

Readme file for Spark Cluster

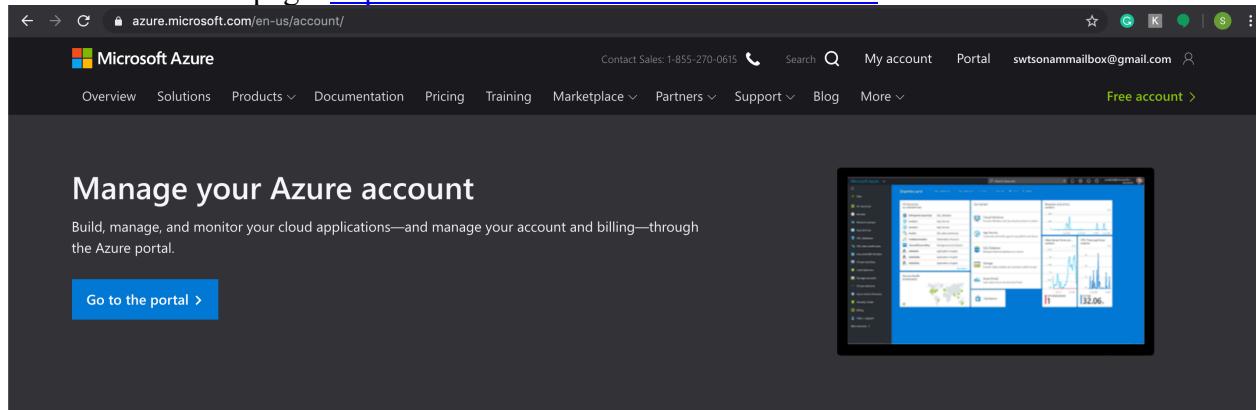
We have two Readme files for this project. One for the local environment and other is for spark cluster environment. Local environment readme file is named as ‘Readme_local.md’ where ‘March2016_Twitter_Processing_Local.ipynb’ notebook is to be executed. This readme file is for the spark cluster setup, 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: [sdaasari3@student.gsu.edu](mailto: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**

The screenshot shows the Microsoft Azure portal homepage. At the top, there is a navigation bar with icons for back, forward, search, and user profile. Below the navigation bar is a blue header bar with the text "Microsoft Azure" and a search bar that says "Search resources, services, and docs (G+)".

The main content area is titled "Azure services" and features a "Create a resource" button with a plus sign icon, which is highlighted with a blue dashed box. Other service icons include HDInsight clusters, All resources, Storage accounts, App Services, Virtual machines, SQL databases, Azure Database for PostgreSQL, Azure Cosmos DB, and More services.

Below the service icons is a section titled "Recent resources" with a table:

NAME	TYPE	LAST VIEWED
sonanscluster4bdp	HDInsight cluster	3 wk ago
clustersonamhdstorage	Storage account	3 wk ago

Below the recent resources is a "Navigate" section with links for Subscriptions, Resource groups, All resources, and Dashboard.

At the bottom left, there is a link to "https://portal.azure.com/#create/hub" and a "Azure mobile app" button. On the right side, there are links for Microsoft Learn, Azure Monitor, Security Center, and Cost Management.

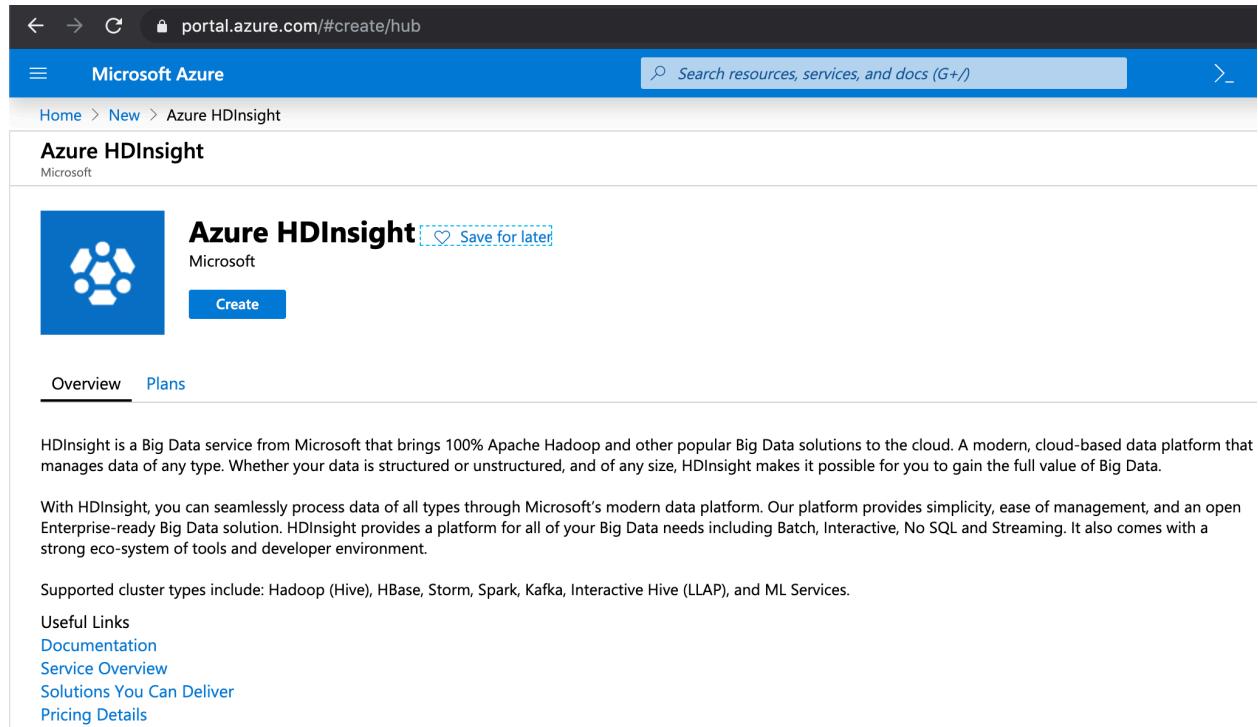
5. From **Analytics**> select **Azure HDInsight**

