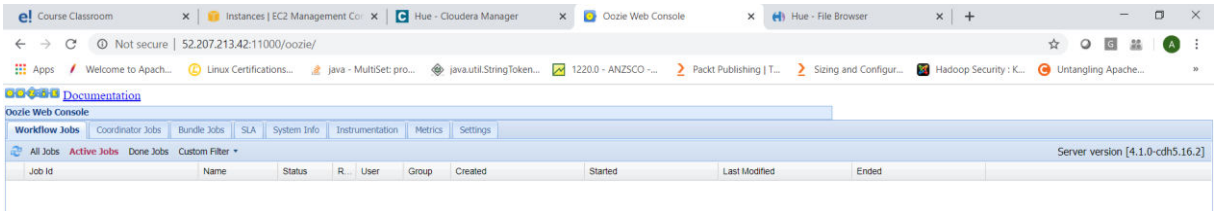
Oozie: Working with workflows in CDH

Add oozie service in your CDH cluster

To enable oozie web console follow instructions here:

https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/admin_oozie_console.html

enable oozie web console by enabling property in Oozie>configuration



Refer : https://oozie.apache.org/docs/4.2.0/index.html

Creating workflow from HUE:

To create a workflow that will run some simple Hive Queries.

Step 1:

From CLI connect to Hive CLI/Beeline (or use Hue) and create a database (mydb) and a table(testtbl2).

Table can be created as follows: [use mydb;create table testtbl2 (id int,name string)]

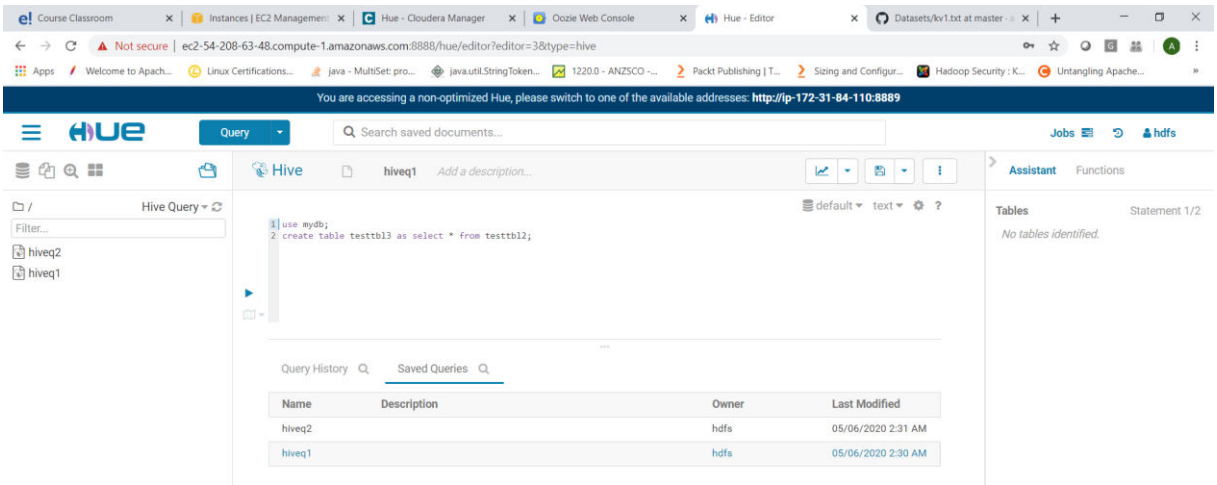
Now load data into this table by using Hue and uploading file into table directory i.e. /user/hive/warehouse/mydb.db/testtbl2/

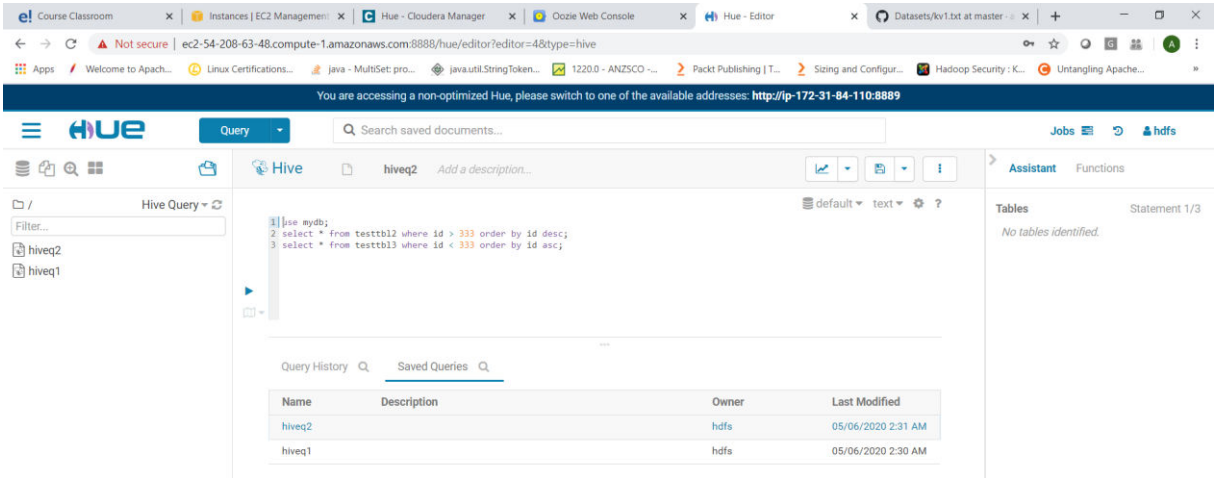
File can be found here: https://github.com/ajaykuma/Datasets/blob/master/kv1.txt

Step 2:

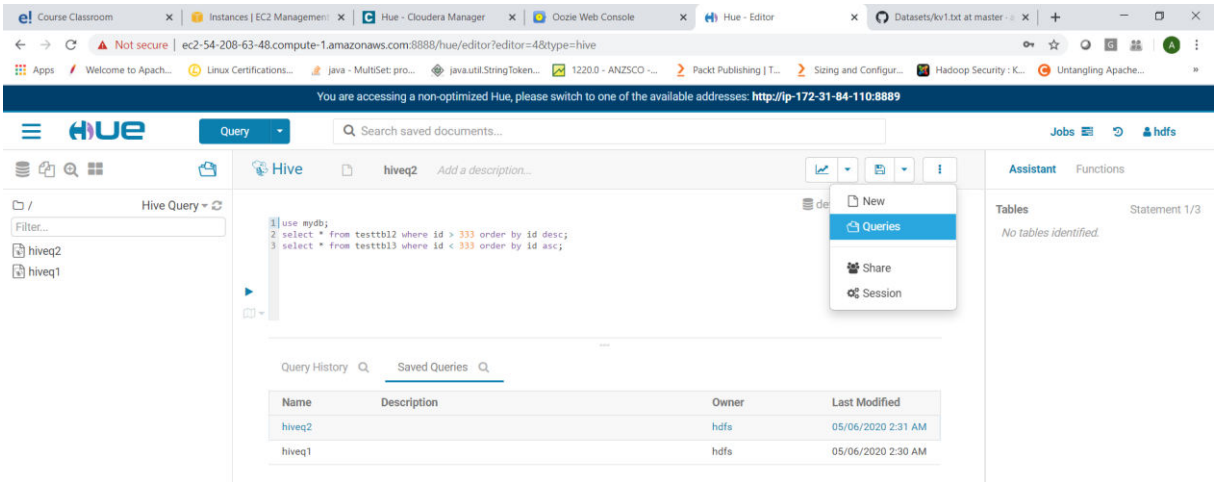
Using Hue create queries and save them.

