**Running A Map Reduce Program Using Oozie**



In Big data projects different **extract/transform/load (ETL) and pre-processing** operations are needed to start the actual processing jobs and *Oozie* is a framework that helps to automate this process and codify this work into repeatable and reusable units or workflows.

In this blog we will  be learning regarding the creation of  a workflow to run a [MapReduce](https://acadgild.com/big-data/big-data-development-training-certification) program using Oozie.

[Get Skilled in Big Data](https://acadgild.com/big-data/big-data-development-training-certification)

So let’s first discuss [Oozie’s basics](https://acadgild.com/big-data/big-data-development-training-certification) and its working details.

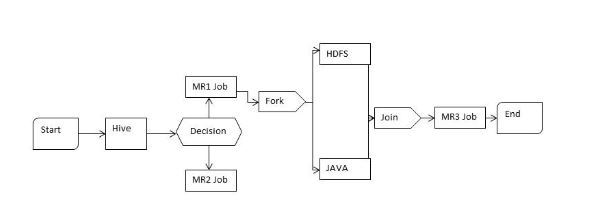
Apache Oozie is an open source project based on Java technology that simplifies the process of creating workflows and managing coordination among jobs.

The ability to combine multiple jobs sequentially into one logical unit of work makes it one of the preferred tool for creating workflow scheduler for jobs.

**Oozie Workflow:**

Oozie provides a structure called workflow engine which stores and runs workflows composed of different types of Hadoop jobs **(MapReduce, Pig, Hive etc.)**

In Oozie terms, a workflow is a Direct Acyclic Graph of **action nodes and control-flow nodes**.



An ***action node*** performs a workflow task, such as moving files in HDFS, running a MapReduce, Pig or Hive job, performing a Sqoop import or running a shell script or Java program.

A ***control-flow node*** includes:

**1.***Starting point and ending point* of a workflow (start , end and fail nodes)

**Start control node:** It is the starting point (first workflow node) for a workflow job

**End control node:** It is the ending point for a workflow job and indicates job has been completed successfully

**Kill control node:** It allows a workflow job to stop/kill itself.

Eg: consider a scenario where more than one actions are started by the workflow job and are in action(running) when the kill node is reached, all the actions that are executing will be stopped at that point of time.

**2.** A structure which control the workflow job execution path( decision , fork and join nodes)

**Decision control node:** It allows a workflow to decide the execution path to follow.

**Fork and Join:** Fork node splits one path of execution into multiple concurrent paths of execution. And Join node will be idle (will halt) until every concurrent execution paths of a previous fork node reach at the join node. (Split an action executing path and after execution merge **all** the executed (split) paths)

**Running an Oozie program:**

To run Oozie program user should have:

***workflow.xml file – It****consists of  all the specific properties and configurations of the job which is referred by oozie at the time of execution.*

***job.properties file –  It****consists of address of name node, job tracker(Resource manager), path of application directory which contains workflow.xml and inside it contains lib directory which contains jar file required to run the job.*

***.jar file –****It consists of library files and input program(Java MR/.pig file/Hbase shell script/.hql(for hive)).*

***input file – Sample dataset to be used.***

***Output directory*** – Output location in HDFS to store output.

**Use the below link to download all the above mentioned files.**

[Files to run MR program in Oozie](https://drive.google.com/open?id=0B_P3pWagdIrrSWhtTVJfeUcwQ1k)

***Oozie execution steps in cloudera:***

**1.** The first step to execute any program in hadoop is that we need to start that component using the required command and the command to start Oozie is:  
***sudo service oozie start***

**2.** Next we need to check whether Oozie is started successfully. We can do it by entering the below mentioned url in web browser.  
***http://localhost:11000***

If Oozie is not started, we will get an error message as shown above. The above image shows that Oozie is not started.

The above image shows that Oozie is started successfully.  
Once Oozie is started then we can continue with the step 3.