The screenshot shows the Microsoft Azure portal's "New" section. The URL in the address bar is "https://portal.azure.com/#create/hub". The search bar at the top contains the text "Azure HDInsight".

The search results list the following options:

- Azure HDInsight
- Unravel for Azure HDInsight
- Starburst Presto for Azure HDInsight
- Azure HDInsight (classic create experience)
- Starburst Presto for Azure HDInsight

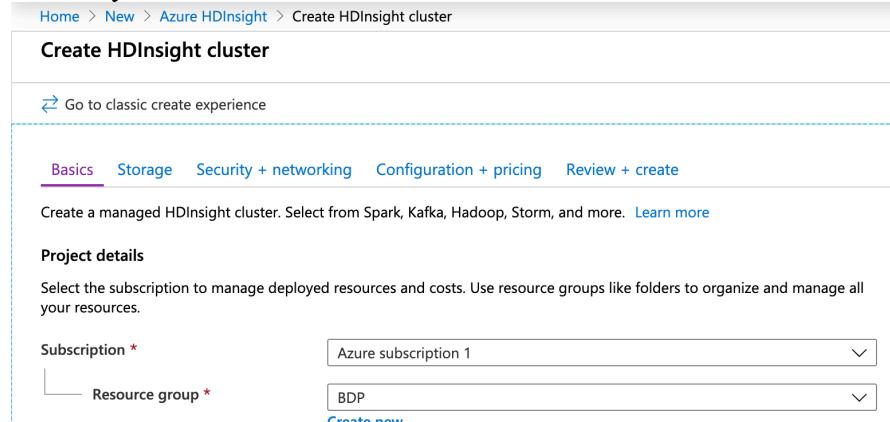
6. Click on Create



The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with back, forward, and search icons, followed by the URL 'portal.azure.com/#create/hub'. Below the URL is the Microsoft Azure logo and a search bar with the placeholder 'Search resources, services, and docs (G+)'. The main content area has a blue header 'Azure HDInsight' with a Microsoft logo. Below the header is a large blue button with a white hexagonal icon containing smaller circles, labeled 'Azure HDInsight'. To its right is a 'Save for later' button with a heart icon. A 'Create' button is also visible. Below the main title, there are two tabs: 'Overview' (which is underlined) and 'Plans'. Under 'Overview', there's a brief description of HDInsight as a Big Data service from Microsoft, mentioning it brings 100% Apache Hadoop and other popular Big Data solutions to the cloud. It highlights its modern, cloud-based nature and ability to manage data of any type. It also mentions supported cluster types like Hadoop (Hive), HBase, Storm, Spark, Kafka, Interactive Hive (LLAP), and ML Services. Below this, there's a section titled 'Useful Links' with links to 'Documentation', 'Service Overview', 'Solutions You Can Deliver', and 'Pricing Details'.

