Oozie - Introduction

Oozie is a Java Web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work and executes them. It Is integrated with the rest of the Hadoop stack. It can execute Hadoop jobs out of the box such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp as well as system specific jobs such as Java programs and shell scripts.

[Oozie - Jobs]

There are two types of Oozie jobs:

- Workflow jobs
- Coordinator jobs

[Oozie - Jobs - Workflow Jobs]

Oozie Workflow jobs are Directed Acyclical Graphs - DAGs, specifying a sequence of actions to execute. DAG is a finite directed graph with no cycles. As shown in the image task 10 can only be executed after task 11 and 3 are executed.

Examples - Almost all task execution systems use DAG. Most Source Control Management Systems implement the revisions as a DAG

[Oozie - Jobs - Coordinator Jobs]

Oozie Coordinator jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability.

Example - Let’s say if we want to take data from HDFS and put to Hive every one hour, we can define a workflow in Oozie to take data from HDFS and put to Hive and run it as a coordinator job.

[Oozie - Use Case]

Let's understand a use case where we will be using Oozie.

Let's say we want to push web server access logs to HDFS and run Spark MLlib recommendation algorithm every day to generate recommendations and display it to the user. To do this, typical steps will be

- Flume agents will be running on web servers and will push access logs to HDFS
- Pig Script will be taking access log from HDFS, clean it and again push the cleaned data to HDFS
- Spark will take cleaned data from HDFS, run the MLlib recommendation algorithm on cleaned data and will push the recommendations to HDFS
- Sqoop will take recommendations from HDFS and will push it to MySQL

- The web server will take recommendations from MySQL and will display it to the end user.

In this use case, Steps 2-7 can be run as Oozie Coordinator jobs daily.

[Oozie - Workflow - XML]

We define Oozie workflow in XML files. Let’s see a sample XML for MapReduce operation. We define map class, reduce class, input and output directories in XML. We can configure similar workflows for other actions like Hive, Pig etc

Running Oozie Workflow From Command Line

Hands On Steps

1. Login to Web Console
2. Copy oozie examples to your home directory in web console: `cp /usr/hdp/current/oozie-client/doc/oozie-examples.tar.gz .`
3. Extract files from tar `tar -zxvf oozie-examples.tar.gz`
4. Edit examples/apps/map-reduce/job.properties and set: `nameNode=hdfs://10.142.1.1:8020 jobTracker=10.142.1.2:8050 queueName=default examplesRoot=examples`
5. Copy the examples directory to HDFS `hadoop fs -copyFromLocal examples`
6. Run the job `oozie job -oozie http://10.142.1.2:11000/oozie -config examples/apps/map-reduce/job.properties -run`
7. Check the job status for the job_id printed in previous step `oozie job -oozie http://10.142.1.2:11000/oozie -info job_id`

Script

Let’s run an Oozie job for MapReduce action. Login to CloudxLab Linux console. Copy Oozie examples to your home directory in the console. Extract files from tar. Edit examples/apps/map-reduce/job.properties and set the value of namenode and jobtracker. We can find the namenode host from Ambari under “HDFS” section.

We will be running examples/apps/map-reduce/workflow.xml in our job. Copy the examples directory to HDFS and run the job using the command displayed on the screen. cxln2.c.thelab-240901.internal:11000 is the host and port where Oozie server is running.

Press enter. We will get the job id in the command prompt. To check the status of job type command displayed on the screen. Job status is "Running".

Running Oozie Workflow From Hue

Oozie - Hands On

```
# In Hue, Get the location of mysql Jar
/data/jars/mysql-connector-java.jar

# Go to Workflows -> Editors

# Click on Create Button on right site

#Drag and drop sqoop1

# Set the password in following command and copy paste
# Also, change the HDFS absolute location
import --connect jdbc:mysql://10.142.1.2:3306/sqoopex --username sqoopuser --password
NHkkP876rp --table widgets --target-dir hdfs:///user/sandeepgiri9034/widgets_import

# Add the files of mysql connector

# Save and Submit

# Open File Browser check if the files are created.
```

[Oozie - Example - Using Hue]

Let’s run an Oozie job using Hue. Hue provides a really nice UI for defining workflows. We will run Sqoop import where we will take data from MySQL widgets table in sqoopex database and push to HDFS.

