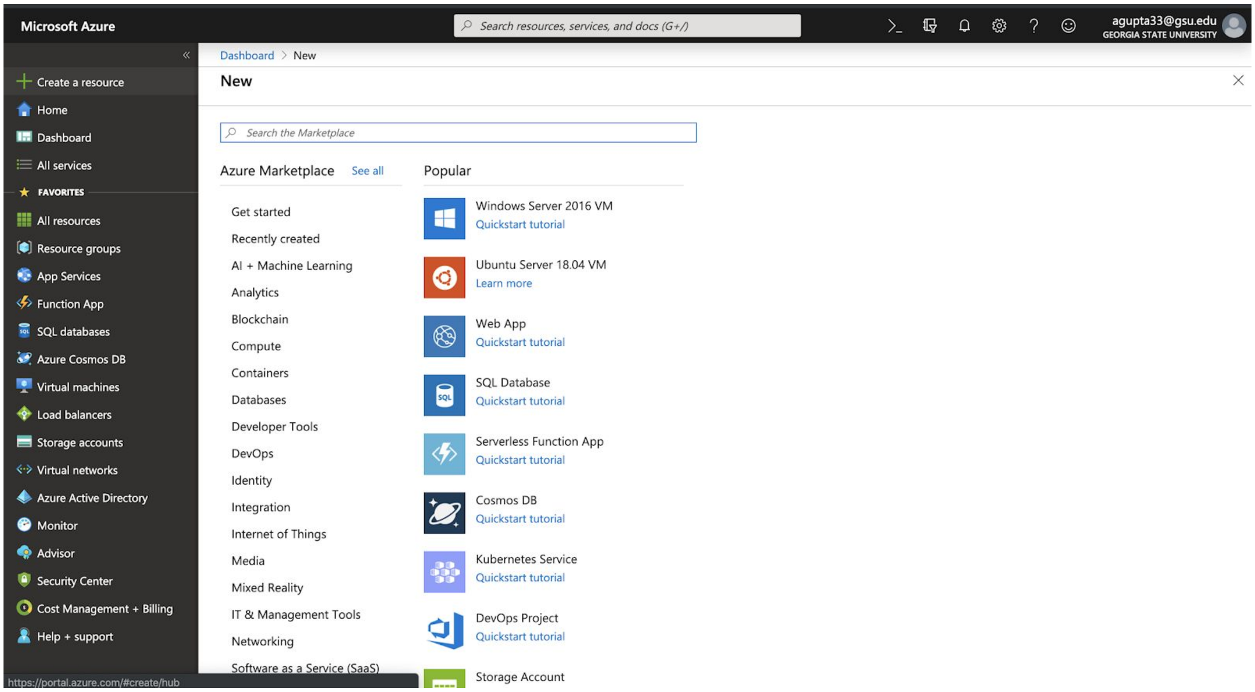
READ ME INSTRUCTIONS

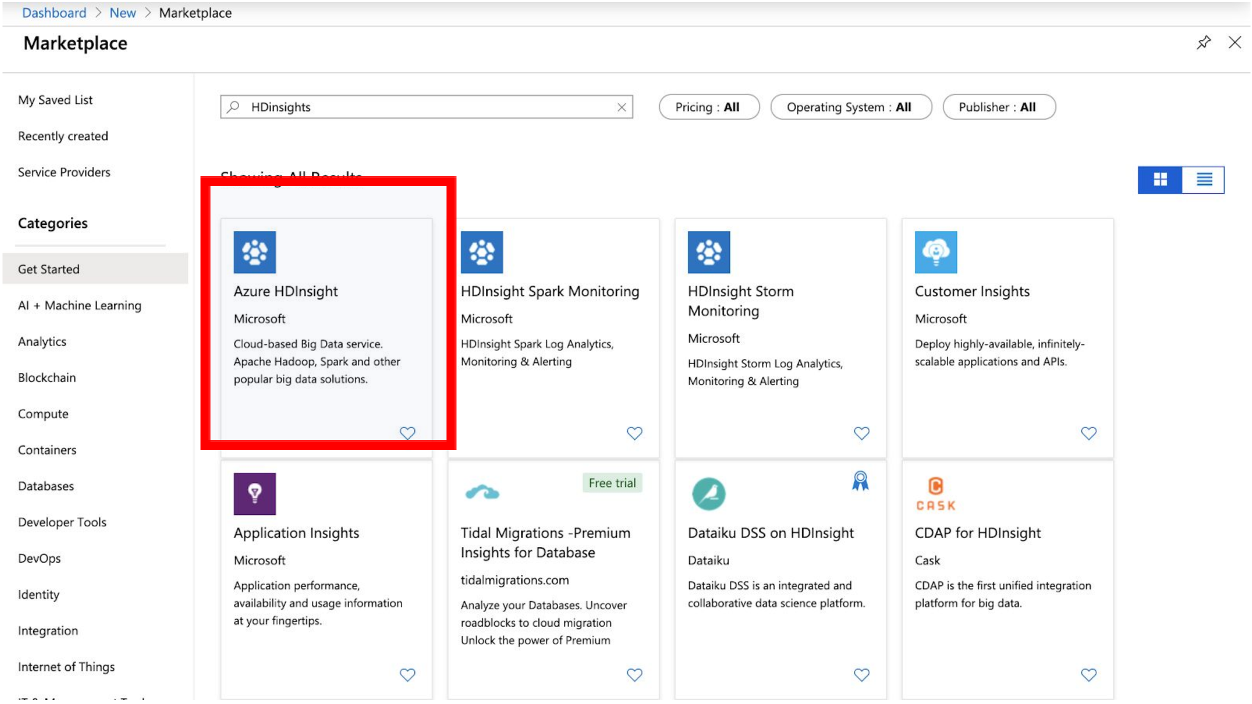
The file is divided into 2 parts.
First part explains how we **can create cluster using azure and run the preprocessing code**.
The second part deals **with installations of the important libraries** which will be required to run the questions.

Part I - CLUSTER CREATION AND PREPROCESSING

Step 1: Download Microsoft Azure Storage Explorer
https://azure.microsoft.com/en-us/features/storage-explorer/
Step 2: Create and Login with your same credentials as you have in Azure Account for HD Insights.
Step 3: login to Microsoft azure account
https://azure.microsoft.com/en-us/
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



Azure HDInsight
Microsoft
Cloud-based Big Data service. Apache Hadoop, Spark and other popular big data solutions.

HDInsight Spark Monitoring
Microsoft
HDInsight Spark Log Analytics, Monitoring & Alerting

HDInsight Storm Monitoring
Microsoft
HDInsight Storm Log Analytics, Monitoring & Alerting

Customer Insights
Microsoft
Deploy highly-available, infinitely-scalable applications and APIs.

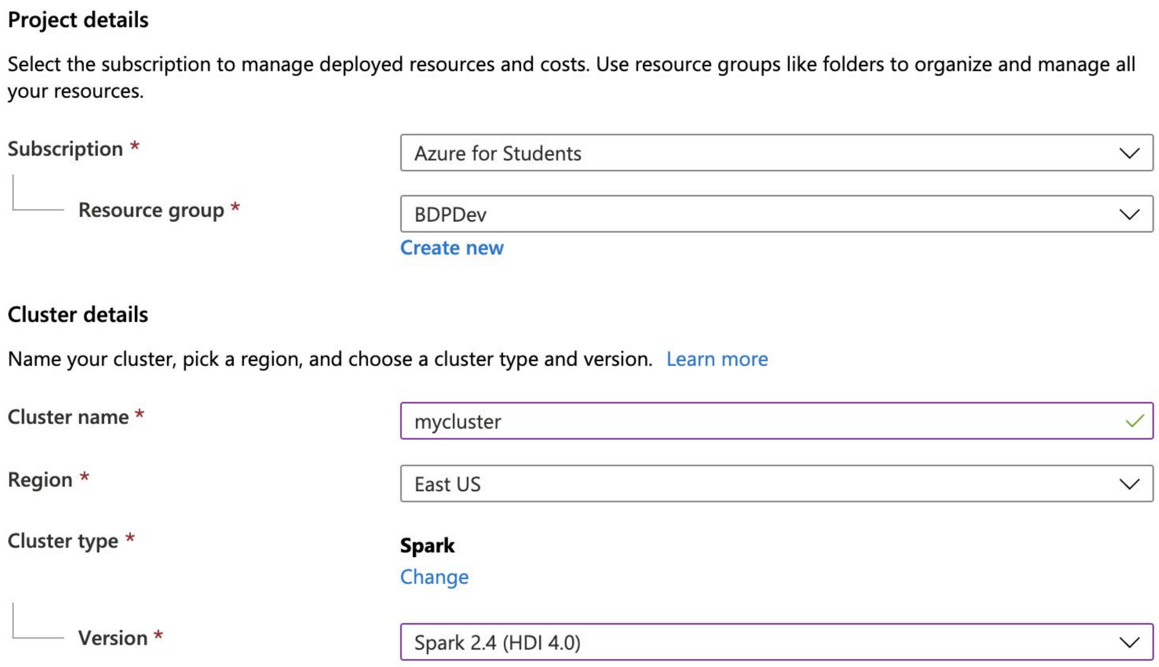
Application Insights
Microsoft
Application performance, availability and usage information at your fingertips.

Tidal Migrations -Premium Insights for Database
tidalmigrations.com
Analyze your Databases. Uncover roadblocks to cloud migration Unlock the power of Premium

Dataiku DSS on HDInsight
Dataiku
Dataiku DSS is an integrated and collaborative data science platform.

