

Runbook for Spark 3.1.1 with Livy & JupyterHub

Date Prepared: Mar 2021

	Spark 3.1.1 with Livy & JupyterHub	 Hewlett Packard Enterprise
--	---	--

Document Information

Project Name	HPE		
Project Owner		Document Version No	0.1
Quality Review Method			
Prepared By	suhe1.shaikh@hpe.com	Preparation Date	23-Mar-2021
Reviewed By	vivek-singh.bhadauriya@hpe.com	Review Date	24-Mar-2021

Table of Contents

1 SUMMARY 4

2 CREATE SPARK 311 CLUSTER..... 5

3 SUBMIT A SPARK JOB FROM LDAP USER 6

 3.1 SUBMITTING SPARK-PI JOB (SPARK PI JAR FROM LOCAL SYSTEM)6

 3.2 SUBMITTING SPARK-PI JOB (SPARK PI JAR FROM DTAP).....7

 3.3 SUBMITTING SPARK-PI JOB IN CLUSTER MODE8

4 SPARK INTERECTIVIE SESSION USING LIVY THROUGH NOTEBOOK..... 11

 4.1 LOGIN TO JUPYTER HUB..... 11

 4.2 PYSARK NOTEBOOK 12

 4.3 SCALA NOTEBOOK..... 13

 4.4 SPARKR NOTEBOOK 15

5 PASS RUNTIME DEPENDENCIES TO SPARK JOB..... 17

 5.1 PYTHON DEPENDENCIES..... 17

 5.2 JAR DEPENDENCIES..... 18

6 KERBEROS TESTING 19

7 SPARK THRIFT SERVER..... 20

	Spark 3.1.1 with Livy & JupyterHub	 Hewlett Packard Enterprise
--	---	--

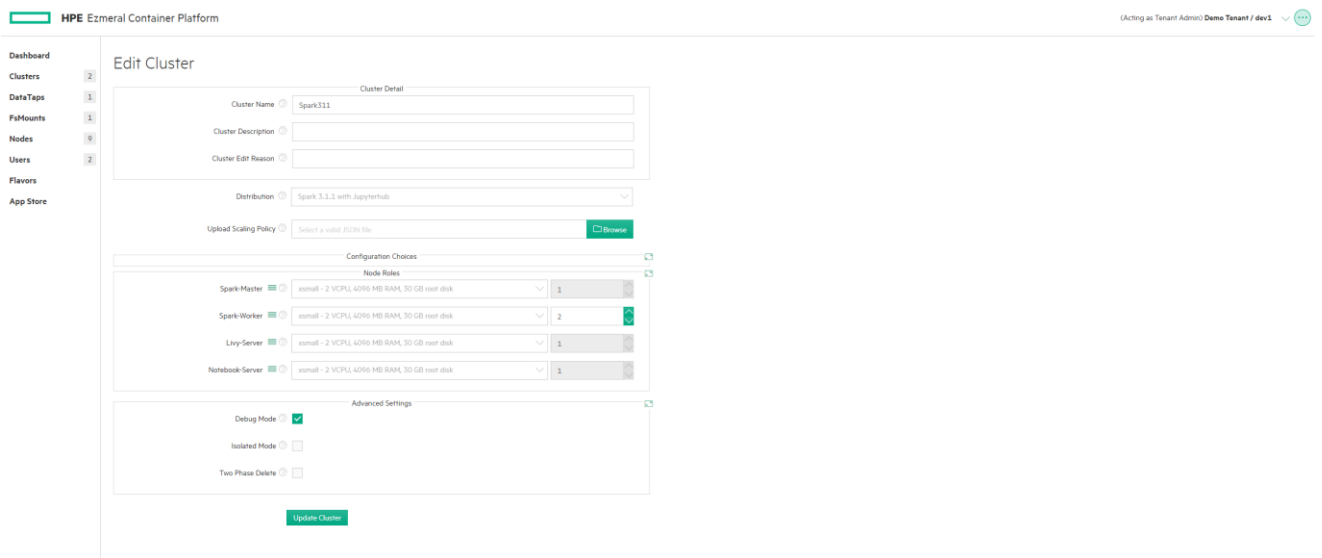
1 SUMMARY

This runbook will demonstrate creating a spark 3.1.1 cluster and different ways to submit Spark jobs.

Bin: <https://hpecp-engineering.s3.us-east-2.amazonaws.com/Spark311/bdcatalog-centos7-ezmeral-spark311-1.0.11.bin>

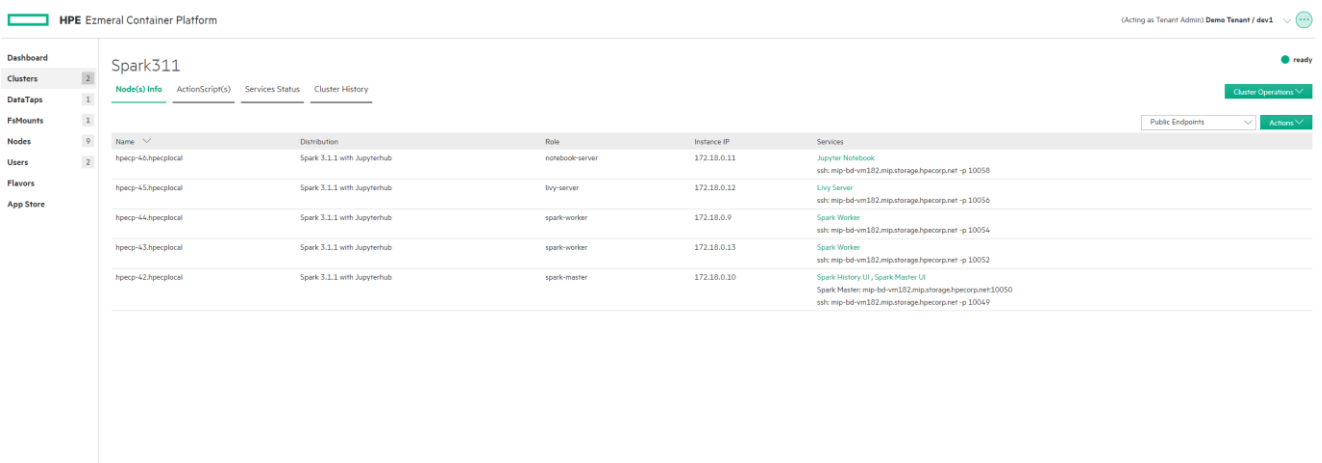
2 CREATE SPARK 311 CLUSTER

Login to HPECP using any LDAP user and create spark 3.1.1 cluster.



