**Oozie Workflow**

**Control Flow**

• Start/end/kill

• Decision

• Fork/join

**Actions**

• Map-reduce

• Pig

• Hdfs

• Sub-workflow

• Java-run custom java code

• To run oozie workflows, two files are needed.

**1. workflow.xml (stored in HDFS)**

• It contains the structure of workflow.

**2. job.properties (stored in local)**

• It contains the configuration properties.

• The Oozie server is designed to work with either MR V1 or YARN. Please note that it cannot work with both simultaneously

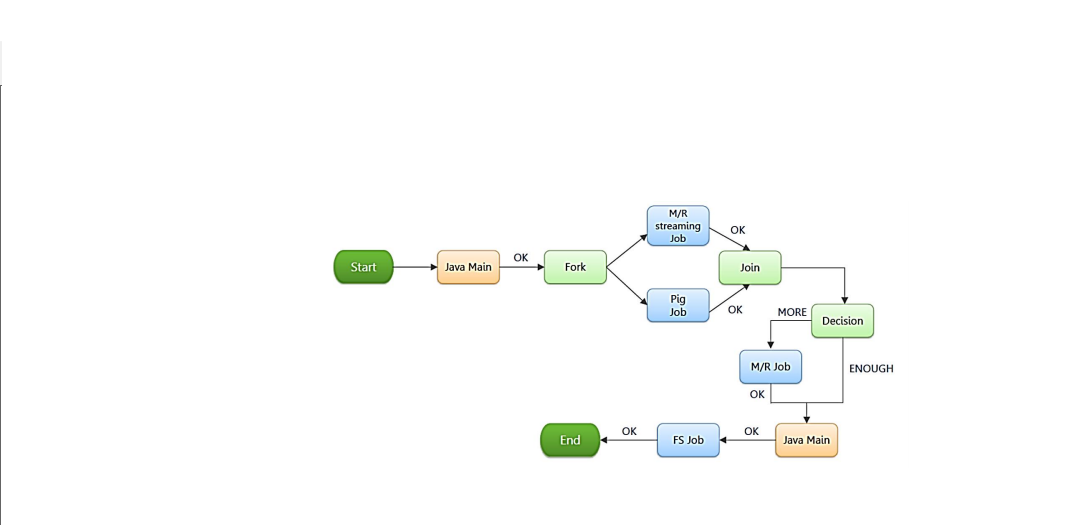
• It can be configured with CATALINA\_BASE variable in /etc/oozie/conf/oozie-env.sh

**Hadoop 1**

• CATALINA\_BASE = /usr/lib/oozie/oozie-server-0.20

**Hadoop 2**

• CATALINA\_BASE=/usr/lib/oozie/oozie-server



**Oozie Sample workflow**

<workflow-app name="Name of Workflow" xmlns="uri:oozie:workflow:0.1">

<start to="Start Node"/>

<action name="Name of Node">

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

<job-tracker>Job Tracker Address</job-tracker>

<name-node>Name Node Address</name-node>

<job-xml>Path of hive-site.xml in HDFS</job-xml>

<configuration>

<property>

<name>Configuration Name</name>

<value>Configuration Value</value>

</property>

</configuration>

<script>Hive Query File</script>

</hive>

<ok to="end"/>

<error to="end"/>

</action>

**Benefits**

1-Oozie is designed to scale in a Hadoop cluster. Each job will be launched from a different datanode. This means that the workflow load will be balanced and no single machine will become overburdened by launching workflows. This also means that the capacity to launch workflows will grow as the cluster grows.

2-Oozie is well integrated with Hadoop security. This is especially important in a kerberized cluster. Oozie knows which user submitted the job and will launch all actions as that user, with the proper privileges. It will handle all the authentication details for the user as well.

3-Oozie is the only workflow manager with built-in Hadoop actions, making workflow development, maintenance and troubleshooting easier.

4-Oozie UI makes it easier to drill down to specific errors in the data nodes. Other systems would require significantly more work to correlate jobtracker jobs with the workflow actions.

5-Oozie is proven to scale in some of the world’s largest clusters. The white paper discusses a deployment at Yahoo! that can handle 1250 job submissions a minute.

6-Oozie gets callbacks from MapReduce jobs so it knows when they finish and whether they hang without expensive polling. No other workflow manager can do this.

7-Oozie Coordinator allows triggering actions when files arrive at HDFS. This will be challenging to implement anywhere else.

8-Oozie is supported by Hadoop vendors. If there is ever an issue with how the workflow manager integrates with Hadoop – you can turn to the people who wrote the code for answers.