CDAP for HDInsight
Cask
CDAP is the first unified integration platform for big data.

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. Please keep the details as same as possible.



Project details

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

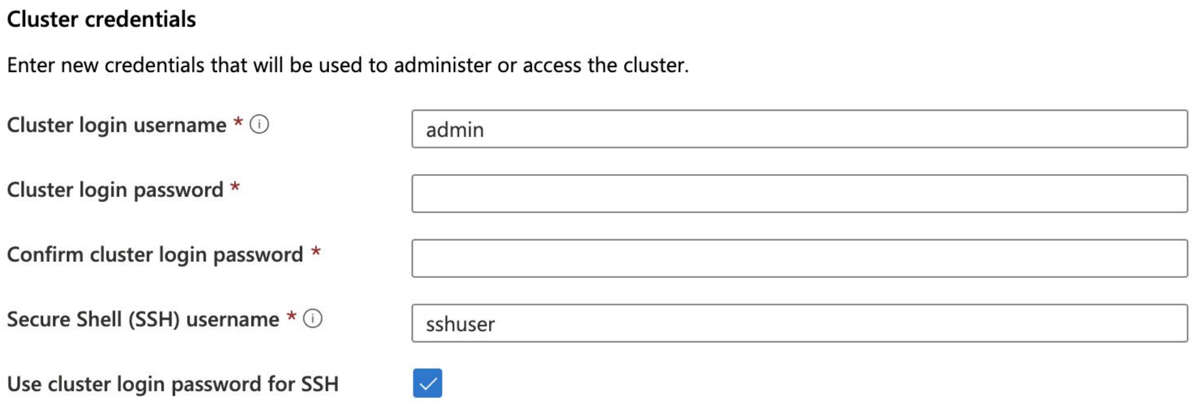
Subscription *	Azure for Students
Resource group *	BDPDev Create new

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. Learn more

Cluster name *	mycluster ✓
Region *	East US
Cluster type *	**Spark** Change
Version *	Spark 2.4 (HDI 4.0)

Step7: Fill in **cluster credentials .** Give the password and do remember the password.



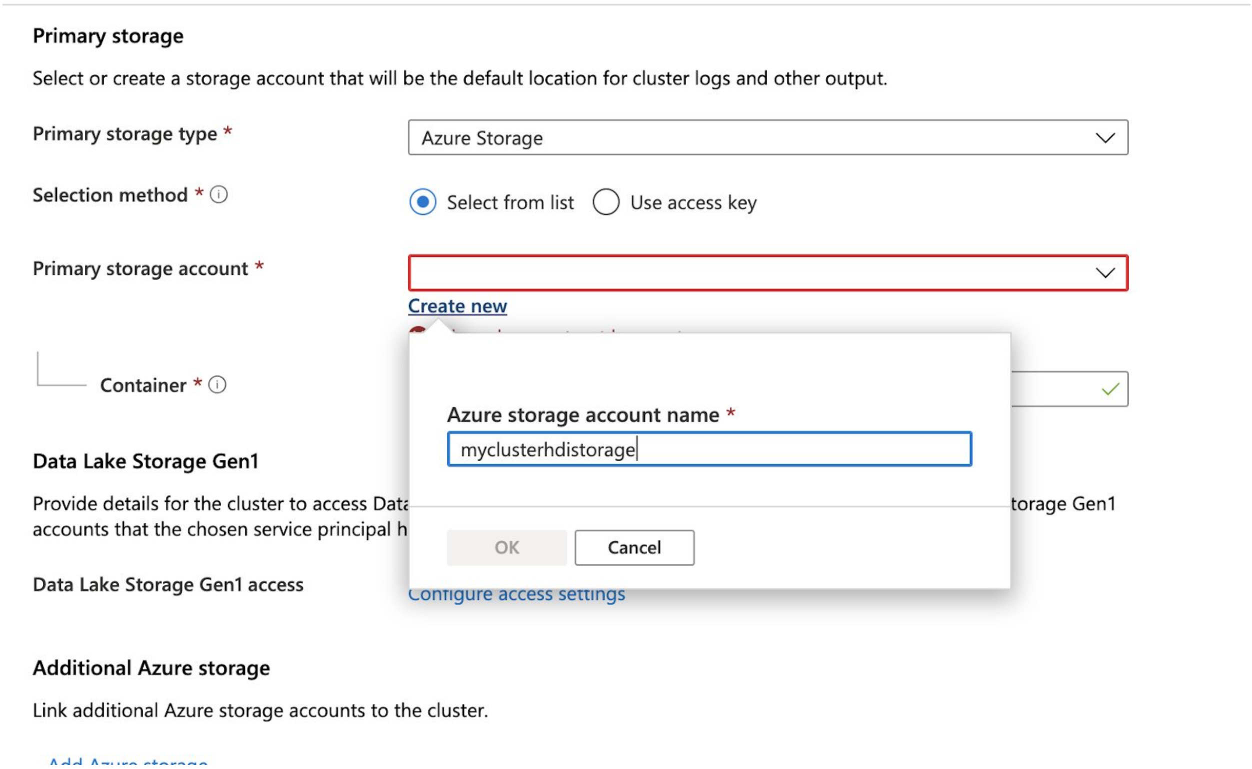
Cluster credentials

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: Once you are on the Storage tab do the following

8.i) click on create new button to create a new storage.



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 *

Container *

Azure storage account name * myclusterhdistorage

Data Lake Storage Gen1

Provide details for the cluster to access Dat… …torage Gen1 accounts that the chosen service principal h…

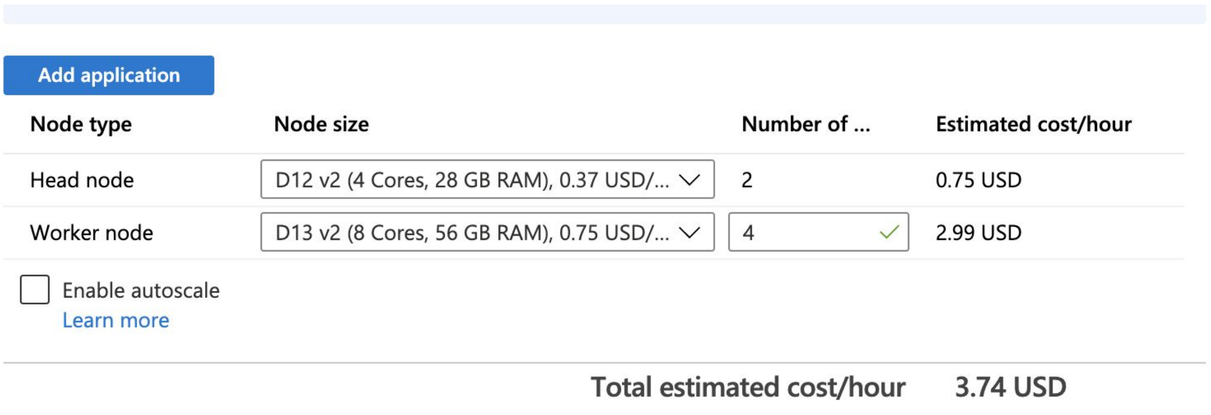
Data Lake Storage Gen1 access

Additional Azure storage

Link additional Azure storage accounts to the cluster.

8.ii) write the **container name.** and move ahead also copy the container name to some where safe as you will need it later.
Click next, There is no need to change anything in Security + networking

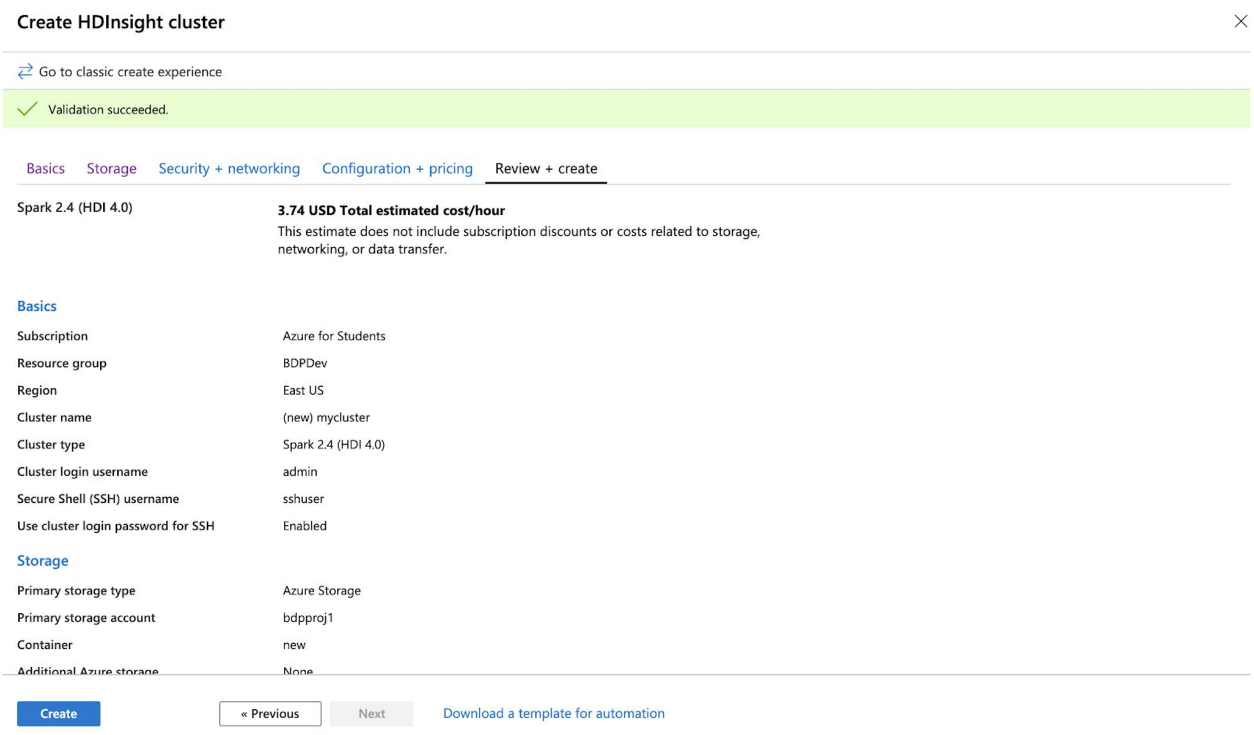
Step 9: Click next then on Configuration + pricing select the configuration of the head node and worker node. And then click next.