The screenshot shows the 'Edit Cluster' form in the HPE Ezmeral Container Platform. The form is titled 'Edit Cluster' and has a sidebar on the left with navigation links: Dashboard, Clusters, DataTaps, FileMounts, Nodes, Users, Flavors, and App Store. The main form area contains several sections: 'Cluster Detail' with fields for Cluster Name (Spark311), Cluster Description, and Cluster Edit Reason; 'Distribution' with a dropdown menu set to 'Spark 3.1.1 with Jupyterhub'; 'Upload Scaling Policy' with a text input and a 'Browse' button; 'Configuration Choices' with a 'Node Roles' section containing four rows: Spark Master, Spark Worker, Livy Server, and Notebook Server, each with a configuration dropdown and a quantity input; and 'Advanced Settings' with checkboxes for Debug Mode (checked), Isolated Mode, and Two Phase Delete. An 'Update Cluster' button is at the bottom.

Wait for the cluster state to change to ready.



The screenshot shows the 'Spark311' cluster details page in the HPE Ezmeral Container Platform. The page has a sidebar on the left with navigation links: Dashboard, Clusters, DataTaps, FileMounts, Nodes, Users, Flavors, and App Store. The main content area shows the cluster name 'Spark311' and a 'ready' status. Below this is a table with columns: Name, Distribution, Role, Instance IP, and Services. The table lists six nodes: hpecp-40.hpecglocal (notebook-server), hpecp-45.hpecglocal (livy-server), hpecp-44.hpecglocal (spark-worker), hpecp-43.hpecglocal (spark-worker), hpecp-42.hpecglocal (spark-master), and hpecp-41.hpecglocal (spark-master). The 'Services' column lists the services running on each node, including Jupyter Notebook, Livy Server, Spark Worker, and Spark Master UI.

Name	Distribution	Role	Instance IP	Services
hpecp-40.hpecglocal	Spark 3.1.1 with Jupyterhub	notebook-server	172.18.0.11	Jupyter Notebook ssh: msp-bd-vm102.mps.storage.hpecorp.net - p 10058
hpecp-45.hpecglocal	Spark 3.1.1 with Jupyterhub	livy-server	172.18.0.12	Livy Server ssh: msp-bd-vm102.mps.storage.hpecorp.net - p 10056
hpecp-44.hpecglocal	Spark 3.1.1 with Jupyterhub	spark-worker	172.18.0.9	Spark Worker ssh: msp-bd-vm102.mps.storage.hpecorp.net - p 10054
hpecp-43.hpecglocal	Spark 3.1.1 with Jupyterhub	spark-worker	172.18.0.13	Spark Worker ssh: msp-bd-vm102.mps.storage.hpecorp.net - p 10052
hpecp-42.hpecglocal	Spark 3.1.1 with Jupyterhub	spark-master	172.18.0.10	Spark History UI, Spark Master UI Spark Master: msp-bd-vm102.mps.storage.hpecorp.net:10050 ssh: msp-bd-vm102.mps.storage.hpecorp.net - p 10049

3 SUBMIT A SPARK JOB FROM LDAP USER

3.1 Submitting Spark-PI Job (Spark PI Jar from Local System)

Login to spark 3.1.1 master node using your LDAP user.

```
[root@mp-hd-vm181 ~]#
[root@mp-hd-vm181 ~]#
[root@mp-hd-vm181 ~]# ssh dev1@172.18.0.10
dev1@172.18.0.10's password:
Last login: Tue Mar 23 03:27:27 2021 from 172.18.0.2
[dev1@hpecp-42 ~]$
[dev1@hpecp-42 ~]$
[dev1@hpecp-42 ~]$
```

Submit a Spark pi job to spark. Locate your main application file on your local file system.


Command:

```
spark-submit --deploy-mode client --class org.apache.spark.examples.SparkPi /usr/lib/spark/spark-3.1.1-bin-hadoop2.7/examples/jars/spark-examples_2.12-3.1.1.jar 100
```

