**Assignment10-1**

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

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. (The shell and ssh actions have been deprecated).

**Workflow**



**Benefits of Oozie:**

* 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. 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.
* 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. Other systems would require significantly more work to correlate jobtracker jobs with the workflow actions**.**
* 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.
* 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.
* Oozie Coordinator allows triggering actions when files arrive at HDFS. This will be challenging to implement anywhere else.
* 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**.**

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

Sqoop is a bulk data transfer tool that allows easy import/export of data from structured datastores such as relational databases, enterprise data warehouses, and NoSQL systems. Using Sqoop, you can provision the data from an external system into HDFS, as well as populate tables in Hive and HBase. Similarly, Sqoop integrates with the workflow coordinator Apache Oozie (incubating), allowing you to schedule and automate import/export tasks. Sqoop uses a connector-based architecture which supports plugins that provide connectivity to additional external systems.

**Workflow:**



**Benefits of Sqoop:**

* 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. ...