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



Create HDInsight cluster

✓ Validation succeeded.

Spark 2.4 (HDI 4.0)

3.74 USD Total estimated cost/hour
This estimate does not include subscription discounts or costs related to storage, networking, or data transfer.

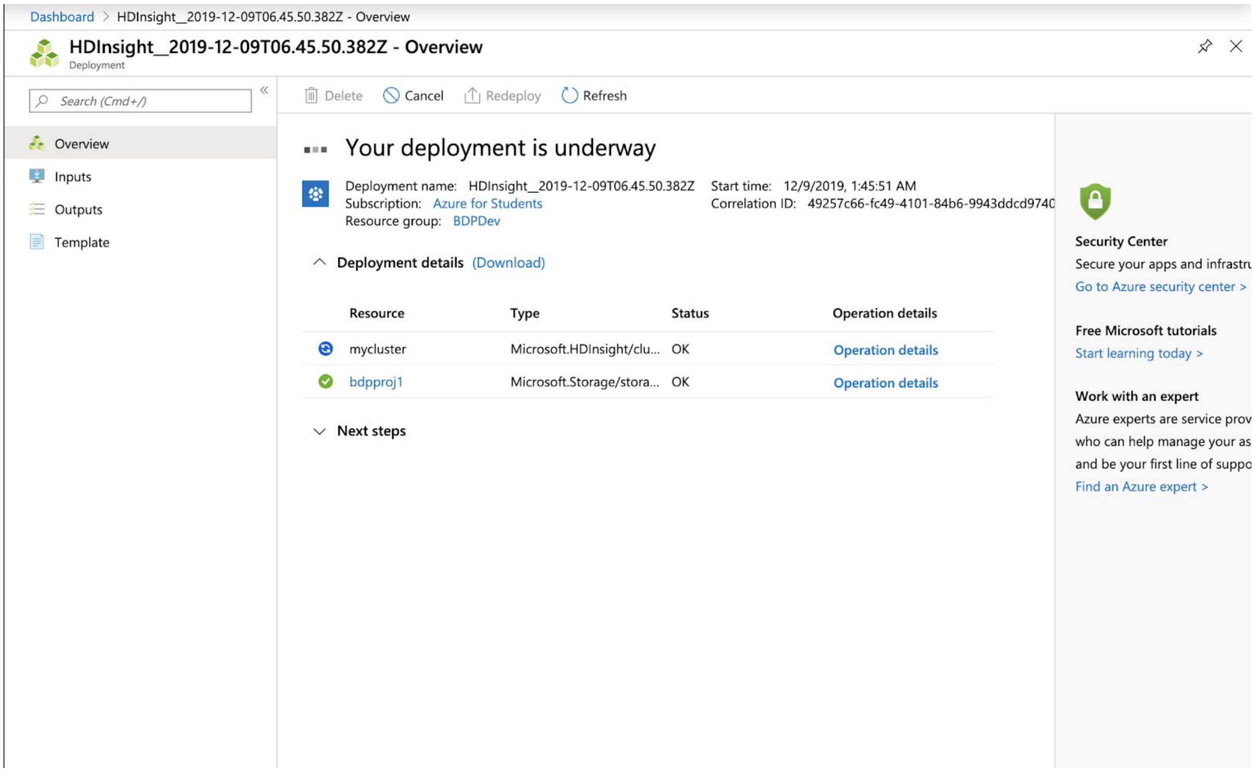
Basics

Subscription	Azure for Students
Resource group	BDPDev
Region	East US
Cluster name	(new) mycluster
Cluster type	Spark 2.4 (HDI 4.0)
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	bdpproj1
Container	new
Additional Azure storage	None

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



Dashboard > HDInsight__2019-12-09T06.45.50.382Z - Overview
HDInsight__2019-12-09T06.45.50.382Z - Overview
Deployment
Search (Cmd+/)
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Refresh

Your deployment is underway

Deployment name: HDInsight__2019-12-09T06.45.50.382Z
Subscription: Azure for Students
Resource group: BDPDev
Start time: 12/9/2019, 1:45:51 AM
Correlation ID: 49257c66-fc49-4101-84b6-9943ddcd9740

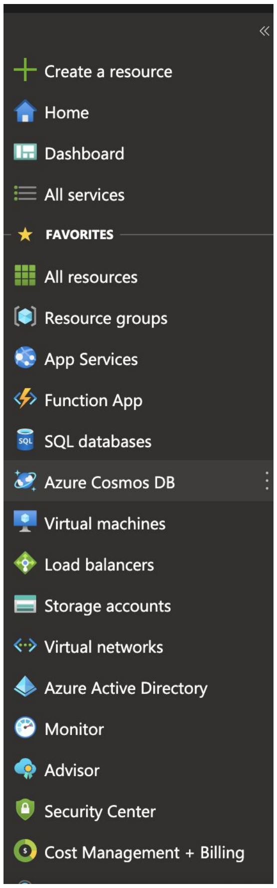
Deployment details (Download)

Resource	Type	Status	Operation details
mycluster	Microsoft.HDInsight/clu...	OK	Operation details
bdpproj1	Microsoft.Storage/stora...	OK	Operation details

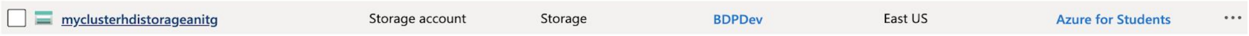
Next steps

Security Center
Secure your apps and infrastru
Go to Azure security center >
Free Microsoft tutorials
Start learning today >
Work with an expert
Azure experts are service prov
who can help manage your as
and be your first line of suppo
Find an Azure expert >

To get your storage account and access key details do the following :

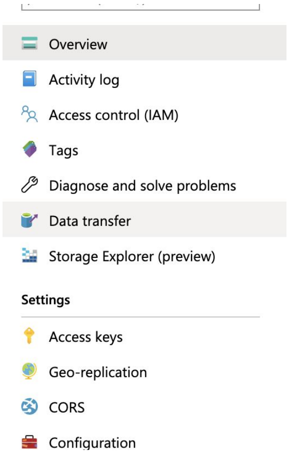


Click on the storage account name on this page

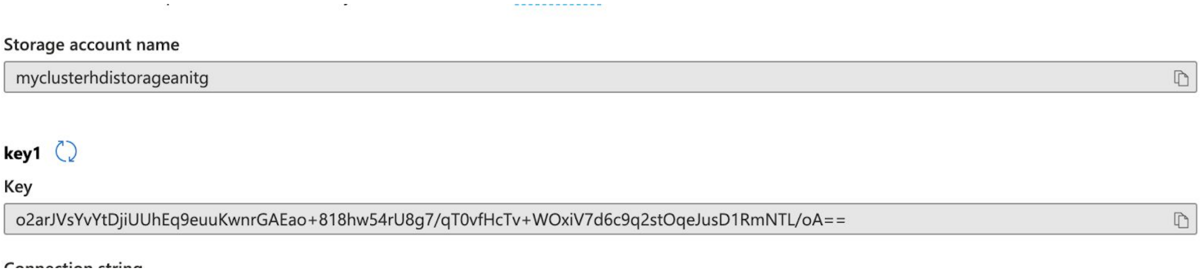


□ mycluserhdistorageanitg	Storage account	Storage	BDPDev	East US	Azure for Students	···

Then click on Access Keys on the left side as shown below.



You need to save **Account Name, Key** and the container name you have given.