Query > Editor > Hive

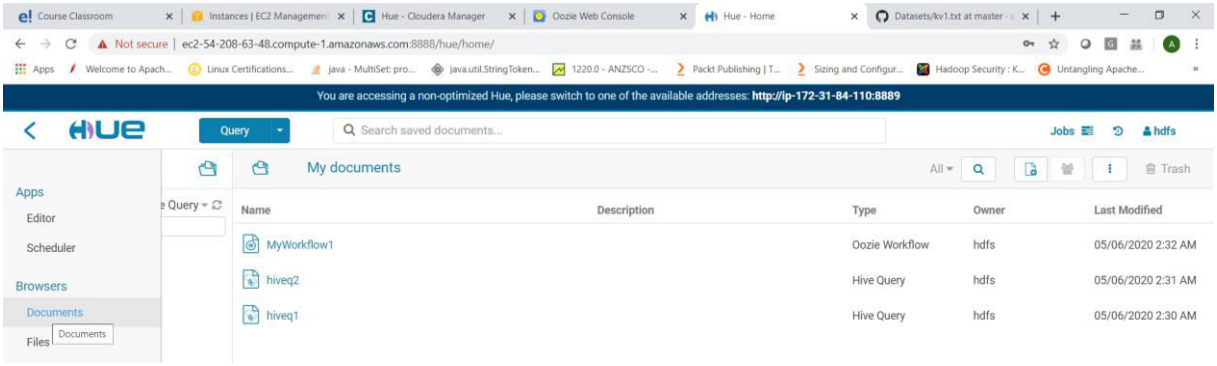




You can find your saved queries in



You can also find queries, workflows etc in

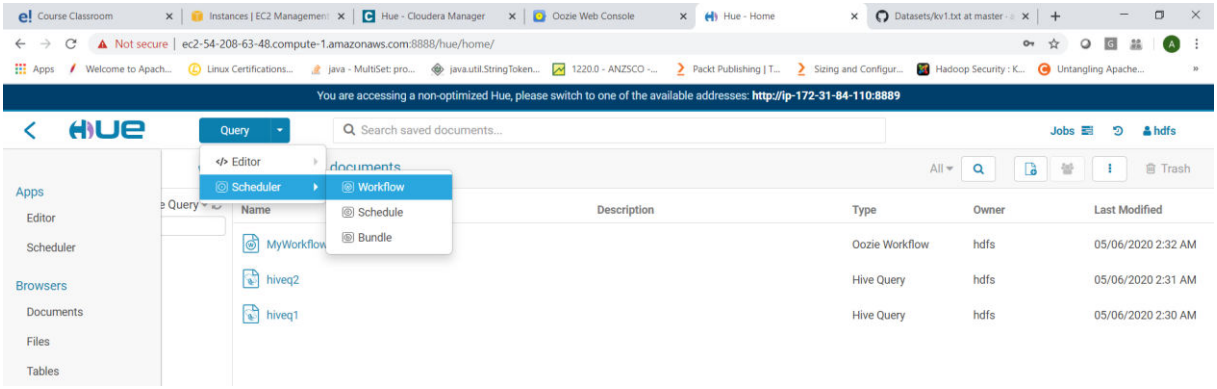


Query> editor> scheduler>workflow

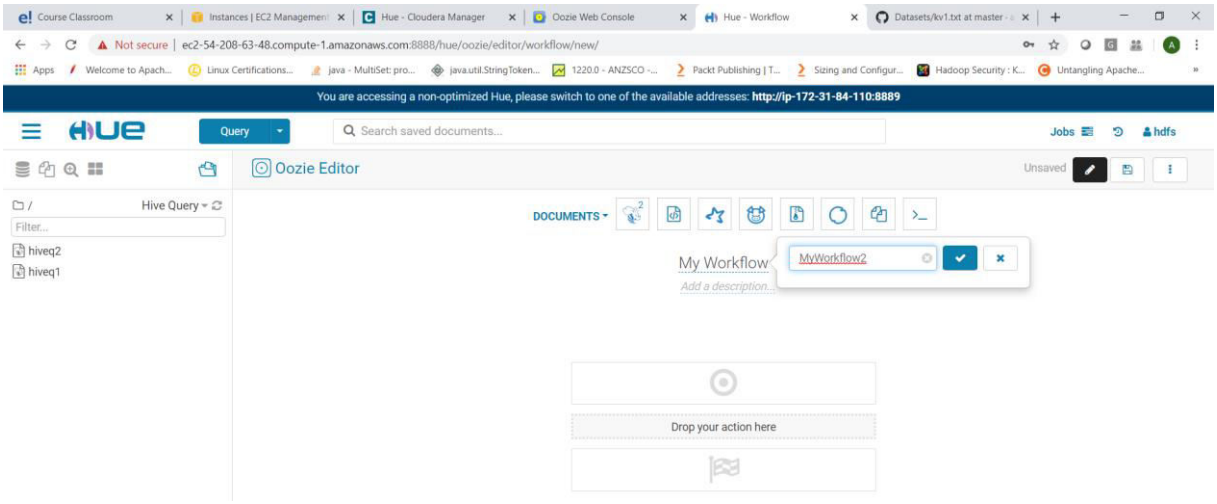
We will create a simple workflow which runs multiple hive queries.

https://oozie.apache.org/docs/4.2.0/DG_Hive2ActionExtension.html

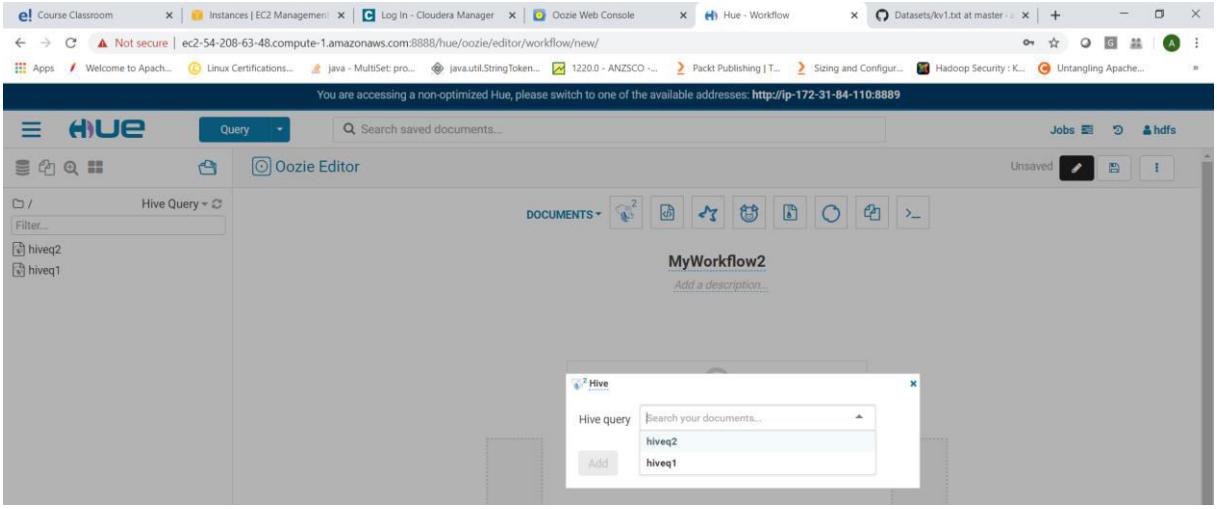
Workflow1



Edit and give a name to your workflow

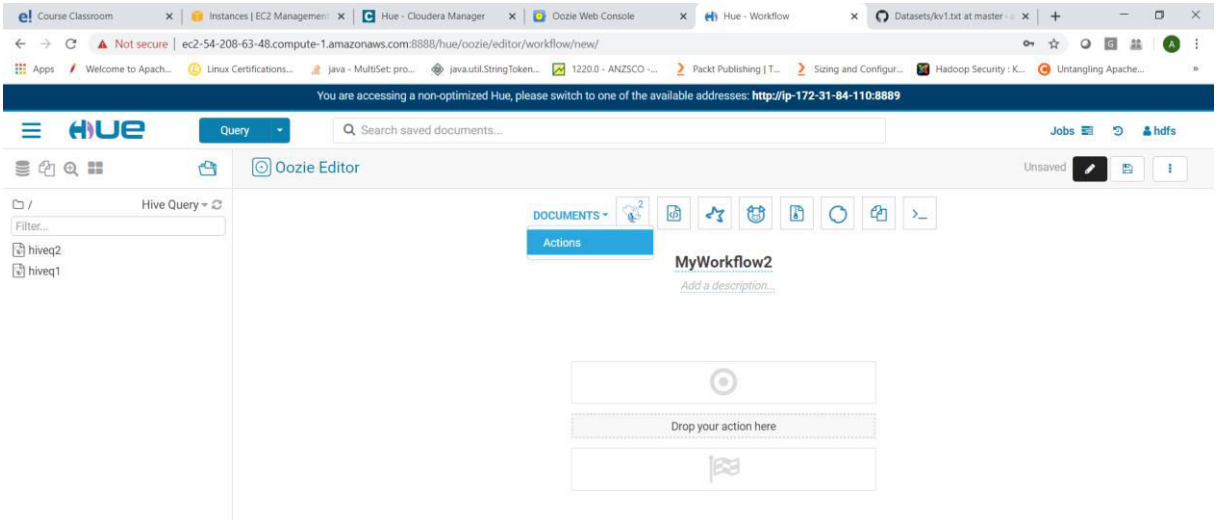


Click on option that says HiveServer2 and drag and drop it in “Drag your action here”

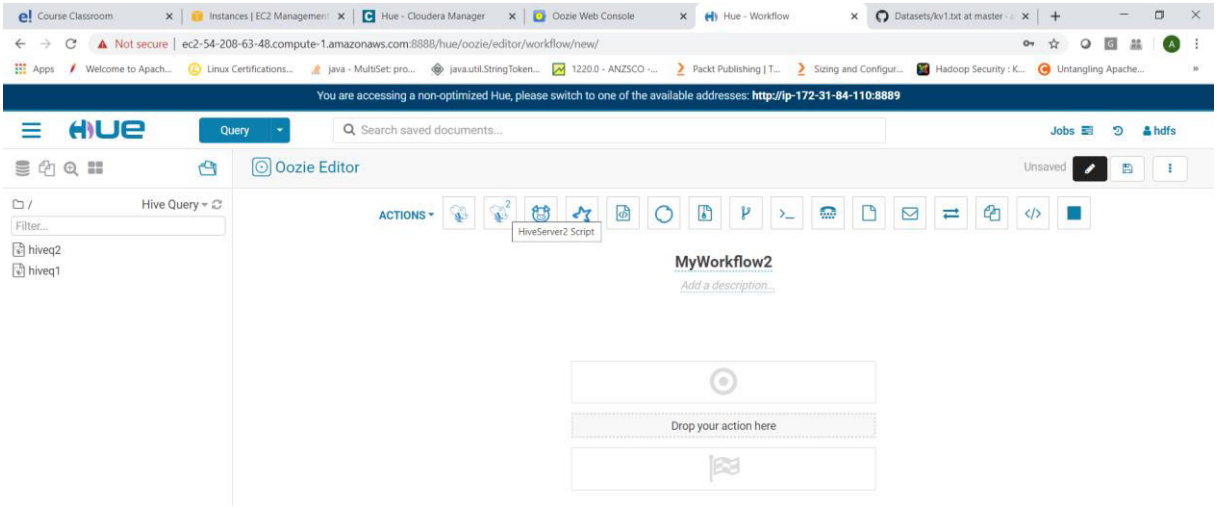


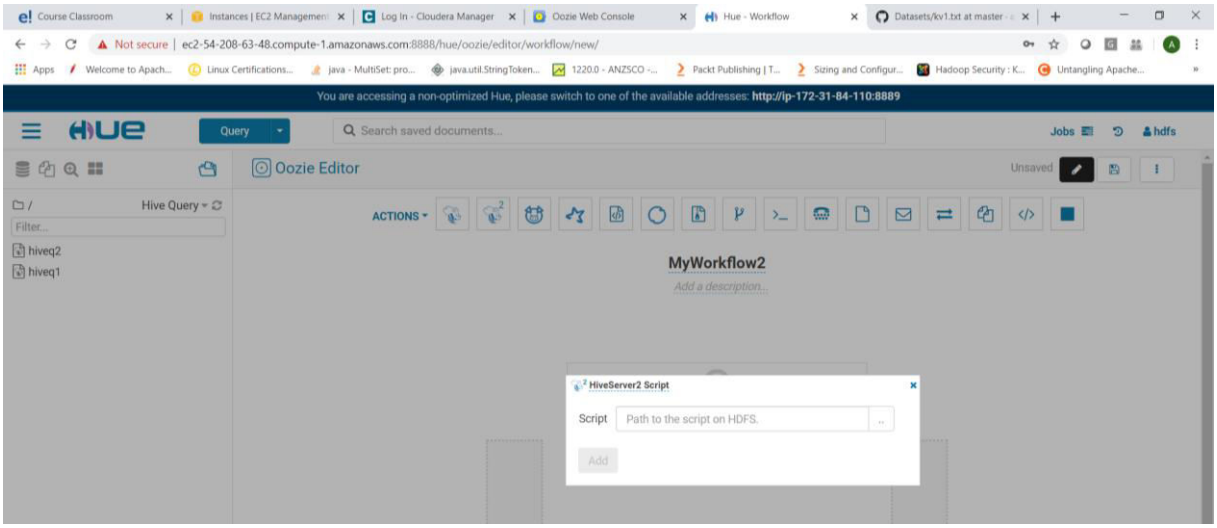
Choose your hiveq1 repeat the same for hiveq2

Optionally you expand options by clicking on actions as shown [This is to be used if you have saved your queries in some folder on HDFS] . If not skip next 3 screenshots.

