

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.

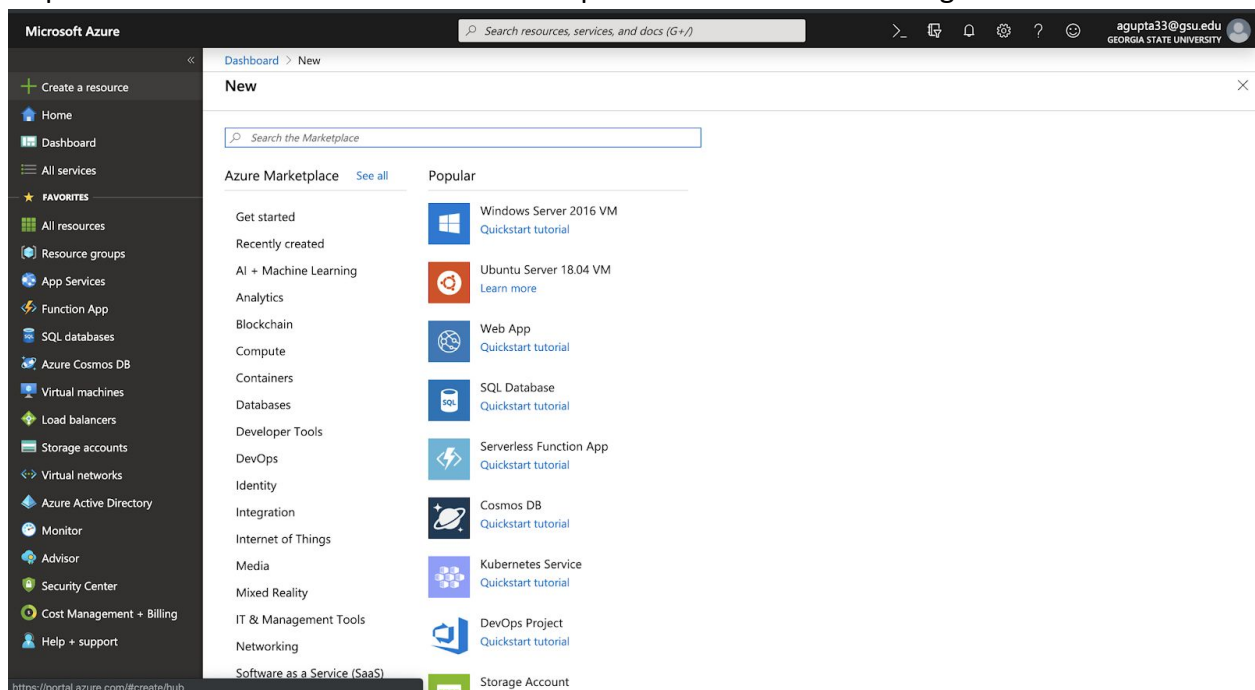
Part I - CLUSTER CREATION AND PREPROCESSING

