

Readme file for Spark Cluster

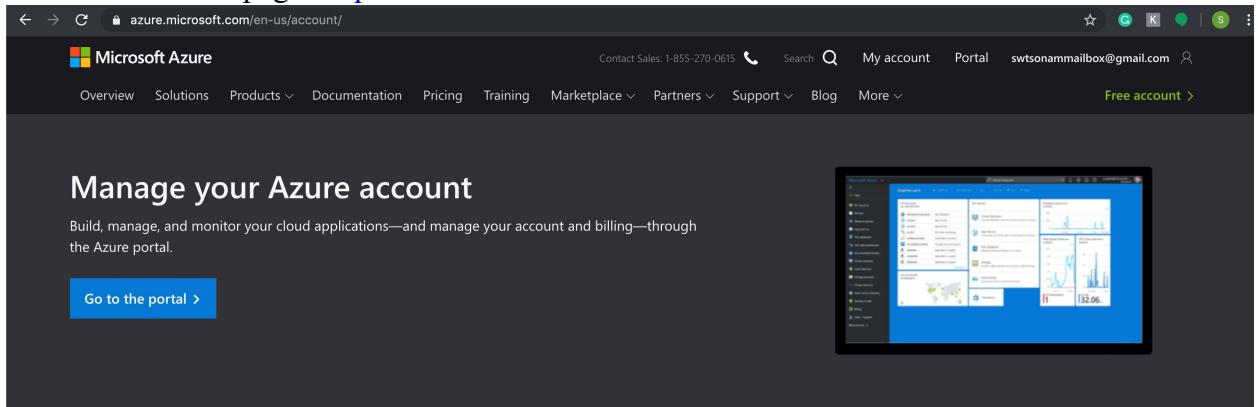
We have two Readme files for this project. One for the local environment and other is for spark cluster environment. Local environment readme file is named as ‘Readme_local.md’ where ‘March2016_Twitter_Processing_Local.ipynb’ notebook is to be executed. This readme file is for the spark cluster setup, 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:
email-ID: sdawani1@student.gsu.edu

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 interface. At the top, there's a navigation bar with icons for back, forward, search, and account information. Below it is the main dashboard area. On the left, there's a sidebar with sections for 'Azure services', 'Recent resources', 'Navigate', and 'Tools'. The 'Azure services' section has a prominent '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, and Azure Cosmos DB. The 'Recent resources' table lists two items: 'sonamscluster4bdp' (HDInsight cluster) and 'clustersonamhdstorage' (Storage account), both viewed 3 weeks ago. The 'Tools' section includes links to Microsoft Learn, Azure Monitor, Security Center, and Cost Management. At the bottom, there's a search bar with the URL 'https://portal.azure.com/#create/hub' and a link to 'Azure mobile app'.

5. From Analytics> select Azure HDInsight

The screenshot shows the 'New' blade in the Microsoft Azure portal. The URL in the address bar is 'https://portal.azure.com/#create/hub'. The search bar at the top contains the text 'Azure HDInsight'. Below the search bar, a list of results is displayed: 'Azure HDInsight' (highlighted in grey), 'Unravel for Azure HDInsight', 'Starburst Presto for Azure HDInsight', 'Azure HDInsight (classic create experience)', and 'Starburst Presto for Azure HDInsight'. The 'Azure HDInsight' result is the first item in the list.

6. Click on Create

The screenshot shows the Microsoft Azure portal at portal.azure.com/#create/hub. The top navigation bar includes back, forward, and search icons. The main header says "Microsoft Azure" and has a search bar with placeholder text "Search resources, services, and docs (G+)".

The page title is "Azure HDInsight" under the Microsoft logo. On the left, there's a blue icon of a hexagonal cluster of nodes. To its right, the text "Azure HDInsight" is displayed with a "Save for later" button. Below this is a "Create" button.

Below the main content area, there are two tabs: "Overview" (which is selected) and "Plans".

The "Overview" section contains the following text:

HDInsight is a Big Data service from Microsoft that brings 100% Apache Hadoop and other popular Big Data solutions to the cloud. A modern, cloud-based data platform that manages data of any type. Whether your data is structured or unstructured, and of any size, HDInsight makes it possible for you to gain the full value of Big Data.

