SMOKE TEST

JupyterHub with Sparkmagic 2.1

Date Prepared: Oct 2019

Document Information

Project Name	**EPIC Accelerator Deployment & Integration Services**		
Project Owner		Document Version No	0.1
Quality Review Method			
Prepared By	Priyanka	Preparation Date	Oct 2019
Reviewed By		Review Date	

Table of Contents

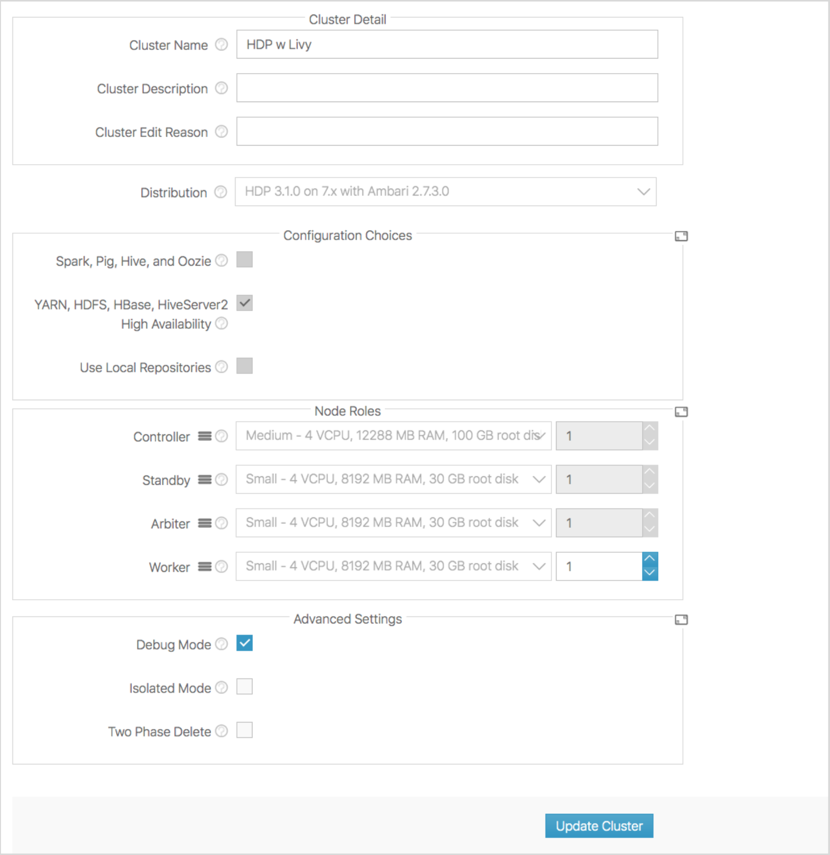
Table of Tables

NO TABLE OF FIGURES ENTRIES FOUND.

1 CONFIGURE AMBARI SPARK2 & LIVY SERVER

In this section, we will configure Ambari Spark2 and Livy Server.

1. Create a Cluster (HDP 3.1.0)



Cluster Detail

Cluster Name: HDP w Livy

Cluster Description:

Cluster Edit Reason:

Distribution: HDP 3.1.0 on 7.x with Ambari 2.7.3.0

Configuration Choices

Spark, Pig, Hive, and Oozie □

YARN, HDFS, HBase, HiveServer2 High Availability ☑

Use Local Repositories □

Node Roles

Controller	Medium - 4 VCPU, 12288 MB RAM, 100 GB root dis	1
Standby	Small - 4 VCPU, 8192 MB RAM, 30 GB root disk	1
Arbiter	Small - 4 VCPU, 8192 MB RAM, 30 GB root disk	1
Worker	Small - 4 VCPU, 8192 MB RAM, 30 GB root disk	1

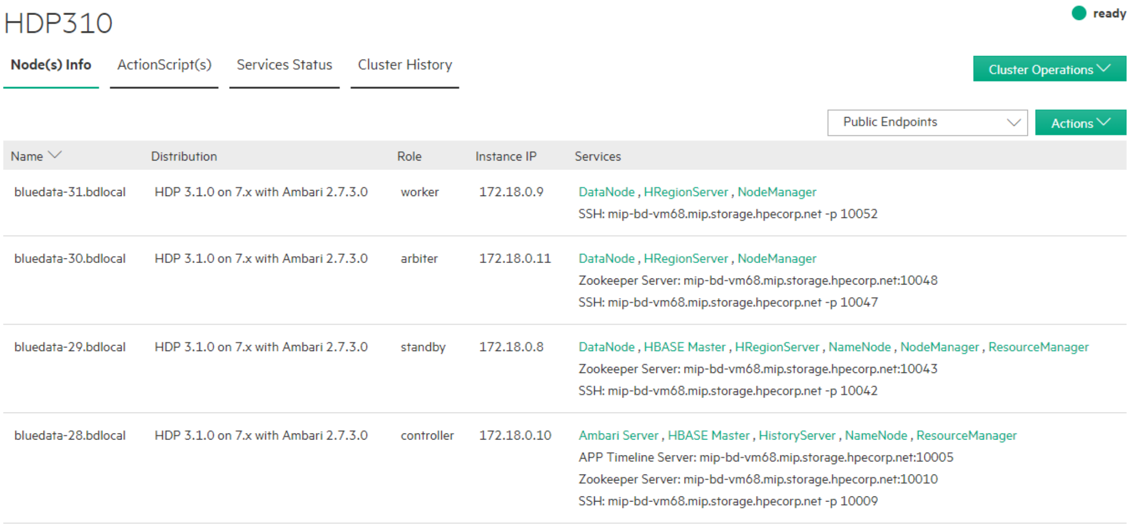
Advanced Settings

Debug Mode ☑

Isolated Mode □

Two Phase Delete □

2. Once cluster is in Ready state, click on **Ambari Server** from the controller role



HDP310

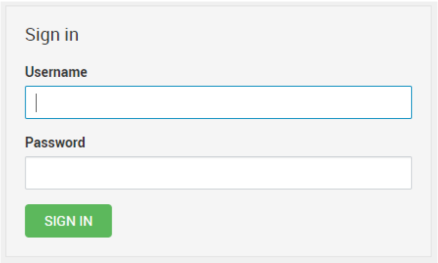
● ready

Node(s) Info | ActionScript(s) | Services Status | Cluster History

Cluster Operations | Public Endpoints | Actions

Name	Distribution	Role	Instance IP	Services
bluedata-31.bdlocal	HDP 3.1.0 on 7.x with Ambari 2.7.3.0	worker	172.18.0.9	DataNode , HRegionServer , NodeManager SSH: mip-bd-vm68.mip.storage.hpecorp.net -p 10052
bluedata-30.bdlocal	HDP 3.1.0 on 7.x with Ambari 2.7.3.0	arbiter	172.18.0.11	DataNode , HRegionServer , NodeManager Zookeeper Server: mip-bd-vm68.mip.storage.hpecorp.net:10048 SSH: mip-bd-vm68.mip.storage.hpecorp.net -p 10047
bluedata-29.bdlocal	HDP 3.1.0 on 7.x with Ambari 2.7.3.0	standby	172.18.0.8	DataNode , HBASE Master , HRegionServer , NameNode , NodeManager , ResourceManager Zookeeper Server: mip-bd-vm68.mip.storage.hpecorp.net:10043 SSH: mip-bd-vm68.mip.storage.hpecorp.net -p 10042
bluedata-28.bdlocal	HDP 3.1.0 on 7.x with Ambari 2.7.3.0	controller	172.18.0.10	Ambari Server , HBASE Master , HistoryServer , NameNode , ResourceManager APP Timeline Server: mip-bd-vm68.mip.storage.hpecorp.net:10005 Zookeeper Server: mip-bd-vm68.mip.storage.hpecorp.net:10010 SSH: mip-bd-vm68.mip.storage.hpecorp.net -p 10009