Output:

```
[dev1@hpecp-42 ~]$
[dev1@hpecp-42 ~]$ spark-submit --deploy-mode client --class org.apache.spark.examples.SparkPi /usr/lib/spark/spark-3.1.1-bin-hadoop2.7/examples/jars/spark-examples_2.12-3.1.1.jar 100
21/03/23 03:28:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/03/23 03:28:39 INFO SparkContext: Running Spark version 3.1.1
21/03/23 03:28:39 INFO ResourceUtils:
21/03/23 03:28:39 INFO ResourceUtils: No custom resources configured for spark.driver.
21/03/23 03:28:39 INFO ResourceUtils:
21/03/23 03:28:39 INFO SparkContext: Submitted application: Spark Pi
21/03/23 03:28:39 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offheap -> name: offheap, amount: 0, script: , vendor: ), task resources: Map(cpu -> name: cpu, amount: 1.0)
21/03/23 03:28:39 INFO ResourceProfile: Limiting resource is cpu
21/03/23 03:28:39 INFO ResourceProfileManager: Added ResourceProfile id: 0
21/03/23 03:28:40 INFO SecurityManager: Changing view acls to: dev1
21/03/23 03:28:40 INFO SecurityManager: Changing modify acls to: dev1
21/03/23 03:28:40 INFO SecurityManager: Changing view acls groups to:
21/03/23 03:28:40 INFO SecurityManager: Changing modify acls groups to:
21/03/23 03:28:40 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(dev1); groups with view permissions: Set(); users with modify permissions: Set(dev1); groups with modify permissions: Set()
21/03/23 03:28:40 INFO Utils: Successfully started service 'sparkDriver' on port 43201.
21/03/23 03:28:40 INFO SparkEnv: Registering MapOutputTracker
21/03/23 03:28:40 INFO SparkEnv: Registering BlockManagerMaster
21/03/23 03:28:40 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
21/03/23 03:28:40 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
21/03/23 03:28:40 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
21/03/23 03:28:40 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-5439164d-5f9f-4dce-970a-273aaf1dfa87
21/03/23 03:28:40 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
21/03/23 03:28:40 INFO SparkEnv: Registering OutputCommitCoordinator
21/03/23 03:28:40 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/03/23 03:28:40 INFO Utils: Successfully started service 'SparkUI' on port 4041.
21/03/23 03:28:40 INFO SparkUI: Bound SparkUI to 0.0.0.0 and started at http://hpecp-42.hpecplocal:4041
21/03/23 03:28:40 INFO SparkContext: Added JAR file:///opt/bluedata/bluedata-dtap.jar at spark://hpecp-42.hpecplocal:43201/jars/bluedata-dtap.jar with timestamp 1616495319869
21/03/23 03:28:40 INFO SparkContext: Added JAR file:/usr/lib/spark/spark-3.1.1-bin-hadoop2.7/examples/jars/spark-examples_2.12-3.1.1.jar at spark://hpecp-42.hpecplocal:43201/jars/spark-examples_2.12-3.1.1.jar with timestamp 1616495319869
21/03/23 03:28:40 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://hpecp-42.hpecplocal:7077...
21/03/23 03:28:40 INFO TransportClientFactory: Successfully created connection to hpecp-42.hpecplocal/172.18.0.10:7077 after 30 ms (0 ms spent in bootstraps)
21/03/23 03:28:40 INFO StandaloneSchedulerBackend: Connected to Spark cluster with app ID app-20210323032840-0019
21/03/23 03:28:40 INFO StandaloneAppClient$ClientEndpoint: Executor added: app-20210323032840-0019/0 on worker-20210319032747-172.18.0.13:42598 (172.18.0.13:42598) with 1 core(s)
21/03/23 03:28:40 INFO StandaloneSchedulerBackend: Granted executor ID app-20210323032840-0019/0 on hostPort 172.18.0.13:42598 with 1 core(s), 1024.0 MiB RAM
21/03/23 03:28:40 INFO StandaloneAppClient$ClientEndpoint: Executor added: app-20210323032840-0019/1 on worker-20210319032746-172.18.0.9:38291 (172.18.0.9:38291) with 1 core(s)
21/03/23 03:28:40 INFO StandaloneSchedulerBackend: Granted executor ID app-20210323032840-0019/1 on hostPort 172.18.0.9:38291 with 1 core(s), 1024.0 MiB RAM
21/03/23 03:28:40 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 33200.
21/03/23 03:28:40 INFO NettyBlockTransferService: Server created on hpecp-42.hpecplocal:33200
21/03/23 03:28:40 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
21/03/23 03:28:40 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, hpecp-42.hpecplocal, 33200, None)
21/03/23 03:28:40 INFO BlockManagerMasterEndpoint: Registering block manager hpecp-42.hpecplocal:33200 with 366.3 MiB RAM, BlockManagerId(driver, hpecp-42.hpecplocal, 33200, None)
21/03/23 03:28:40 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, hpecp-42.hpecplocal, 33200, None)
21/03/23 03:28:40 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, hpecp-42.hpecplocal, 33200, None)
21/03/23 03:28:40 INFO StandaloneAppClient$ClientEndpoint: Executor updated: app-20210323032840-0019/1 is now RUNNING
21/03/23 03:28:40 INFO StandaloneAppClient$ClientEndpoint: Executor updated: app-20210323032840-0019/0 is now RUNNING
21/03/23 03:28:41 INFO SingleEventLogFileWriter: Logging events to dtap://TenantStorage/tmp/spark/app-20210323032840-0019.inprogress
21/03/23 03:28:41 INFO StandaloneSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
21/03/23 03:28:41 INFO SparkContext: Starting job: reduce at SparkPi.scala:38
21/03/23 03:28:41 INFO DAGScheduler: Got job 0 (reduce at SparkPi.scala:38) with 100 output partitions
21/03/23 03:28:41 INFO DAGScheduler: Final stage: ResultStage 0 (reduce at SparkPi.scala:38)
21/03/23 03:28:41 INFO DAGScheduler: Parents of final stage: List()
21/03/23 03:28:41 INFO DAGScheduler: Missing parents: List()
21/03/23 03:28:41 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[] at map at SparkPi.scala:34), which has no missing parents
```

Spark History Server UI:

 **History Server**

Event log directory: dtap://TenantStorage/tmp/spark

Last updated: 2021-03-23 15:59:34

Client local time zone: Asia/Calcutta

Show entries

Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.1.1	app-20210323032840-0019	Spark Pi	2021-03-23 15:58:39	2021-03-23 15:58:44	5 s	dev1	2021-03-23 15:58:43	Download

3.2 Submitting Spark-PI Job (Spark PI Jar from DTAP)

After logging into the spark master, copy the **spark-examples_2.12-3.1.1.jar** to DTAP.

First we will copy the jar to DTAP.

Command:

```
hadoop fs -put /usr/lib/spark/spark-3.1.1-bin-hadoop2.7/examples/jars/spark-examples_2.12-3.1.1.jar dtap://TenantStorage/
```

```
[dev1@hpecp-42 ~]$ hadoop fs -put /usr/lib/spark/spark-3.1.1-bin-hadoop2.7/examples/jars/spark-examples_2.12-3.1.1.jar dtap://TenantStorage/
[dev1@hpecp-42 ~]$
[dev1@hpecp-42 ~]$
[dev1@hpecp-42 ~]$ hadoop fs -ls dtap://TenantStorage/
Found 8 items
-rw-r--r--  3 dev1 Domain Users    1527168 2021-03-23 03:34 dtap://TenantStorage/code
drwxrwxrwx - dev1 Domain Users      0 2021-03-22 06:00 dtap://TenantStorage/elkdata
drwxrwxrwx - dev1 Domain Users      0 2021-03-22 06:00 dtap://TenantStorage/elklogs
drwxrwxrwx - dev1 Domain Users      0 2021-03-19 03:55 dtap://TenantStorage/input
drwxrwxrwx - dev1 Domain Users      0 2021-03-19 03:39 dtap://TenantStorage/jar
drwxrwxrwx - dev1 Domain Users      0 2021-03-19 05:22 dtap://TenantStorage/output
-rw-r--r--  3 dev1 Domain Users    1527168 2021-03-23 03:39 dtap://TenantStorage/spark-examples_2.12-3.1.1.jar
drwxr-xr-x - dev1 Domain Users      0 2021-03-21 23:08 dtap://TenantStorage/tmp
[dev1@hpecp-42 ~]$
```

Submit a spark pi job to spark where main application file is within DTAP.