With HDInsight, you can seamlessly process data of all types through Microsoft's modern data platform. Our platform provides simplicity, ease of management, and an open Enterprise-ready Big Data solution. HDInsight provides a platform for all of your Big Data needs including Batch, Interactive, No SQL and Streaming. It also comes with a strong eco-system of tools and developer environment.

Supported cluster types include: Hadoop (Hive), HBase, Storm, Spark, Kafka, Interactive Hive (LLAP), and ML Services.

Useful Links:

- Documentation
- Service Overview
- Solutions You Can Deliver
- Pricing Details

7. In Basics

- Give any resource name and cluster name

The screenshot shows the "Create HDInsight cluster" wizard. The top navigation bar shows the path: Home > New > Azure HDInsight > Create HDInsight cluster.

The main heading is "Create HDInsight cluster". Below it is a link to "Go to classic create experience".

The "Basics" tab is selected. The sub-headings are "Project details" and "Subscription *".

The "Project details" section asks to select a subscription to manage deployed resources and costs. It shows "Azure subscription 1" selected. The "Subscription *" field is also labeled "Subscription *".

The "Resource group *" section shows "BDP" selected in a dropdown, with a "Create new" link below it.

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

➤ Version: **Spark 2.4 (HDI 4.0)**

[Create new](#)

Cluster details

Name your cluster, pick a region, and choose

Cluster name *

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)

Spark 2.4 (HDI 4.0)

Spark 2.3 (HDI 3.6)

Region *

Cluster type *

Version *

Cluster credentials

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

➤ Fill in your usernames and passwords as necessary

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 *

Password and confirm password must match.

Secure Shell (SSH) username *

sshuser

Use cluster login password for SSH

[Review + create](#)

« Previous

Next: Storage »

➤ 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

Azure storage account name * cluterforbdphdistorage

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.

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

Create new

Container *

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 *

 ✓

Access key *

 ✓

- 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.

ⓘ This configuration will use 40 of 250 available cores in the East US region.
[View cores usage](#)

[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 Azure portal. At the top, there's a green success message: 'Validation succeeded.' Below it, the 'Configuration + pricing' tab is selected. A summary table shows a total estimated cost of \$3.74 USD per hour for a Spark 2.4 (HDI 4.0) cluster. The configuration details include the subscription (Azure subscription 1), resource group (BDP), region (East US), cluster name (new CluterForBDP), cluster type (Spark 2.4 (HDI 4.0)), and login credentials (admin, sshuser). The storage type is set to 'Azure Storage'. At the bottom, there are 'Create', 'Previous', 'Next', and 'Download a template for automation' buttons.

The screenshot shows the 'HDInsight_2019-12-09T21.56.56.715Z - Overview' page. It indicates that the deployment is underway. Deployment details show the name, subscription, and resource group. The deployment status table is empty. There are sections for 'Deployment details' 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 HDInsight cluster overview page. The main message is "Your deployment is complete". Deployment details include name: HDInsight_2019-12-09T21.56.56.715Z, subscription: Azure subscription 1, resource group: BDP, start time: 12/9/2019, 4:56:57 PM, correlation ID: 3b19d4a1-ba6c-4ffe-a997-903018aa1932. There are sections for "Deployment details" (Download) and "Next steps". A "Go to resource" button is present. On the right, there are links to Security Center, Free Microsoft tutorials, and Work with an expert.

Once the cluster is created,

The screenshot shows the Microsoft Azure CluterForBDP cluster overview page. The cluster status is Running, located in East US, using Spark 2.4 (HDI 4.0). It has an URL: https://CluterForBDP.azurehdinsight.net and a Getting started link: Quickstart. The left sidebar includes options like 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. The Cluster size section shows 6 nodes.