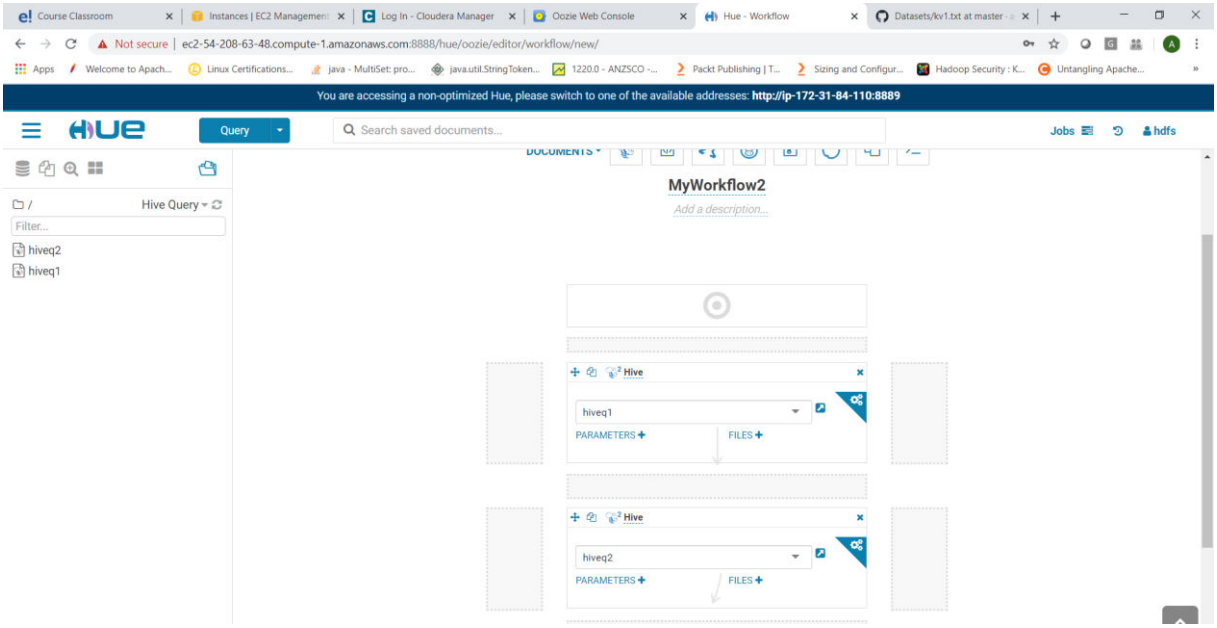


Click on second option, ie HiveServer2 Script and drag it to box which says “Drop your action here”



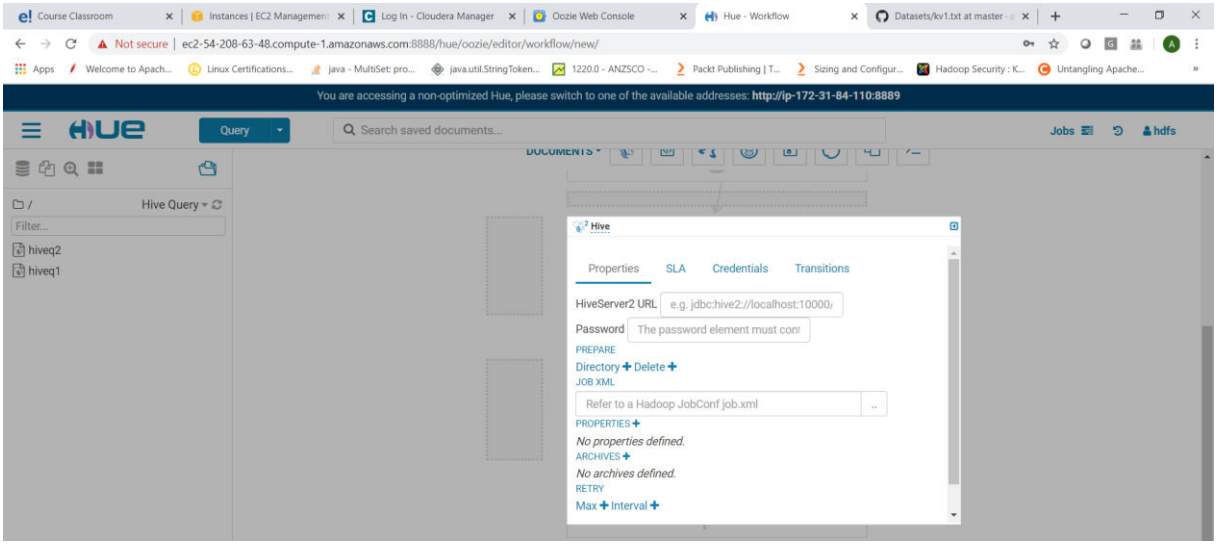


Now your workflow would look like this:



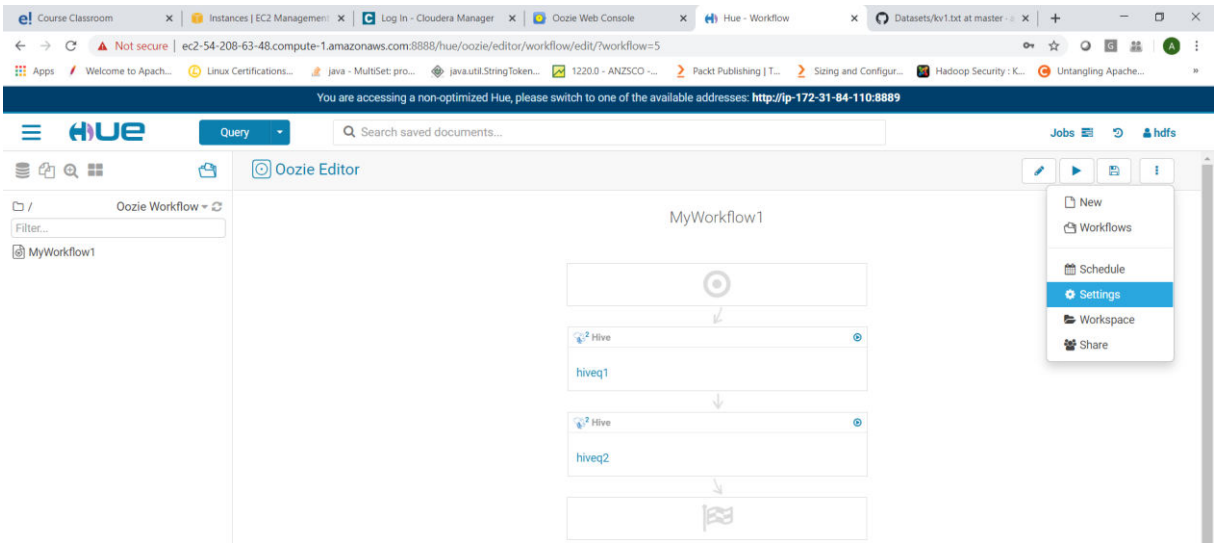
We can edit settings here or add files/parameters as needed.