Step 1: Download Microsoft Azure Storage Explorer

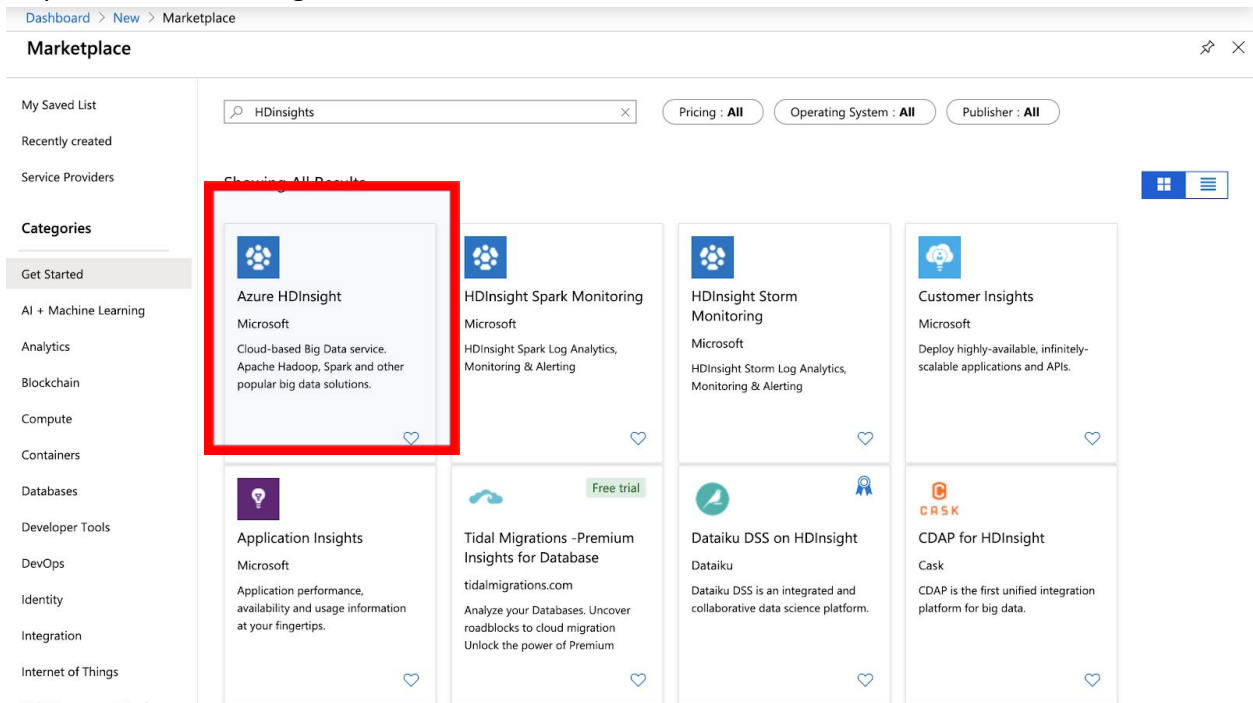
Step2: login with your same credentials as you have in hdinstights.

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

Project details

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

Subscription *

Resource group *

[Create new](#)

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name *

Region *

Cluster type * **Spark**
[Change](#)

Version *

Step7: Fill in **cluster credentials**

Cluster credentials

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

Cluster login username * ⓘ

Cluster login password *

Confirm cluster login password *

Secure Shell (SSH) username * ⓘ

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 *

Selection method * ⓘ ☒ Select from list ☐ Use access key

Primary storage account *
[Create new](#)

Container * ⓘ

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 | <input type="text" value="D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/..."/> | 2 | 0.75 USD |
| Worker node | <input type="text" value="D13 v2 (8 Cores, 56 GB RAM), 0.75 USD/..."/> | 4 <input checked="" type="checkbox"/> | 2.99 USD |

☐ Enable autoscale [Learn more](#)

Total estimated cost/hour 3.74 USD

Create HDInsight cluster



[Go to classic create experience](#)

✓ Validation succeeded.

[Basics](#) [Storage](#) [Security + networking](#) [Configuration + pricing](#) [Review + create](#)

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 |

Create

« Previous

Next »

[Download a template for automation](#)

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

Overview

Inputs

Outputs

Template

Delete Cancel Redeploy Refresh

■ ■ ■ Your deployment is underway

Deployment name: HDInsight__2019-12-09T06.45.50.382Z Start time: 12/9/2019, 1:45:51 AM

Subscription: [Azure for Students](#) Correlation ID: 49257c66-fc49-4101-84b6-9943ddcd9740

Resource group: [BDPDev](#)

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 infrastr

[Go to Azure security center >](#)

Free Microsoft tutorials

[Start learning today >](#)