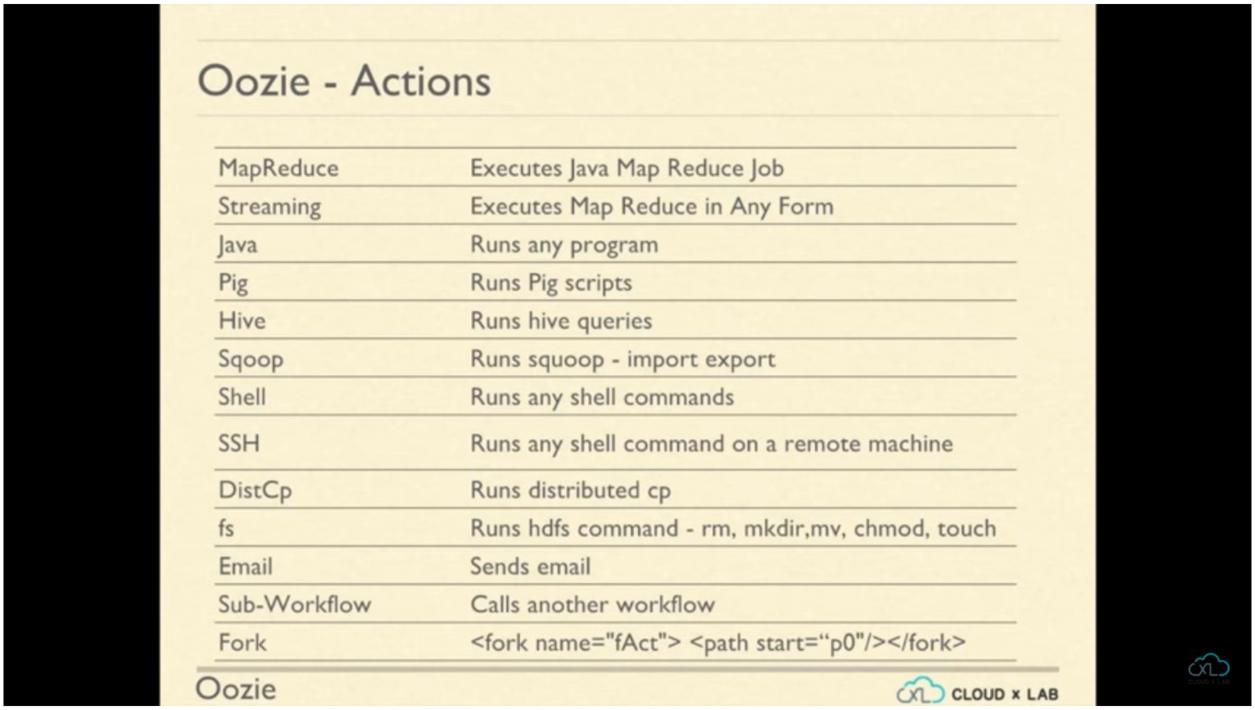
Login to Hue. Click on “Workflows”, and then “Editors". It will open a dashboard with our previous workflows. Click on “Create”. Select action as “sqoop1” and drop it to the editor. Type in command displayed on the screen. ip-172-31-13-154 is the host where MySQL server is running. We can find MySQL host, username and password from “MySQL Credentials” section under “My Lab”. We will be storing the data in widgets_import directory in HDFS under our home directory.

Click on "Add"

Now specify MySQL connector JAR which is located at /user/oozie/share/lib/sqoop. Click on “Files” and add the MySQL connector JAR

Let's name our workflow as “Sqoop import”. Press Enter. Click on “save” and then click on “Submit”. Click on “Submit” when the confirmation dialog popups.