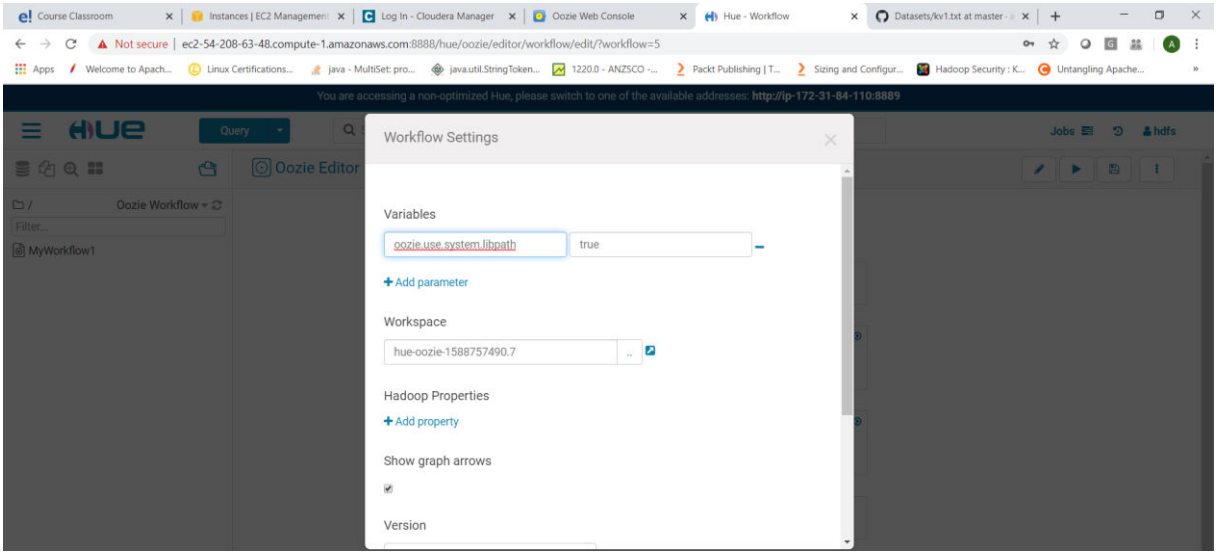
If we click on settings it shows

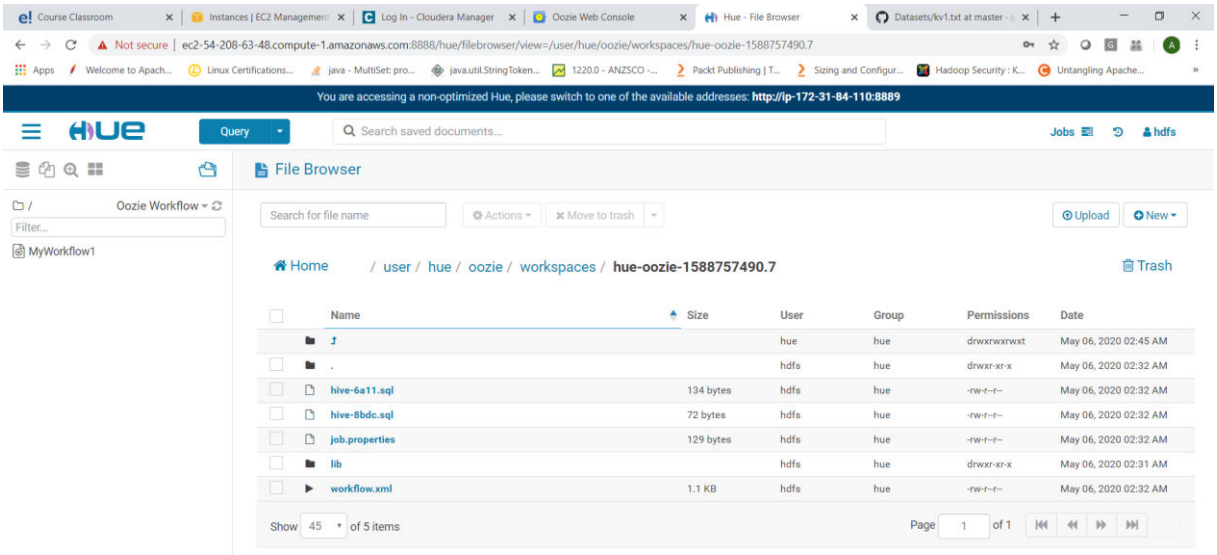


Now your workflow is ready.

We can see the workflow by

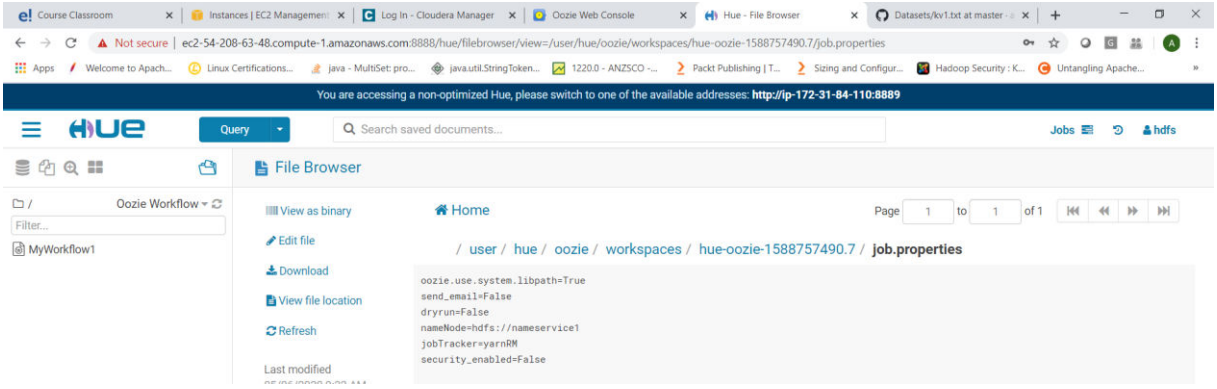




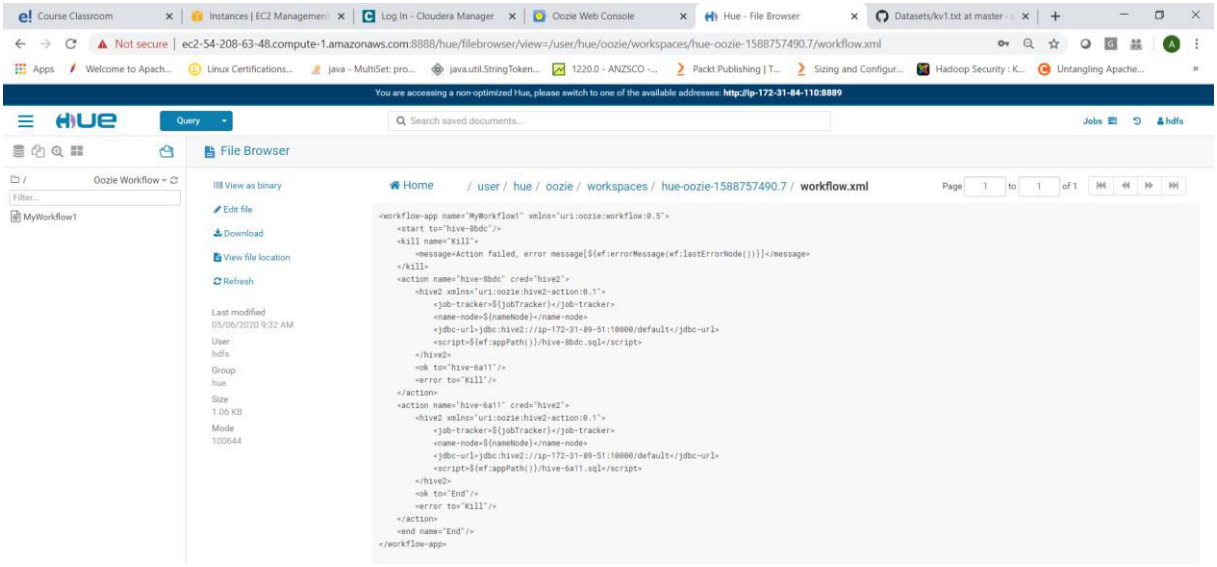


This shows our queries that we added to workflow.