7. In Basics

- Give any resource name and cluster name



The screenshot shows the 'Create HDInsight cluster' wizard. At the top, there's a breadcrumb navigation: 'Home > New > Azure HDInsight > Create HDInsight cluster'. Below the breadcrumb is a title 'Create HDInsight cluster' with a 'Go to classic create experience' link. There are five tabs at the top of the form: 'Basics' (which is underlined), 'Storage', 'Security + networking', 'Configuration + pricing', and 'Review + create'. The 'Basics' tab is active. The form contains a section for 'Project details' with a note about selecting a subscription to manage resources. It includes fields for 'Subscription *' (set to 'Azure subscription 1') and 'Resource group *' (set to 'BDP'). There's also a 'Create new' link for creating a new resource group.

- In Select cluster Type, select the Spark cluster

The screenshot shows the 'Create HDInsight cluster' wizard. On the left, there's a 'Cluster details' section with fields for 'Cluster name' (CluterForBDP), 'Region' (East US), and 'Cluster type' (set to 'Select cluster type'). Below that is a 'Cluster credentials' section. On the right, a sidebar titled 'Select cluster type' lists several options with 'Select' buttons: Hadoop, Spark, Kafka, HBase, Interactive Query, and Storm. The 'Spark' option is currently selected.

- Version: Spark 2.4 (HDI 4.0)

The screenshot shows the 'Create HDInsight cluster' wizard. In the 'Cluster details' section, the 'Version' dropdown is open, displaying a list of available versions: Spark 1.6.3 (HDI 3.5), Spark 2.1 (HDI 3.6), Spark 2.2 (HDI 3.6), Spark 2.3 (HDI 3.6), Spark 2.3 (HDI 4.0), and Spark 2.4 (HDI 4.0). The 'Spark 2.4 (HDI 4.0)' option is highlighted.

- Fill in your usernames and passwords as necessary

The screenshot shows the 'Create HDInsight cluster' wizard. In the 'Cluster credentials' section, there are fields for 'Cluster login username' (admin), 'Cluster login password' (redacted), 'Confirm cluster login password' (redacted), 'Secure Shell (SSH) username' (sshuser), and a checked checkbox for 'Use cluster login password for SSH'. A note says 'Password and confirm password must match.'

- 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

Microsoft Azure

Search resources, services, and docs (G)

Home > New > Azure HDInsight > Create HDInsight cluster

Create HDInsight cluster

Go to classic create experience

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 * [Create new](#)

Container * [Data Lake Storage Gen1](#)

Provide details for the cluster to access Data Lake Storage accounts that the chosen service principal has permission to access.

Azure storage account name * cluterforbdphdistorage

OK Cancel

- Let the container as it is
- 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 * [\(New\) cluterforbdphdistorage](#)

Container * [Create new](#)

cluterforbdp-2019-12-09t21-13-35-254z

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

Additional Azure storage

Link additional Azure storage accounts to the cluster.

Account name

[Add Azure storage](#)

- Select ‘Access key’ as selection method:

Storage accounts

Selection method * ⓘ
 My subscriptions Access key

Storage account name *

Access key *

- Use the below details:

Storage account name: bdptwittersparkstorage

Access Key:

MYeM1QzOpYKkSQv3CCyKuLG5pKHkskBA5D2p0wlPmD1yQHtdfnRv0j
X+mEtAzspsv7DnGc6OaQw15lY9qlZB0w==

Storage accounts

Selection method * ⓘ

My subscriptions Access key

Storage account name *

bdptwittersparkstorage 

Access key *

MYeM1QzOpYKkSQv3CCyKuLG5pKHkskB... 

- After clicking Select, you will find the storage added
Additional Azure storage

Link additional Azure storage accounts to the cluster.

Account name
bdptwittersparkstorage
Add Azure storage

- 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:

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.

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

[Review + create](#) [« Previous](#) [Next: Review + create »](#)

11. Click **Review+Create**

12. Once review the details, then Create

The screenshot shows the 'Create HDInsight cluster' wizard in the Microsoft Azure portal. At the top, there's a blue header bar with the Microsoft Azure logo and a search bar. Below the header, the breadcrumb navigation shows 'Home > New > Azure HDInsight > Create HDInsight cluster'. The main title is 'Create HDInsight cluster'. A green success message 'Validation succeeded.' is displayed. Below the message, there are tabs for 'Basics', 'Storage', 'Security + networking', 'Configuration + pricing', and 'Review + create'. The 'Review + create' tab is selected. Under 'Spark 2.4 (HDI 4.0)', it shows a total estimated cost of '3.74 USD Total estimated cost/hour'. A note states: 'This estimate does not include subscription discounts or costs related to storage, networking, or data transfer.' The 'Basics' section contains the following configuration details:

Subscription	Azure subscription 1
Resource group	BDP
Region	East US
Cluster name	(new) CluterForBDP
Cluster type	Spark 2.4 (HDI 4.0)
Cluster login username	admin
Secure Shell (SSH) username	sshuser
Use cluster login password for SSH	Enabled

The 'Storage' section shows 'Primary storage type' as 'Azure Storage'. At the bottom, there are buttons for 'Create', '<< Previous', 'Next >>', and 'Download a template for automation'.