Storage account name

myclusterhdistorageanitg

key1

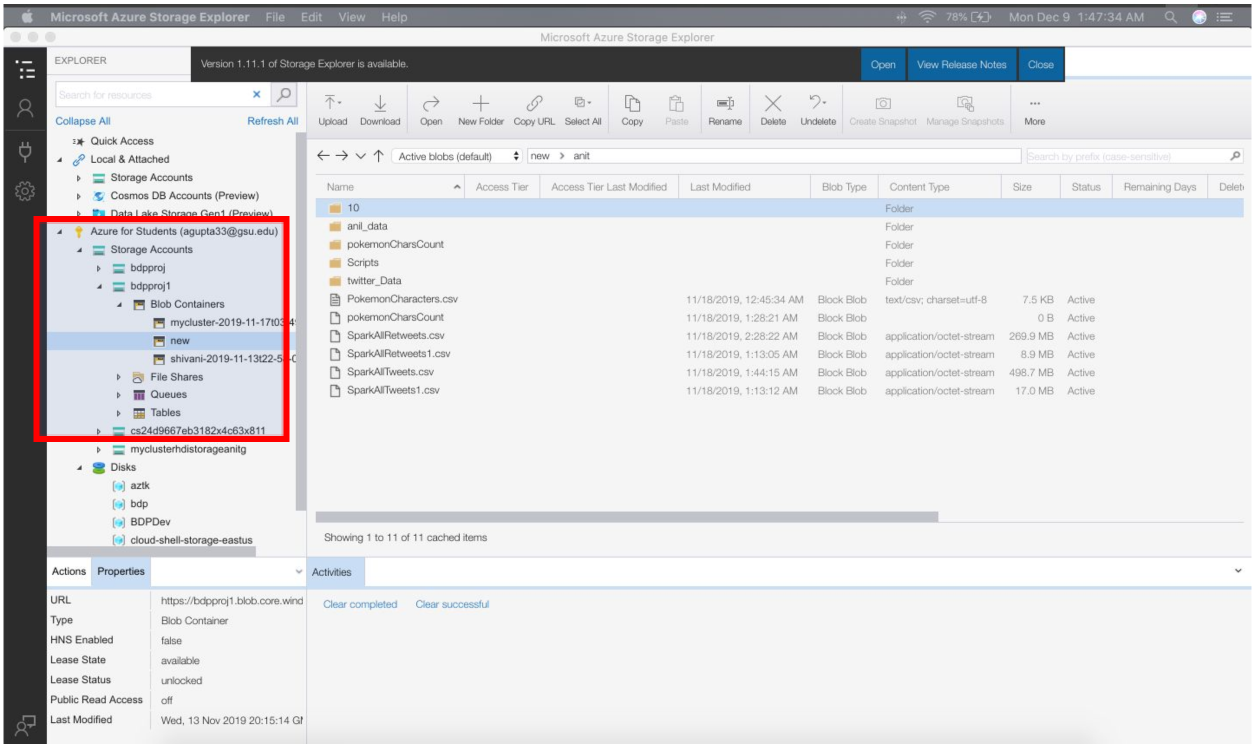
Key

o2arJVsYvYtDjiUUhEq9euuKwnrGAEao+818hw54rU8g7/qT0vfHcTv+WOxiV7d6c9q2stOqeJusD1RmNTL/oA==

Connection string

AZURE STORAGE EXPLORER

Step 10: After the cluster is created go to Azure Storage Explorer and refresh you will see your storage account as below :



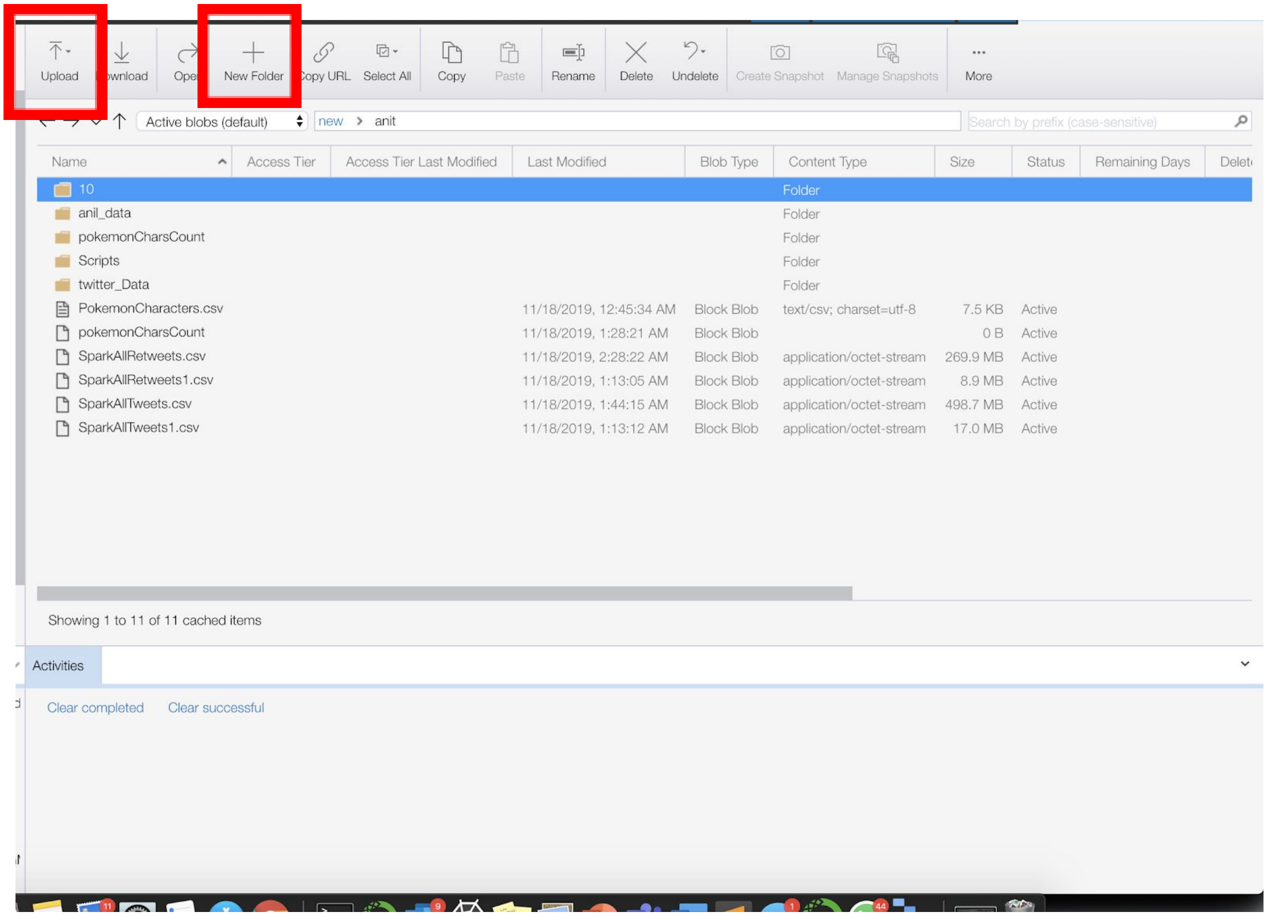
Version 1.11.1 of Storage Explorer is available.

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Status	Remaining Days	Delet
10					Folder				
anil_data					Folder				
pokemonCharsCount					Folder				
Scripts					Folder				
twitter_Data					Folder				
PokemonCharacters.csv			11/18/2019, 12:45:34 AM	Block Blob	text/csv; charset=utf-8	7.5 KB	Active		
pokemonCharsCount			11/18/2019, 1:28:21 AM	Block Blob		0 B	Active		
SparkAllRetweets.csv			11/18/2019, 2:28:22 AM	Block Blob	application/octet-stream	269.9 MB	Active		
SparkAllRetweets1.csv			11/18/2019, 1:13:05 AM	Block Blob	application/octet-stream	8.9 MB	Active		
SparkAllTweets.csv			11/18/2019, 1:44:15 AM	Block Blob	application/octet-stream	498.7 MB	Active		
SparkAllTweets1.csv			11/18/2019, 1:13:12 AM	Block Blob	application/octet-stream	17.0 MB	Active		

Showing 1 to 11 of 11 cached items

URL	https://bdpproj1.blob.core.wind
Type	Blob Container
HNS Enabled	false
Lease State	available
Lease Status	unlocked
Public Read Access	off
Last Modified	Wed, 13 Nov 2019 20:15:14 GI

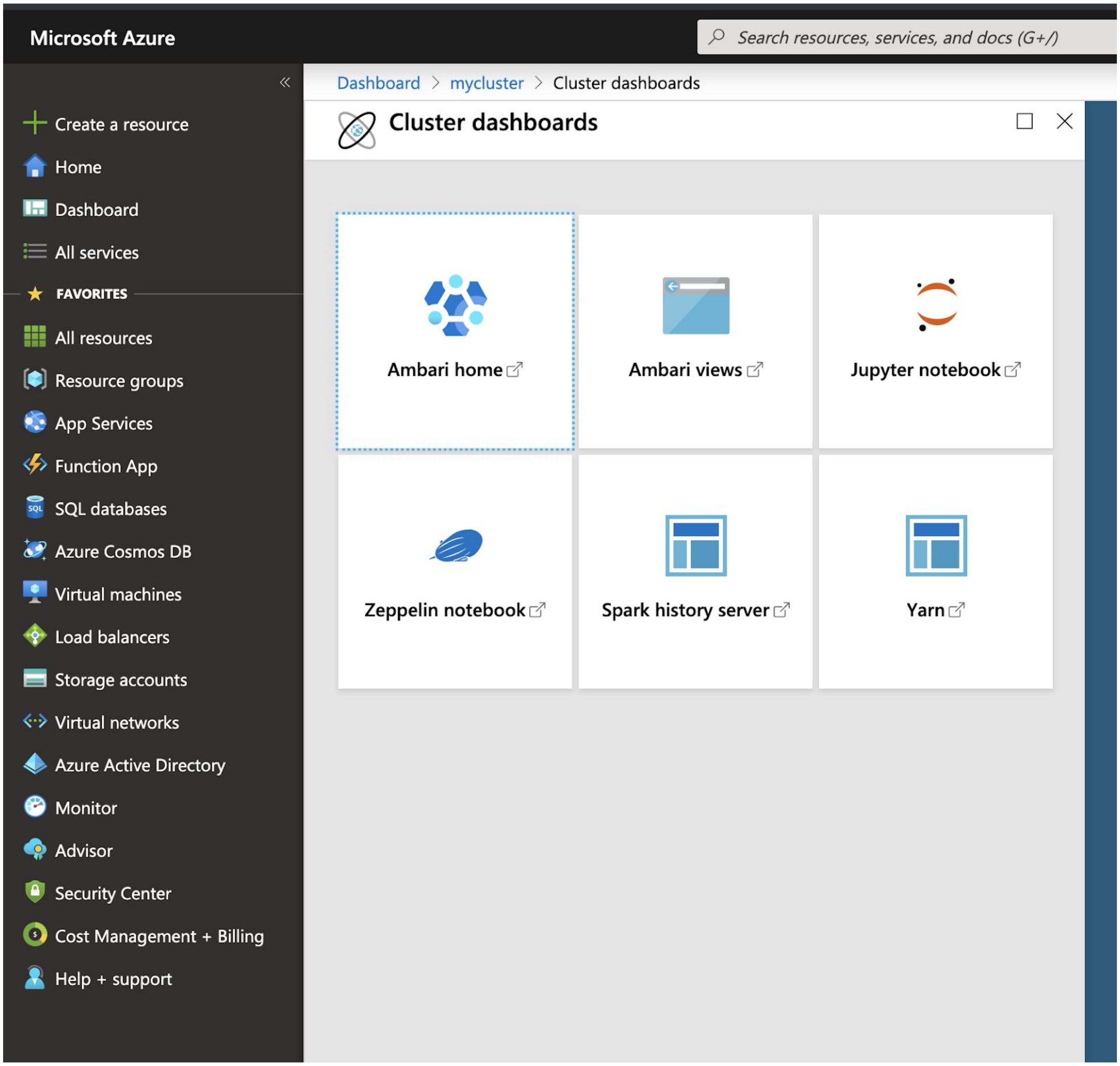
Step 11: Then in your container first create a folder and then upload the **october months data**. from your local machine to the storage and upload **bash.sh** from the zipped folder