job.properties shows

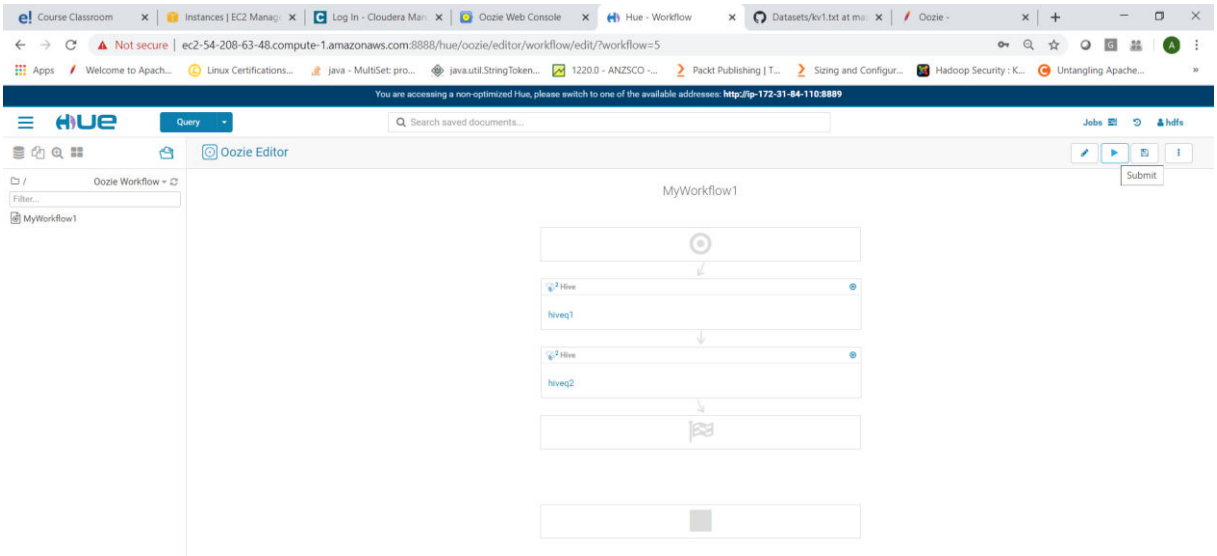


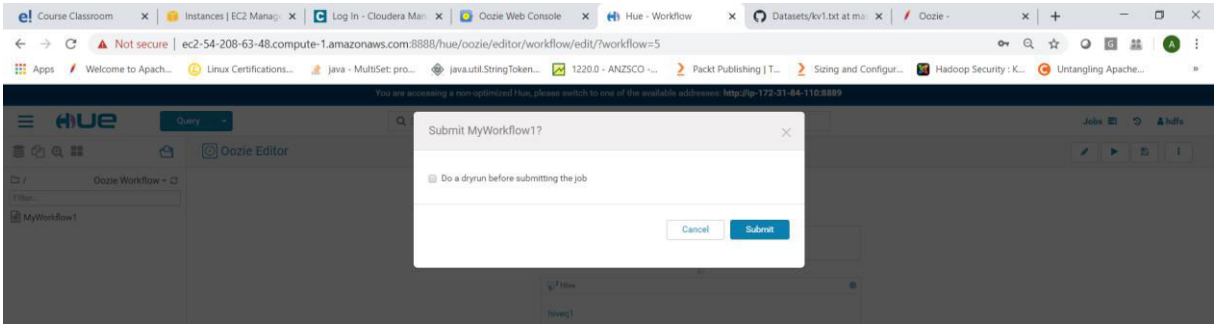
workflow.xml

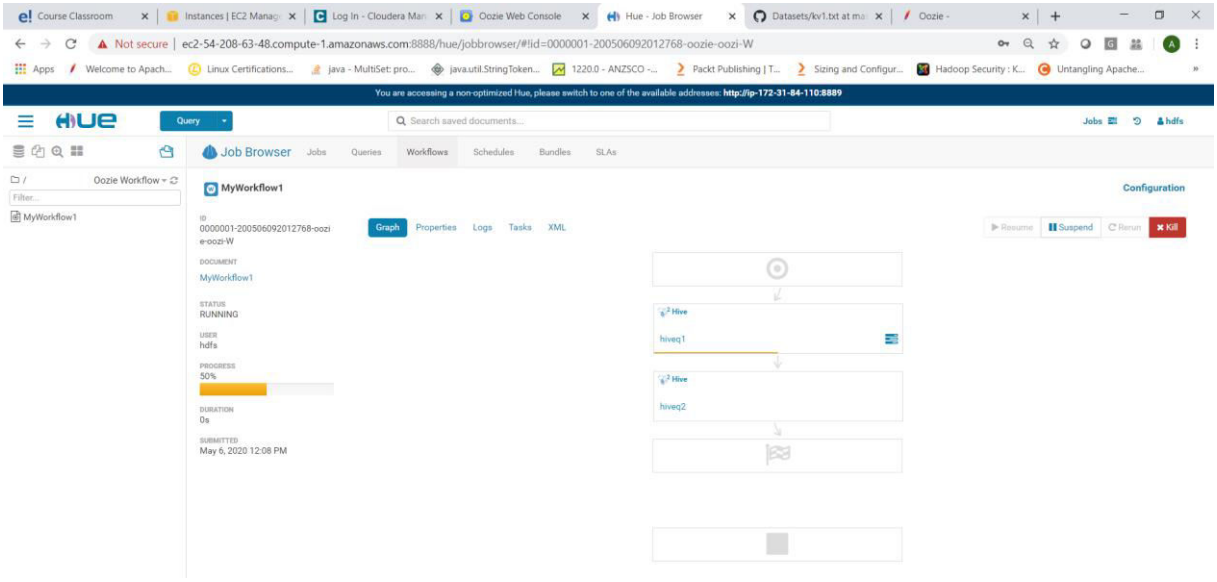


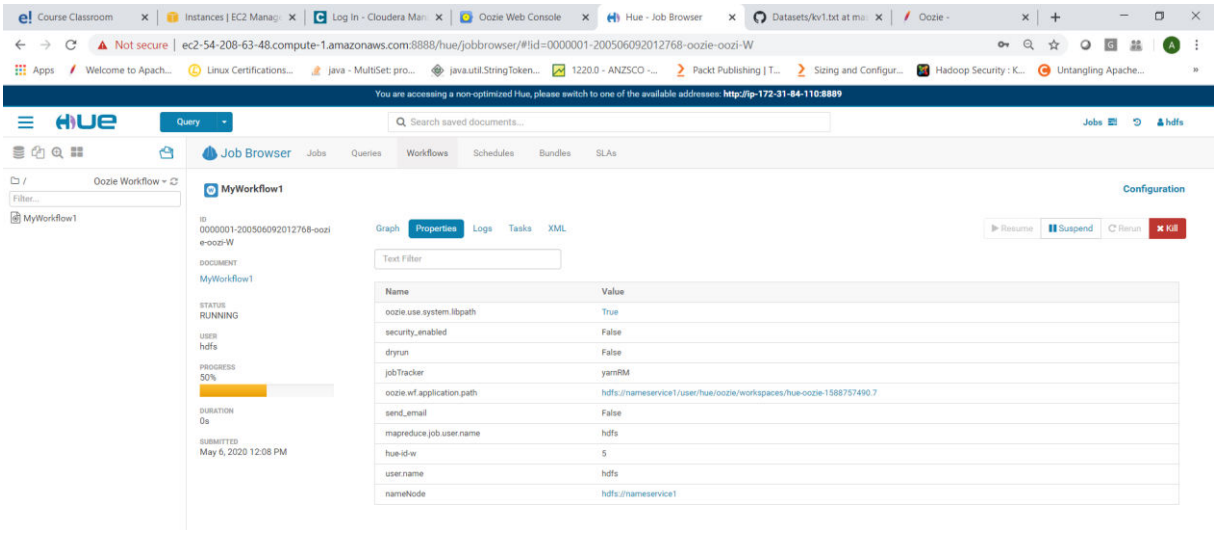
Refer <workflow.xml>

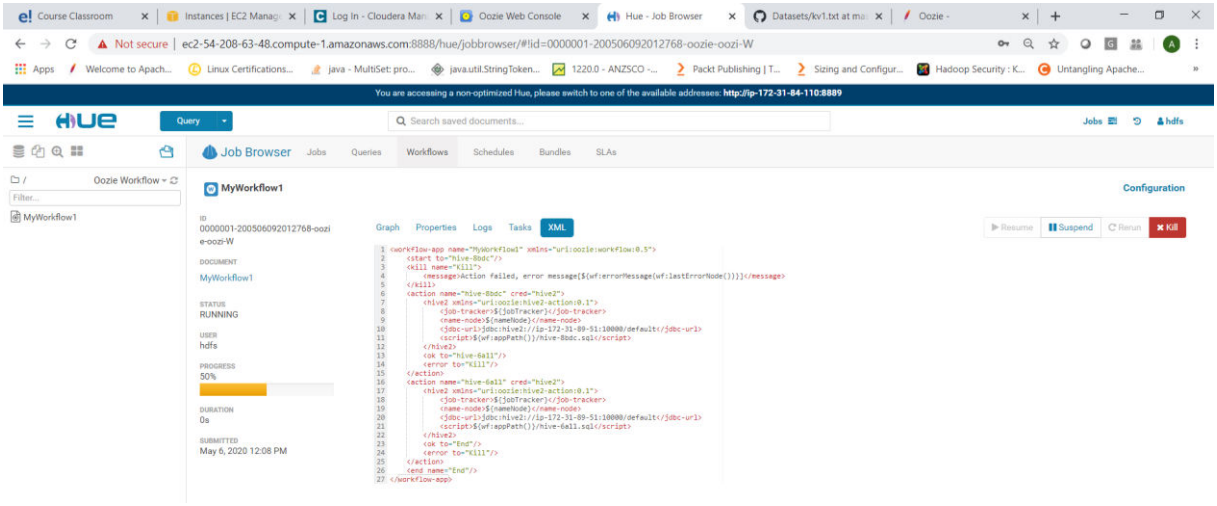
We can submit this workflow



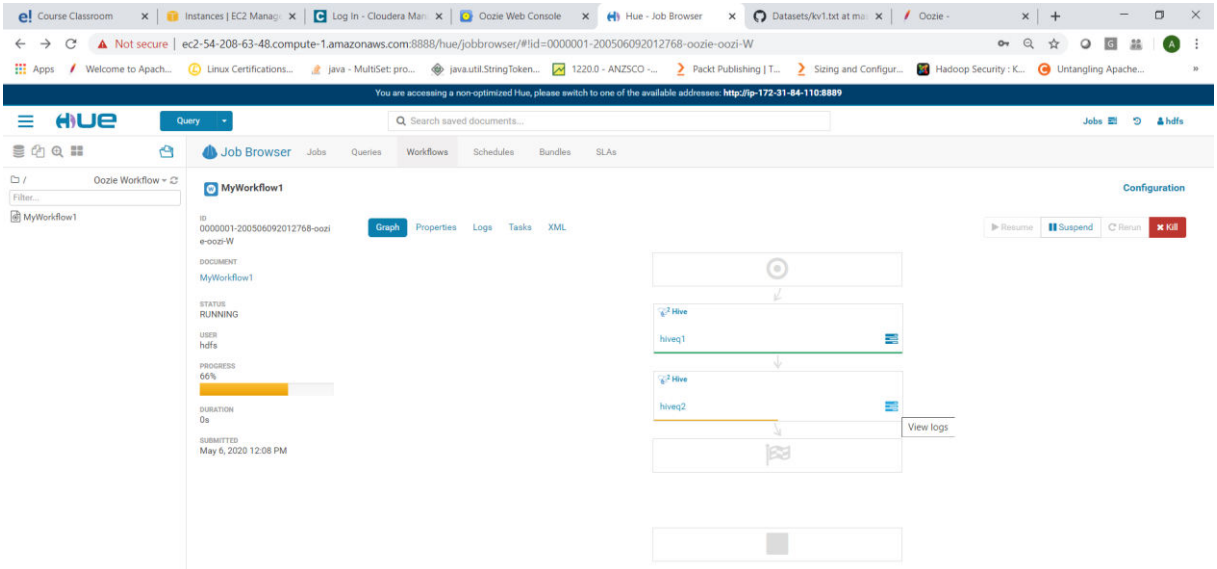

Submit MyWorkflow1?
Do a dryrun before submitting the job
Cancel
Submit

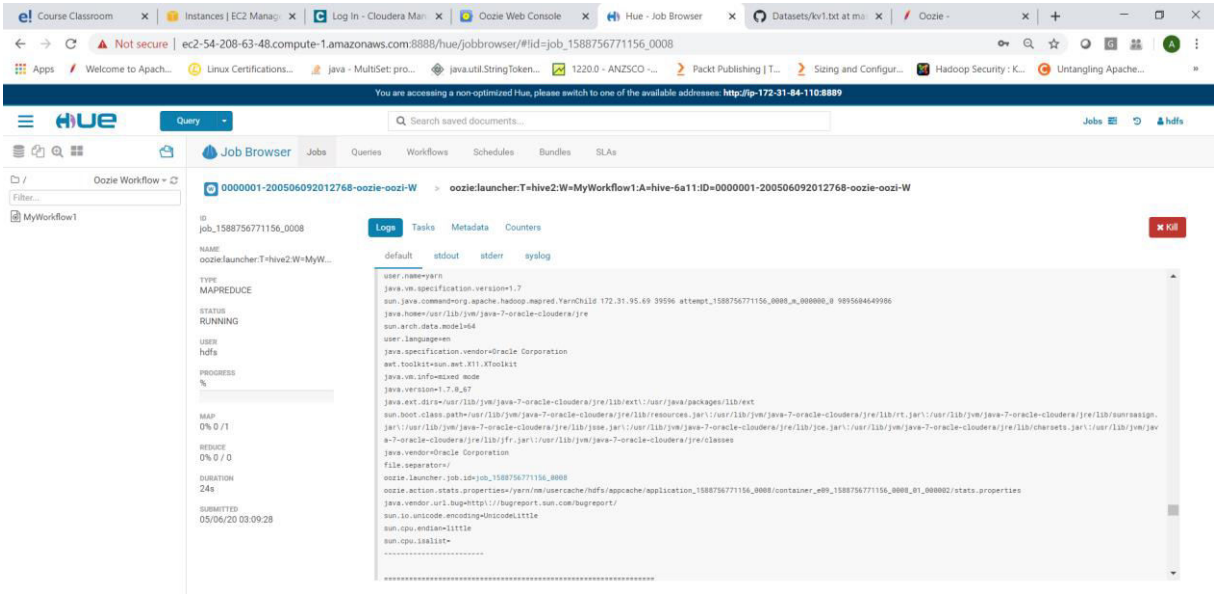

Job Browser
MyWorkflow1
Configuration
Graph
Properties
Logs
Tasks
XML
Suspend
Kill
STATUS
RUNNING
USER
hdfs
PROGRESS
50%
DURATION
0s
SUBMITTED
May 6, 2020 12:08 PM
Hive
hiveq1
Hive
hiveq2

