**ASSIGNMENT 34.1**

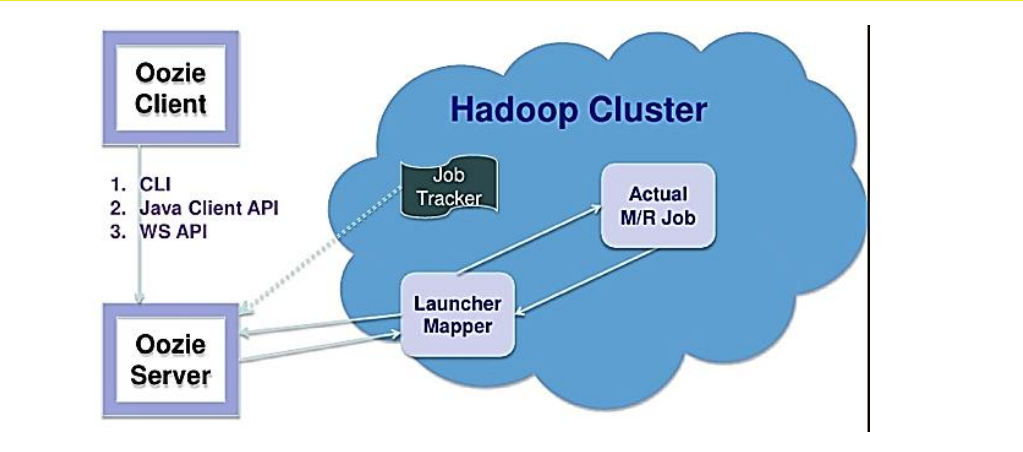
**The complete structure and the working of “Oozie Workflow scheduler “**

1. Apache Oozie is a Java Web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work.

2. It is integrated with the Hadoop stack, with YARN as its architectural center, and supports Hadoop jobs for Apache MapReduce, Apache Pig, Apache Hive, and Apache Sqoop. Oozie can also schedule jobs specific to a system, like Java programs or shell scripts.

3. Apache Oozie is a tool for Hadoop operations that allows cluster administrators to build complex data transformations out of multiple component tasks. This provides greater control over jobs and also makes it easier to repeat those jobs at predetermined intervals.

4. At its core, Oozie helps administrators derive more value from Hadoop.

****

**WORK FLOW**

1. Oozie runs as a service in the cluster and clients submit workflow definitions for immediate or later processing.

2. Oozie workflow consists of action nodes and control-flow nodes.

3. An action node represents a workflow task, e.g., moving files into HDFS, running a MapReduce, Pig or Hive jobs, importing data using Sqoop or running a shell script of a program written in Java.

4. A control-flow node controls the workflow execution between actions by allowing constructs like conditional logic wherein different branches may be followed depending on the result of earlier action node.

5. Start Node, End Node and Error Node fall under this category of nodes.

6. Start Node, designates start of the workflow job.

7. End Node, signals end of the job.

8. Error Node, designates an occurrence of error and corresponding error message to be printed.

9. At the end of execution of workflow, HTTP callback is used by Oozie to update client with the workflow status. Entry-to or exit-from an action node may also trigger callback.

