

# Bigdata Assignment 4.1

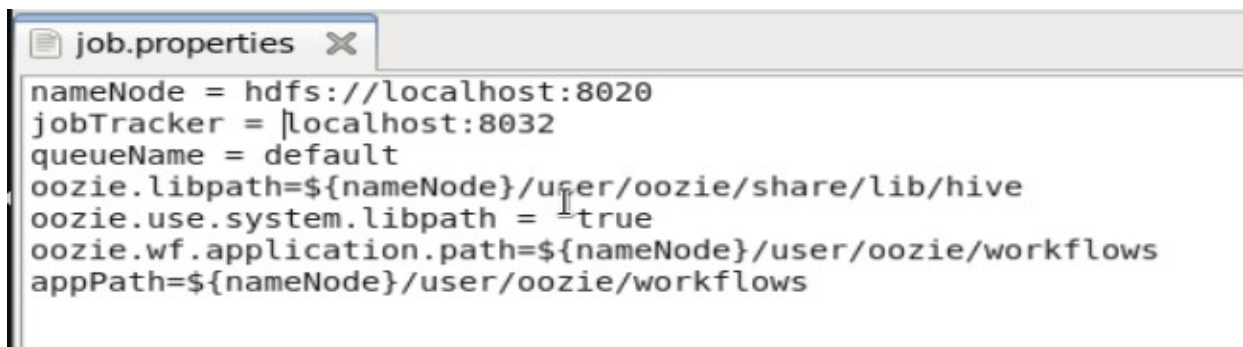
## Oozie Job Scheduling in Hive

Solution -

To schedule Hive job using Oozie, we need to write a Hive-action. The Oozie job will consist of mainly three things.

- 1.workflow.xml
- 2.job.properties
- 3.Hive script

- Created job.properties file inside /home/cloudera/RITESH directory using gedit job.properties command. This file consists of all the variables definitions that are used in workflow.xml.



```
nameNode = hdfs://localhost:8020
jobTracker = localhost:8032
queueName = default
oozie.libpath=${nameNode}/user/oozie/share/lib/hive
oozie.use.system.libpath = true
oozie.wf.application.path=${nameNode}/user/oozie/workflows
appPath=${nameNode}/user/oozie/workflows
```

Line 1: In nameNode variable, assigning address of namenode

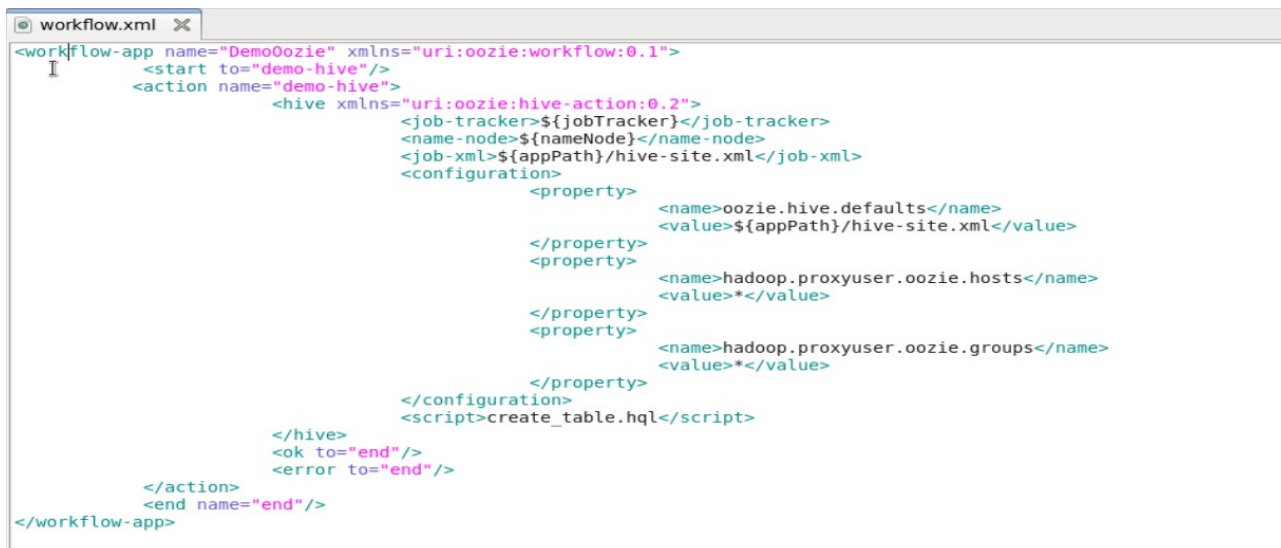
Line 2: In jobTracker variable, assigning address of resource manager

Line 3: queueName variable stores default value

Line 4: oozie.libpath stores that path where all hive .jar files are present

Line 5: oozie.use.system.libpath is set to true so that path specified at Line 4 is picked.