The screenshot shows the 'HDInsight_2019-12-09T21.56.56.715Z - Overview' page in the Microsoft Azure portal. The top navigation bar includes 'Search resources, services, and docs (G+)', a back arrow, and a gear icon. The main title is 'HDInsight_2019-12-09T21.56.56.715Z - Overview'. On the left, a sidebar menu has 'Overview' selected. The main content area displays a message: 'Your deployment is underway'. It shows deployment details: Deployment name: HDInsight_2019-12-09T21.56.56.715Z, Subscription: Azure subscription 1, Resource group: BDP. It also shows start time: 12/9/2019, 4:56:57 PM and Correlation ID: 3b19d4a1-ba6c-4ffe-a997-903018aa1932. There are sections for 'Deployment details' (with a download link) and 'Next steps'.

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.

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information. Below the navigation bar, the URL is 'HDInsight_2019-12-09T21.56.56.715Z - Overview'. On the left, there is a sidebar with 'Overview', 'Inputs', 'Outputs', and 'Template' options. The main content area displays a message: 'Your deployment is complete' with a green checkmark icon. Deployment details include: Deployment name: HDInsight_2019-12-09T21.56.56.715Z, Subscription: Azure subscription 1, Resource group: BDP. Deployment started at 12/9/2019, 4:56:57 PM, with Correlation ID: 3b19d4a1-babc-4ffe-a997-903018aa1932. Below this, there are sections for 'Deployment details' (with a 'Download' link) and 'Next steps' (with a 'Go to resource' button). To the right, there are links to 'Security Center', 'Free Microsoft tutorials', and 'Work with an expert'. A note states: 'Secure your apps and infrastructure', 'Start learning today', and 'Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support'.

Once the cluster is created,

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information. Below the navigation bar, the URL is 'HDInsight_2019-12-09T21.56.56.715Z - Overview > CluterForBDP'. On the left, there is a sidebar with 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Quick start', 'Tools', 'Settings', 'Cluster size', 'Quota limits', 'SSH + Cluster login', 'Data Lake Storage Gen1', 'Storage accounts', 'Applications', 'Script actions', 'External metastores', and 'HDInsight partner' options. The main content area displays cluster details: Resource group (change): BDP, Status: Running, Location: East US, Subscription (change): Azure subscription 1, Subscription ID: 28cad354-fa1d-4c9b-b7e9-4c4f0c4b4b41, and Tags (change): Click here to add tags. To the right, there are links to 'Learn more' (Documentation), 'Cluster type, HDI version' (Spark 2.4 (HDI 4.0)), 'URL' (https://CluterForBDP.azurehdinsight.net), and 'Getting started' (Quickstart). Below this, there are sections for 'Cluster dashboards' (Ambari home, Ambari views, Zeppelin notebook, Jupyter notebook, Spark history server, Yarn) and 'Cluster size' (6 nodes).

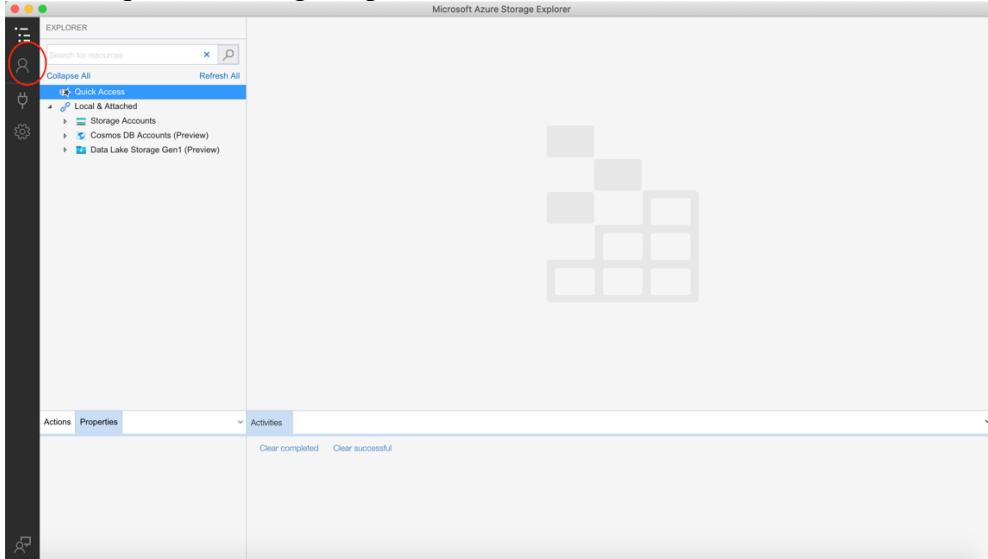
we have to update some settings on cluster according to the requirement of the project. The instructions for the same are listed under **Instructions to update cluster setting and installing external libraries**