3. It will navigate you to the Ambari Login page, login using:
   a. Username: admin
   b. Password: admin



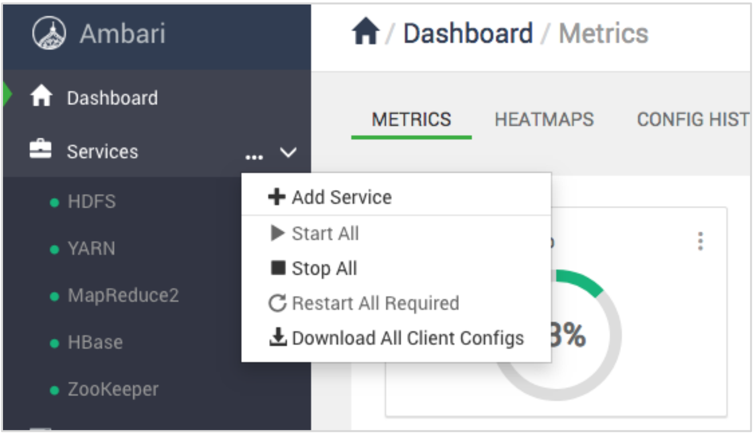
Sign in

Username

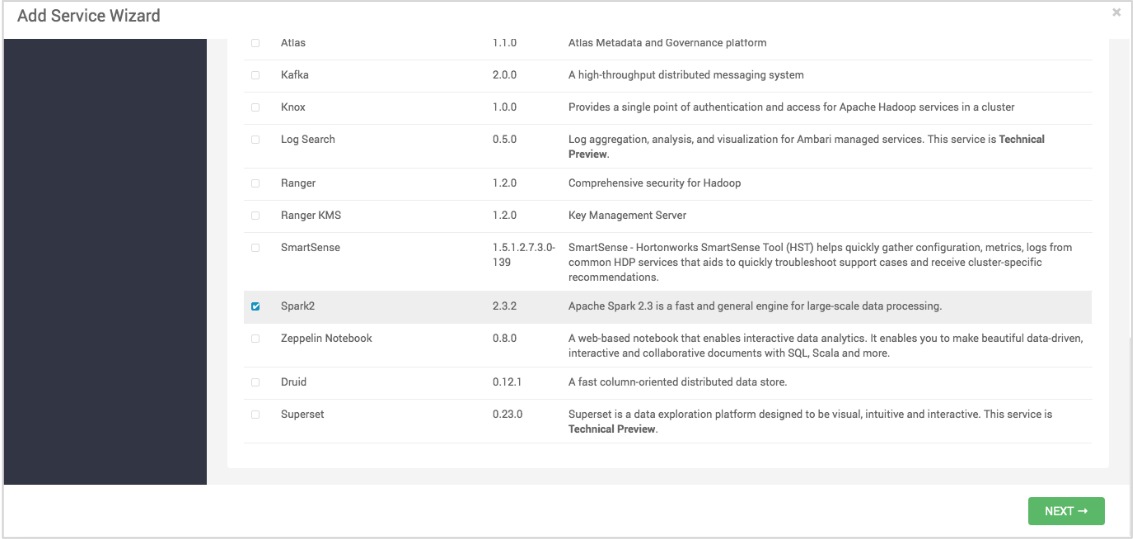
Password

SIGN IN

4. From the left-hand panel, click on **Services** and click on **Add Service**



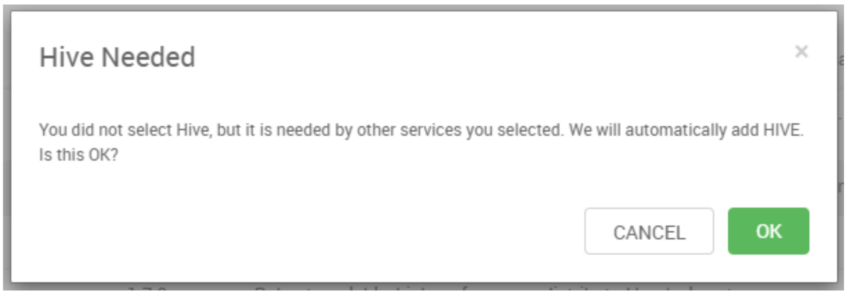
5. From the Add Service Wizard, search for Spark2 and select it, then click on **NEXT**



Add Service Wizard

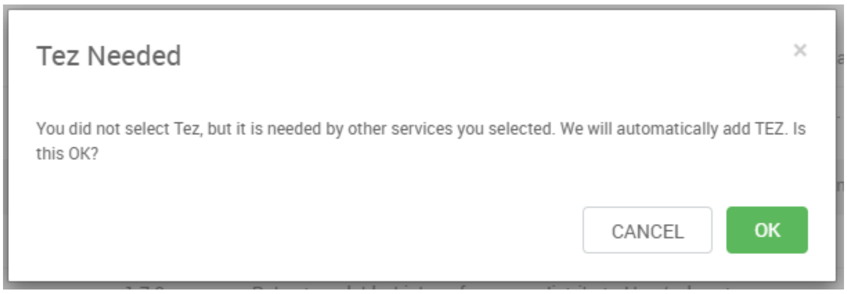
□	Atlas	1.1.0	Atlas Metadata and Governance platform
□	Kafka	2.0.0	A high-throughput distributed messaging system
□	Knox	1.0.0	Provides a single point of authentication and access for Apache Hadoop services in a cluster
□	Log Search	0.5.0	Log aggregation, analysis, and visualization for Ambari managed services. This service is **Technical Preview**.
□	Ranger	1.2.0	Comprehensive security for Hadoop
□	Ranger KMS	1.2.0	Key Management Server
□	SmartSense	1.5.1.2.7.3.0-139	SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
☑	Spark2	2.3.2	Apache Spark 2.3 is a fast and general engine for large-scale data processing.
□	Zeppelin Notebook	0.8.0	A web-based notebook that enables interactive data analytics. It enables you to make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.
□	Druid	0.12.1	A fast column-oriented distributed data store.
□	Superset	0.23.0	Superset is a data exploration platform designed to be visual, intuitive and interactive. This service is **Technical Preview**.