Command:

```
spark-submit --deploy-mode client --class org.apache.spark.examples.SparkPi dtap://TenantStorage/spark-examples_2.12-3.1.1.jar 100
```


Spark 3.1.1 with Livy & JupyterHub



Output:

```
[dev]@hpecp-42 ~$ spark-submit --deploy-mode client --class org.apache.spark.examples.SparkPi dtap://TenantStorage/spark-examples_2.12-3.1.1.jar 100
21/03/23 03:41:09 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/03/23 03:41:10 INFO SparkContext: Running Spark version 3.1.1
21/03/23 03:41:10 INFO ResourceUtils: =====
21/03/23 03:41:10 INFO ResourceUtils: No custom resources configured for spark.driver.
21/03/23 03:41:10 INFO ResourceUtils: =====
21/03/23 03:41:10 INFO SparkContext: Submitted application: Spark Pi
21/03/23 03:41:10 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: Cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpu -> name: cpu, amount: 1.0)
21/03/23 03:41:10 INFO ResourceProfile: Limiting resource is cpu
21/03/23 03:41:10 INFO ResourceProfileManager: Added ResourceProfile id: 0
21/03/23 03:41:10 INFO SecurityManager: Changing view acls to: dev1
21/03/23 03:41:10 INFO SecurityManager: Changing modify acls to: dev1
21/03/23 03:41:10 INFO SecurityManager: Changing view acls groups to:
21/03/23 03:41:10 INFO SecurityManager: Changing modify acls groups to:
21/03/23 03:41:10 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(dev1); groups with view permissions: Set(); users with modify permissions: Set(dev1); groups with modify permissions: Set()
21/03/23 03:41:10 INFO Utils: Successfully started service 'sparkDriver' on port 39193.
21/03/23 03:41:10 INFO SparkEnv: Registering MapOutputTracker
21/03/23 03:41:10 INFO SparkEnv: Registering BlockManagerMaster
21/03/23 03:41:10 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
21/03/23 03:41:10 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
21/03/23 03:41:10 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
21/03/23 03:41:10 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-8ce76d41-4194-4c47-b123-8852fd7b6610
21/03/23 03:41:10 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
21/03/23 03:41:10 INFO SparkEnv: Registering OutputCommitCoordinator
21/03/23 03:41:11 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/03/23 03:41:11 INFO Utils: Successfully started service 'SparkUI' on port 4041.
21/03/23 03:41:11 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://hpecp-42.hpecplocal:4041
21/03/23 03:41:11 INFO SparkContext: Added JAR file:///opt/bluedata/bluedata-dtap.jar at spark://hpecp-42.hpecplocal:39193/jars/bluedata-dtap.jar with timestamp 1616496070425
21/03/23 03:41:11 INFO SparkContext: Added JAR dtap://TenantStorage/spark-examples_2.12-3.1.1.jar at dtap://TenantStorage/spark-examples_2.12-3.1.1.jar with timestamp 1616496070425
21/03/23 03:41:11 INFO TransportClientFactory: Successfully created connection to hpecp-42.hpecplocal/172.18.0.10:7077 after 32 ms (0 ms spent in bootstraps)
21/03/23 03:41:11 INFO StandaloneSchedulerBackend: Connected to Spark cluster with app ID app-20210323034111-0020
21/03/23 03:41:11 INFO StandaloneAppClient$ClientEndpoint: Executor added: app-20210323034111-0020/0 on worker-20210319032747-172.18.0.13-42598 (172.18.0.13:42598) with 1 core(s)
21/03/23 03:41:11 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://hpecp-42.hpecplocal:7077...
21/03/23 03:41:11 INFO StandaloneAppClient$ClientEndpoint: Executor added: app-20210323034111-0020/1 on worker-20210319032746-172.18.0.9-38291 (172.18.0.9:38291) with 1 core(s)
21/03/23 03:41:11 INFO StandaloneSchedulerBackend: Granted executor ID app-20210323034111-0020/1 on hostPort 172.18.0.9:38291 with 1 core(s), 1024.0 MiB RAM
21/03/23 03:41:11 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44745.
21/03/23 03:41:11 INFO NettyBlockTransferService: Server created on hpecp-42.hpecplocal/44745
21/03/23 03:41:11 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
21/03/23 03:41:11 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, hpecp-42.hpecplocal, 44745, None)
21/03/23 03:41:11 INFO BlockManagerMasterEndpoint: Registering block manager hpecp-42.hpecplocal:44745 with 366.3 MiB RAM, BlockManagerId(driver, hpecp-42.hpecplocal, 44745, None)
21/03/23 03:41:11 INFO BlockManagerMaster: Registered block manager BlockManagerId(driver, hpecp-42.hpecplocal, 44745, None)
21/03/23 03:41:11 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, hpecp-42.hpecplocal, 44745, None)
```

Spark History Server UI:

 **History Server**

Event log directory: dtap://TenantStorage/tmp/spark

Last updated: 2021-03-23 16:12:35

Client local time zone: Asia/Calcutta

Show 20 entries

Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.1.1	app-20210323034111-0020	Spark Pi	2021-03-23 16:11:10	2021-03-23 16:11:16	6 s	dev1	2021-03-23 16:11:14	Download
3.1.1	app-20210323032840-0019	Spark Pi	2021-03-23 15:58:39	2021-03-23 15:58:44	5 s	dev1	2021-03-23 15:58:43	Download

3.3 Submitting Spark-PI Job in Cluster Mode

Upload the **spark-terasort-1.2-SNAPSHOT-jar-with-dependencies.jar** to DTAP from the HPEC Web UI.

You can get the Jar from below link:

Jar: <https://hpecp-engineering.s3.us-east-2.amazonaws.com/Spark311/spark-terasort-1.2-SNAPSHOT-jar-with-dependencies.jar>

Spark 3.1.1 with Livy & JupyterHub



Dashboard

Clusters2

DataTaps1

FileMounts1

Nodes7

Users2

Flavors

App Store

TenantStorage Browser

dtap://TenantStorage/spark-terasort-1.2-SNAPSHOT-jar-with-dependencies.jar

dependencies.zip

tera

spark-terasort-1.2-SNAPSHOT-jar-with-dependencies.jar

etlgen

DataTap Details

Host	Path	Details
16.143.20.91,16.0.8.175,16.143.21.251	/ext/tp/Tenant-2180a	Type: mapr Cluster Name: d001 CLDB Port: 7222 Ticket File: hdp-service-ticket Ticket User: mapr Ticket Type: service Impersonation Enabled: False MapR Tenant Volume: False

Here we are running TeraGen spark job in cluster mode and our main application will reside in DTAP.