Instructions to upload the data into Azure blob storage account:

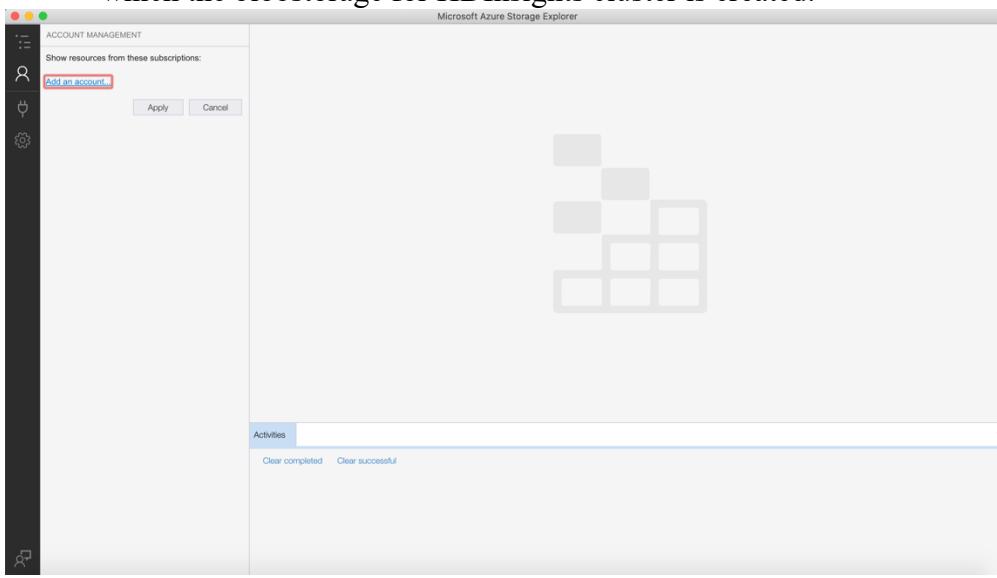
Please find below the link to download Microsoft Azure Storage Explorer:
<https://azure.microsoft.com/en-us/features/storage-explorer/>

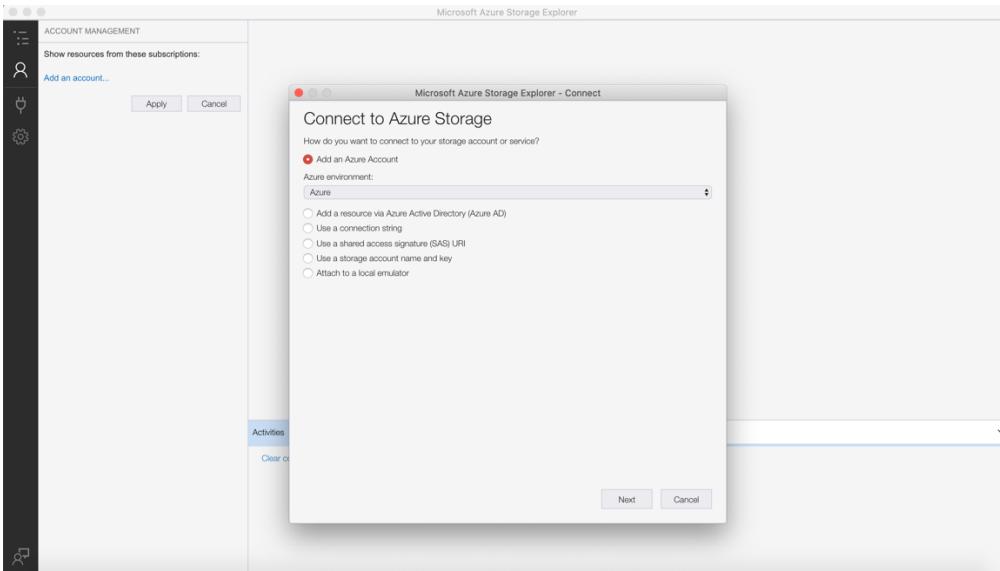
Once the Storage explorer is downloaded and installed, follow the below instructions:
The necessary parts are highlighted in red in below screenshots.

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.