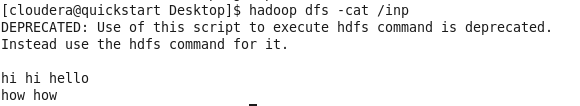
**3.** We have to copy the input dataset into hdfs to perform WordCount(MR) on it.  
command to put the input dataset into hdfs is:

***hadoop fs -put ~/Downloads/wordcount/inp /***

https://s3.amazonaws.com/acadgildsite/wordpress_images/bigdatadeveloper/Beginner%27s+guide+for+Oozie/store+input+file+in+hdfs.png

we can check the contents of the input file using command.

***hadoop fs -cat /inp***



Here  our input file name is **“inp”**

***4.***Next we have to create an application directory in HDFS, where it will hold the workflow.xml file and lib directory(within lib directory we should store .jar file which conatins MR program) .  
command to create application directory in hdfs is :

***https://i2.wp.com/s3.amazonaws.com/acadgildsite/wordpress_images/bigdatadeveloper/Beginner%27s+guide+for+Oozie/Creating+sample(Application)+directory+in+hdfs.PNG?zoom=0.9024999886751175&resize=538%2C33&ssl=1***Here our application directory name is **“sample”**

***5.*** Next we have to Create an workflow.xml file which contains all the job related configurations and copy this workflow.xml file into application directory, which is in hdfs.

***hadoop dfs -put ~/Downloads/wordcount/workflow.xml  /sample/***

https://i1.wp.com/s3.amazonaws.com/acadgildsite/wordpress_images/bigdatadeveloper/Beginner%27s+guide+for+Oozie/store+workflow.xml+file+in+hdfs+application+directory.PNG?zoom=0.9024999886751175&resize=766%2C36&ssl=1

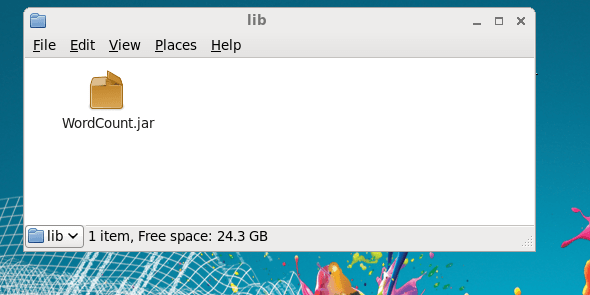
The below image represents the workflow.xml file. Where we should include workflow app name, job tracker & name node port number, input file name, output directory name, input & output file type to run an Oozie program.

***6.***Next we have to copy the .jar file into lib directory then we should store this lib directory in application directory which is in hdfs.

***hadoop fs –put ‘/home/cloudera/Downloads/wordcount/lib’ /sample/***

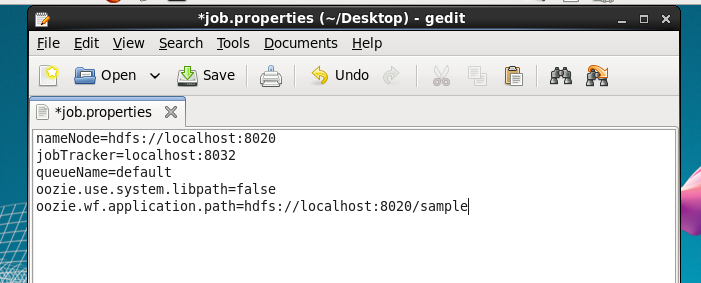
https://i0.wp.com/s3.amazonaws.com/acadgildsite/wordpress_images/bigdatadeveloper/Beginner%27s+guide+for+Oozie/store+lib+directory+in+hdfs+application+directory.PNG?zoom=0.9024999886751175&resize=756%2C28&ssl=1

The below image shows WordCount.jar file in lib directory.



Note:The directory which is holding .jar file in application directory should be named as **lib**.