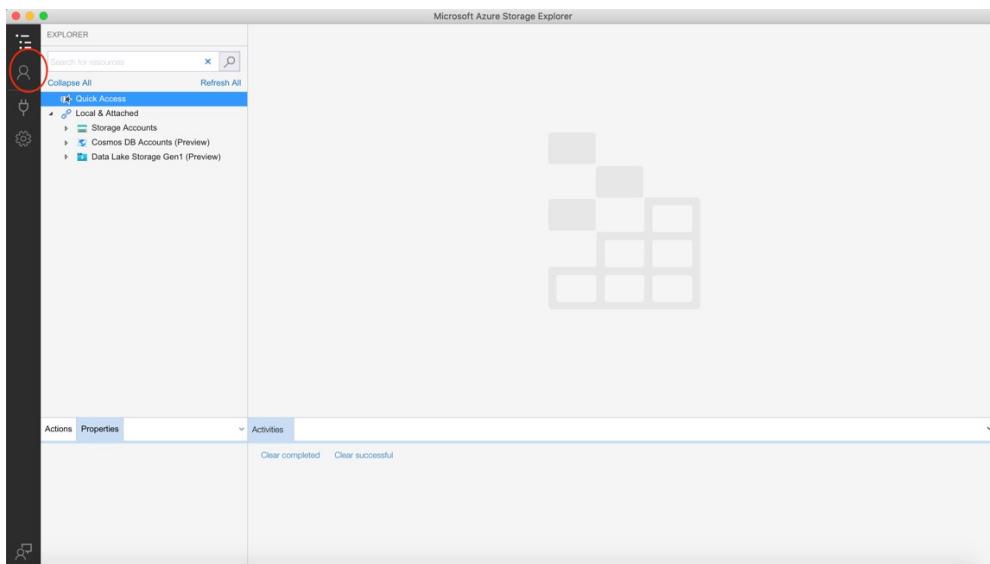
we have to update some settings on cluster according to the requirement of the project. The instructions for the same are listed under **Instructions to update cluster setting and installing external libraries**

Instructions to upload the data into Azure blob storage account:

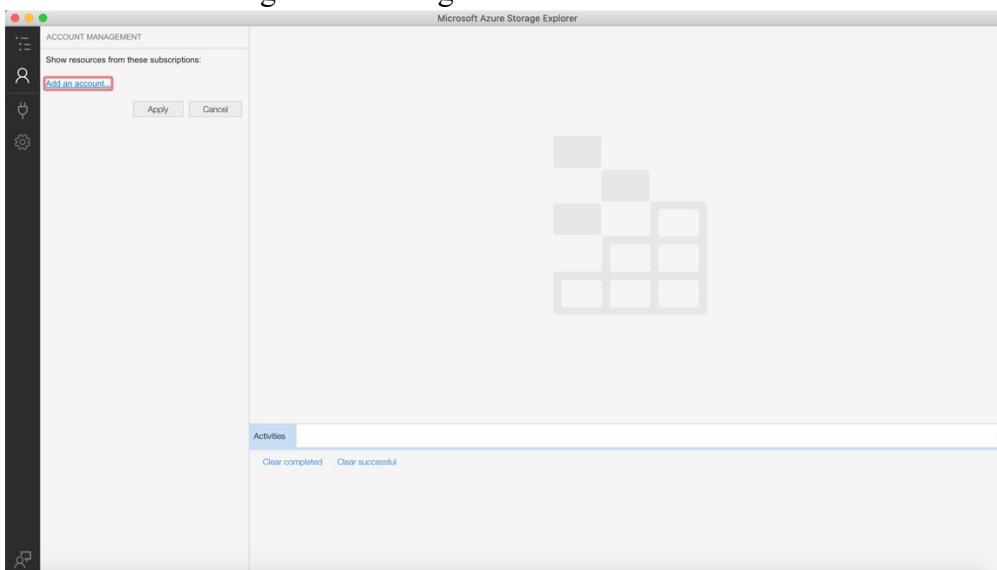
Please find below the link to download Microsoft Azure Storage Explorer: <https://azure.microsoft.com/en-us/features/storage-explorer/>