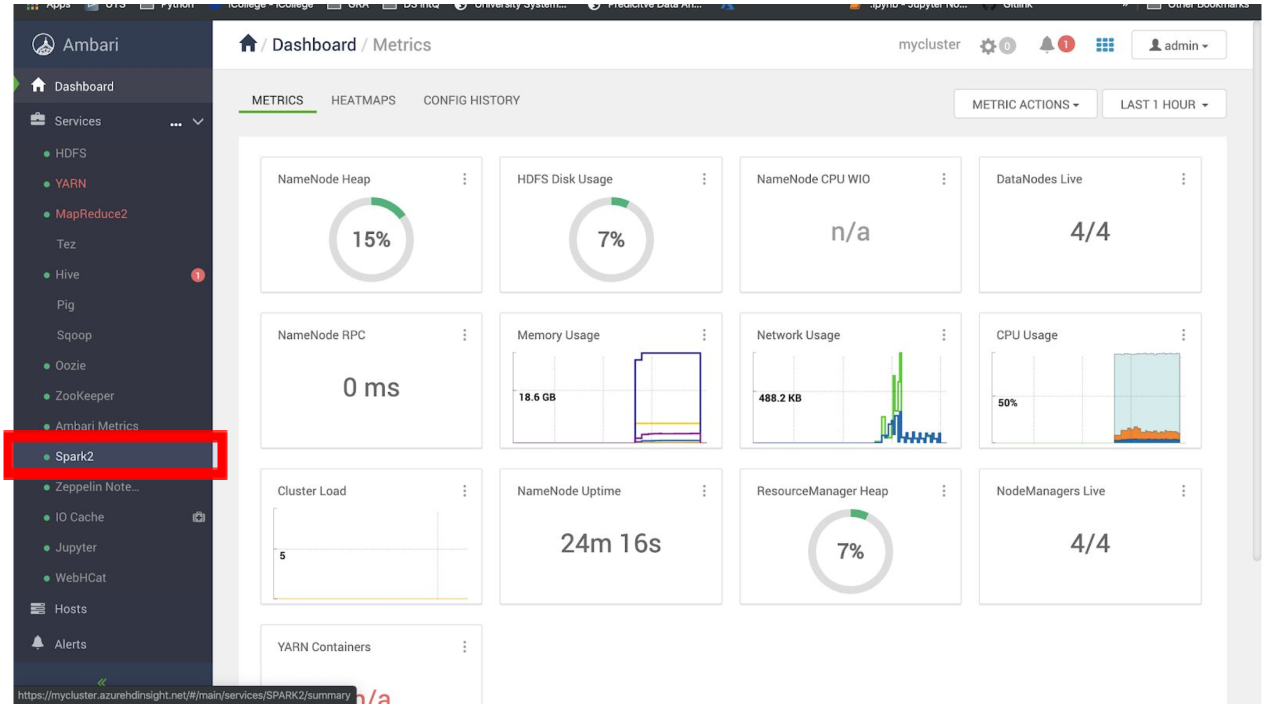
Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Status	Remaining Days	Delet
10					Folder				
anil_data					Folder				
pokemonCharsCount					Folder				
Scripts					Folder				
twitter_Data					Folder				
PokemonCharacters.csv			11/18/2019, 12:45:34 AM	Block Blob	text/csv; charset=utf-8	7.5 KB	Active		
pokemonCharsCount			11/18/2019, 1:28:21 AM	Block Blob		0 B	Active		
SparkAllRetweets.csv			11/18/2019, 2:28:22 AM	Block Blob	application/octet-stream	269.9 MB	Active		
SparkAllRetweets1.csv			11/18/2019, 1:13:05 AM	Block Blob	application/octet-stream	8.9 MB	Active		
SparkAllTweets.csv			11/18/2019, 1:44:15 AM	Block Blob	application/octet-stream	498.7 MB	Active		
SparkAllTweets1.csv			11/18/2019, 1:13:12 AM	Block Blob	application/octet-stream	17.0 MB	Active		

CLUSTER DASHBOARD

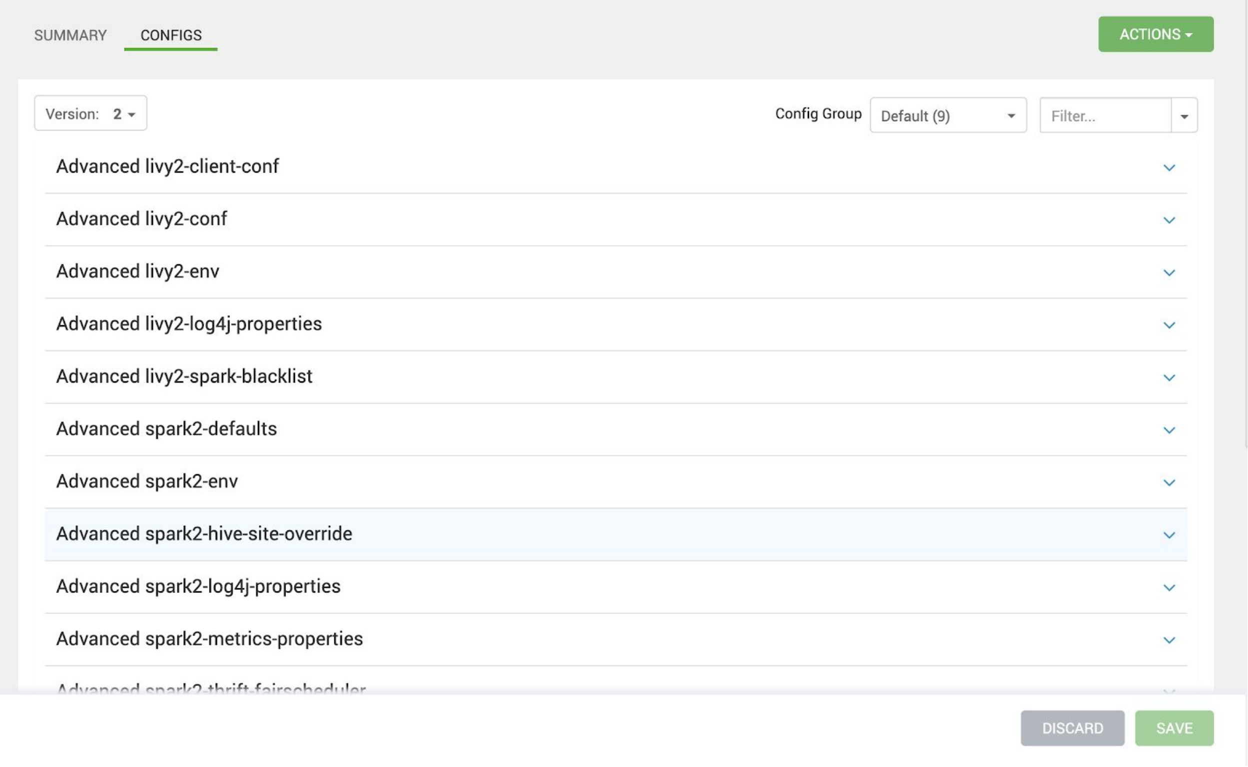
Step 12: Once the cluster is setup go-to resource and click on Ambari home