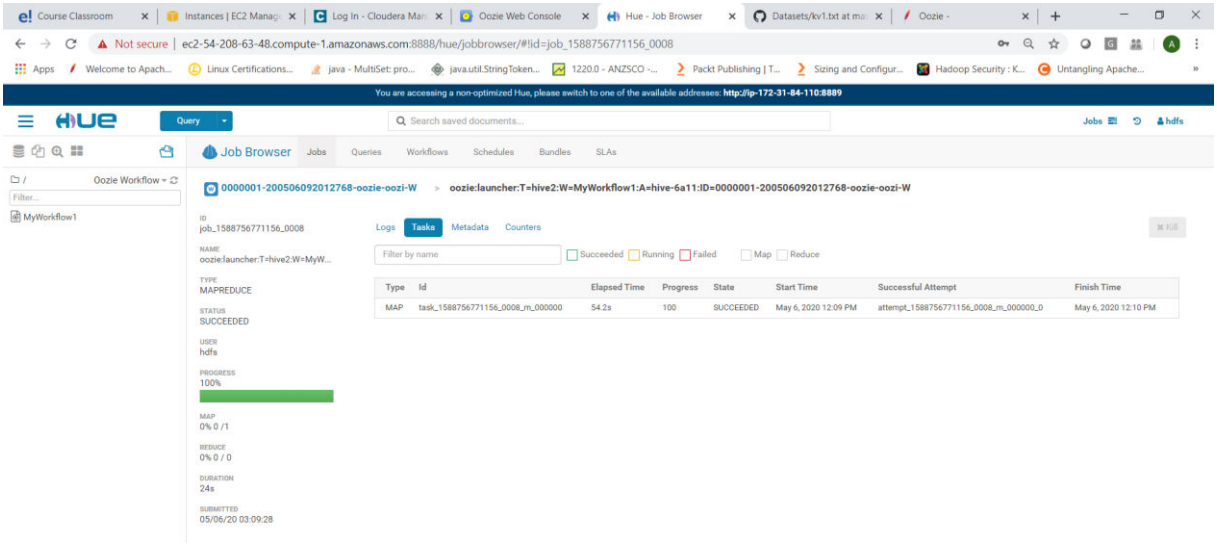

Job Browser
MyWorkflow1
Configuration
Graph
Properties
Logs
Tasks
XML
Suspend
Kill
STATUS
RUNNING
USER
hdfs
PROGRESS
50%
DURATION
0s
SUBMITTED
May 6, 2020 12:08 PM
Name
Value
oozie.use.system.libpath
True
security_enabled
False
dryrun
False
jobTracker
yarnRM
oozie.wf.application.path
hdfs://nameservice1/user/hue/oozie/workspaces/hue-oozie-1588757490.7
send_email
False
mapreduce.job.user.name
hdfs
hue-id-w
5
user.name
hdfs
nameNode
hdfs://nameservice1