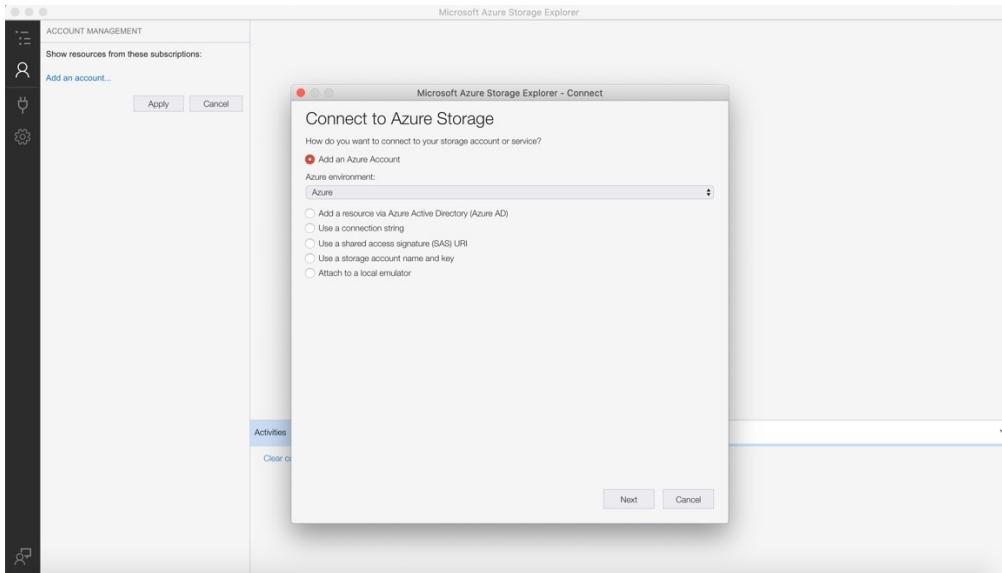
Once the Storage explorer is downloaded and installed, follow the below instructions: The necessary parts are highlighted in red in below screenshots.

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.



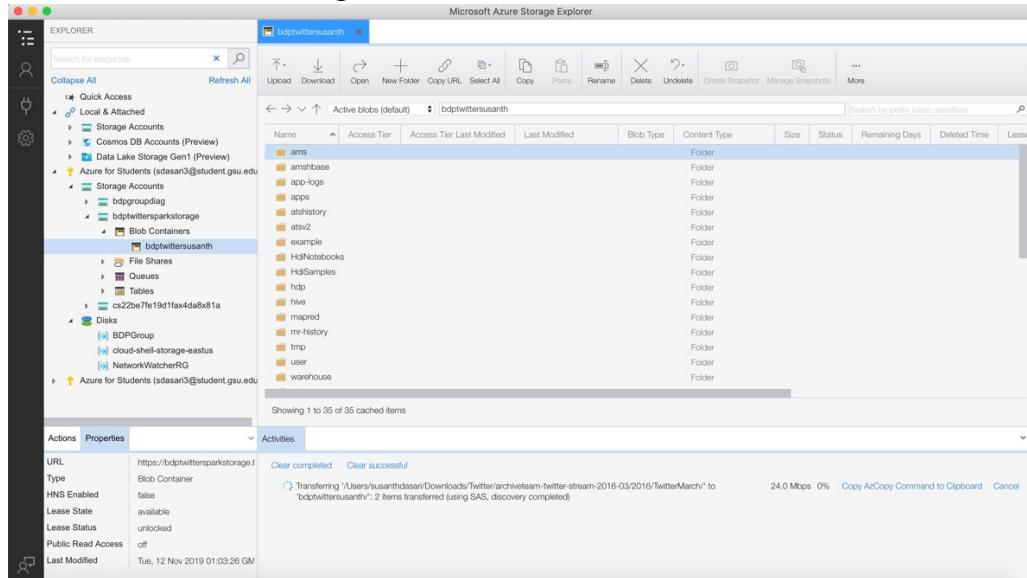


- Once the account is added, please navigate to the blobstorage created for HDInsights cluster as shown below.

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						
ethtestory				Folder						
atsv2				Folder						
example				Folder						
HdNotebooks				Folder						
HdSamples				Folder						
hdp				Folder						
hive				Folder						
mrsped				Folder						
mr-history				Folder						
tmp				Folder						
user				Folder						
warehouse				Folder						

- 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.