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

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Status	Remaining Days	Deleted Time	Lease
ams				Folder						
amshbase				Folder						
app-logs				Folder						
apps				Folder						
ahistory				Folder						
atav2				Folder						
example				Folder						
HdNotebooks				Folder						
HdSamples				Folder						
hdp				Folder						
hive				Folder						
mapred				Folder						
mr-history				Folder						
tmp				Folder						
user				Folder						
warehouse				Folder						

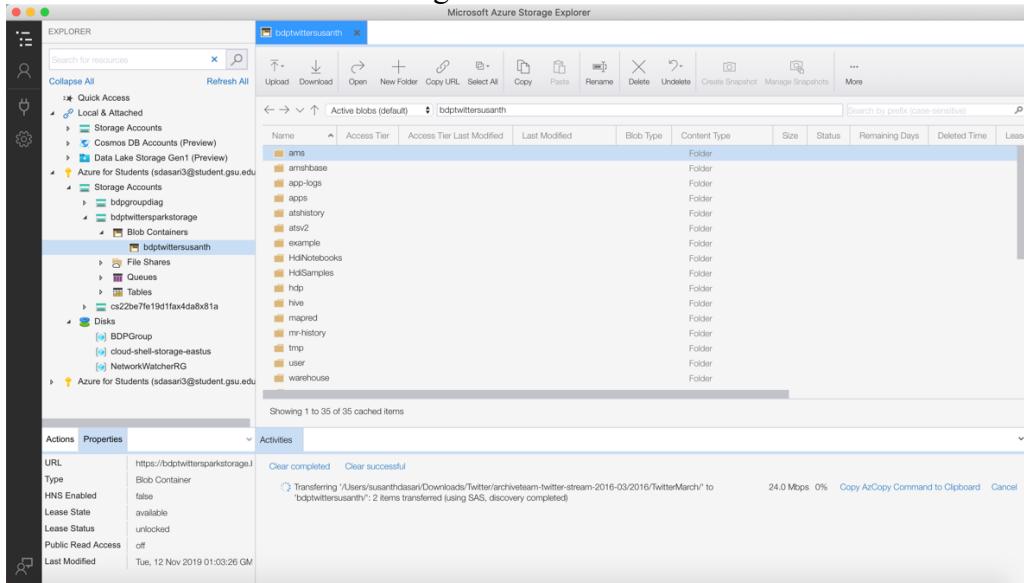
Showing 1 to 35 of 35 cached items

Actions Properties Activities

URL: https://bdptwittersparkstorage.blob.core.windows.net/bdptwittersparkstorage
Type: Blob Container
HNS Enabled: false
Lease State: available
Lease Status: unlocked
Public Read Access: off
Last Modified: Tue, 12 Nov 2019 01:03:26 GM

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



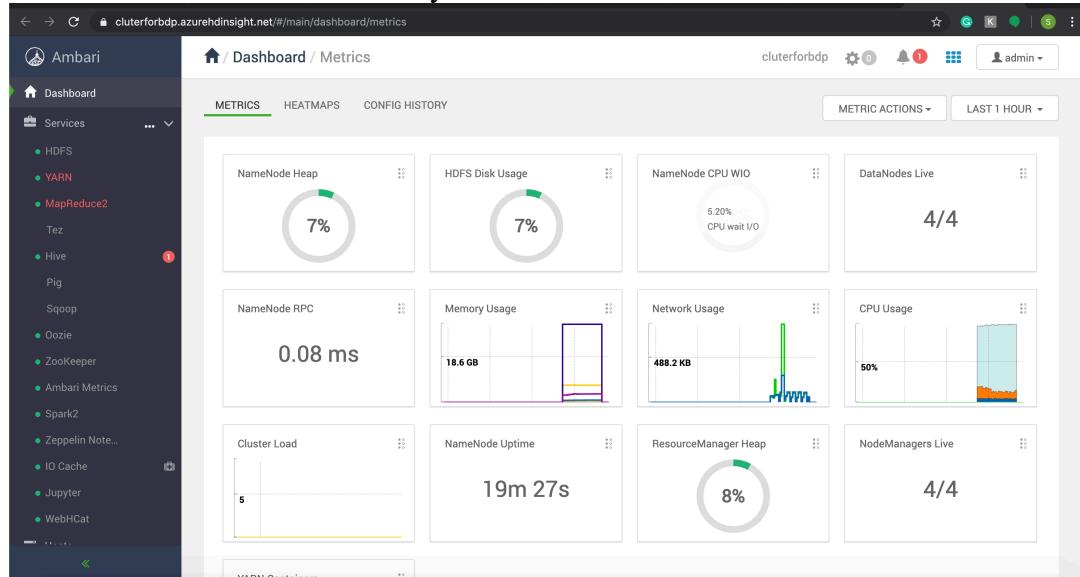
Instructions to update cluster setting and installing external libraries:

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

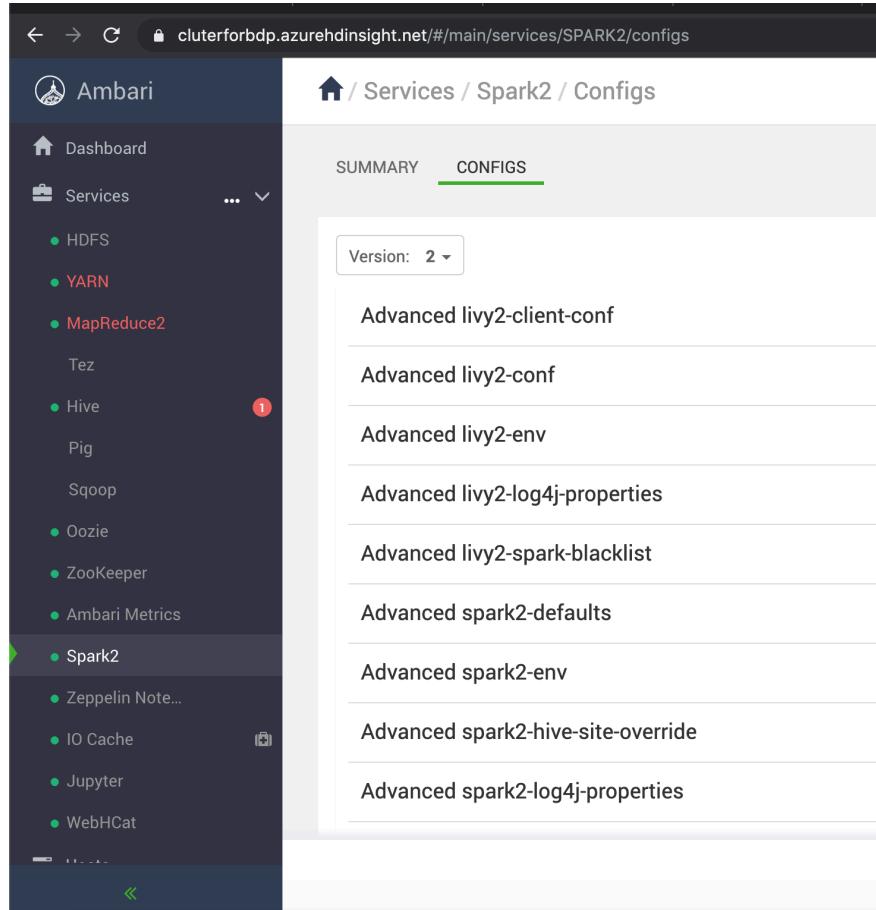
Dashboard will look like:

The screenshot shows the Microsoft Azure portal's HDInsight cluster overview page. The left sidebar includes options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Tools, Cluster size, Quota limits, SSH + Cluster login, Data Lake Storage Gen1, Storage accounts, Applications, Script actions, External metastores, and HDInsight partner. The main content area displays cluster details: Resource group (BDP), Status (Running), Location (East US), Subscription (Azure subscription 1), Subscription ID (28cad354-fa1d-4c9b-b7e9-4c4f0c4b4b41), and Tags (Click here to add tags). It also shows cluster management interfaces (Ambari home, Ambari views, Zeppelin notebook, Jupyter notebook, Spark history server, Yarn) and a summary of 6 nodes with columns for Type, Size, Cores, and Nodes. A 'Move' button is available for the cluster.

- 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

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	zk0-cluter.hbffc3aeabuu5gnemgt2fmvixe.bx.internal.cloudapp.net:2181,zk1-cluter.hbffc3a
livy.server.session.timeout	3600000
livy.spark.master	yarn-cluster

livy.server.session.timeout
Time in milliseconds on how long Livy will wait before timing out an idle session.
Default is one hour.

DISCARD SAVE

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

Configurations

Highly Recommended Configurations (17)

Please review the following recommended changes, and click on the property name to change its value.

Type	Service	Property	Current Value	Description
Error	HDFS	hadoop.proxyuser.hdfs.groups		Value should be set for hadoop.proxyuser.hdfs.groups
Error	HDFS	hadoop.proxyuser.yarn.hosts		Value should be set for hadoop.proxyuser.yarn.hosts
Error	HDFS	hadoop.proxyuser.hdfs.hosts		Value should be set for hadoop.proxyuser.hdfs.hosts
Error	HDFS	hadoop.proxyuser.root.hosts		Value should be set for hadoop.proxyuser.root.hosts
Error	HDFS	hadoop.proxyuser.root.groups		Value should be set for hadoop.proxyuser.root.groups
Warning	YARN	yarn.nodemanager.linux-container-executor.cgroups.hierarchy	/yarn	yarn.nodemanager.linux-container-executor.cgroups.hierarchy and yarn_hierarchy should always have same value yarn.nodemanager.linux-container-executor.cgroups.hierarchy and yarn_hierarchy should always have same value Name of the Cgroups hierarchy under which all YARN jobs will be launched
Warning	YARN	yarn.scheduler.maximum-allocation-mb	51200	Values greater than 47616MB are not recommended The maximum allocation for every container request at the RM, in MBs. Memory requests higher than this won't take effect, and will get capped to this value.
Warning	YARN	yarn.scheduler.maximum-allocation-vcores	15	Value is greater than the recommended maximum of 6