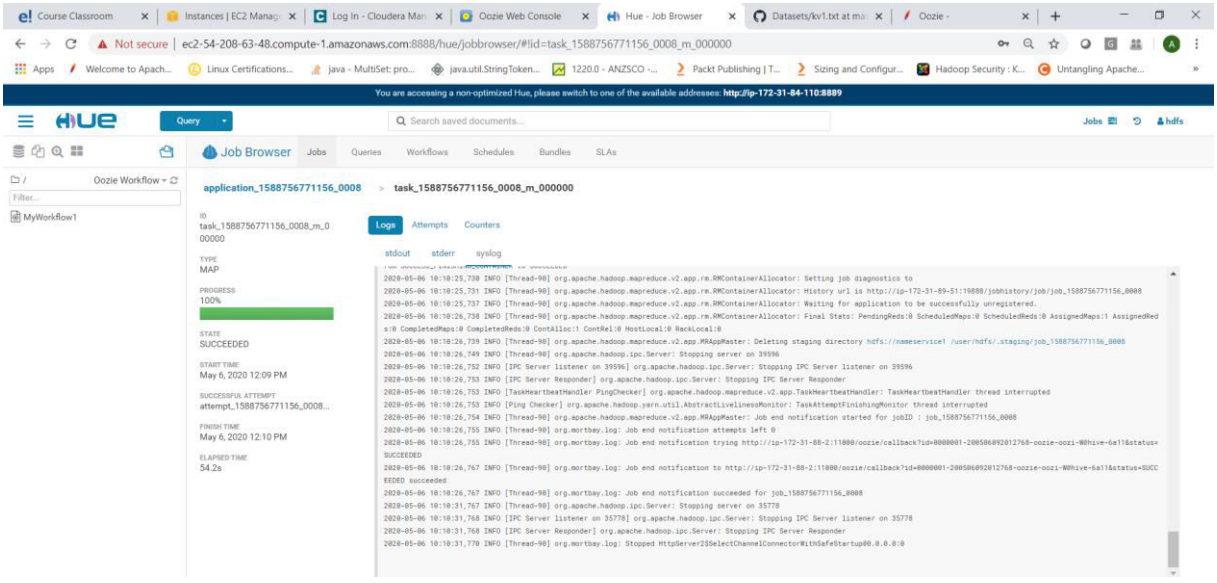
We can also monitor this from

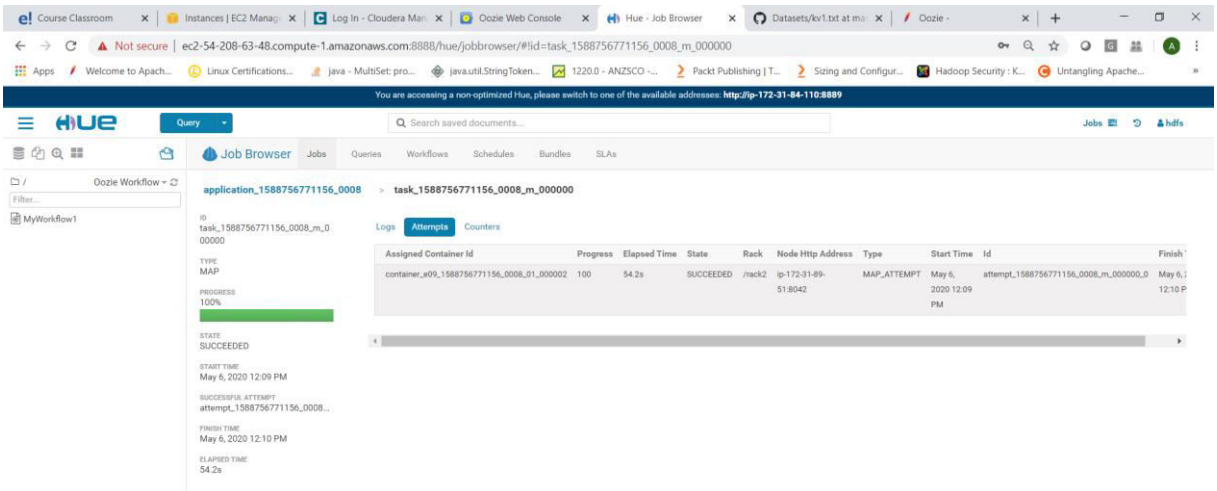


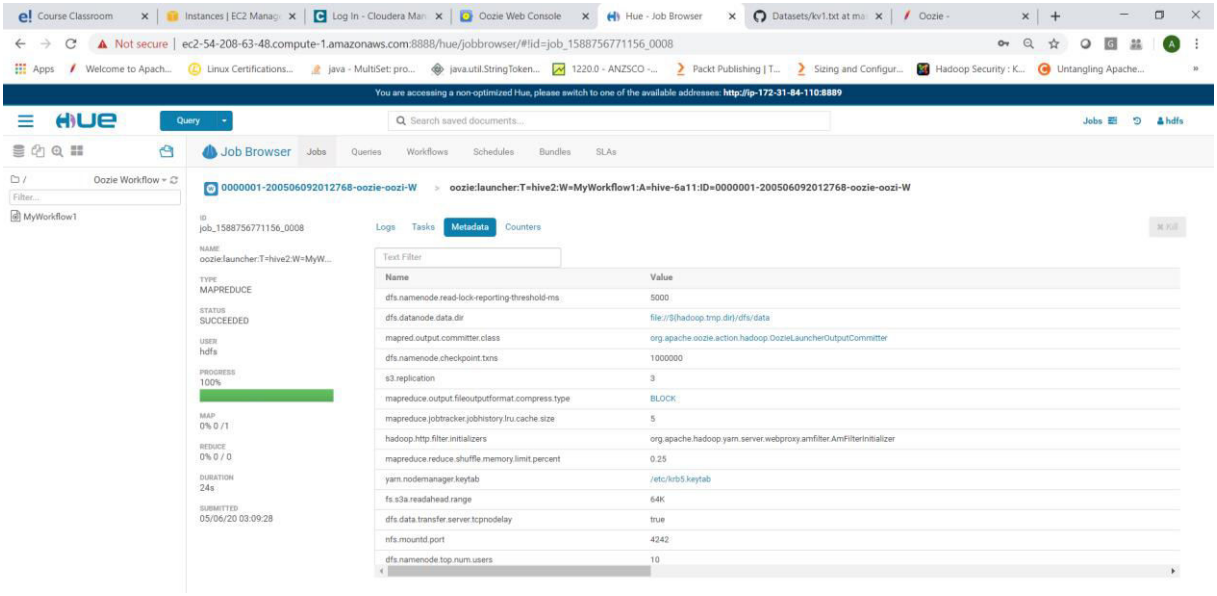




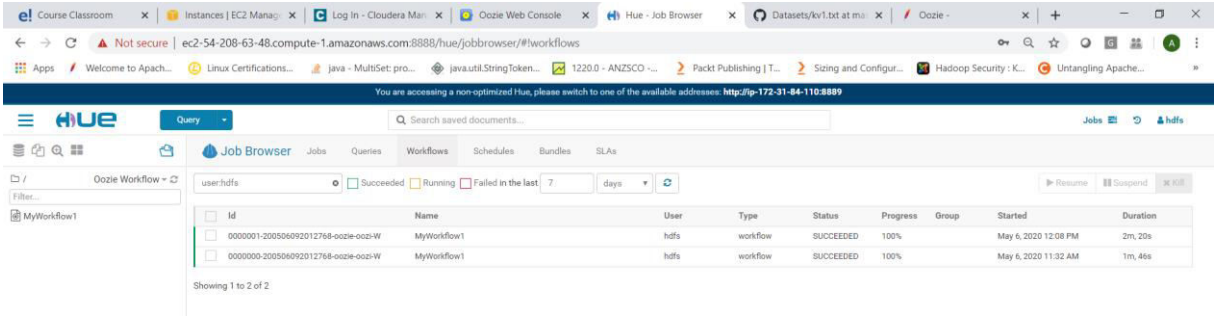
Click on Task shows



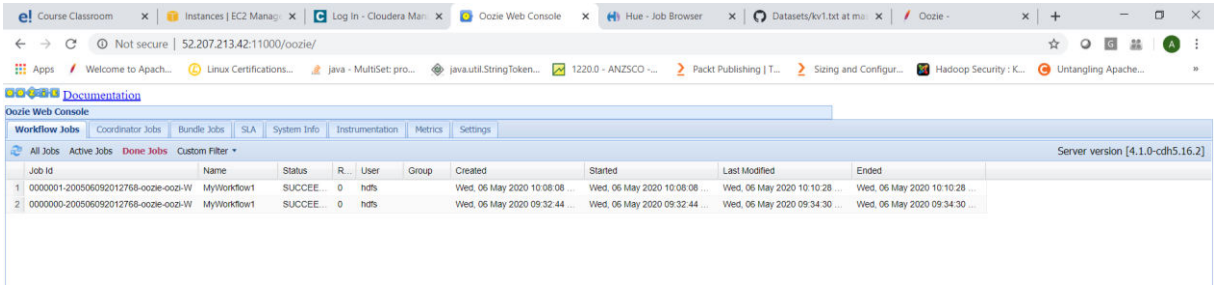




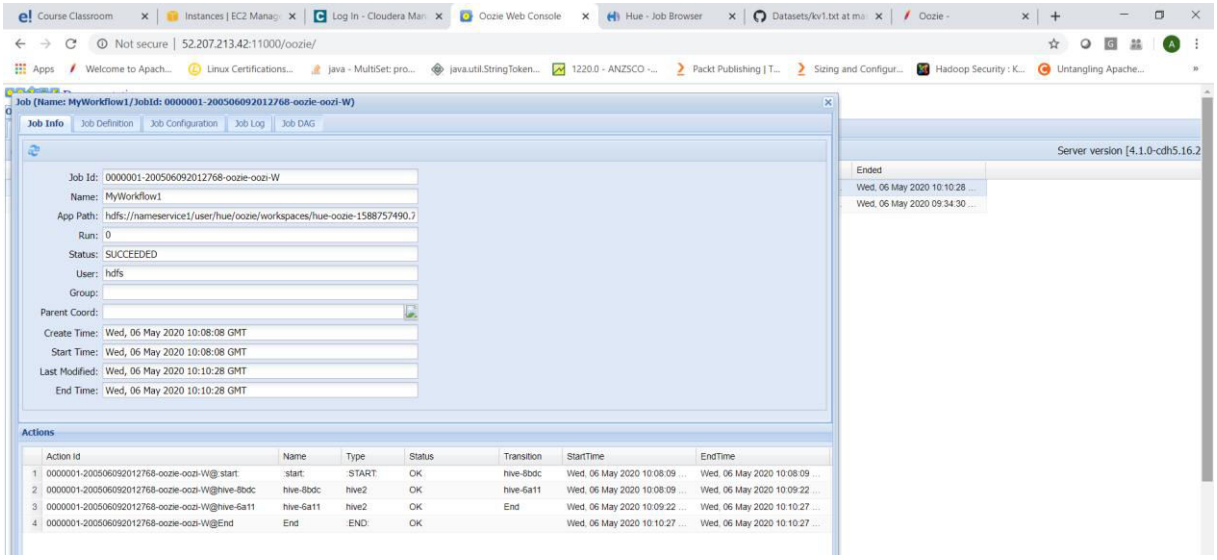
Looking at all workflows



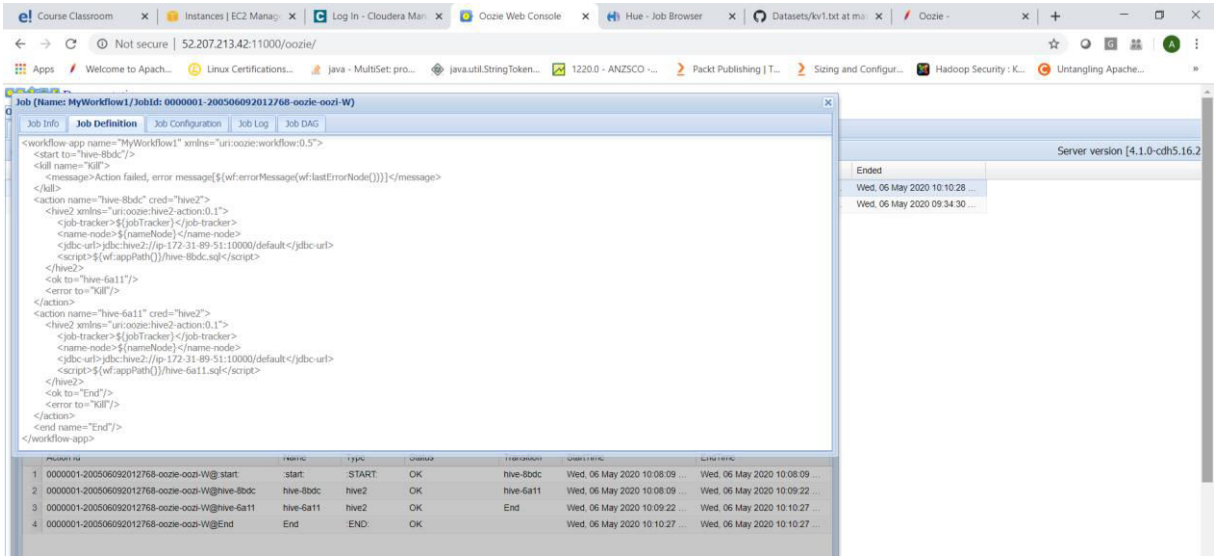
From oozie console, check Workflow Jobs> Done Jobs . Optionally you can also explore other tabs.

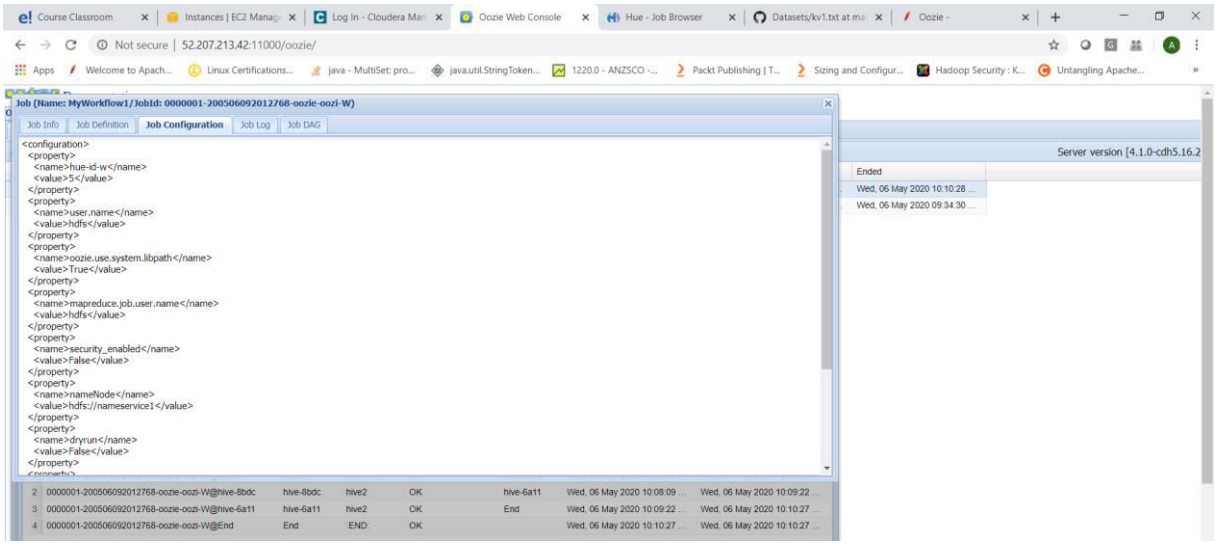


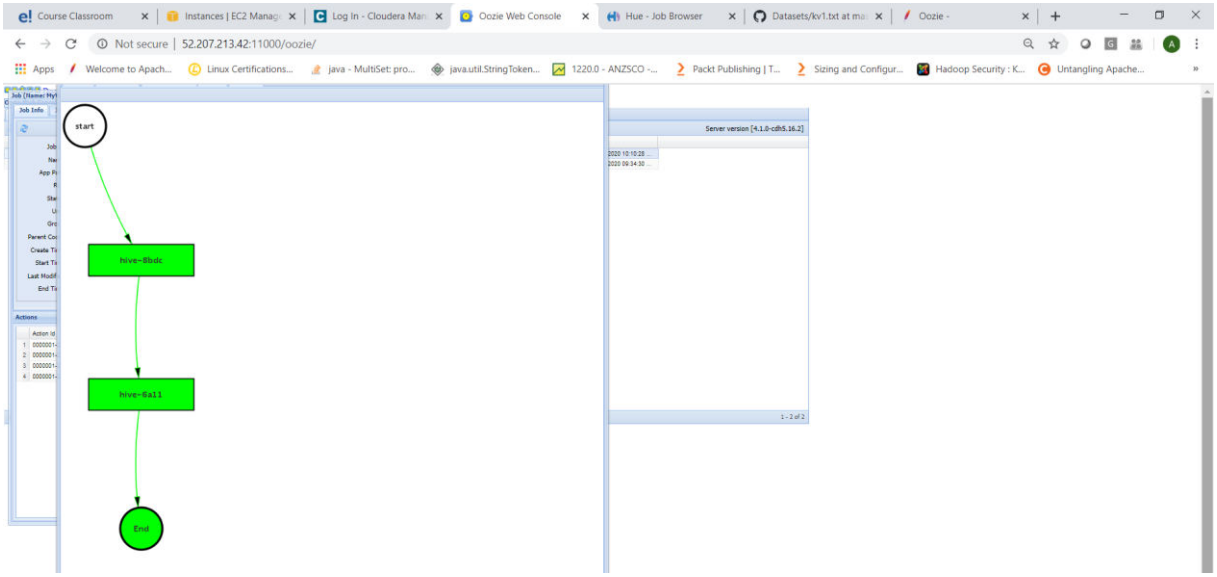
Double click on Job ID



Explore other tabs







Now check from Hive CLI, as a result of your query, you should new table

```
hive> show tables;
OK
testtbl
testtbl2
Time taken: 0.195 seconds, Fetched: 2 row(s)
hive> show tables;
OK
testtbl
testtbl2
testtbl3
Time taken: 0.018 seconds, Fetched: 3 row(s)
hive> drop table testtbl3;
OK
Time taken: 0.242 seconds
hive> show tables;
OK
testtbl
testtbl2
testtbl3
Time taken: 0.017 seconds, Fetched: 3 row(s)
hive>
```

We can run this workflow from command line too.