6. Hive needed and Tez needed prompt will pop-up, click on **Ok** for both



Hive Needed

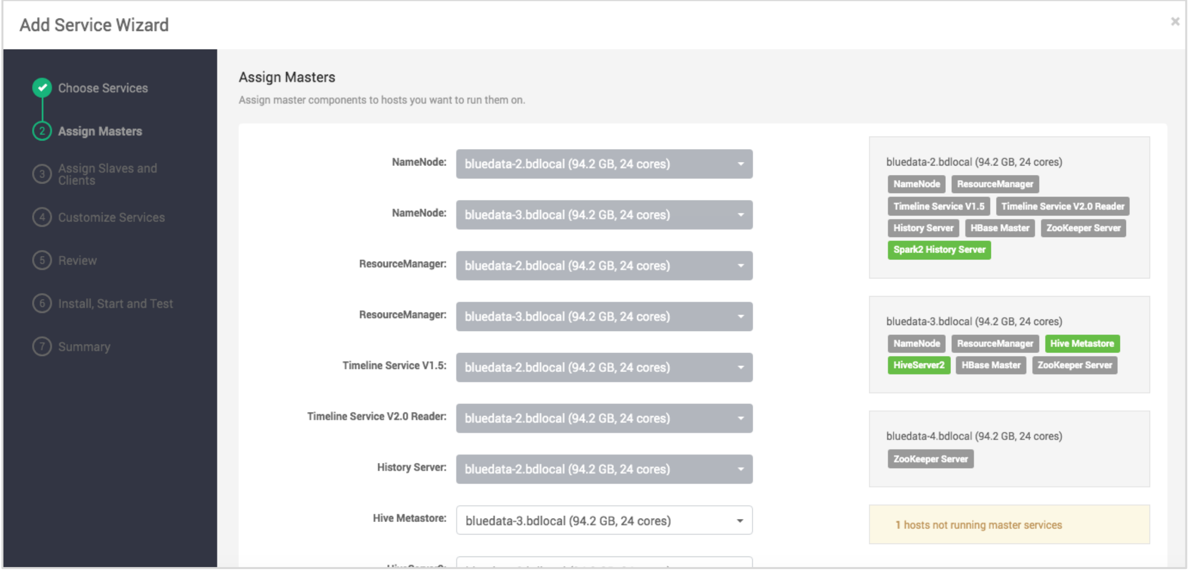
You did not select Hive, but it is needed by other services you selected. We will automatically add HIVE. Is this OK?



Tez Needed

You did not select Tez, but it is needed by other services you selected. We will automatically add TEZ. Is this OK?

7. Click on **NEXT** and accept all default in Assign Masters page. If Spark2 History Server is not adding on controller, manually map it to controller node.



Add Service Wizard

1. Choose Services
2. Assign Masters
3. Assign Slaves and Clients
4. Customize Services
5. Review
6. Install, Start and Test
7. Summary

Assign Masters

Assign master components to hosts you want to run them on.

NameNode: bluedata-2.bdlocal (94.2 GB, 24 cores)
NameNode: bluedata-3.bdlocal (94.2 GB, 24 cores)
ResourceManager: bluedata-2.bdlocal (94.2 GB, 24 cores)
ResourceManager: bluedata-3.bdlocal (94.2 GB, 24 cores)
Timeline Service V1.5: bluedata-2.bdlocal (94.2 GB, 24 cores)
Timeline Service V2.0 Reader: bluedata-2.bdlocal (94.2 GB, 24 cores)
History Server: bluedata-2.bdlocal (94.2 GB, 24 cores)
Hive Metastore: bluedata-3.bdlocal (94.2 GB, 24 cores)

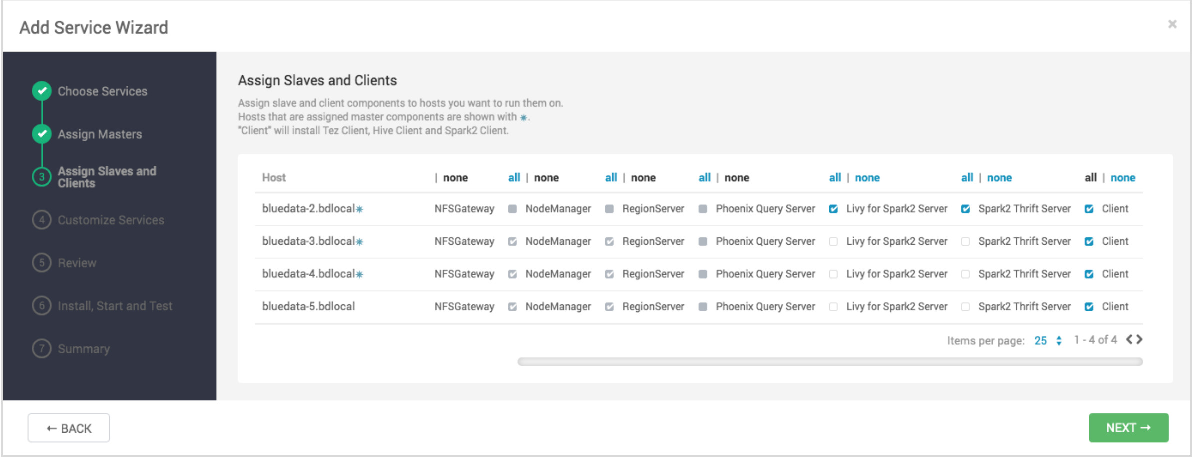
bluedata-2.bdlocal (94.2 GB, 24 cores)
NameNode, ResourceManager, Timeline Service V1.5, Timeline Service V2.0 Reader, History Server, HBase Master, ZooKeeper Server, Spark2 History Server

bluedata-3.bdlocal (94.2 GB, 24 cores)
NameNode, ResourceManager, Hive Metastore, HiveServer2, HBase Master, ZooKeeper Server

bluedata-4.bdlocal (94.2 GB, 24 cores)
ZooKeeper Server

1 hosts not running master services

8. Check on **Livy for Spark2 Server** and **Spark2 Thrift Server** on Controller host and click on **NEXT**



Add Service Wizard

1. Choose Services
2. Assign Masters
3. Assign Slaves and Clients
4. Customize Services
5. Review
6. Install, Start and Test
7. Summary

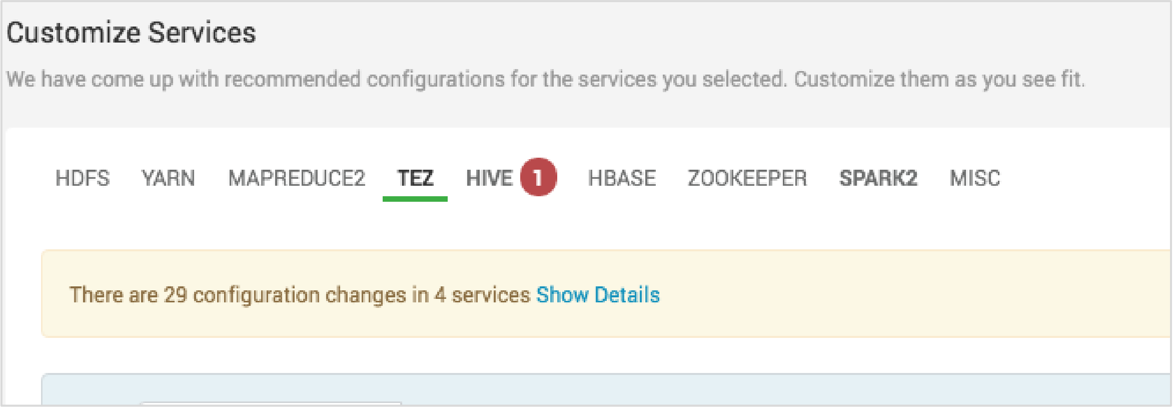
Assign Slaves and Clients

Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with ✱.
"Client" will install Tez Client, Hive Client and Spark2 Client.

