**ASSIGNMENT-10.1**

1. The workflow of Oozie and its Benefits

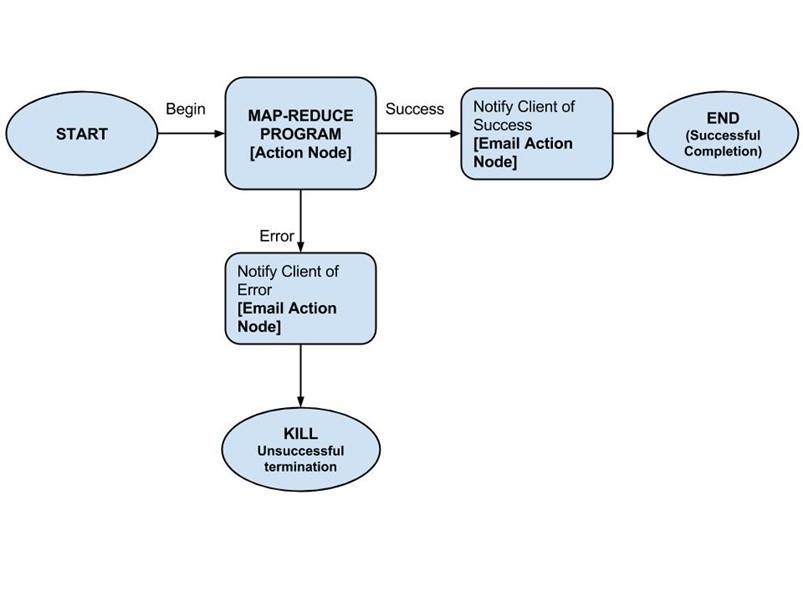
Apache Oozie is a workflow scheduler for Hadoop. It is a system which runs workflow of dependent jobs. Here, users are permitted to create Directed Acyclic Graphs of workflows, which can be run in parallel and sequentially in Hadoop.

It consists of two parts:

* **Workflow engine** : Responsibility of a workflow engine is to store and run workflows composed of Hadoop jobs e.g., MapReduce, Pig, Hive.
* **Coordinator engine**: It runs workflow jobs based on predefined schedules and availability of data.

Oozie is scalable and can manage timely execution of thousands of workflows (each consisting of dozens of jobs) in a Hadoop cluster.

**Workflow Diagram**



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

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

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.

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.

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

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

**End Node,** signals end of the job.

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

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.

**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**

Main purpose of using Oozie is to manage different type of jobs being processed in Hadoop system.

Dependencies between jobs are specified by user in the form of Directed Acyclic Graphs. Oozie consumes this information and takes care of their execution in correct order as specified in a workflow. That way user's time to manage complete workflow is saved. In addition, Oozie has a provision to specify frequency of execution of a particular job.

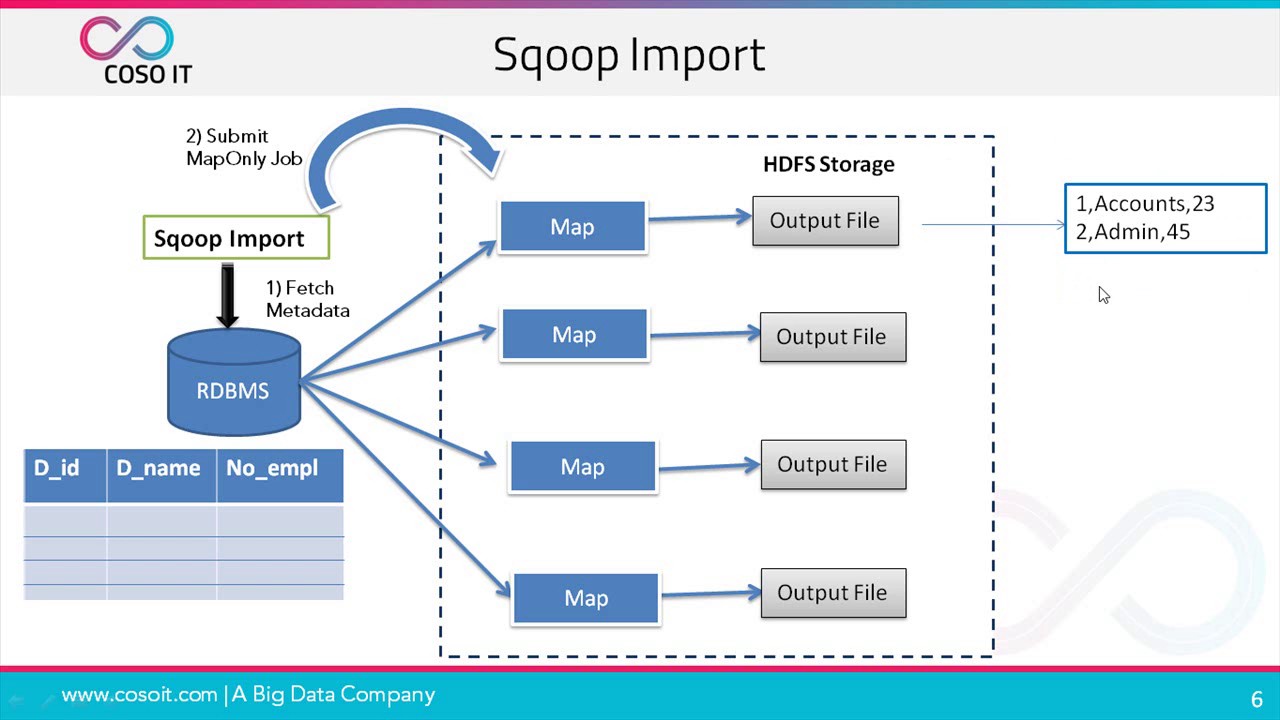
**FEATURES OF OOZIE**

* Oozie has client API and command line interface which can be used to launch, control and monitor job from Java application.
* Using its Web Service APIs one can control jobs from anywhere.
* Oozie has provision to execute jobs which are scheduled to run periodically.
* Oozie has provision to send email notifications upon completion of jobs.

1. **The workflow of Sqoop and its Benefits**

Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and external datastores such as relational databases, enterprise data warehouses.

Sqoop is used to import data from external datastores into Hadoop Distributed File System or related Hadoop eco-systems like Hive and HBase. Similarly, Sqoop can also be used to extract data from Hadoop or its eco-systems and export it to external datastores such as relational databases, enterprise data warehouses. Sqoop works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres etc.



**Sqoop Import**

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.

**Sqoop Export**

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter.

**Advantages of Sqoop**

Below are the advantages of Apache Sqoop, which is also the reason for choosing this technology in this layer.

* + 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.
  + Helps to integrate with sequential data from the mainframe. This helps not only to limit the usage of the mainframe, but also reduces the high cost in executing certain jobs using mainframe hardware.