CANCEL 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

Custom livy2-conf

livy.server.session.state-retain.sec	180000000
livy.server.yarn.app-lookup-timeout	2m

Add Property ...

Parameters to add:

1. Under ‘Custom Spark2-defaults’

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

spark.sql.files.maxPartitionBytes 1073741824
spark.sql.join.preferSortMergeJoin false
spark.ui.enhancement.enabled true
spark.yarn.access.hadoopFileSystems hdfs://mycluster
spark.yarn.appMasterEnv.PYSPARK_K3_PYTHON /usr/bin/anaconda/envs/py35/bin/python3
spark.yarn.appMasterEnv.PYSPARK_PYTHON /usr/bin/anaconda/bin/python
spark.yarn.containerLauncherMaxThreads 25
spark.yarn.jars local:///usr/hdp/current/spark2-client/jars/*
spark.yarn.preserve.staging.files false
spark.yarn.scheduler.heartbeat.interval-ms 5000
spark.yarn.submit.file.replication 3

- Fill the details as:

Key: spark.sql.broadcastTimeout

Value: 6000

Add Property

Type	spark2-defaults.xml		
Key	spark.sql.broadcastTimeout		
Value	6000		
Property Type	PASSWORD USER GROUP TEXT		

CANCEL ADD

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

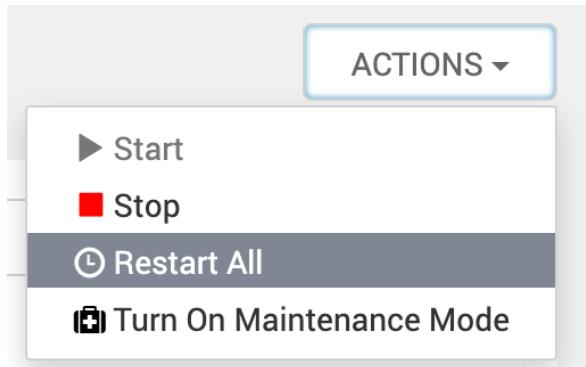
Key: MappingKernelManager.cull_idle_timeoutInt

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'

The screenshot shows the Microsoft Azure portal interface for managing an HDInsight cluster. At the top, there's a blue header bar with the Microsoft Azure logo and a navigation menu. Below it, the URL shows 'Home > HDInsight_2019-12-09T21.56.56'. The main content area is titled 'CluterForBDP' and 'HDInsight cluster'. A search bar at the top says 'Search (Cmd+ /)'. On the left, there's a sidebar with several options: Overview (selected), Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Tools, Cluster size, Quota limits, SSH + Cluster login, Data Lake Storage Gen1, Storage accounts, Applications, Script actions (selected), External metastores, and HDInsight partner.

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

The screenshot shows the Microsoft Azure portal interface for managing an HDInsight cluster named 'CluterForBDP'. The left sidebar lists various cluster management options like Tags, Diagnose and solve problems, Quick start, Tools, Settings, Cluster size, Quota limits, SSH + Cluster login, Data Lake Storage Gen1, Storage accounts, Applications, Script actions (which is selected), External metastores, HDInsight partner, Properties, Locks, and Export template. The main content area shows 'PERSISTED SCRIPT ACTIONS' with a table header 'Name' and 'Roles' and a message 'No script actions found'. Below it is 'SCRIPT ACTION HISTORY' with a table header 'Status', 'Name', and 'Date' and a message 'No script actions found'. On the right, the 'Submit script action' dialog is open. It has a 'Script type' dropdown set to 'Custom', a 'Name' input field, a 'Bash script URI' input field containing 'https://bdptwittersparkstorage.blob.core.windows.net/bdptwittersusanth/pip3-install-packages-bash.sh', a 'Node type(s)' section with checkboxes for Head, Worker, and Zookeeper, and a 'Parameters' section with a large text area. A note at the bottom states 'Persist this script action to rerun when new nodes are added to the cluster.' A 'Create' button is at the bottom right.

- 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.