Host	| none	all | none	all | none	all | none	all | none	all | none	all | none
bluedata-2.bdlocal✱	NFSGateway	■ NodeManager	■ RegionServer	■ Phoenix Query Server	☑ Livy for Spark2 Server	☑ Spark2 Thrift Server	☑ Client
bluedata-3.bdlocal✱	NFSGateway	☑ NodeManager	☑ RegionServer	■ Phoenix Query Server	☐ Livy for Spark2 Server	☐ Spark2 Thrift Server	☑ Client
bluedata-4.bdlocal✱	NFSGateway	☑ NodeManager	☑ RegionServer	■ Phoenix Query Server	☐ Livy for Spark2 Server	☐ Spark2 Thrift Server	☑ Client
bluedata-5.bdlocal	NFSGateway	☑ NodeManager	☑ RegionServer	■ Phoenix Query Server	☐ Livy for Spark2 Server	☐ Spark2 Thrift Server	☑ Client

Items per page: 25 1 - 4 of 4

9. A error will occur on Hive, click on it

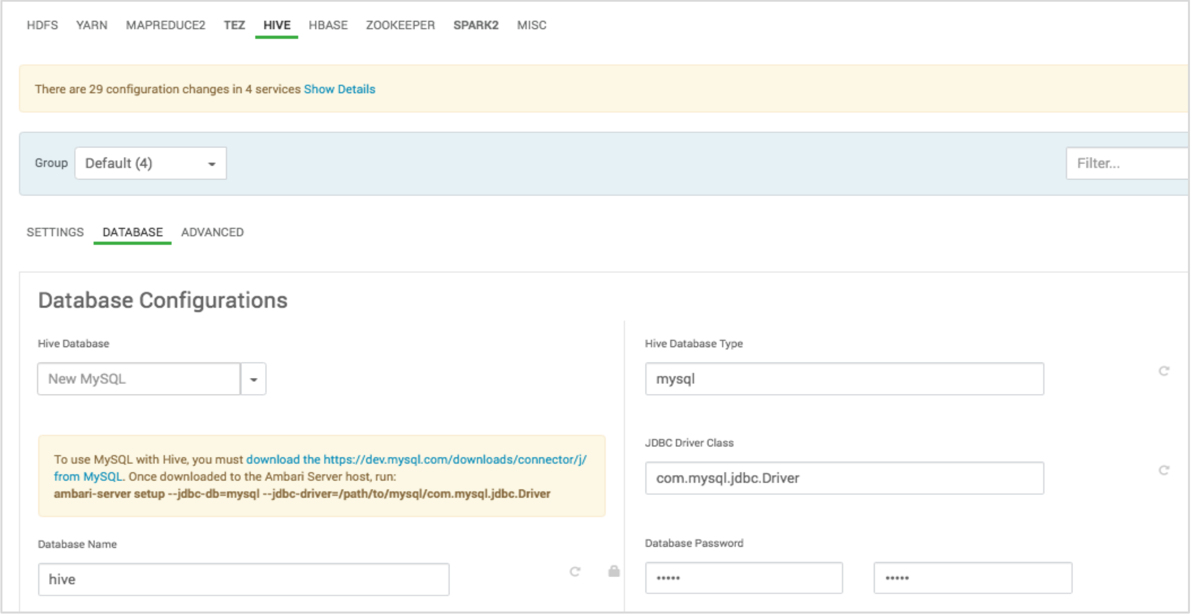


Customize Services

We have come up with recommended configurations for the services you selected. Customize them as you see fit.

There are 29 configuration changes in 4 services Show Details

10. You will get a page like below. Click on **DATABASE** tab and provide any password for **Database Password**



There are 29 configuration changes in 4 services Show Details

Group Default (4)

Filter...

Database Configurations

Hive Database: New MySQL

To use MySQL with Hive, you must download the https://dev.mysql.com/downloads/connector/j/ from MySQL. Once downloaded to the Ambari Server host, run:
ambari-server setup --jdbc-db=mysql --jdbc-driver=/path/to/mysql/com.mysql.jdbc.Driver