Work with an expert

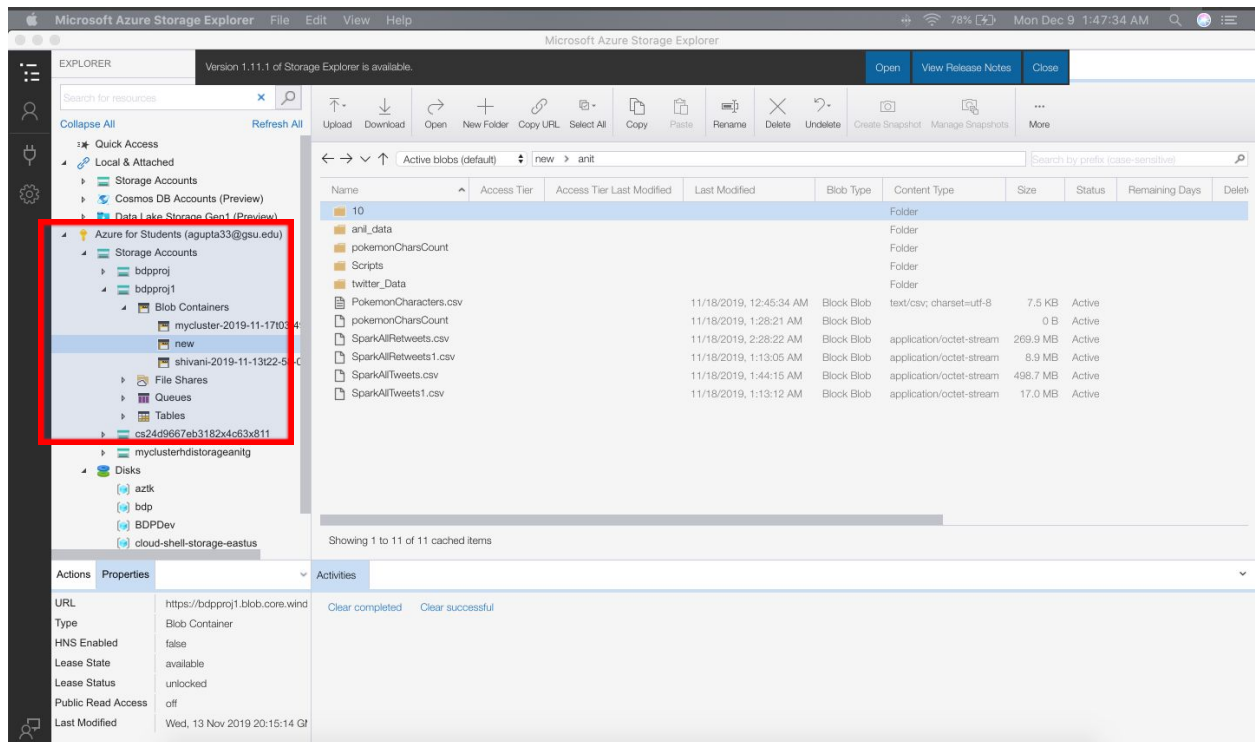
Azure experts are service prov

who can help manage your as