***7.***Create job.properties file in local file system where it should contain the below commands to run the Oozie workflow job



– We have to update nameNode and jobTracker address and port numbers  
– We have to update the queueName , since we are not scheduling multiple jobs in this program we have set queueName to default(FIFO).  
– We have to set oozie.use.system.libpath=false because we dont require any additional jar files in the lib directories to run this Java mapreduce program since we have already exported required jar files before saving as it to .jar file. But we must be enusre that, if we are performing pig, Hive, HBase programs we must include additional required jar files in the lib directory or set oozie.use.system.libpath=true.  
– We should specify the path of application directory.

**8.** Now we can Execute the Oozie workflow by using the below command

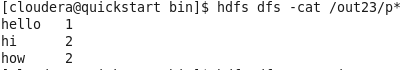
***oozie job -oozie http://localhost:11000/oozie -config ~/Desktop/job.properties -run***

https://s3.amazonaws.com/acadgildsite/wordpress_images/bigdatadeveloper/Beginner%27s+guide+for+Oozie/oozie_execution.PNG

In the above command we should give the full path of job.properties in the local file system

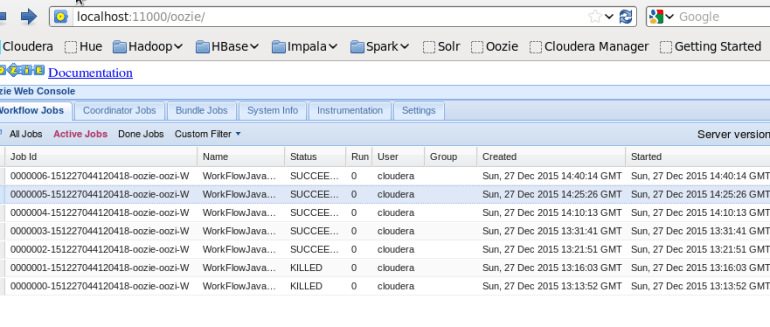
**9.** The final output of Oozie workflow can be found in hdfs output directory. Here the output directory is out23 and ‘p’ represents the output part file in the output directory.

hadoop fs –cat /out23/p\*

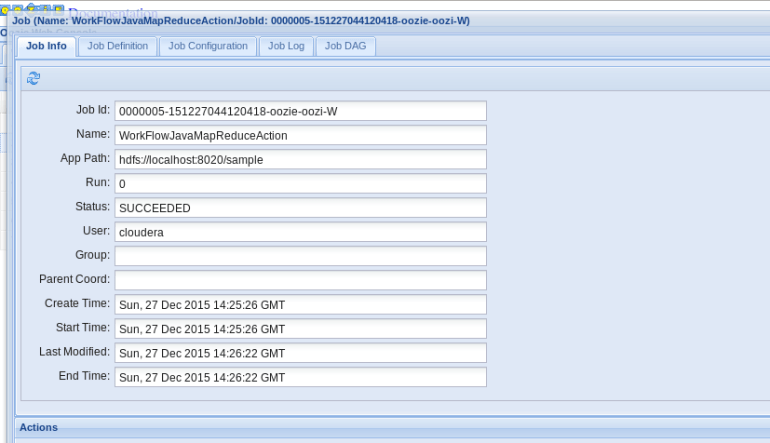


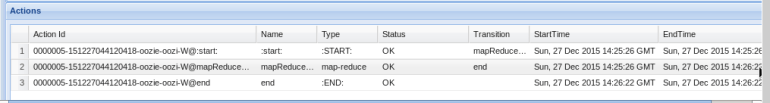
***We can also check Oozie Job status and detailed information of workflow on Oozie web UI:***

⦁ Enter the url localhost:11000 in the web browser  
⦁ There you can find submitted workflow jobs complete information with Job Id, Name, Status and many more.



⦁ You can select the particular Job Id where you can find job status and job control flow in Action part





This completes Oozie guide for Beginners.