**Assignment 10.1:**

1. **Explain in brief the workflow of Oozie and its benefits**

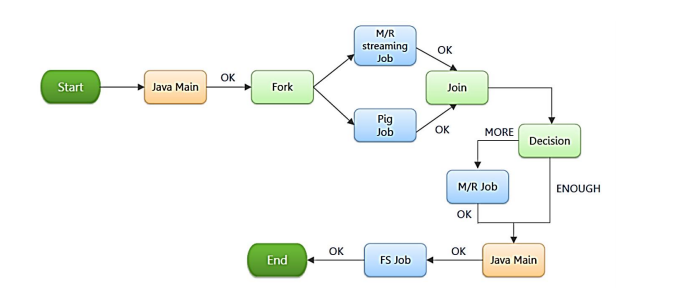
Oozie Workflow Nodes:

1. Control flow:
   1. Start / end / kill
   2. Decision
   3. Fork / Join
2. Actions
   1. Map-reduce
   2. Pig
   3. Hdfs
   4. Sub-workflow
   5. Java-run custom java code

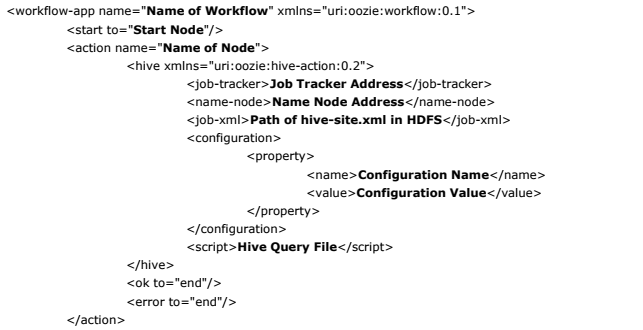
Two files are needed to run oozie workflows:

1. Workflow.xml (stored in HDFS)
   * + It contains the structure of workflow.
2. Job.properties (stored in local)
   * + It contains the configuration properties.

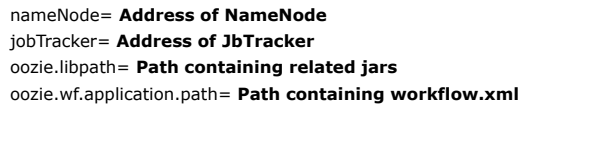
**Oozie workflow:**



**Oozie Sample workflow:**



**Oozie sample properties file:**



[Types of Nodes in Apache Oozie:](http://www.besthadooptraining.in/)

**Action Node –** It represents the workflow jobs and jobs program are written in java

**Control Flow Node –** It used to controls the workflow jobs between actions

**Start Node –** It used to starts the jobs execution

**End Node –** It used to stops the jobs execution

**Error Node –** If any error occurs while execution of job error node prints the error message

**Benefits of Oozie:**

Oozie is mainly used to manage the [hadoop jobs in HDFS](http://www.credosystemz.com/training-in-chennai/best-bigdata-training-in-chennai/) and it combines the multiple jobs in particular order to achieve the big task. It is the open source framework and used to make multiple hadoop jobs. Oozie supports the jobs in mapreduce, hive and hdfs also. In Oozie job workflow based on Directed Acylic Graph and it contains two nodes for managing the jobs that nodes are action and control flow nodes.

Advantages of Oozie is it integrate with [hadoop stack](http://www.credosystemz.com/training-in-chennai/best-bigdata-training-in-chennai/) and also support mapreduce and hdfs jobs.

1. **Explain in brief the workflow of Sqoop and its benefits**

Sqoop allows easy import and export of data from structured data stores such as

relational databases, enterprise data warehouses, and NoSQL systems.

Using Sqoop, we can provision the data from external system on to HDFS, and populate

tables in Hive and HBase.

Sqoop integrates with Oozie, allowing you to schedule and automate import and export

tasks.

Sqoop uses a connector based architecture which supports plugins that provide

connectivity to new external systems.

The following command is used to import all data from a table called ORDERS from a

MySQL database:

***sqoop import --connect jdbc:mysql://localhost/acmedb --table ORDERS --username test --***

***password \*\*\*\****

In this command the various options specified are as follows:

**import:** This is the sub-command that instructs Sqoop to initiate an import.

**--connect <connect string>**

Connection String contains type of RDBMS Server, RDBMS Sever IP Address, port,

database name and other optional arguments

**--username <user name>**

**--password <password>:**

These are connection parameters that are used to connect with the database. This is no

different from the connection parameters that you use when connecting to the database

via a JDBC connection.

**--table <table name>:**

This parameter specifies the table which will be imported.

The import is done in two steps:

* In the first Step Sqoop introspects the database to gather the necessary metadata for the data being imported.
* The second step is a map-only Hadoop job that Sqoop submits to the cluster.
* It is this job that does the actual data transfer using the metadata captured in the previous step.