Database Name: hive

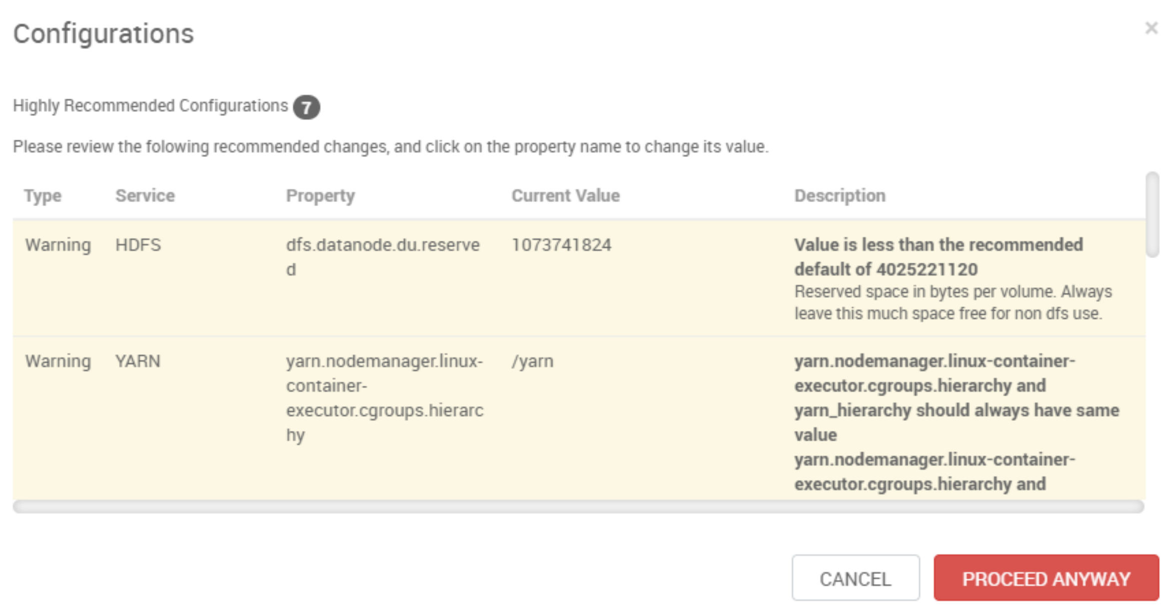
Hive Database Type: mysql

JDBC Driver Class: com.mysql.jdbc.Driver

Database Password: ••••• •••••

11. Click on **Next**

12. Ignore the Configuration warning. Click on **Proceed Anyway**



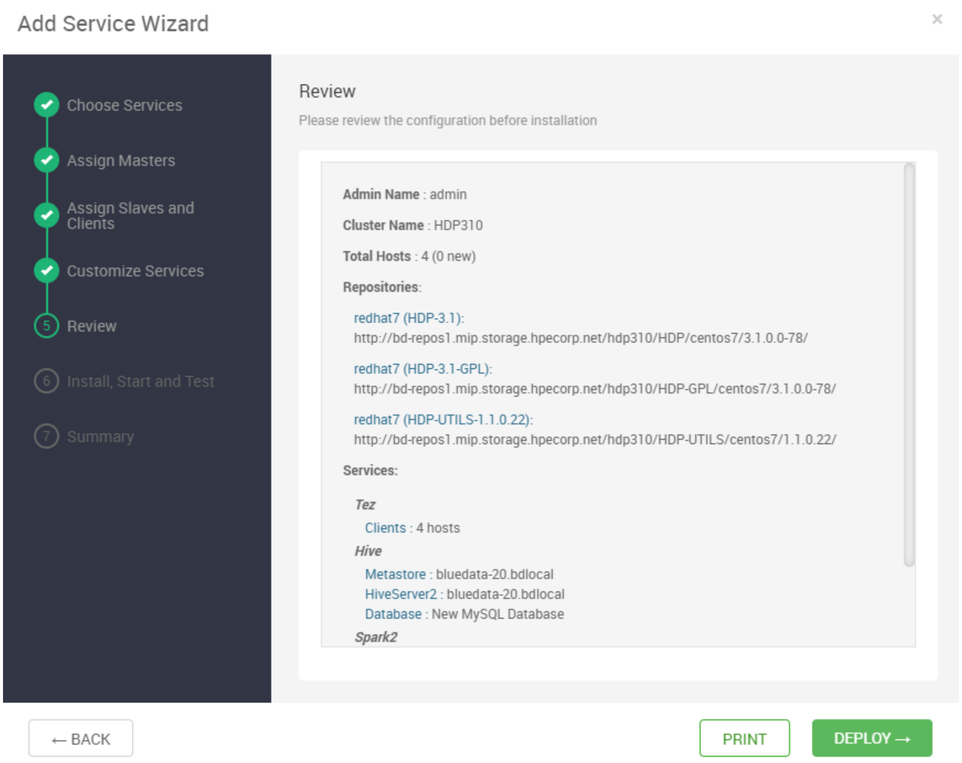
Configurations

Highly Recommended Configurations 7

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

Type	Service	Property	Current Value	Description
Warning	HDFS	dfs.datanode.du.reserved	1073741824	**Value is less than the recommended default of 4025221120** Reserved space in bytes per volume. Always leave this much space free for non dfs use.
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**

13. A Review page will come, click on **DEPLOY**



Add Service Wizard

Review

Please review the configuration before installation

Admin Name : admin

Cluster Name : HDP310

Total Hosts : 4 (0 new)

Repositories:

redhat7 (HDP-3.1):
http://bd-repos1.mip.storage.hpecorp.net/hdp310/HDP/centos7/3.1.0.0-78/

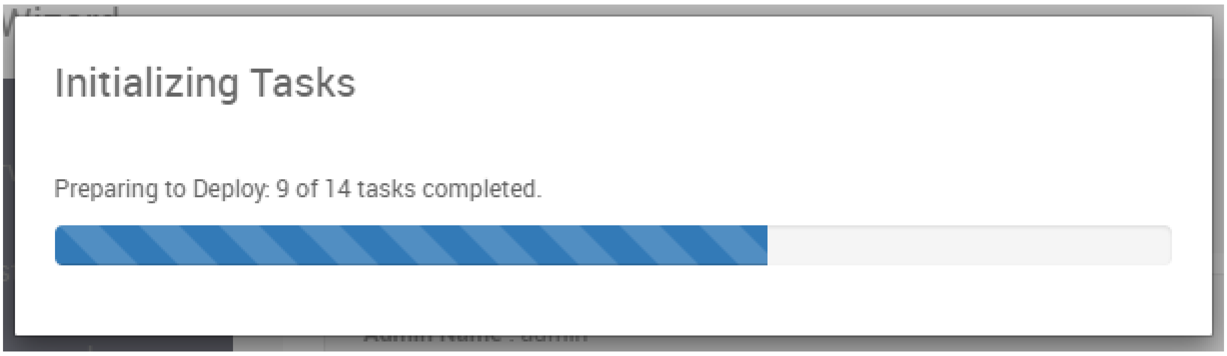
redhat7 (HDP-3.1-GPL):
http://bd-repos1.mip.storage.hpecorp.net/hdp310/HDP-GPL/centos7/3.1.0.0-78/

redhat7 (HDP-UTILS-1.1.0.22):
http://bd-repos1.mip.storage.hpecorp.net/hdp310/HDP-UTILS/centos7/1.1.0.22/

Services:

Tez
Clients : 4 hosts
Hive
Metastore : bluedata-20.bdlocal
HiveServer2 : bluedata-20.bdlocal
Database : New MySQL Database
Spark2

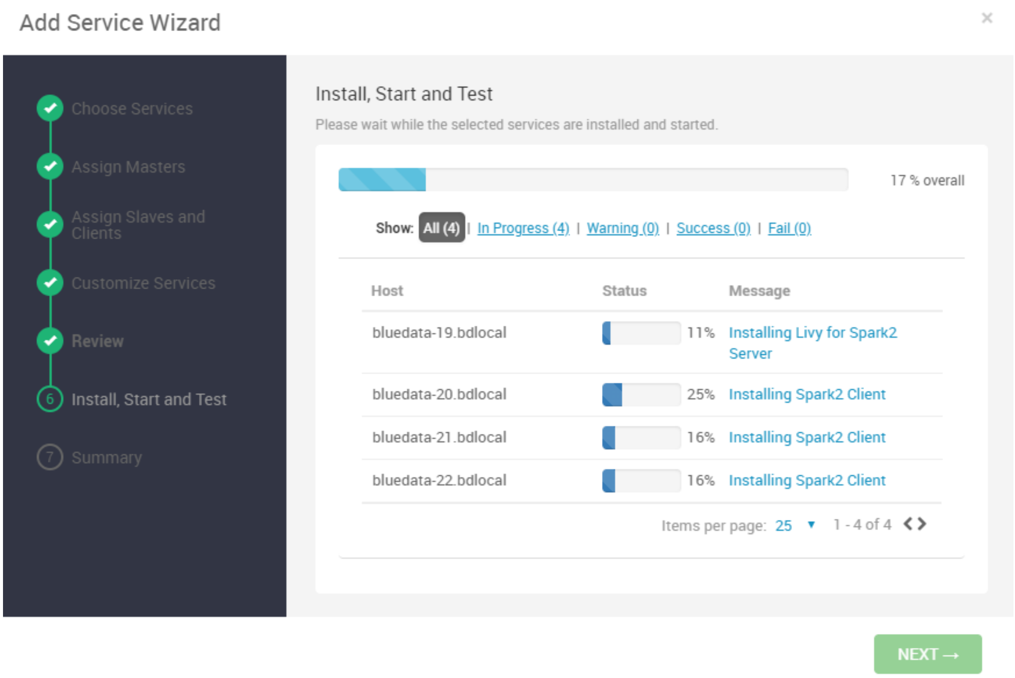
14. Initializing Tasks process starts now



Initializing Tasks

Preparing to Deploy: 9 of 14 tasks completed.

15. You will get a Install, Start and Test page, wait till all components get installed



Add Service Wizard

×
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
6 Install, Start and Test
7 Summary

Install, Start and Test

Please wait while the selected services are installed and started.