Step 13: Then go to Spark 2

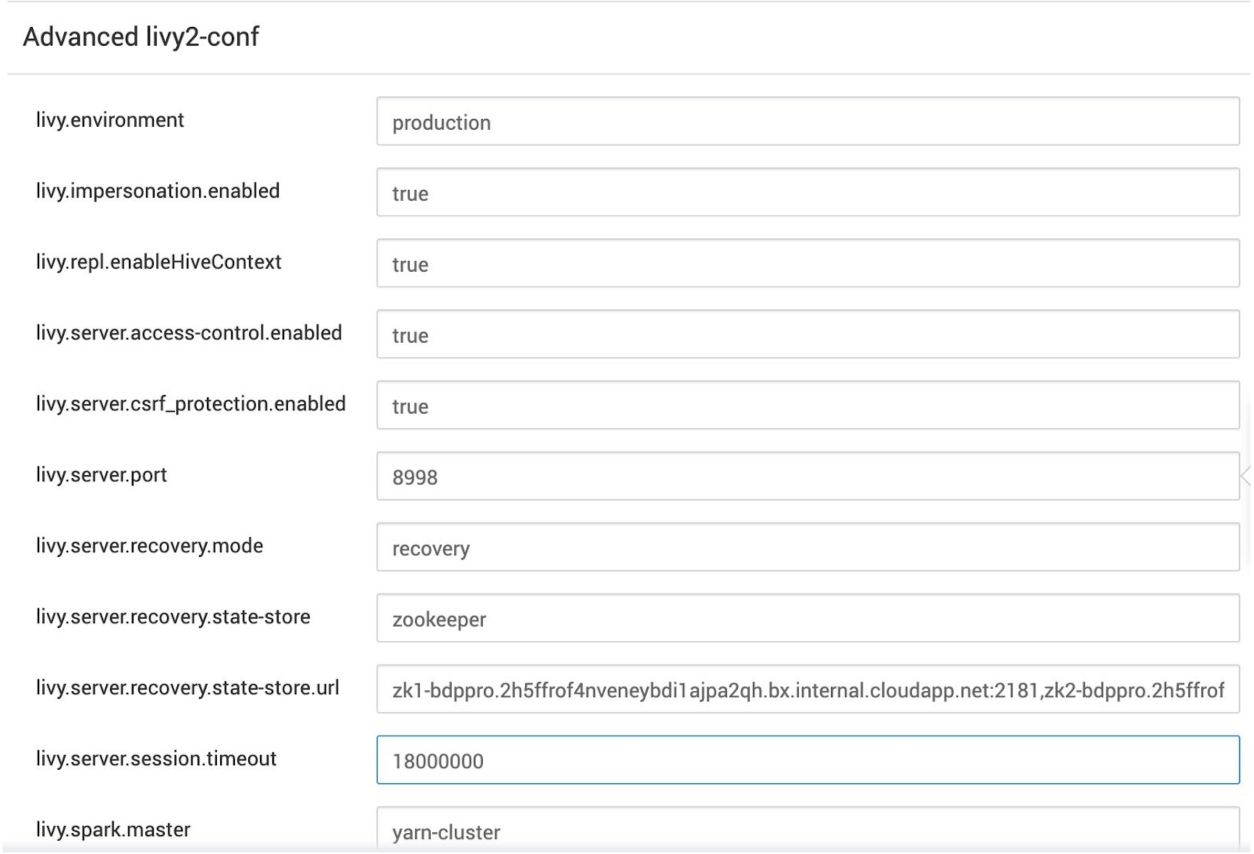


STEP 14: Click on config and change the below-mentioned parameters


SUMMARY
CONFIGS
ACTIONS
Version: 2
Config Group
Default (9)
Filter...
Advanced livy2-client-conf
Advanced livy2-conf
Advanced livy2-env
Advanced livy2-log4j-properties
Advanced livy2-spark-blacklist
Advanced spark2-defaults
Advanced spark2-env
Advanced spark2-hive-site-override
Advanced spark2-log4j-properties
Advanced spark2-metrics-properties
Advanced spark2-thrift-fairscheduler
DISCARD
SAVE

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

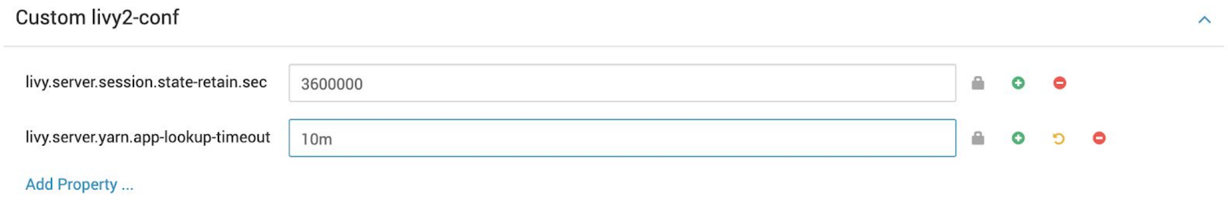
Step 15. i) Update livy.server.session.timeout from 36000000 to 180000000: make sure you put right number of zeros. (we did this mistake every single time we created the clusters)



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

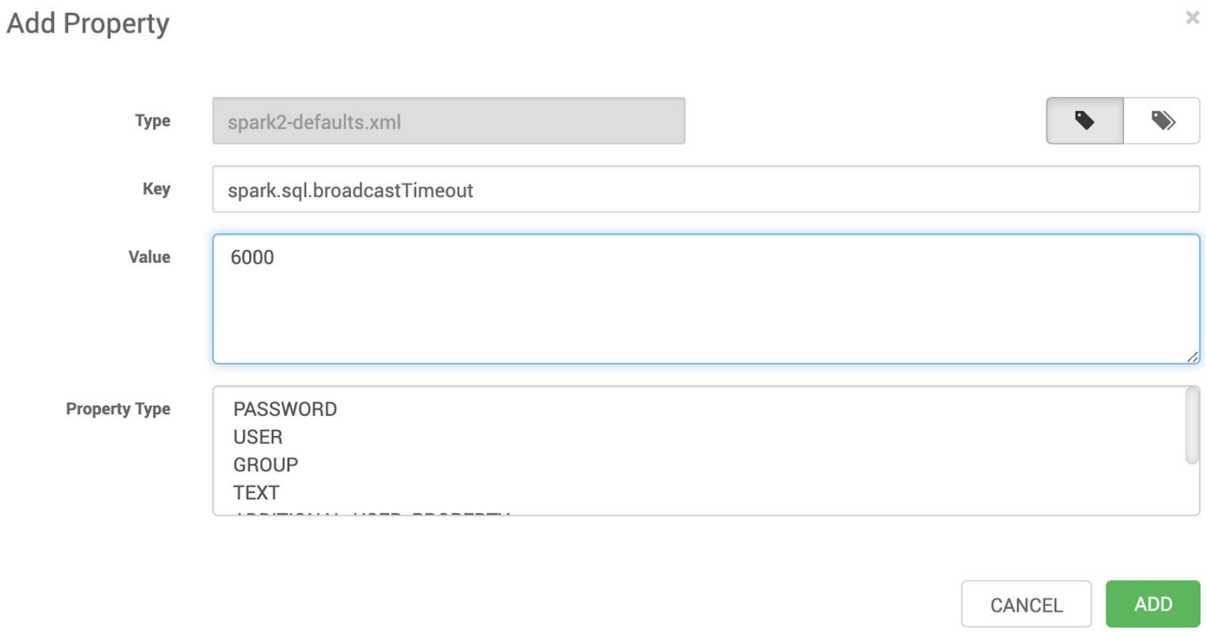
Step 15.ii) Change livy.server.yarn.app-loolup-timeout from 2m to 10m as shown below