Command:

```
spark-submit --deploy-mode cluster --class com.github.ehiggs.spark.terasort.TeraGen dtap://TenantStorage/spark-terasort-1.2-SNAPSHOT-jar-with-dependencies.jar "1G" dtap://TenantStorage/tera_gen
```

Output:

```
[root@hpecp-42 bluedata]#
[root@hpecp-42 bluedata]# spark-submit --deploy-mode cluster --class com.github.ehiggs.spark.terasort.TeraGen dtap://TenantStorage/spark-terasort-1.2-SNAPSHOT-jar-with-dependencies.jar "1G" dtap://TenantStorage/tera_gen
21/03/24 00:43:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/03/24 00:43:15 INFO SecurityManager: Changing view acls to: root
21/03/24 00:43:15 INFO SecurityManager: Changing modify acls to: root
21/03/24 00:43:15 INFO SecurityManager: Changing view acls groups to:
21/03/24 00:43:15 INFO SecurityManager: Changing modify acls groups to:
21/03/24 00:43:15 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); groups with view permissions: Set(); users with modify permissions: Set(root); groups with modify permissions: Set()
21/03/24 00:43:16 INFO Utils: Successfully started service 'driverClient' on port 46231.
21/03/24 00:43:16 INFO TransportClientFactory: Successfully created connection to hpecp-42.hpecplocal/172.18.0.10:7077 after 29 ms (0 ms spent in bootstraps)
21/03/24 00:43:16 INFO ClientEndpoint: ... waiting before polling master for driver state
21/03/24 00:43:16 INFO ClientEndpoint: Driver successfully submitted as driver-20210324004316-0023
21/03/24 00:43:21 INFO ClientEndpoint: State of driver-20210324004316-0023 is RUNNING
21/03/24 00:43:21 INFO ClientEndpoint: Driver running on 172.18.0.9:38291 (worker-20210319032746-172.18.0.9-38291)
21/03/24 00:43:21 INFO ClientEndpoint: spark-submit not configured to wait for completion, exiting spark-submit JVM.
21/03/24 00:43:21 INFO ShutdownHookManager: Shutdown hook called
21/03/24 00:43:21 INFO ShutdownHookManager: Deleting directory /tmp/spark-3fd163a3-4768-4fa4-8087-77388dbed39b
[root@hpecp-42 bluedata]#
```

Spark History Server UI:

spark 3.1.1

History Server

Event log directory: dtap://TenantStorage/tmp/spark

Last updated: 2021-03-24 13:13:59

Client local time zone: Asia/Calcutta

Show 20 entries

Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.1.1	app-20210324004319-0055	TeraGen (1000MB)	2021-03-24 13:13:18	2021-03-24 13:13:32	14 s	spark	2021-03-24 13:13:31	Download

Note: When submitting job in cluster mode with any user, the **spark** user will start the driver on worker node.

DTAP UI:

HPE Ezmeral Container Platform

(Acting as Tenant Admin) Demo Tenant / admin

Dashboard

Clusters2

DataTaps1

FsMounts1

Nodes7

Users2

Flavors

App Store

TenantStorage Browser

dtap://TenantStorage/tena_gen

dependencies.zip

tmp

spark-tenant-1.2-SNAPSHOT.jar-with-dependencies.jar

tena_gen

spark-n-00000

_SUCCESS

spark-n-00001

spark-tenant-1.2-SNAPSHOT.jar-with-dependencies.jar.crc

elklogs

DataTap Details

Host	Path	Details
16.143.20.91,16.0.8.175,16.143.23.251	jenhqp/tenant-286co	Type: mapr Cluster Name: drc1 CLDB Port: 7222 Ticket File: hdp-service-ticket Ticket User: mapr Ticket Type: service Impersonation Enabled: False MapR Tenant Volume: False

4 SPARK INTERRECTIVIE SESSION USING LIVY THROUGH NOTEBOOK.

4.1 Login to Jupyter Hub.

Login to JupyterHub service using LDAP user account and password.

 jupyterhub

Sign in

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

Username:

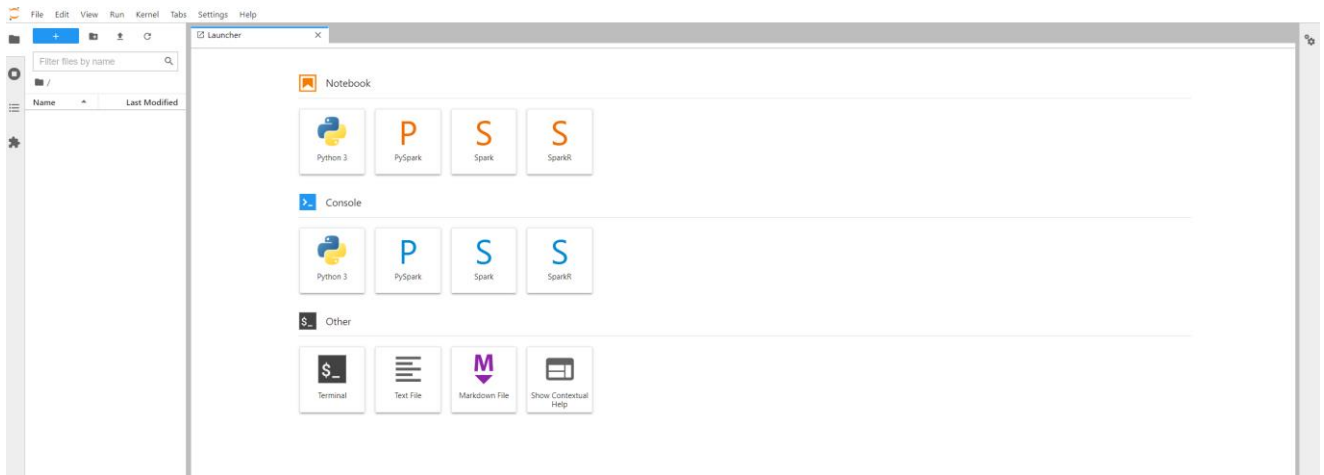
dev1

Password:

Sign in

4.2 PySpark Notebook

Click on PySpark Notebook and create a notebook.



