**Oozie Exercise One**

**Import Dummy Data into Mysql**

**Scenario: Update the movie lens Dataset in hadoop from mysql.**

**Step: we will delete complete directory and table of movie-lens dataset and we will create again using oozie workflow and sqoop**

1. Create file by the name of oldmovies.sql to do so open notepad or any text editor like "sublime" then put the following scrip inside file and save it by the name of oldmovies.sql file

DROP TABLE movies;

CREATE EXTERNAL TABLE movies (movie\_id INT, title STRING, release DATE) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/maria\_dev/movies/';

INSERT OVERWRITE DIRECTORY '${OUTPUT}' SELECT \* FROM movies WHERE release < '1940-01-01' ORDER BY release;

1. Create Ozzie workflow.xml . To do so open notepad or any text editor like "sublime" then put the following XML scrip inside the file and save it by the name of workflow.xml

<?xml version="1.0" encoding="UTF-8"?>

<workflow-app xmlns="uri:oozie:workflow:0.2" name="old-movies">

<start to="sqoop-node"/>

<action name="sqoop-node"> <!--node name-->

<sqoop xmlns="uri:oozie:sqoop-action:0.2">

<job-tracker>${jobTracker}</job-tracker> <!--job tracking node is http://sandbox.hortonworks.com:8050 which is define in job properties-->

<name-node>${nameNode}</name-node> <!--name node is http://sandbox.hortonworks.com:8020 which is define in job properties-->

<prepare>

<delete path="${nameNode}/user/maria\_dev/movies"/>

</prepare>

<configuration>

<property>

<name>mapred.job.queue.name</name>

<value>${queueName}</value>

</property>

</configuration>

<command>import --connect jdbc:mysql://localhost/movielens --driver com.mysql.jdbc.Driver --table movies -target-dir /user/maria\_dev/movies -m 1</command> <!-- get data from mysql using sqoop and store data at /user/maria\_dev/movies directory -->

</sqoop>

<ok to="hive-node"/> <!-- if everything is ok then go to hive node -->

<error to="fail"/>

</action>

<action name="hive-node">

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

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

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

<prepare>

<delete path="${nameNode}/user/maria\_dev/oldmovies"/> <!--if oldmovies directory is present in hadoop then this line will delet oldmovies directory because when oldmovies.sql file will create this create directry again for new dataset -->

</prepare>

<configuration>

<property>

<name>mapred.job.queue.name</name>

<value>${queueName}</value>

</property>

</configuration>

<script>oldmovies.sql</script>

<param>OUTPUT=/user/maria\_dev/oldmovies</param> <!-- this output varibale is define inside the oldmovies.sql file -->

</hive>

<ok to="end"/>

<error to="fail"/>

</action>

<kill name="fail">

<message>Sqoop failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>

</kill>

<end name="end"/>

</workflow-app>

1. Create configuration file by the name of job.properties to do so open notepad or any text editor like "sublime" and put the following scrip in it and save it by the name of "job.properties" .

***Note configuration file extension must be ".properties" extension.***

nameNode=hdfs://sandbox.hortonworks.com:8020

jobTracker=http://sandbox.hortonworks.com:8050

queueName=default

oozie.use.system.libpath=true

oozie.wf.application.path=${nameNode}/user/maria\_dev/oozie\_movies/workflow.xml

1. Upload workflow.xml and oldmovies.sql files inside the HDFS using Ambari (File view) under the following path inside

user/maria\_dev/oozie\_movies"

1. Upload job.properties file inside VMware local directory e.g Documents/foldername/job.properties

**Run Oozie Job**

1. We need to install oozie connector for mysql to do so type below command on terminal

hadoop fs -put /usr/share/java/mysql-connector-java.jar /user/oozie/share/lib/lib-20171110144231/sqoop

**// NOTE "lib\_20171110144231" this folder must be different for every computer. You can get this folder name from ambare For that Goto ambari --> Click on File view --> then**

**goto user/oozie/share/lib**

1. Press Oozie button-> click on "service action" button then Press Restart all button.
2. Execute oozie job to do so type below command on terminal :

oozie job -oozie http://localhost:11000/oozie -config .Documents /job.properties -run

1. Type your sandbox ip on the browser : IP:110000/oozie (e.g 192.168.567.0:11000)/oozie

then You can see oozie view and oozie job)

Reference:

https://oozie.apache.org/docs/3.1.3-incubating/DG\_CommandLineTool.html

https://www.slideshare.net/martyhall/hadoop-tutorial-oozie