17 % overall

Show: All (4) | In Progress (4) | Warning (0) | Success (0) | Fail (0)

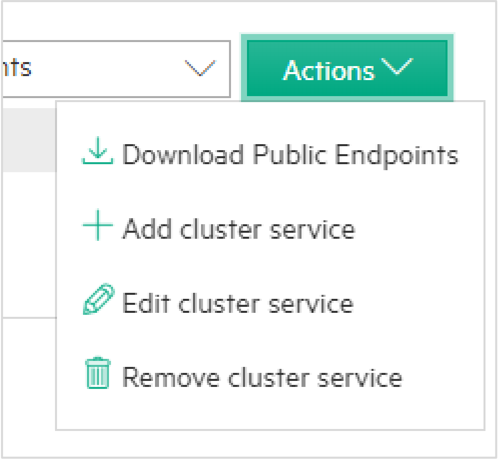
Host	Status	Message
bluedata-19.bdlocal	11%	Installing Livy for Spark2 Server
bluedata-20.bdlocal	25%	Installing Spark2 Client
bluedata-21.bdlocal	16%	Installing Spark2 Client
bluedata-22.bdlocal	16%	Installing Spark2 Client

Items per page: 25 ▾ 1 - 4 of 4 < >
NEXT →

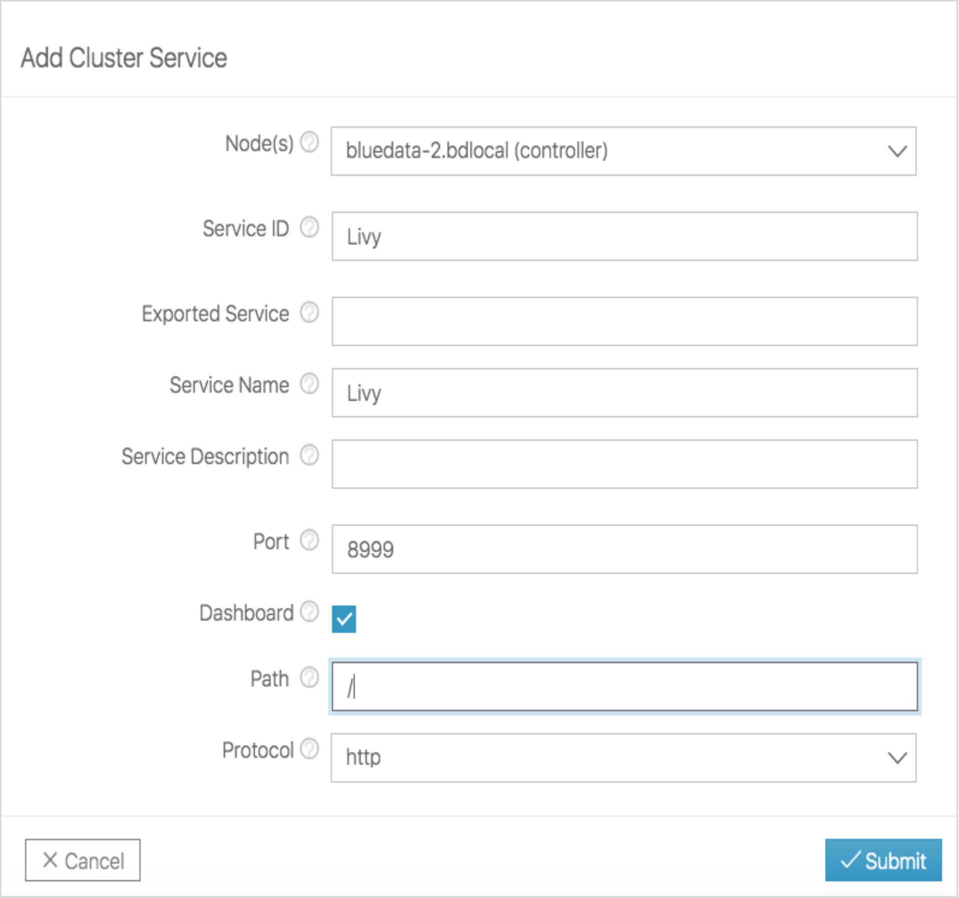
16. Once all components are installed, click on **Next** > **Complete** > **Restart** > **all Required services**

17. In order to get Livy server on EPIC cluster as Service, navigate back to EPIC cluster page, click on **Actions**

18. A drop-down menu will appear, click on **Add cluster service**



19. Provide the following details and click on **Submit**



Add Cluster Service

Node(s): bluedata-2.bdlocal (controller)

Service ID: Livy

Exported Service:

Service Name: Livy

Service Description:

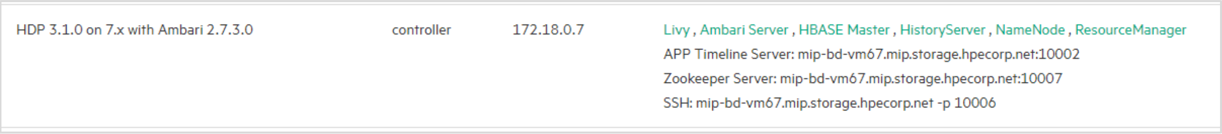
Port: 8999

Dashboard: ☑

Path: /

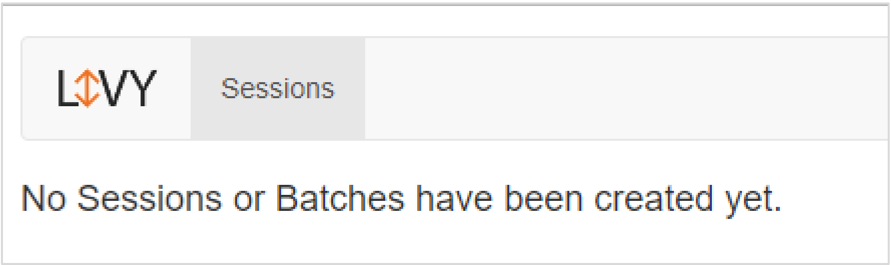
Protocol: http

20. Once submitted, you can see Livy service on Controller



HDP 3.1.0 on 7.x with Ambari 2.7.3.0	controller	172.18.0.7	Livy , Ambari Server , HBASE Master , HistoryServer , NameNode , ResourceManager APP Timeline Server: mip-bd-vm67.mip.storage.hpecorp.net:10002 Zookeeper Server: mip-bd-vm67.mip.storage.hpecorp.net:10007 SSH: mip-bd-vm67.mip.storage.hpecorp.net -p 10006

21. Click on Livy, it will navigate you to a new window



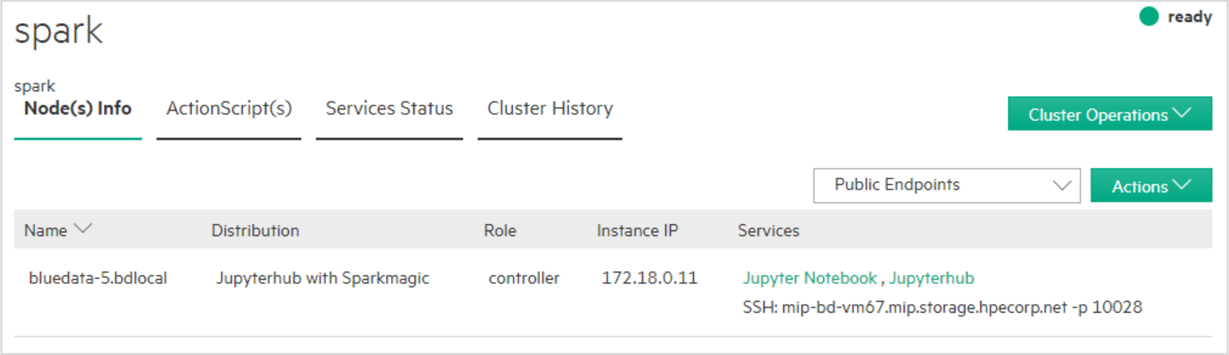
No Sessions or Batches have been created yet.