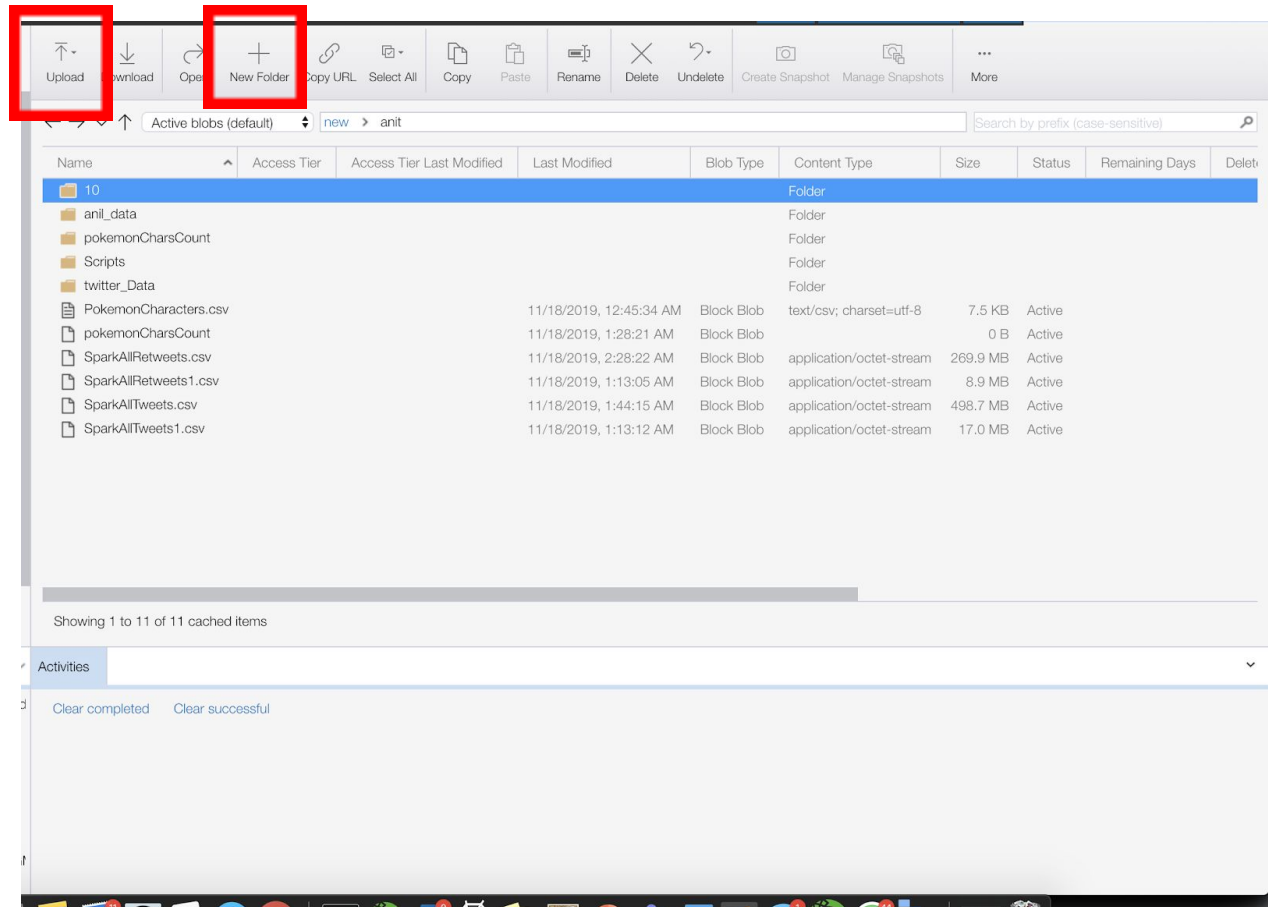
and be your first line of supp

[Find an Azure expert >](#)