Line 6 & 7: oozie.wf.application.path and appPath store the path where all dependent files are present like workflow.xml, hive-script, hive-site.xml



```

<workflow-app name="DemoOozie" xmlns="uri:oozie:workflow:0.1">
  <start to="demo-hive"/>
  <action name="demo-hive">
    <hive xmlns="uri:oozie:hive-action:0.2">
      <job-tracker>${jobTracker}</job-tracker>
      <name-node>${nameNode}</name-node>
      <job-xml>${appPath}/hive-site.xml</job-xml>
      <configuration>
        <property>
          <name>oozie.hive.defaults</name>
          <value>${appPath}/hive-site.xml</value>
        </property>
        <property>
          <name>hadoop.proxyuser.oozie.hosts</name>
          <value>*</value>
        </property>
        <property>
          <name>hadoop.proxyuser.oozie.groups</name>
          <value>*</value>
        </property>
      </configuration>
      <script>create_table.hql</script>
    </hive>
    <ok to="end"/>
    <error to="end"/>
  </action>
  <end name="end"/>
</workflow-app>

```

- Created workflow.xml file inside /home/cloudera/RITESH directory using gedit workflow.xml command. This is the place where we write our Oozie action. It contains all the details of files, scripts required to schedule and run Oozie job. As the name suggests, it is an XML file where we need to mention the details in a proper tag.

The first line creates a workflow app and we assign a name (according to our convenience) to recognize the job.

**`<workflow-app name="DemoOozie">`**

Indicates, we are creating a workflow app whose name is 'DemoOozie'. All the other properties will remain inside this main tag.

**`<start to="demo-hive"/>`**

**`<action name="demo-hive">`**

First tag `<start to/>` gives a name to hive action (i.e. 'demo-hive') and when `<action name>` matches with `<start to/>` then it starts oozie job.

**`<hive xmlns="uri:oozie:hive-action:0.2">`**

The line above is very important as, it says what kind of action we are going to run. It can be a MR action, or a Pig action, or Hive. Here, name is specified as Hive-action.

**`<job-tracker>${jobTracker}</job-tracker>`**

**`<name-node>${nameNode}</name-node>`**

**`<job-xml>${appPath}/hive-site.xml</job-xml>`**

All the above tags point to the variable where job-tracker, NameNode, and Hive-site.xml are present. The exact declaration of these variables is done in Job.properties file.

**<script>create\_table.hql</script>**

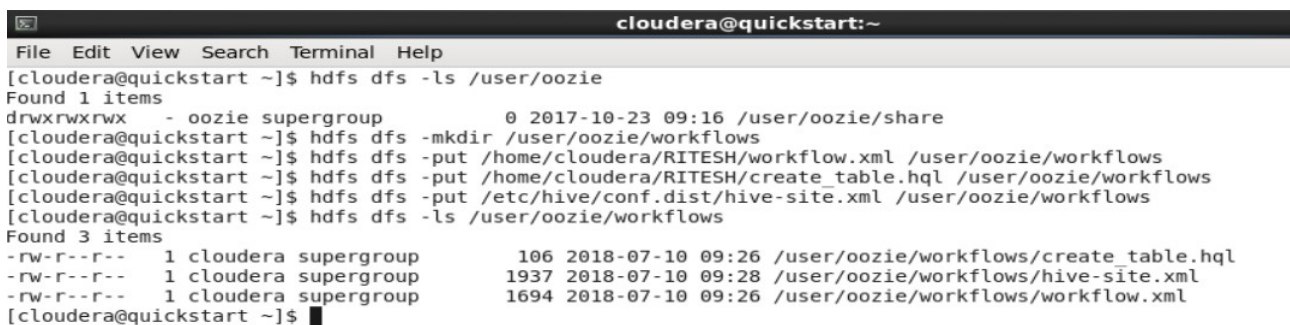
In this tag we need to write the exact name of script file (here, it is a Hive script file i.e. create\_table.hql) which will be looked for and the query will get executed.