Our job has started running now. We'll Wait for some time to get it executed. To see the data let’s go to the widgets_import directory in HDFS. We can see multiple files inside widgets_import containing data from widgets table.

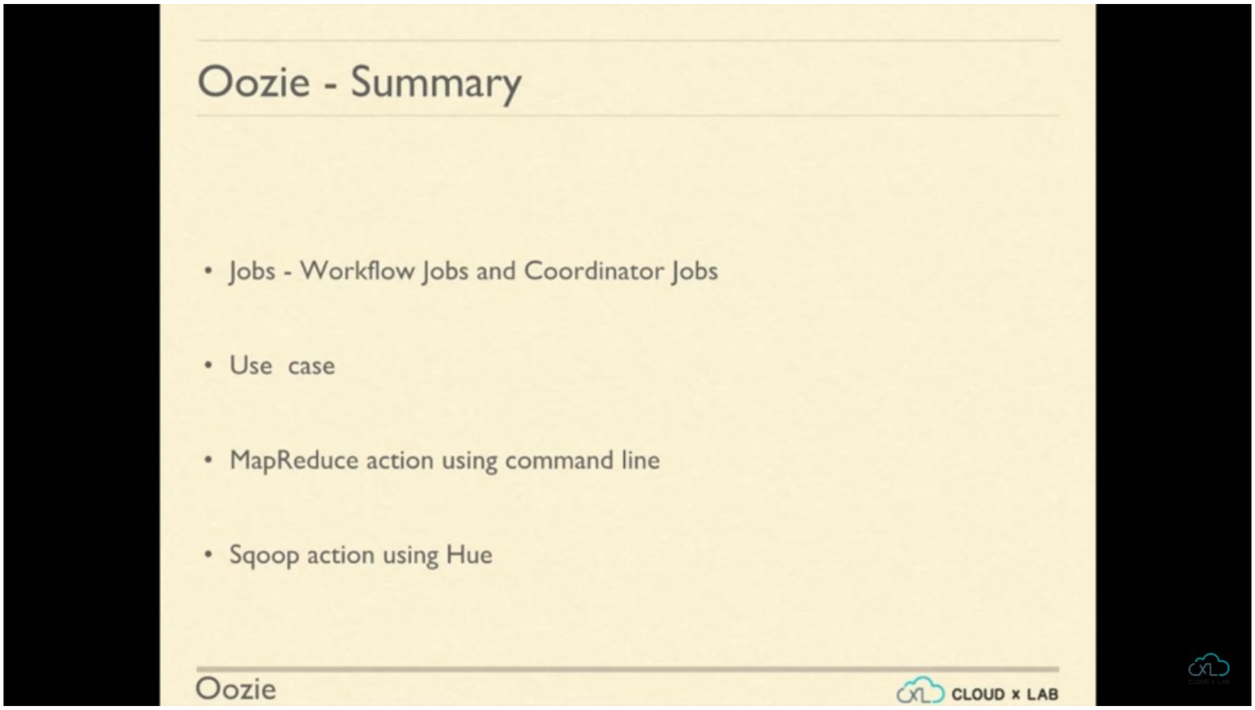


Oozie - Actions

MapReduce	Executes Java Map Reduce Job
Streaming	Executes Map Reduce in Any Form
Java	Runs any program
Pig	Runs Pig scripts
Hive	Runs hive queries
Sqoop	Runs squoop - import export
Shell	Runs any shell commands
SSH	Runs any shell command on a remote machine
DistCp	Runs distributed cp
fs	Runs hdfs command - rm, mkdir,mv, chmod, touch
Email	Sends email
Sub-Workflow	Calls another workflow
Fork	<fork name="fAct"> <path start=“p0"/></fork>

Oozie

CLOUD x LAB



Oozie - Summary

- Jobs - Workflow Jobs and Coordinator Jobs
- Use case
- MapReduce action using command line
- Sqoop action using Hue

Oozie

CLOUD x LAB