Lab: Running an Oozie Workflow

|  |  |
| --- | --- |
| **Objective:** | Deploy and run an Oozie workflow |
| **Successful Outcome:** | You will run an Oozie job that executes a Pig script and a Hive script. |
| **Before You Begin:** | SSH into hadoop-master. |

1. View the Raw Data
   1. On hadoop-master, change directories to the oozielab folder:

# cd ~/labs/oozielab/

* 1. Unzip the archive in the oozielab folder, which contains a file named whitehouse\_visits.txt that is quite large:

# unzip whitehouse\_visits.zip

* 1. View the contents of this file:

# tail whitehouse\_visits.txt

This publicly available data contains records of visitors to the White House in Washington, D.C.

1. Load the Data into HDFS
   1. Make a new directory in HDFS named whitehouse. (If you already have a whitehouse folder in HDFS, delete it first):

# hadoop fs -rm -R whitehouse

# hadoop fs -mkdir whitehouse

* 1. Use the put command in the Grunt shell to copy the whitehouse\_visits.txt file the whitehouse folder in HDFS, renaming the file visits.txt. (Be sure to enter this command on a single line):

# hadoop fs -put whitehouse\_visits.txt whitehouse/visits.txt

* 1. Use the ls command to verify the file was uploaded successfully:

# hadoop fs -ls whitehouse

Found 1 items

-rw-r--r-- 3 root root 183292235 whitehouse/visits.txt

1. Configure Oozie User Permissions
   1. In Ambari, go to the Services page and click the Stop All button (or use the shutdown script in the /root/scripts folder).
   2. Go the HDFS page in Ambari, then scroll down and expand the Custom core-site.xml section.
   3. The Oozie workflow you defined is going to be executed by the root user, so root needs permission to communicate with the Oozie server. Add root to the hadoop.proxyuser.oozie.groups property:

Description: Macintosh HD:Users:rich:Desktop:Screen Shot 2013-11-11 at 8.37.54 AM.png

* 1. Click the Add Property... link and add two properties. Assign the hadoop.proxyuser.root.hosts property to \* and also the hadoop.proxyuser.root.groups:



* 1. Click the Save button to save your changes to the HDFS config.
  2. Restart all dependent services as marked in Services tab.

1. Deploy the Oozie Workflow
   1. SSH into **hadoop-master**.
   2. View the file workflow.xml in /root/labs/oozielab.
   3. How many actions are in this workflow? \_\_\_\_\_\_\_\_\_\_\_\_\_
   4. Which action will execute first? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   5. If the first action is successful, which action will execute next? \_\_\_\_\_\_\_\_\_\_\_\_
   6. To deploy this workflow, we need a directory in HDFS:

# hadoop fs -mkdir congress

* 1. Put congress\_visits.hive and whitehouse.pig from the oozielab folder into the new congress folder in HDFS.
  2. Also, put workflow.xml into the congress folder.
  3. If you look at the Hive action in workflow.xml, you will notice that it references a file named hive-site.xml within the <job-xml> tag. This file represents the settings Oozie needs to connect to your Hive instance, and the file needs to be deployed in HDFS (using a relative path to the workflow directory). Put hive-site.xml into the congress directory:

# hadoop fs -put /etc/hive/conf/hive-site.xml congress

* 1. Verify you have four files now in your congress folder in HDFS:

# hadoop fs -ls congress

Found 4 items

-rw-r--r-- 3 root root 429 congress/congress\_visits.hive

-rw-r--r-- 3 root root 3509 congress/hive-site.xml

-rw-r--r-- 3 root root 580 congress/whitehouse.pig

-rw-r--r-- 3 root root 1623 congress/workflow.xml

1. Define the OOZIE\_URL Environment Variable
   1. Although not required, you can simplify oozie commands by defining the OOZIE\_URL environment variable. From the command line, enter the following command:

# export OOZIE\_URL=http://hadoop-master:11000/oozie

1. Define the Job Properties
   1. View the contents of job.properties in oozielab.
   2. Notice the oozie.wf.application.path property points to the congress folder in HDFS. This property is how you denote which Oozie job is going to execute.
   3. Make sure the resourceManager and nameNode properties are defined properly.
2. Run the Workflow
   1. From the oozielab folder, run the workflow with the following command:

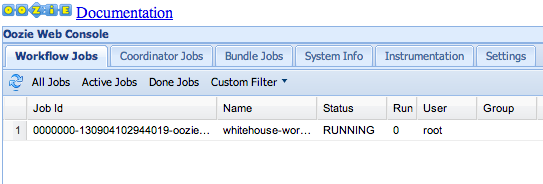
# oozie job -config job.properties -run

If successful, the job ID should be displayed at the command prompt.

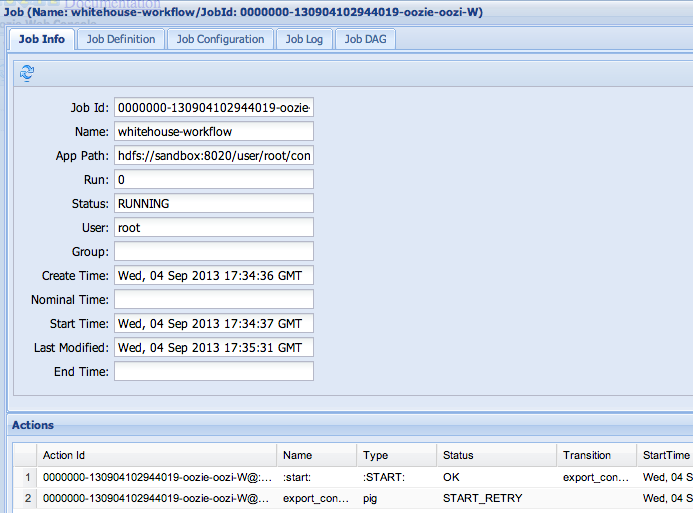
1. Monitor the Workflow
   1. Point your Web browser to the Oozie Web Console:

http://*hadoop-master*:11000/oozie/

You should see your Oozie job in the list of Workflow Jobs:



* 1. Double-click on the Job Id to view the Job Info page:



Notice you can view the status of each Action within the workflow.

1. Verify the Results
   1. Once the Oozie job is completed successfully, start the Hive Shell.
   2. Run a select statement on congress\_visits and verify the table is populated:

hive> select \* from congress\_visits;

...

WATERS MAXINE 12/8/2010 17:00 POTUS OEOB MEMBERS OF CONGRESS AND CONGRESSIONAL STAFF

WATT MEL 12/8/2010 17:00 POTUS OEOB MEMBERS OF CONGRESS AND CONGRESSIONAL STAFF

WEGNER DAVID L 12/8/2010 16:46 12/8/2010 17:00 POTUS OEOB MEMBERS OF CONGRESS AND CONGRESSIONAL STAFF

WILLOUGHBY JEANNE P 12/8/2010 17:07 12/8/2010 17:00 POTUS OEOB MEMBERS OF CONGRESS AND CONGRESSIONAL STAFF

WILSON ROLLIE E 12/8/2010 16:49 12/8/2010 17:00 POTUS OEOB MEMBERS OF CONGRESS AND CONGRESSIONAL STAFF

YOUNG DON 12/8/2010 17:00 POTUS OEOB MEMBERS OF CONGRESS AND CONGRESSIONAL STAFF

MCCONNELL MITCH 12/14/2010 9:00 POTUS WH MEMBER OF CONGRESS MEETING WITH POTUS.

Time taken: 1.082 seconds, Fetched: 102 row(s)

**RESULT**: You have just executed an Oozie workflow that consists of a Pig script followed by a Hive script.

**ANSWERS:**

Step 4.2: Two

Step 4.3: The Pig action named export\_congress

Step 4.4: The Hive action named define\_congress\_table