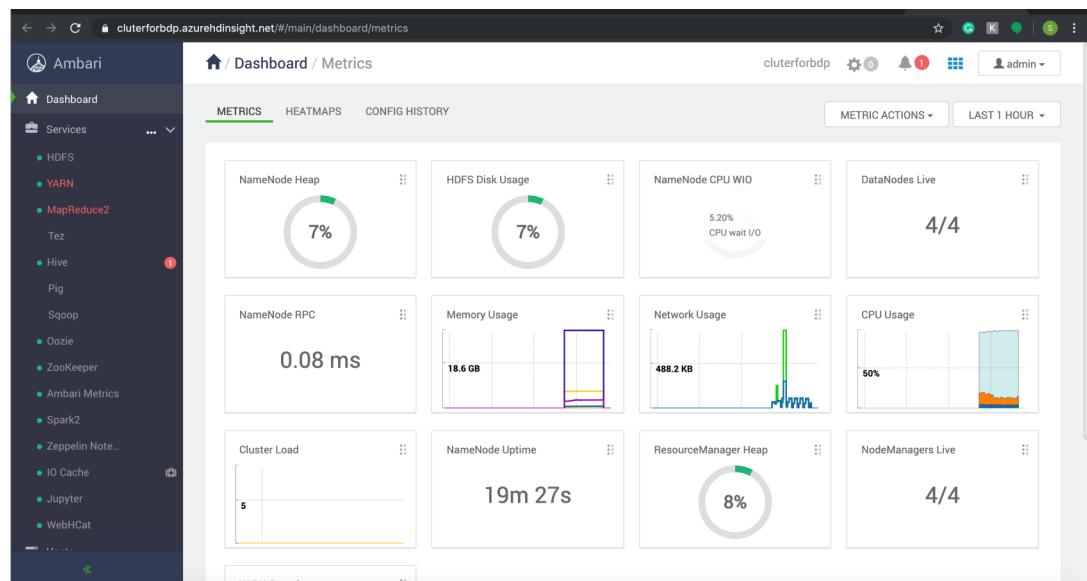


Instructions to update cluster setting and installing external libraries:

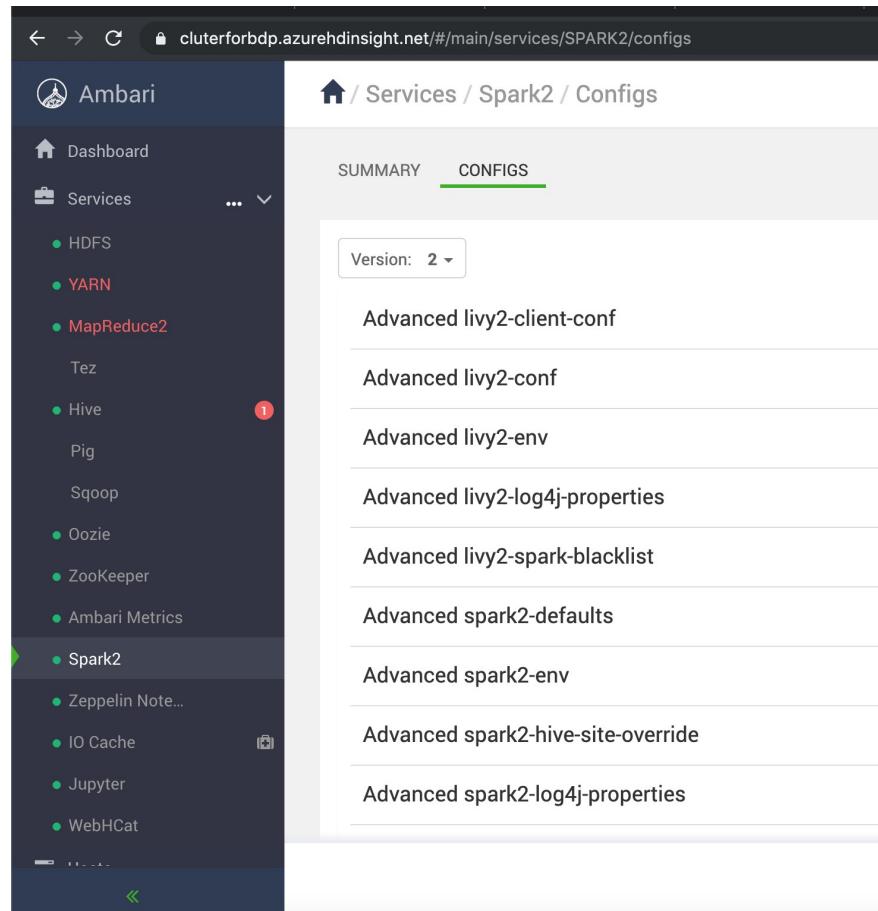
So we will be starting with dashboard to add and update some configuration parameters