Custom livy2-conf

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

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



Add Property

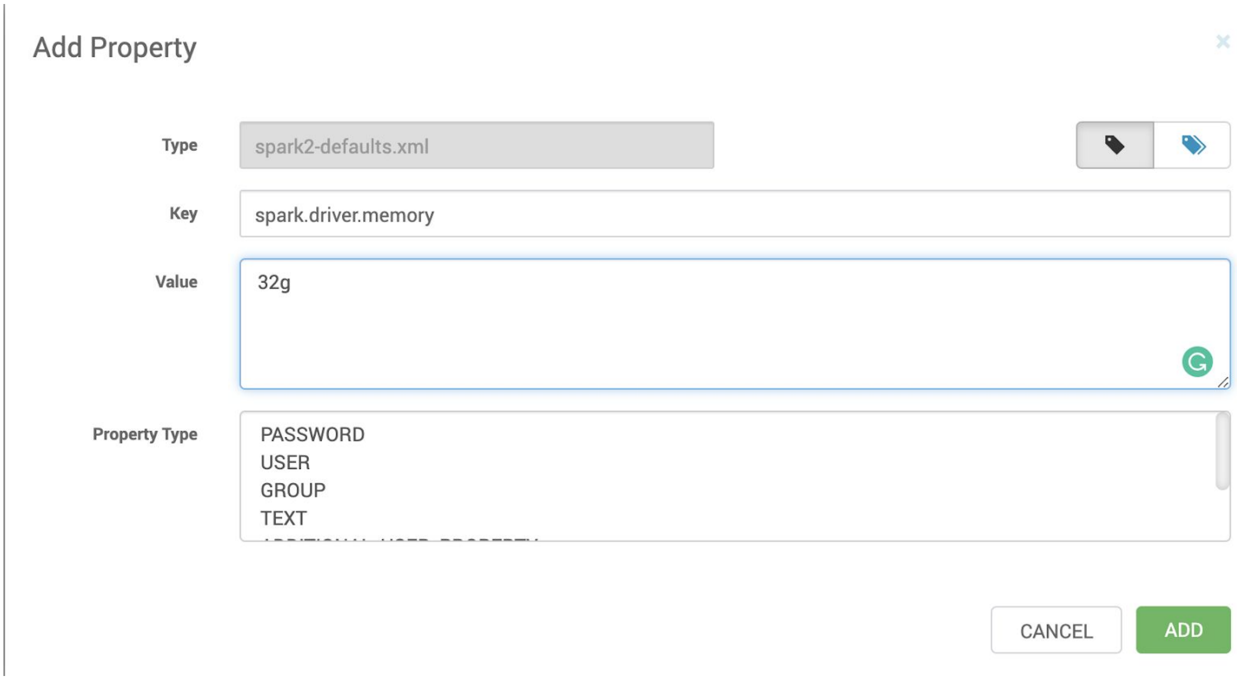
Type: spark2-defaults.xml

Key: spark.sql.broadcastTimeout

Value: 6000

Property Type: PASSWORD, USER, GROUP, TEXT

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



Add Property

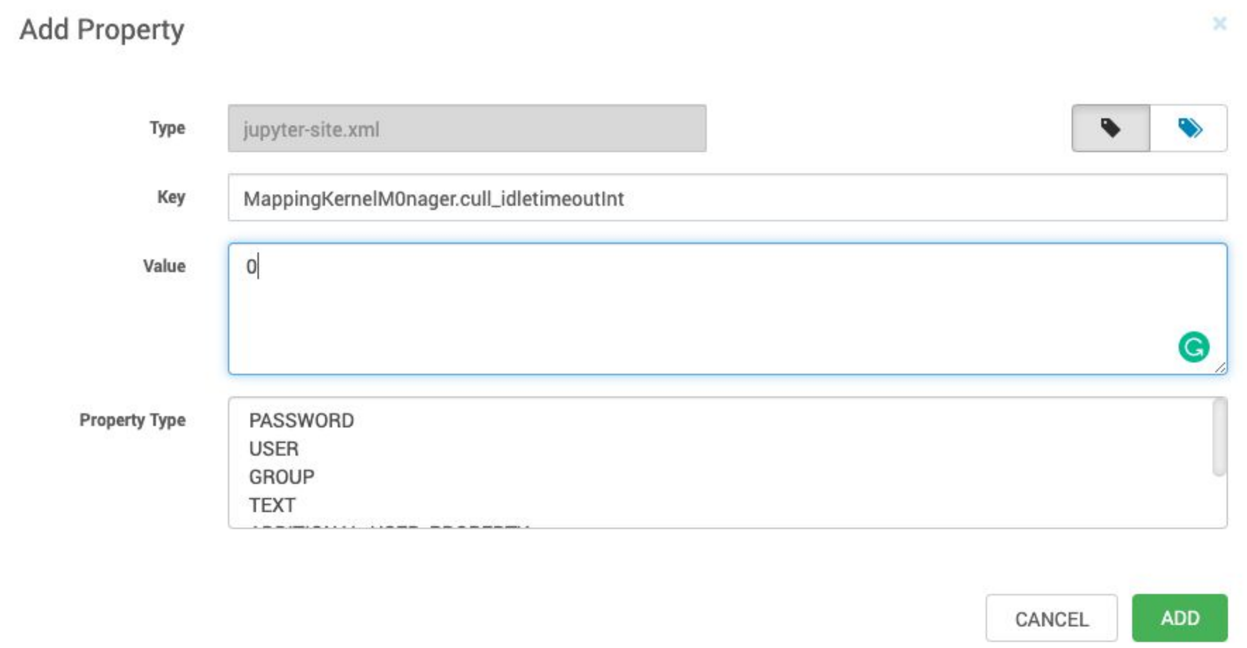
Type: spark2-defaults.xml

Key: spark.driver.memory

Value: 32g

Property Type: PASSWORD, USER, GROUP, TEXT

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

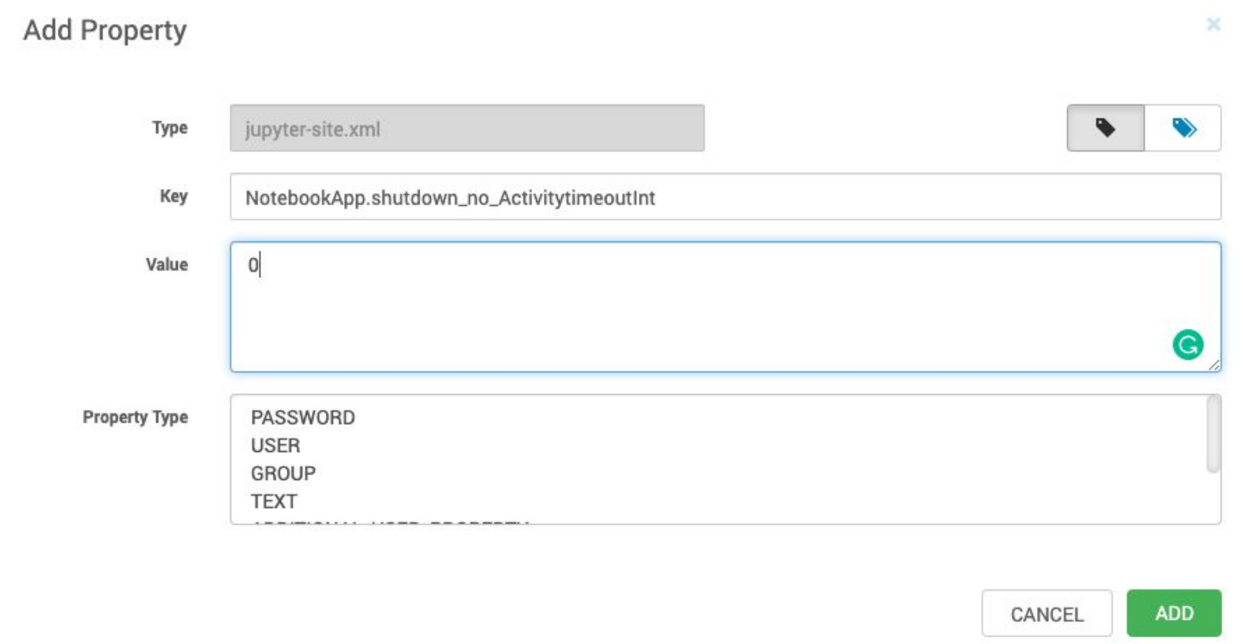

Add Property

Type: jupyter-site.xml

Key: MappingKernelM0nager.cull_idletimeoutInt

Value: 0

Property Type: PASSWORD, USER, GROUP, TEXT


Add Property

Type: jupyter-site.xml

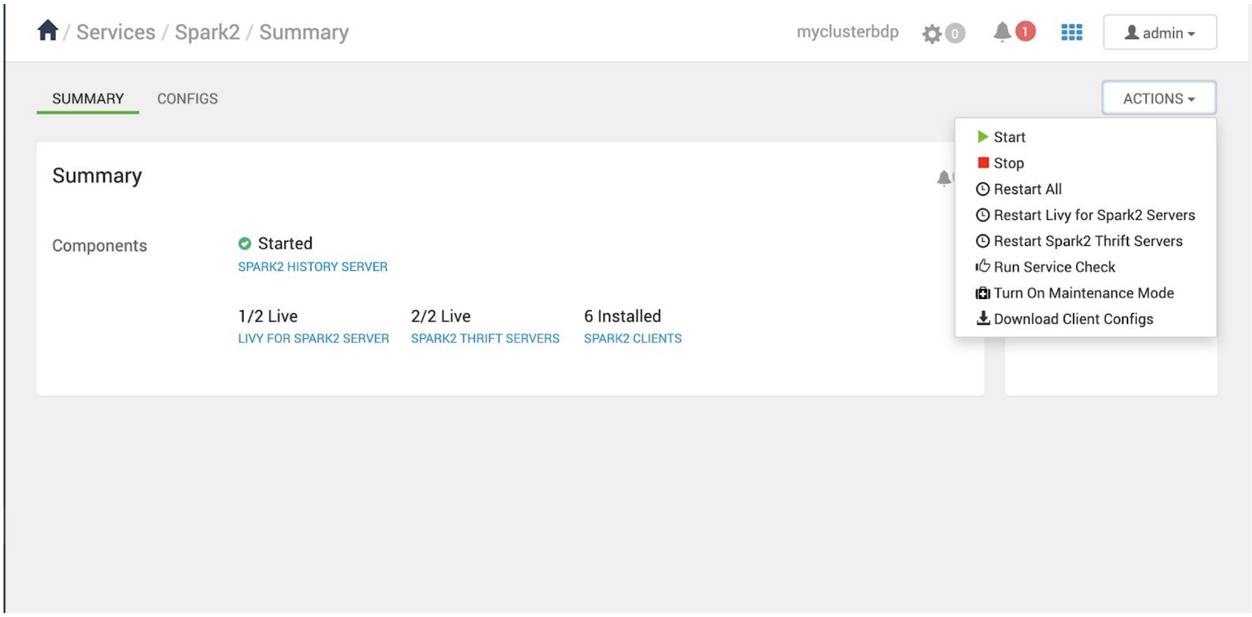
Key: NotebookApp.shutdown_no_ActivitytimeoutInt

Value: 0

Property Type: PASSWORD, USER, GROUP, TEXT

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.