- Created hive script i.e. create\_table.hql which we want to schedule in Oozie, inside /home/cloudera/RITESH directory using gedit create\_table.hql command.



```
use default;
create table hive_oozie(
id INT,
name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

- Created workflows directory in hdfs inside /user/oozie, and have put workflow.xml, create\_table.hql and hive-site.xml inside /user/oozie/workflows location using below commands



```
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -ls /user/oozie
Found 1 items
drwxrwxrwx - oozie supergroup          0 2017-10-23 09:16 /user/oozie/share
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/oozie/workflows
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/RITESH/workflow.xml /user/oozie/workflows
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/RITESH/create_table.hql /user/oozie/workflows
[cloudera@quickstart ~]$ hdfs dfs -put /etc/hive/conf.dist/hive-site.xml /user/oozie/workflows
[cloudera@quickstart ~]$ hdfs dfs -ls /user/oozie/workflows
Found 3 items
-rw-r--r--  1 cloudera supergroup      106 2018-07-10 09:26 /user/oozie/workflows/create_table.hql
-rw-r--r--  1 cloudera supergroup     1937 2018-07-10 09:28 /user/oozie/workflows/hive-site.xml
-rw-r--r--  1 cloudera supergroup      1694 2018-07-10 09:26 /user/oozie/workflows/workflow.xml
[cloudera@quickstart ~]$
```

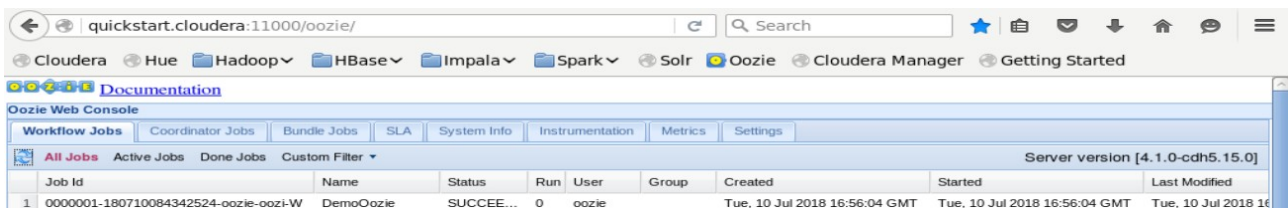
- Checked whether hive\_oozie table already exists inside default database or not

```
hive> show databases;
OK
default
Time taken: 0.911 seconds, Fetched: 1 row(s)
hive> use default;
OK
Time taken: 0.059 seconds
hive> show tables;
OK
Time taken: 0.223 seconds
hive>
```

- Ran Oozie job by using the below command.

```
cloudera@quickstart:~/RITESH
File Edit View Search Terminal Help
[cloudera@quickstart RITESH]$ sudo -u oozie oozie job -oozie http://127.0.0.1:11000/oozie -config job.properties -run
job: 0000001-180710084342524-oozie-oozi-W
[cloudera@quickstart RITESH]$
```

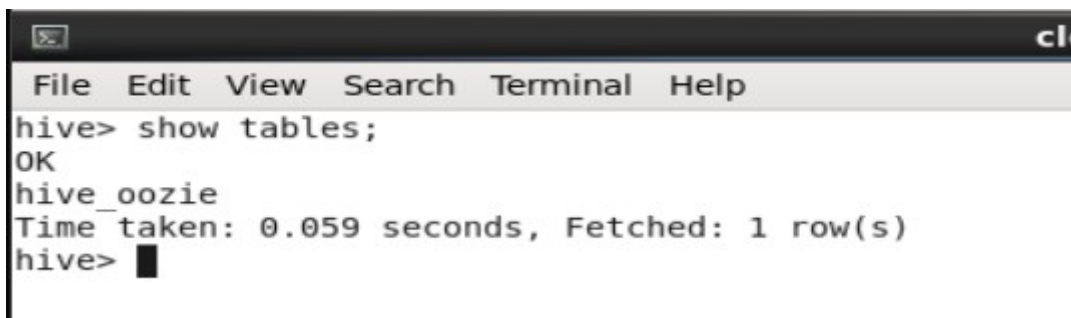
After running the job, checked the status of job in Oozie web console.



The screenshot shows the Oozie Web Console interface. The top navigation bar includes links for Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. The main content area displays the 'Oozie Web Console' with tabs for Workflow Jobs, Coordinator Jobs, Bundle Jobs, SLA, System Info, Instrumentation, Metrics, and Settings. Under 'Workflow Jobs', there are sub-tabs for All Jobs, Active Jobs, and Done Jobs. A table lists the job details for '0000001-180710084342524-oozie-oozi-W', showing it as 'SUCCE...' (Successful) with a status of '0' and user 'oozie'. The table also shows the job was created and started on 'Tue, 10 Jul 2018 16:56:04 GMT'.

Job Id	Name	Status	Run	User	Group	Created	Started	Last Modified
0000001-180710084342524-oozie-oozi-W	DemoOozie	SUCCE...	0	oozie		Tue, 10 Jul 2018 16:56:04 GMT	Tue, 10 Jul 2018 16:56:04 GMT	Tue, 10 Jul 2018 16:56:04 GMT

- Below screenshot shows that hive\_oozie table has been created successfully.



```
File Edit View Search Terminal Help
hive> show tables;
OK
hive_oozie
Time taken: 0.059 seconds, Fetched: 1 row(s)
hive>
```