2 TESTING SPARKMAGIC

In this section, we will test Sparkmagic.

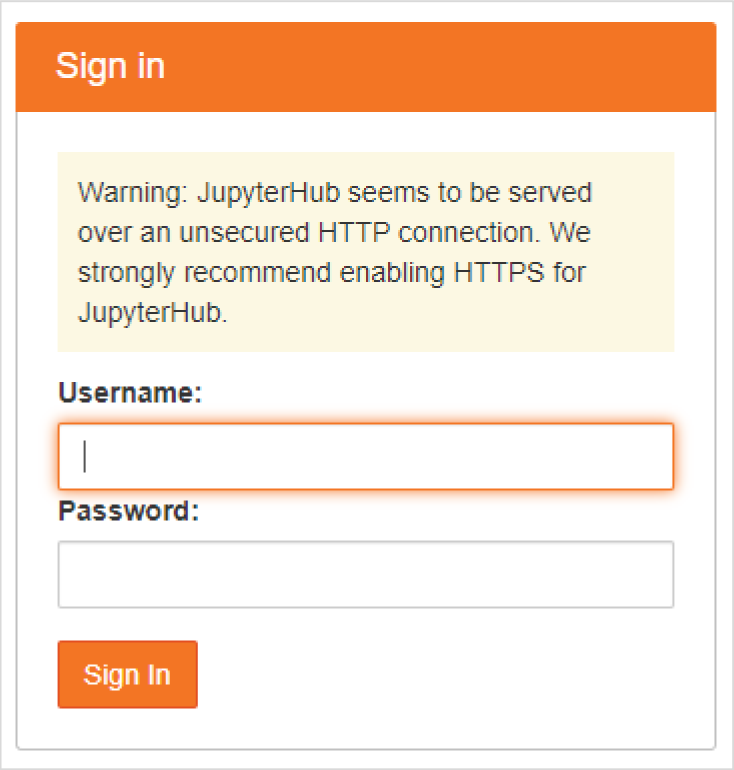
2.1 Login to JupyterHub

1. From JupyterHub with Sparkmagic cluster, click on **JupyterHub** service



Name	Distribution	Role	Instance IP	Services
bluedata-5.bdlocal	Jupyterhub with Sparkmagic	controller	172.18.0.11	Jupyter Notebook , Jupyterhub SSH: mip-bd-vm67.mip.storage.hpecorp.net -p 10028

2. It will navigate you to JupyterHub login page, login using your credentials



Sign in

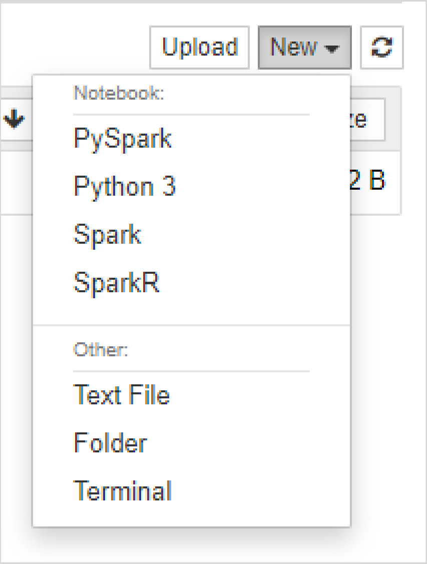
Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub.

Username:

Password:

2.2 Create new Notebook – PySpark

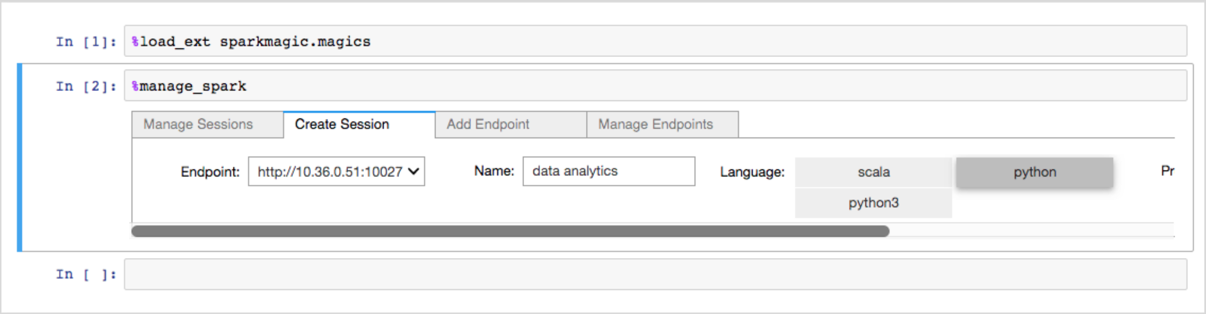
1. Click on **New**, a drop-down menu will appear, click on **PySpark**. It will navigate you to a new Jupyter Notebook



2. Execute the below command:

```
%load_ext sparkmagic.magics
```

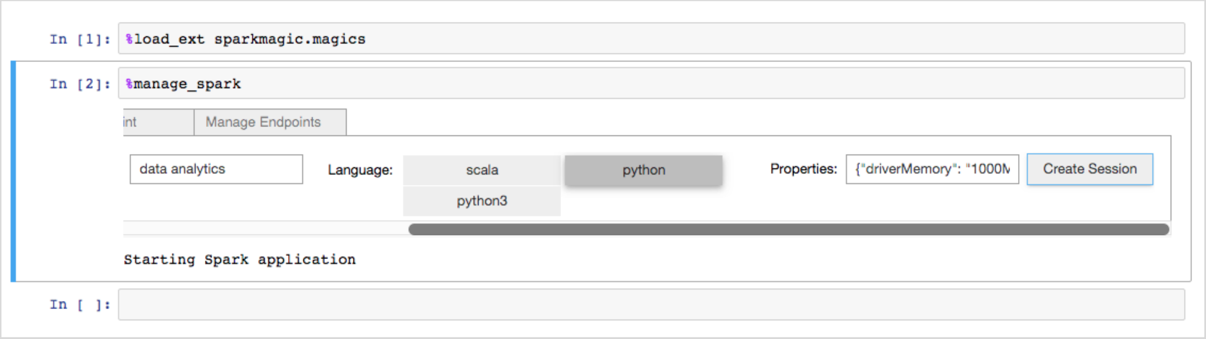
```
%manage_spark
```



```
In [1]: %load_ext sparkmagic.magics
In [2]: %manage_spark
```

Endpoint: http://10.36.0.51:10027
Name: data analytics
Language: python

```
In [ ]:
```

