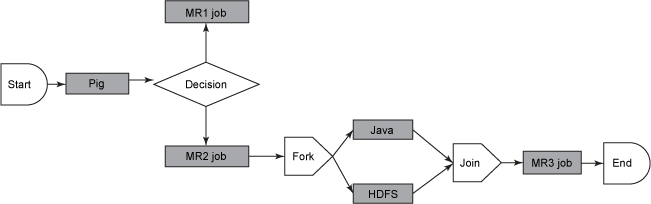
SESSION 10 ASSIGNMNET 1

The Workflow Of Oozie And Its Benefits

**Oozie** is an open source project based on Java™ technology that simplifies the process of creating workflows and managing coordination among jobs. In principle, Oozie offers the ability to combine multiple jobs sequentially into one logical unit of work.

**WorkFlow** -> An Oozie workflow is a collection of actions arranged in a directed acyclic graph (DAG). This graph can contain two types of nodes: control nodes and action nodes. Control nodes, which are used to define job chronology, provide the rules for beginning and ending a workflow and control the workflow execution path with possible decision points known as fork and join nodes. Action nodes are used to trigger the execution of tasks. In particular, an action node can be a MapReduce job, a Pig application, a file system task, or a Java application.



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 or YARN but cannot work with both simultaneously.

**Benefits** ->

* 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.
* Oozie is well integrated with Hadoop security. 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.
* Oozie is the only workflow manager with built-in Hadoop actions, making workflow development, maintenance and troubleshooting easier.
* Oozie UI makes it easier to drill down to specific errors in the data nodes.
* 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.

The Workflow Of Sqoop And Its Benefits

**Sqoop** is a command-line interface application for transferring data between relational databases and Hadoop. Sqoop supports incremental loads of a single table or a free form SQL query as well as saved jobs which can be run multiple times to import updates made to a database since the last import. Imports can also be used to populate tables in Hive or HBase.

**WorkFlow ->**

****

**Architecture**

On execution of Sqoop command internally Map Reduce jobs are fired with **zero** reducers as there is no shuffle operation involved only transfer of data happens from database. By default no of mappers induced by sqoop is 4(can be changed).

Sqoop is used to transfer data from database to hdfs using JDBC connection through which it gets access to metastore which has metadata of mysql tables and using which it import/export data from/to mysql/HDFS.

**Benefits ->**

* Allows the transfer of data with a variety of structured data stores like Postgres, Oracle, Teradata, and so on.
* Since the data is transferred and stored in Hadoop, Sqoop allows us to offload certain processing done in the ETL (Extract, Load and Transform) process into low-cost, fast, and effective Hadoop processes.
* Sqoop can execute the data transfer in parallel, so execution can be quick and more cost effective.