Dashboard will look like:

- 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	<input type="button" value=""/>
livy.impersonation.enabled	true	<input type="button" value=""/>
livy.repl.enableHiveContext	true	<input type="button" value=""/>
livy.server.access-control.enabled	true	<input type="button" value=""/>
livy.server.csrf_protection.enabled	true	<input type="button" value=""/>
livy.server.port	8998	<input type="button" value=""/>
livy.server.recovery.mode	recovery	<input type="button" value=""/>
livy.server.recovery.state-store	zookeeper	<input type="button" value=""/>
livy.server.recovery.state-store.url	zk0-cluter.hbyyc3aaebuu5gnemgt2fmvixe.bx.internal.cloudapp.net:2181,zk1-cluter.hbyyc3a	<input type="button" value=""/>
livy.server.session.timeout	3600000	<input type="button" value=""/>
livy.spark.master	yarn-cluster	<input type="button" value=""/>

livy.server.session.timeout

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

- 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
Warning	YARN	yarn.scheduler.maximum-allocation-mb	51200	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-vcores	15	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.

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	<input type="button" value=""/>
livy.server.yarn.app-lookup-timeout	2m	<input type="button" value=""/>

[Add Property ...](#)

Parameters to add:

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

The screenshot shows the Ambari UI for managing configurations. The left sidebar has a tree view with 'Spark2' selected. The main area shows configuration properties for 'spark.sql', 'spark.yarn', and 'spark.yarn.scheduler'. At the bottom right, there are 'DISCARD' and 'SAVE' buttons.

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

The screenshot shows the 'Add Property' dialog. It has fields for 'Type' (set to 'spark2-defaults.xml'), 'Key' (set to 'spark.sql.broadcastTimeout'), and 'Value' (set to '6000'). A 'Property Type' dropdown menu is open, showing options: PASSWORD, USER, GROUP, and TEXT. At the bottom right, there are 'CANCEL' and 'ADD' buttons.

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

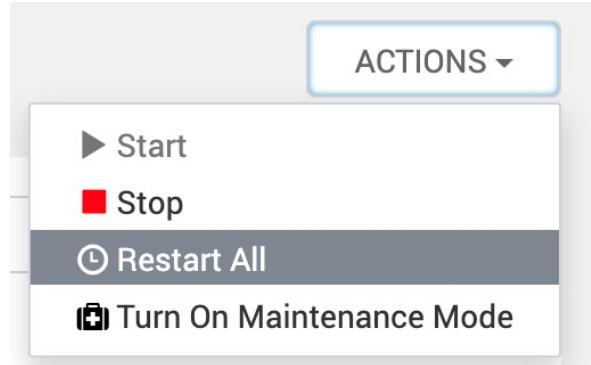
4. Being in Jupyter config

- 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. At the top, there's a blue header bar with the 'Microsoft Azure' logo and a navigation menu icon. Below the header, the URL 'Home > HDInsight_2019-12-09T21.56.56.' is displayed. The main content area features a cluster icon and the name 'CluterForBDP' followed by 'HDInsight cluster'. A search bar with the placeholder 'Search (Cmd+ /)' is present. On the left, a sidebar menu lists various cluster management options under 'Overview' and 'Settings' categories. The 'Script actions' option is highlighted with a light gray background.

- 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
- HDInsight partner

- 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.
- Check the 'Head' and 'Worker' check boxes

URI: <https://bdptwittersparkstorage.blob.core.windows.net/bdptwittersusanth/pip3install-packages-bash.sh>

This script will install required libraries for the project.