Code:

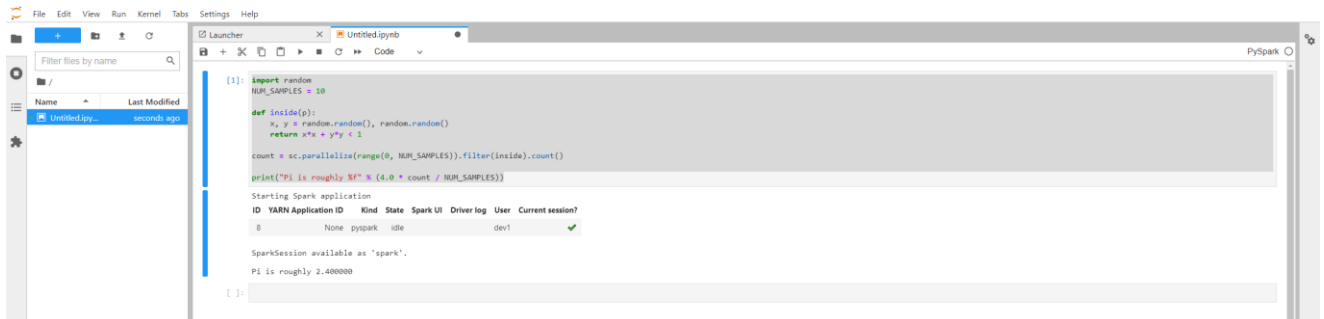
```
import random

NUM_SAMPLES = 10

def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1

count = sc.parallelize(range(0, NUM_SAMPLES)).filter(inside).count()

print("Pi is roughly %f" % (4.0 * count / NUM_SAMPLES))
```

Notebook:**Livy:**

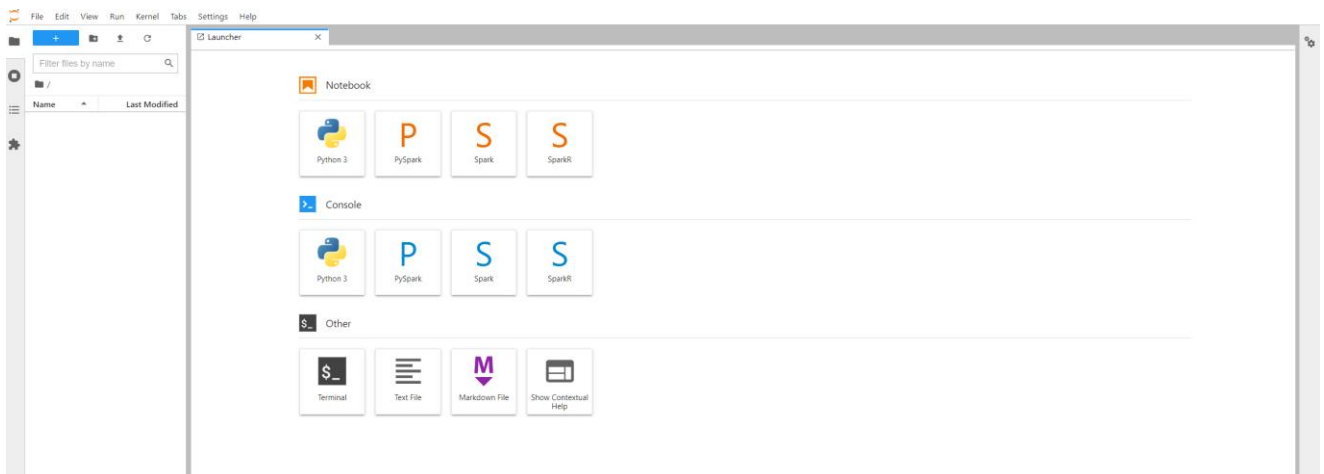
LIVY Sessions Session 8									
Session 8 Proxy User: dev1 Session Kind: pyspark State: idle Logs: session									
Statements									
Show 10 entries									
ID	Execution Code	Execution State	Progress	Execution Status	Output	Started	Completed	Duration	
0	spark	available	100%	ok	<gyspark.sql.session.SparkSession object at 0x7f71a0b990d0>	2021-03-23 16:34:57	2021-03-23 16:34:57	1 ms	
1	<pre>import random NUM_SAMPLES = 10 def inside(p): x, y = random.random(), random.random() return x*x + y*y < 1 count = sc.parallelize(range(0, NUM_SAMPLES)).filter(inside).count() print("Pi is roughly %f" % (4.0 * count / NUM_SAMPLES))</pre>	available	100%	ok	Pi is roughly 2.400000	2021-03-23 16:34:58	2021-03-23 16:35:00	2 s	

Showing 1 to 2 of 2 entries

Previous 1 Next

4.3 Scala Notebook

Click on Spark Notebook and create a notebook.

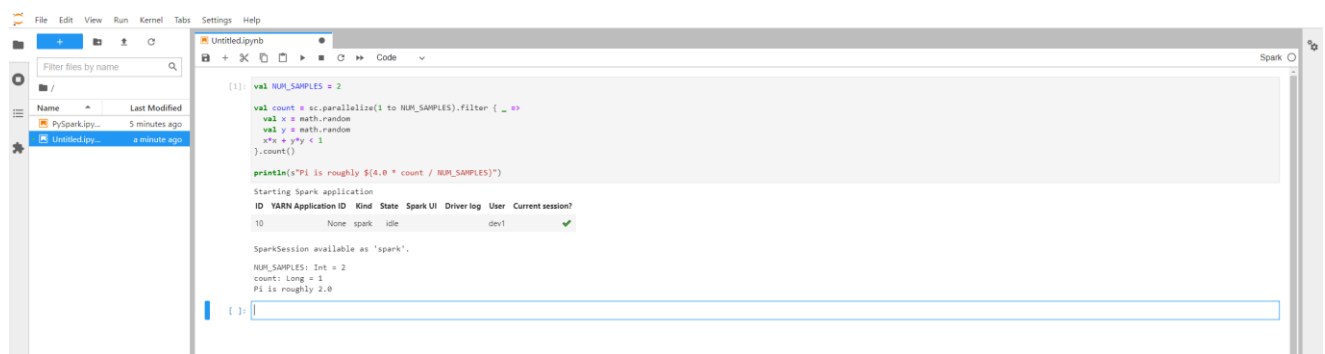


Code:

```
val NUM_SAMPLES = 2

val count = sc.parallelize(1 to NUM_SAMPLES).filter { _ =>
    val x = math.random
    val y = math.random
    x*x + y*y < 1
}.count()

println(s"Pi is roughly ${4.0 * count / NUM_SAMPLES}")
```