Workflow1-Command line

```
hdfs@ip-172-31-88-2:~
hdfs@ip-172-31-88-2:~$ hdfs dfs -ls /
Found 6 items
-rw-r--r--   2 hdfs  supergroup    1054212 2020-05-04 08:00 /auction.csv
drwxr-xr-x   - hbase hbase               0 2020-05-06 09:19 /hbase
drwxr-xr-x   - hdfs  supergroup          0 2020-05-04 06:35 /mydata
drwxr-xr-x   - hdfs  supergroup          0 2020-05-04 06:51 /system
drwxrwxrwt   - hdfs  supergroup          0 2020-05-04 06:27 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2020-05-04 06:31 /user
hdfs@ip-172-31-88-2:~$ hdfs dfs -ls /user/hdfs
Found 4 items
drwx------   - hdfs hdfs          0 2020-05-06 10:07 /user/hdfs/.Trash
drwx------   - hdfs hdfs          0 2020-05-06 10:10 /user/hdfs/.staging
drwxr-xr-x   - hdfs hdfs          0 2020-05-06 09:34 /user/hdfs/hue-oozie-1588757490.7
drwxr-xr-x   - hdfs hdfs          0 2020-05-06 10:10 /user/hdfs/oozie-oozi
hdfs@ip-172-31-88-2:~$ hdfs dfs -get /user/hdfs/hue-oozie-1588757490.7 .
hdfs@ip-172-31-88-2:~$ ls
hue-oozie-1588757490.7
hdfs@ip-172-31-88-2:~$ mv hue-oozie-1588757490.7 workflow1
hdfs@ip-172-31-88-2:~$ ls
workflow1
hdfs@ip-172-31-88-2:~$ ls workflow1/
hive-6a11.sql  hive-8bdc.sql  job.properties  lib  workflow.xml
hdfs@ip-172-31-88-2:~$
```

```
hdfs@ip-172-31-88-2:~$ logout
root@ip-172-31-88-2:~# su - hdfs
hdfs@ip-172-31-88-2:~$ oozie admin -oozie http://172.31.88.2:11000/oozie -status
System mode: NORMAL
hdfs@ip-172-31-88-2:~$ oozie admin -oozie http://172.31.88.2:11000/oozie --version
Oozie server build version: 4.1.0-cdh5.16.2
hdfs@ip-172-31-88-2:~$ export OOZIE_URL="http://172.31.88.2:11000/oozie"
hdfs@ip-172-31-88-2:~$ oozie admin -status
System mode: NORMAL
hdfs@ip-172-31-88-2:~$
```

```
hdfs@ip-172-31-88-2:~/workflow1$ ls lib
hdfs@ip-172-31-88-2:~/workflow1$ vi job.properties
hdfs@ip-172-31-88-2:~/workflow1$ vi workflow.xml
hdfs@ip-172-31-88-2:~/workflow1$ cd ..
hdfs@ip-172-31-88-2:~$ ls workflow1/
hive-6a11.sql  hive-8bdc.sql  job.properties  lib  workflow.xml
hdfs@ip-172-31-88-2:~$ hdfs dfs -mkdir /Workflow
hdfs@ip-172-31-88-2:~$ hdfs dfs -put workflow1/*.sql /Workflow
hdfs@ip-172-31-88-2:~$ hdfs dfs -put workflow1/job.properties /Workflow
hdfs@ip-172-31-88-2:~$ hdfs dfs -put workflow1/lib /Workflow/
hdfs@ip-172-31-88-2:~$ hdfs dfs -put workflow1/workflow.xml /Workflow
hdfs@ip-172-31-88-2:~$ hdfs dfs -ls /Workflow
Found 5 items
-rw-r--r--   2 hdfs supergroup        134 2020-05-06 11:05 /Workflow/hive-6a11.sql
-rw-r--r--   2 hdfs supergroup         72 2020-05-06 11:05 /Workflow/hive-8bdc.sql
-rw-r--r--   2 hdfs supergroup        177 2020-05-06 11:05 /Workflow/job.properties
drwxr-xr-x   - hdfs supergroup          0 2020-05-06 11:06 /Workflow/lib
-rw-r--r--   2 hdfs supergroup       1088 2020-05-06 11:06 /Workflow/workflow.xml
```

```
hdfs@ip-172-31-88-2:~$ ls workflow1/
hive-6a11.sql  hive-8bdc.sql  job.properties  lib  workflow.xml
hdfs@ip-172-31-88-2:~$ vi workflow1/job.properties
hdfs@ip-172-31-88-2:~$ hdfs dfs -ls /Workflow
Found 5 items
-rw-r--r--   2 hdfs supergroup        134 2020-05-06 11:05 /Workflow/hive-6a11.sql
-rw-r--r--   2 hdfs supergroup         72 2020-05-06 11:05 /Workflow/hive-8bdc.sql
-rw-r--r--   2 hdfs supergroup        177 2020-05-06 11:05 /Workflow/job.properties
drwxr-xr-x   - hdfs supergroup          0 2020-05-06 11:06 /Workflow/lib
-rw-r--r--   2 hdfs supergroup       1088 2020-05-06 11:06 /Workflow/workflow.xml
hdfs@ip-172-31-88-2:~$ oozie job -config workflow1/job.properties -run
job: 0000002-200506092012768-oozie-oozi-W
hdfs@ip-172-31-88-2:~$
hdfs@ip-172-31-88-2:~$
```

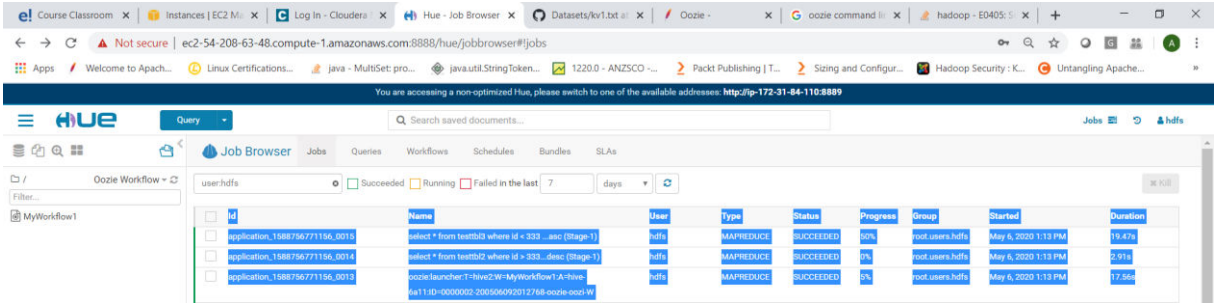
Job.properties contains

```
oozie.use.system.libpath=true
send_email=False
dryrun=False
nameNode=hdfs://nameservice1
jobTracker=yarnRM
security_enabled=False
oozie.wf.application.path=${nameNode}/Workflow
```

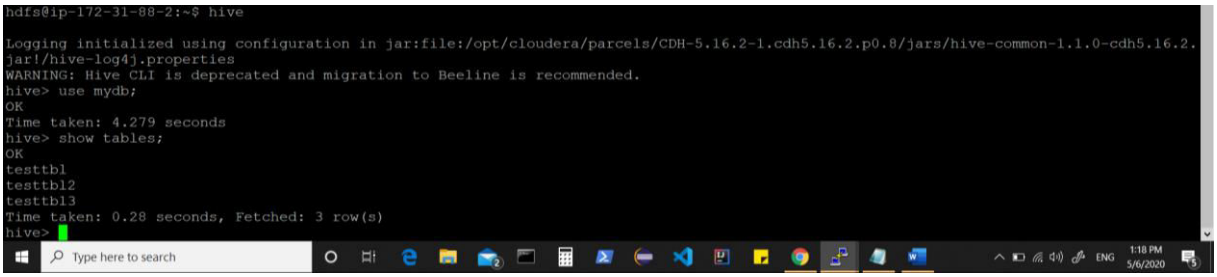
Workflow.xml contains

```
<workflow-app name="MyWorkflow1" xmlns="uri:oozie:workflow:0.5">
    <start to="hive-8bdc"/>
    <kill name="Kill">
        <message>Action failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
    </kill>
    <action name="hive-8bdc" cred="hive2">
        <hive2 xmlns="uri:oozie:hive2-action:0.1">
            <job-tracker>${jobTracker}</job-tracker>
            <name-node>${nameNode}</name-node>
            <jdbc-url>jdbc:hive2://ip-172-31-89-51:10000/default</jdbc-url>
            <script>${wf:appPath()}/hive-8bdc.sql</script>
        </hive2>
        <ok to="hive-6a11"/>
        <error to="Kill"/>
    </action>
    <action name="hive-6a11" cred="hive2">
        <hive2 xmlns="uri:oozie:hive2-action:0.1">
            <job-tracker>${jobTracker}</job-tracker>
            <name-node>${nameNode}</name-node>
            <jdbc-url>jdbc:hive2://ip-172-31-89-51:10000/default</jdbc-url>
            <script>${wf:appPath()}/hive-6a11.sql</script>
        </hive2>
        <ok to="End"/>
        <error to="Kill"/>
    </action>
    <end name="End"/>
</workflow-app>
```

Looking job run from hue



Finally verifying if hive new table was created



Workflow2

As previous example, we will create a workflow which will run a mapreduce job for wordcount.

Workflow3

We will create a workflow which will run a mapreduce job followed by a hive query.

Refer the files in folders.

https://github.com/ajaykuma/Hadoop_Spark_Setup/tree/master/Oozie_Setup