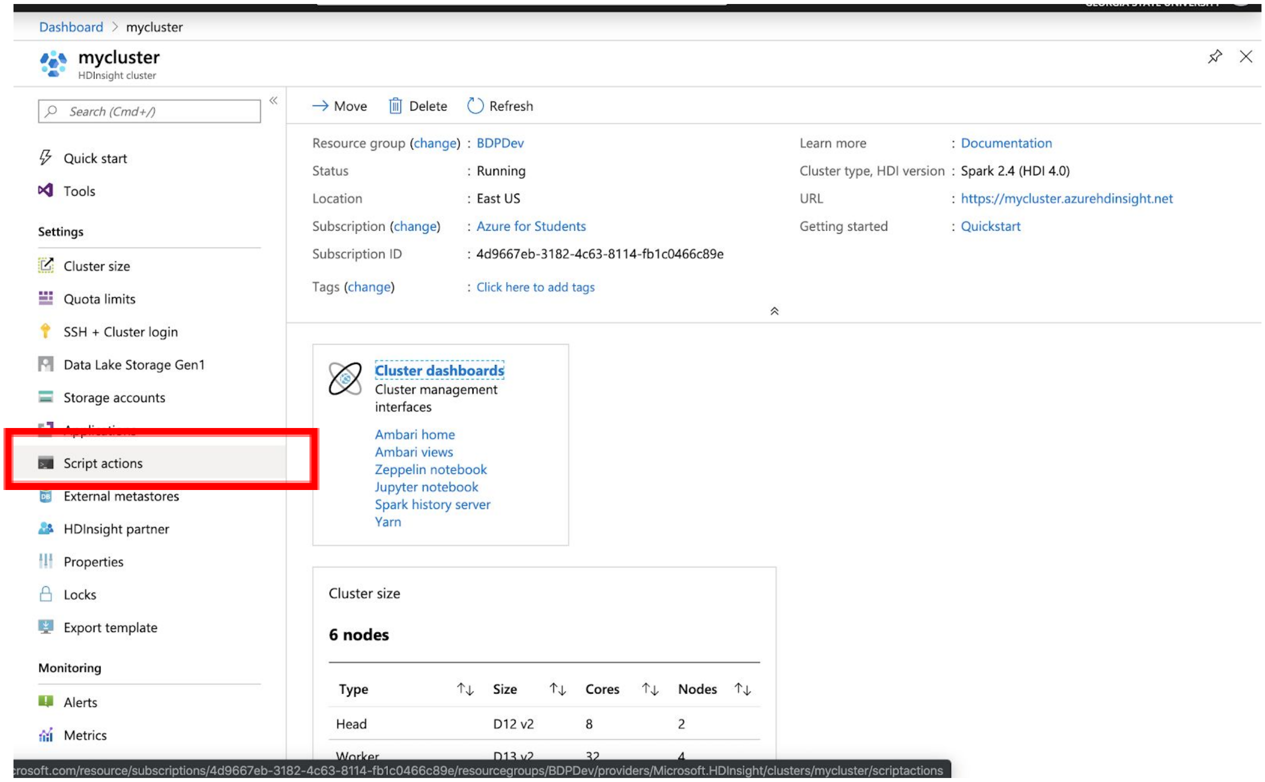
Started SPARK2 HISTORY SERVER

1/2 Live LIVY FOR SPARK2 SERVER

2/2 Live SPARK2 THRIFT SERVERS

6 Installed SPARK2 CLIENTS

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



Resource group (change) : BDPDev
Status : Running
Location : East US
Subscription (change) : Azure for Students
Subscription ID : 4d9667eb-3182-4c63-8114-fb1c0466c89e
Tags (change) : Click here to add tags

Learn more : Documentation
Cluster type, HDI version : Spark 2.4 (HDI 4.0)
URL : https://mycluster.azurehdinsight.net
Getting started : Quickstart

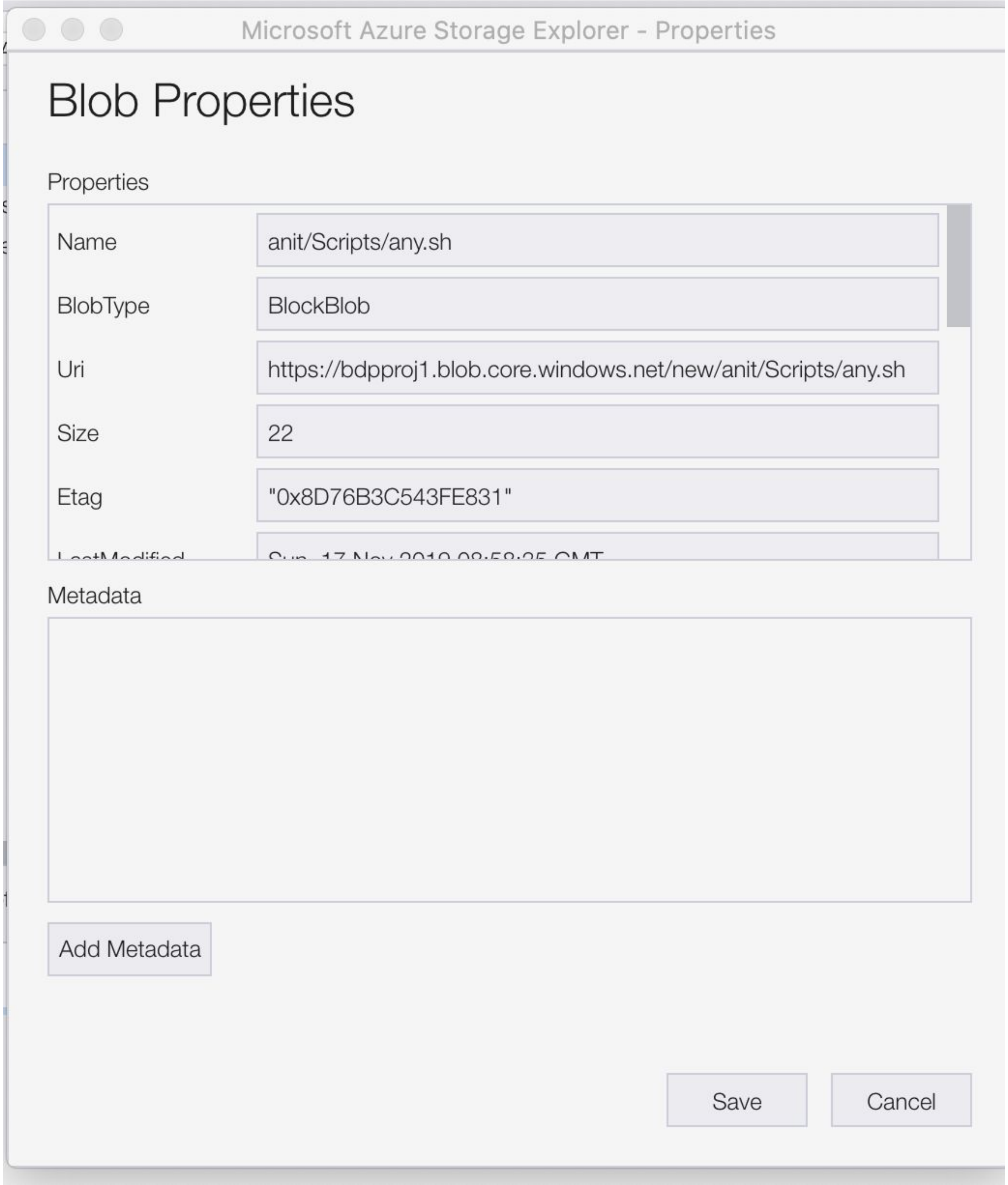
Cluster size

6 nodes

Type	Size	Cores	Nodes
Head	D12 v2	8	2
Worker	D13 v2	32	4

rosoft.com/resource/subscriptions/4d9667eb-3182-4c63-8114-fb1c0466c89e/resourcegroups/BDPDev/providers/Microsoft.HDInsight/clusters/mycluster/scriptactions

Step) Get bash script URI from azure storage account explorer by right clicking on your file and clicking properties there you can find the URI



Microsoft Azure Storage Explorer - Properties

Blob Properties

Properties

Name	anit/Scripts/any.sh
BlobType	BlockBlob
Uri	https://bdpproj1.blob.core.windows.net/new/anit/Scripts/any.sh
Size	22
Etag	"0x8D76B3C543FE831"
LastModified	Sun, 17 Nov 2019 08:58:35 GMT

Metadata

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.

SparkAllWords.ipynb file changes that you need to do as per the storage Account.

You would need to add your paths and storage account details in the file as we have saved them above.
You need to update below details as per you saved above.

```
------------------------------------------------------------------------------------------------------------------------
accountName = "bdpproj1"
accountKey=
"qDRRza9BkroEhjC8rOJcEZ70WU9fFbXOylwcjtWT6mjGd1AVZ9O2dgd+s0+m5vzL39Fix+TvxDwSABtvadjQEg==
"
   containerName = "new"
------------------------------------------------------------------------------------------------------------------------
```

You need to set the path of the folder that you created in Step 11()

```
------------------------------------------------------------------------------------------------------------------------
blobService.create_blob_from_text('new/anit',file_name, output)
------------------------------------------------------------------------------------------------------------------------
```

You need to add path to your data set here and add * if you want an entire folder to be read.

```
------------------------------------------------------------------------------------------------------------------------
dataset_path = "wasbs:///anit/10/01/*/*"
------------------------------------------------------------------------------------------------------------------------
```

Once done you should download the generated CSVs from the azure storage explorer and put it where you keep the final code for all the codes.

PART II - Question Specific installations

Run the below commands on the local spark, not on Azure spark since all the rest of the code is to be run on local spark.

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
```