Notebook:**Livy:**

LIVY Sessions Session 10

Session 10
Proxy User: dev1
Session Kind: spark
State: idle
Logs: session

Statements

Show 10 entries

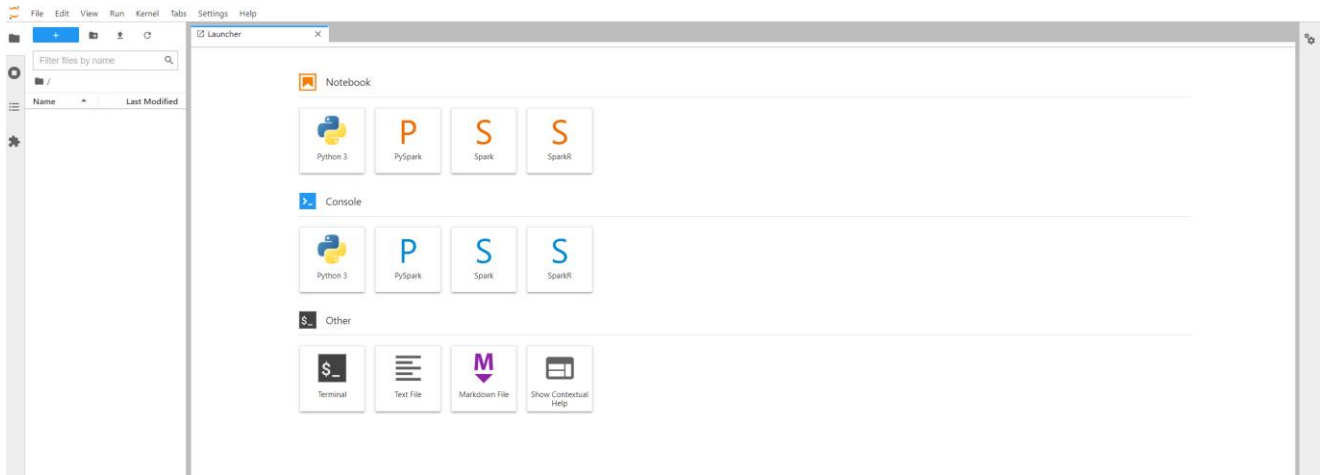
Id	Execution Code	Execution State	Progress	Execution Status	Output	Started	Completed	Duration
0	spark	available	100%	ok	res0: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@98c336c	2021-03-23 16:41:15	2021-03-23 16:41:15	0.2 s
1	<pre>val NUM_SAMPLES = 2 val count = sc.parallelize(1 to NUM_SAMPLES).filter { _ => val x = math.random val y = math.random x*x + y*y < 1 }.count() println(s"Pi is roughly \${4.0 * count / NUM_SAMPLES}")</pre>	available	100%	ok	<pre>NUM_SAMPLES: Int = 2 count: Long = 1 Pi is roughly 2.0</pre>	2021-03-23 16:41:16	2021-03-23 16:41:18	2 s

Showing 1 to 2 of 2 entries

Previous 1 Next

4.4 SparkR Notebook

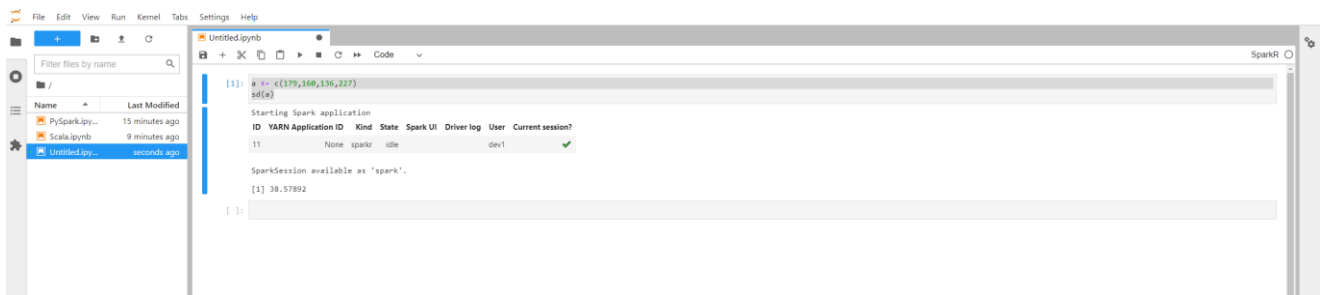
Click on SparkR Notebook and create a notebook.




Code:

```
a <- c(179,160,136,227)
sd(a)
```

Notebook:



Livy:

 Sessions Session 11

Session 11
Proxy User: dev1
Session Kind: sparkr
State: idle
Logs:
[session](#)

Statements

Show entries

Search:

Id	Execution Code	Execution State	Progress	Execution Status	Output	Started	Completed	Duration
0	spark	available	<div>100%</div>	ok	Java ref type org.apache.spark.sql.SparkSession id 1	2021-03-23 16:52:02	2021-03-23 16:52:02	33 ms
1	a <- c(179,168,136,227) sd(a)	available	<div>100%</div>	ok	[1] 38.57892	2021-03-23 16:52:02	2021-03-23 16:52:02	2 ms

Showing 1 to 2 of 2 entries

Previous **1** Next

5 PASS RUNTIME DEPENDENCIES TO SPARK JOB

5.1 Python Dependencies

Here we are packaging python dependencies and will pass those dependencies to PySpark job at run time.

Below are the commands to package python dependencies in a zip file and upload the zip file to DTAP.

In **requirements.txt** file we just added as single python package openpyxl (python package for reading and writing excel file.)

Note: This work fine with some python package but for some package it will fail.

```
pip3 install -t dependencies -r requirements.txt
cd dependencies/
zip -r ../dependencies.zip .
cd ..
hadoop fs -put dependencies.zip dtap://TenantStorage/
hadoop fs -ls dtap://TenantStorage/
```

PySpark Job:

In the job we are adding dependencies.zip file to our program using **sc.addPyFile()** method. In the logs, it will display all object of openpyxl python package. If you comment line 6 where we are adding dependencies.zip and re execute the job import openpyxl line will through.