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



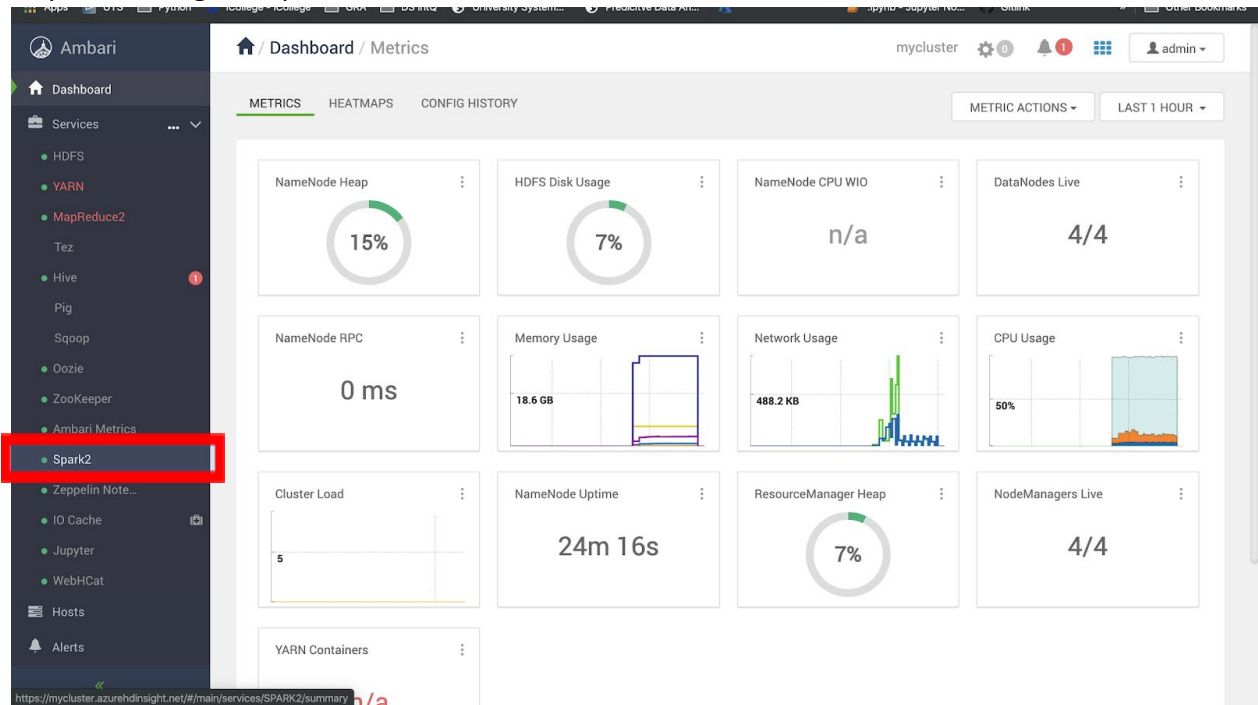
Step 11: Here you can see that you October data with the as folder 10 if you are using our storage account or you will need to upload your own data folder from your local machine. That can be simply done by creating a new folder and then by clicking on the upload button.



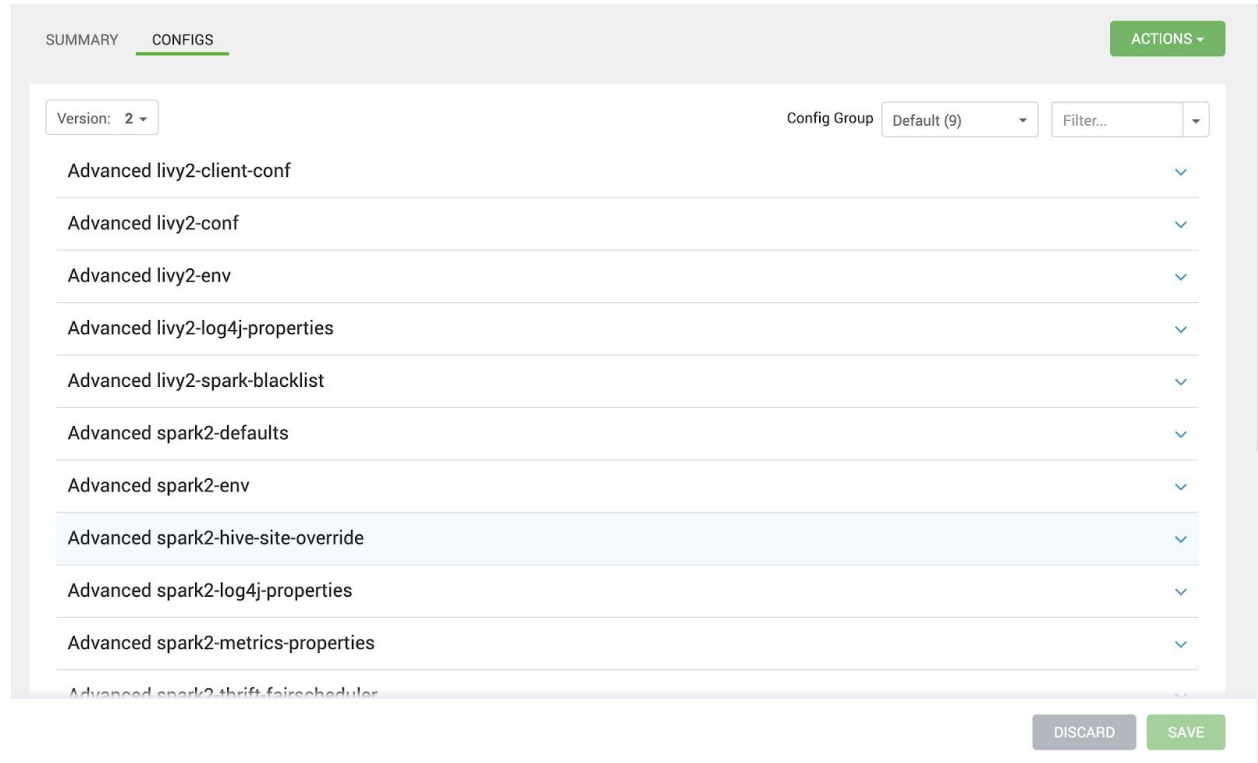
Step 12: Once the cluster is setup go-to resource and click on Ambari home