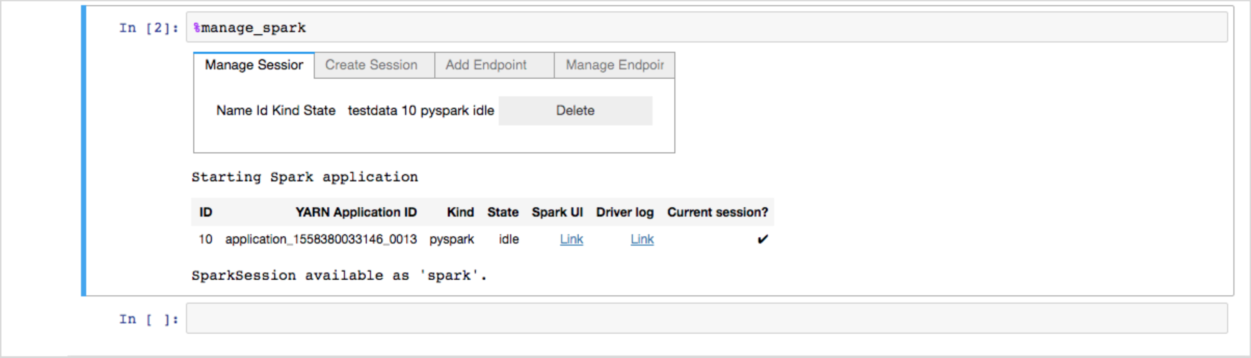


```
In [1]: %load_ext sparkmagic.magics
In [2]: %manage_spark
```

```
Starting Spark application
In [ ]:
```

Note: If you don’t see Endpoint created automatically, you can add Endpoint manually or we can define in config.json (Here, Endpoint URL is: **<Livy_server_URL>** from the EPIC Cluster)

3. Click on **Create Session** (You may have to scroll right to see the option), in some time Spark session will be available



```
In [2]: %manage_spark
```

```
Starting Spark application
```

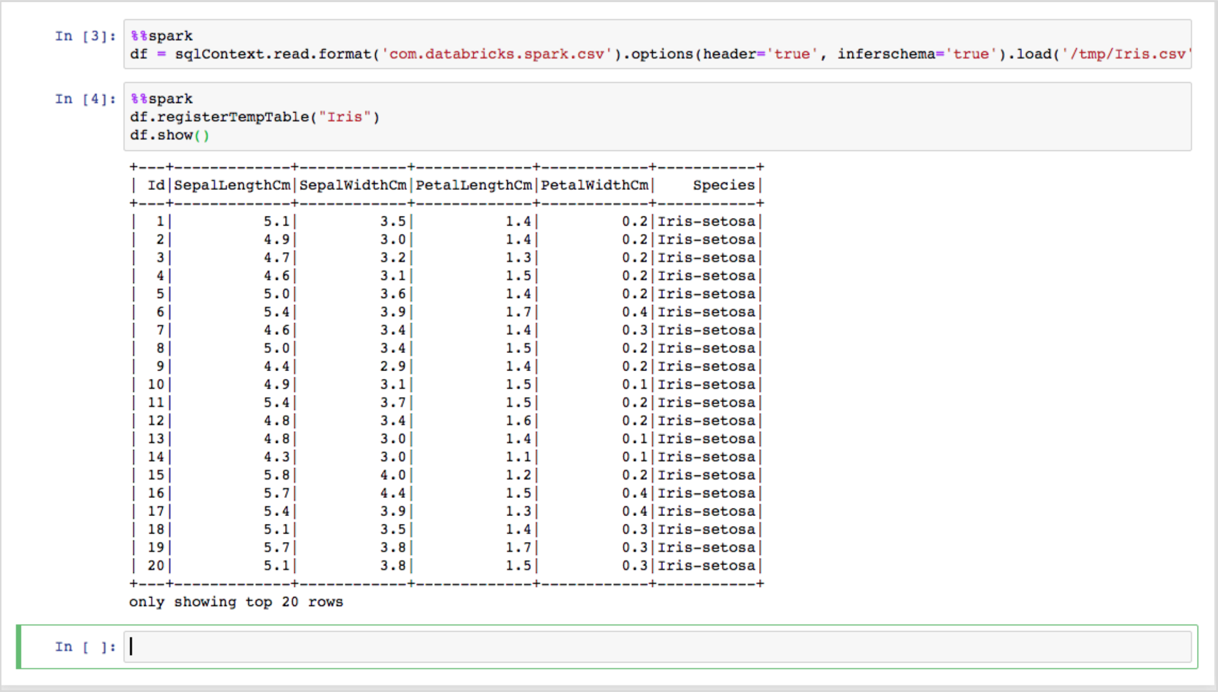
ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
10	application_1558380033146_0013	pyspark	idle	Link	Link	✔

```
SparkSession available as 'spark'.
In [ ]:
```

4. Use sample PySpark code to load the data from HDFS

```
%%spark
df =
sqlContext.read.format('com.databricks.spark.csv').options(head
er='true', inferschema='true').load('/tmp/Iris.csv')
```

```
%%spark
df.registerTempTable("Iris")
df.show()
```



```
In [3]: %%spark
        df = sqlContext.read.format('com.databricks.spark.csv').options(header='true', inferschema='true').load('/tmp/Iris.csv'

In [4]: %%spark
        df.registerTempTable("Iris")
        df.show()

        +---+-------------+------------+-------------+------------+-----------+
        | Id|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|    Species|
        +---+-------------+------------+-------------+------------+-----------+
        |  1|          5.1|         3.5|          1.4|         0.2|Iris-setosa|
        |  2|          4.9|         3.0|          1.4|         0.2|Iris-setosa|
        |  3|          4.7|         3.2|          1.3|         0.2|Iris-setosa|
        |  4|          4.6|         3.1|          1.5|         0.2|Iris-setosa|
        |  5|          5.0|         3.6|          1.4|         0.2|Iris-setosa|
        |  6|          5.4|         3.9|          1.7|         0.4|Iris-setosa|
        |  7|          4.6|         3.4|          1.4|         0.3|Iris-setosa|
        |  8|          5.0|         3.4|          1.5|         0.2|Iris-setosa|
        |  9|          4.4|         2.9|          1.4|         0.2|Iris-setosa|
        | 10|          4.9|         3.1|          1.5|         0.1|Iris-setosa|
        | 11|          5.4|         3.7|          1.5|         0.2|Iris-setosa|
        | 12|          4.8|         3.4|          1.6|         0.2|Iris-setosa|
        | 13|          4.8|         3.0|          1.4|         0.1|Iris-setosa|
        | 14|          4.3|         3.0|          1.1|         0.1|Iris-setosa|
        | 15|          5.8|         4.0|          1.2|         0.2|Iris-setosa|
        | 16|          5.7|         4.4|          1.5|         0.4|Iris-setosa|
        | 17|          5.4|         3.9|          1.3|         0.4|Iris-setosa|
        | 18|          5.1|         3.5|          1.4|         0.3|Iris-setosa|
        | 19|          5.7|         3.8|          1.7|         0.3|Iris-setosa|
        | 20|          5.1|         3.8|          1.5|         0.3|Iris-setosa|
        +---+-------------+------------+-------------+------------+-----------+
        only showing top 20 rows

In [ ]:
```

Note: You should have **Iris.csv** in controller Node.

Note: In order to use the curl command to submit jobs in Notebook use the ! (Bang) in the beginning.