Content of PySparkJob.py

```
from pyspark import SparkContext

sc = SparkContext()

#Adding Dependencies
sc.addPyFile("dtap://TenantStorage/dependencies.zip")

#Loading Python Package
import openpyxl
print(dir(openpyxl))
```

Spark Submit:

```
spark-submit PySparkJob.py
```

Output:

```
[dev]@hpecp-42 ~$ spark-submit PySparkJob.py
21/03/23 04:46:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/03/23 04:46:28 INFO SparkContext: Running Spark version 3.1.1
21/03/23 04:46:28 INFO ResourceUtils: =====
21/03/23 04:46:28 INFO ResourceUtils: No custom resources configured for spark.driver.
21/03/23 04:46:28 INFO SparkContext: Submitted application: PySparkJob.py
21/03/23 04:46:28 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
21/03/23 04:46:28 INFO ResourceProfileManager: Added ResourceProfile id: 0
21/03/23 04:46:28 INFO SecurityManager: Changing view acls to: dev1
21/03/23 04:46:28 INFO SecurityManager: Changing modify acls to: dev1
21/03/23 04:46:28 INFO SecurityManager: Changing view acls groups to:
21/03/23 04:46:28 INFO SecurityManager: Changing modify acls groups to:
21/03/23 04:46:28 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(dev1); groups with view permissions: Set(); users with modify permissions: Set(dev1); groups with modify permissions: Set()
21/03/23 04:46:28 INFO Utils: Successfully started service 'sparkDriver' on port 46343.
21/03/23 04:46:28 INFO SparkEnv: Registering MapOutputTracker
21/03/23 04:46:28 INFO SparkEnv: Registering BlockManagerMaster
21/03/23 04:46:28 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
21/03/23 04:46:28 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
21/03/23 04:46:28 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
21/03/23 04:46:28 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-5e7ec6c9-64bd-407c-b6fc-4d88eaac0111
21/03/23 04:46:28 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
21/03/23 04:46:28 INFO SparkEnv: Registering OutputCommitCoordinator
21/03/23 04:46:29 INFO Utils: Successfully started service 'SparkUI' on port 4040.
21/03/23 04:46:29 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://hpecp-42.hpecplocal:4040
21/03/23 04:46:29 INFO SparkContext: Added JAR file:///opt/bluedata/bluedata-dtap.jar at spark://hpecp-42.hpecplocal:46343/jars/bluedata-dtap.jar with timestamp 1616499988226
21/03/23 04:46:29 INFO StandaloneAppClientClientEndpoint: Connecting to master spark://hpecp-42.hpecplocal:7077...
21/03/23 04:46:29 INFO TransportClientFactory: Successfully created connection to hpecp-42.hpecplocal/172.18.0.10:7077 after 32 ms (0 ms spent in bootstraps)
21/03/23 04:46:29 INFO StandaloneSchedulerBackend: Connected to Spark cluster with app ID app-20210323044629-0031
21/03/23 04:46:29 INFO StandaloneAppClientClientEndpoint: Executor added: app-20210323044629-0031/0 on worker-20210319032746-172.18.0.9-38291 (172.18.0.9:38291) with 2 core(s)
21/03/23 04:46:29 INFO StandaloneSchedulerBackend: Granted executor ID app-20210323044629-0031/0 on hostPort 172.18.0.9:38291 with 2 core(s), 1024.0 MiB RAM
21/03/23 04:46:29 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 37948.
21/03/23 04:46:29 INFO NettyBlockTransferService: Server created on hpecp-42.hpecplocal:37948
21/03/23 04:46:29 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
21/03/23 04:46:29 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, hpecp-42.hpecplocal, 37948, None)
21/03/23 04:46:29 INFO BlockManagerMasterEndpoint: Registering block manager hpecp-42.hpecplocal:37948 with 366.3 MiB RAM, BlockManagerId(driver, hpecp-42.hpecplocal, 37948, None)
21/03/23 04:46:29 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, hpecp-42.hpecplocal, 37948, None)
21/03/23 04:46:29 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, hpecp-42.hpecplocal, 37948, None)
21/03/23 04:46:29 INFO StandaloneAppClientClientEndpoint: Executor updated: app-20210323044629-0031/0 is now RUNNING
21/03/23 04:46:30 INFO SingleEventLogFileWriter: Logging events to dtap://TenantStorage/tmp/spark/app-20210323044629-0031.inprogress
21/03/23 04:46:30 INFO StandaloneSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
21/03/23 04:46:30 INFO SparkContext: Added file dtap://TenantStorage/dependencies.zip at dtap://TenantStorage/dependencies.zip with timestamp 1616499990301
21/03/23 04:46:30 INFO Utils: Fetching dtap://TenantStorage/dependencies.zip to /tmp/spark-710e056c-0ab0-4afc-b6ee-caccc0d80d87/userFiles-1dc9ffc0-18ca-40a0-9115-a73f5c710b7f/fetchFileTemp7187739464295905343.tmp
['DEFUSEDOMP', 'LOXL', 'NUMPY', 'Workbook', '__author__', '__author_email__', '__builtins__', '__cached__', '__doc__', '__file__', '__license__', '__loader__', '__maintainer_email__', '__name__', '__package__', '__path__', '__spec__', '__url__', '__version__', '__constants__', 'Cell', 'Chart', 'chartsheet', 'comments', 'compat', 'constants', 'descriptors', 'drawing', 'formatting', 'formula', 'load_workbook', 'open', 'packaging', 'pivot', 'reader', 'styles', 'utils', 'workbook', 'worksheet', 'writer', 'xml']
21/03/23 04:46:30 INFO SparkContext: Invoking stop() from shutdown hook
21/03/23 04:46:30 INFO SparkUI: Stopped Spark web UI at http://hpecp-42.hpecplocal:4040
21/03/23 04:46:30 INFO StandaloneSchedulerBackend: Shutting down all executors
21/03/23 04:46:30 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
21/03/23 04:46:30 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
21/03/23 04:46:30 INFO MemoryStore: MemoryStore cleared
```

5.2 Jar Dependencies

Working on it.

	Spark 3.1.1 with Livy & JupyterHub	 Hewlett Packard Enterprise
--	------------------------------------	---

6 KERBEROS TESTING

Pending

	Spark 3.1.1 with Livy & JupyterHub	 Hewlett Packard Enterprise
--	------------------------------------	---

7 SPARK THRIFT SERVER

As of now not include in the bin.