The screenshot shows the Microsoft Azure portal interface. On the left is a dark sidebar with navigation options: 'Create a resource', 'Home', 'Dashboard', 'All services', 'FAVORITES', 'All resources', 'Resource groups', 'App Services', 'Function App', 'SQL databases', 'Azure Cosmos DB', 'Virtual machines', 'Load balancers', 'Storage accounts', 'Virtual networks', 'Azure Active Directory', 'Monitor', 'Advisor', 'Security Center', 'Cost Management + Billing', and 'Help + support'. The main area has a breadcrumb trail: 'Dashboard > mycluster > Cluster dashboards'. Below this is a window titled 'Cluster dashboards' containing a 2x3 grid of tiles. The tiles are: 'Ambari home' (highlighted with a dashed blue border), 'Ambari views', 'Jupyter notebook', 'Zeppelin notebook', 'Spark history server', and 'Yarn'. Each tile features an icon and a link icon in the top right corner.

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

| | |
|---|--|
| <code>livy.environment</code> | <input type="text" value="production"/> |
| <code>livy.impersonation.enabled</code> | <input type="text" value="true"/> |
| <code>livy.repl.enableHiveContext</code> | <input type="text" value="true"/> |
| <code>livy.server.access-control.enabled</code> | <input type="text" value="true"/> |
| <code>livy.server.csrf_protection.enabled</code> | <input type="text" value="true"/> |
| <code>livy.server.port</code> | <input type="text" value="8998"/> |
| <code>livy.server.recovery.mode</code> | <input type="text" value="recovery"/> |
| <code>livy.server.recovery.state-store</code> | <input type="text" value="zookeeper"/> |
| <code>livy.server.recovery.state-store.url</code> | <input type="text" value="zk1-bdppro.2h5ffrof4nveneybdi1ajpa2qh.bx.internal.cloudapp.net:2181,zk2-bdppro.2h5ffrof"/> |
| <code>livy.server.session.timeout</code> | <input type="text" value="18000000"/> |
| <code>livy.spark.master</code> | <input type="text" value="yarn-cluster"/> |

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

Custom livy2-conf



| | | |
|---|--------------------------------------|---|
| <code>livy.server.session.state-retain.sec</code> | <input type="text" value="3600000"/> |    |
| <code>livy.server.yarn.app-lookup-timeout</code> | <input type="text" value="10m"/> |     |

[Add Property ...](#)

Step 15.iii) then in Custom 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 | |

CANCEL ADD

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 | |



CANCEL ADD

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

Add Property

Type

jupyter-site.xml




Key

MappingKernelMOnager.cull_idletimeoutInt

Value

0



Property Type

PASSWORD
USER
GROUP
TEXT



CANCEL

ADD

Add Property

Type

jupyter-site.xml




Key

NotebookApp.shutdown_no_ActivitytimeoutInt

Value

0



Property Type

PASSWORD
USER
GROUP
TEXT

CANCEL

ADD

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

Dashboard > mycluster

mycluster
HDInsight cluster

Search (Cmd+/)

Move Delete Refresh

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 dashboards
Cluster management interfaces
Ambari home
Ambari views
Zeppelin notebook
Jupyter notebook
Spark history server
Yarn

Cluster size
6 nodes

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

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 PYSARK_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
