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

Apache Oozie is a Java Web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work. 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. 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. At its core, Oozie helps administrators derive more value from Hadoop.

There are two basic types of Oozie jobs:

* Oozie Workflow jobs are Directed Acyclical Graphs (DAGs), specifying a sequence of actions to execute. The Workflow job has to wait
* Oozie Coordinator jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability.

Oozie Bundle provides a way to package multiple coordinator and workflow jobs and to manage the lifecycle of those jobs.

In many places Oozie is used for scheduling jobs .It has configurable xml file in which you can schedule jobs…

E.g Workflow.xml file:

1. <action name="identity-MR">
2. <map-reduce>
3. <job-tracker>localhost:8032</job-tracker>
4. <name-node>http://hdfs://localhost:8020</name-node>
5. <prepare>
6. <delete path="/user/joe/data/output"/>
7. </prepare>
8. <configuration>
9. <property>
10. <name>mapred.mapper.class</name> <value>org.apache.hadoop.mapred.lib.IdentityMapper</value>
11. </property>
12. <property>
13. <name>mapred.reducer.class</name> <value>org.apache.hadoop.mapred.lib.IdentityReducer</value>
14. </property>
15. <property>
16. <name>mapred.input.dir</name>
17. <value>/user/joe/data/input</value>
18. </property>
19. <property>
20. <name>mapred.output.dir</name>
21. <value>/user/joe/data/input</value>
22. </property>
23. </configuration>
24. </map-reduce>
25. <ok to="success"/>
26. <error to="fail"/>
27. </action>

Above code is used for scheduling Mapreduce job.

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. Oozie contains following three types of jobs

1. Workflow jobs – It used to represents the sequence of jobs executed.

2. Coordinator Jobs – It contains workflow jobs and it triggered by time

3. Bundle Jobs – It contains the workflow and coordinator jobs

[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

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

Sqoop component is used for importing data from external sources into related Hadoop components like HDFS, HBase or Hive. It can also be used for exporting data from Hadoop o other external structured data stores. Sqoop parallelized data transfer, mitigates excessive loads, allows data imports, efficient data analysis and copies data quickly.

Sqoop Use Case-

Online Marketer [Coupons.com](http://Coupons.com) uses Sqoop component of the [Hadoop](https://www.dezyre.com/Hadoop-Training-online/19) ecosystem to enable transmission of data between Hadoop and the IBM Netezza data warehouse and pipes backs the results into Hadoop using Sqoop.

Efficiently transfers bulk data between Apache Hadoop and structured datastores

Apache Sqoop efficiently transfers bulk data between Apache Hadoop and structured datastores such as relational databases. Sqoop helps offload certain tasks (such as ETL processing) from the EDW to Hadoop for efficient execution at a much lower cost. Sqoop can also be used to extract data from Hadoop and export it into external structured datastores. Sqoop works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB

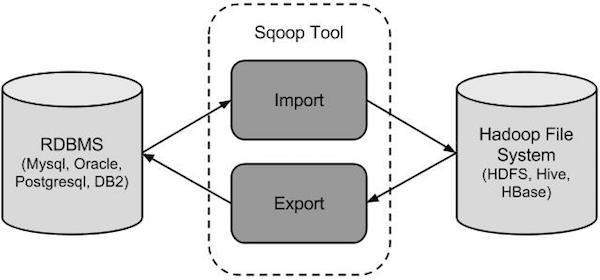
What Sqoop Does

Apache Sqoop does the following to integrate bulk data movement between Hadoop and structured datastores:

| Function | Benefit |
| --- | --- |
| Import sequential datasets from mainframe | Satisfies the growing need to move data from mainframe to HDFS​ |
| Import direct to ORCFiles | ​Improved compression and light-weight indexing for improved query performance |
| Data imports | Moves certain data from external stores and EDWs into Hadoop to optimize cost-effectiveness of combined data storage and processing |
| Parallel data transfer | For faster performance and optimal system utilization |
| Fast data copies | From external systems into Hadoop |
| Efficient data analysis | Improves efficiency of data analysis by combining structured data with unstructured data in a schema-on-read data lake |
| Load balancing | Mitigates excessive storage and processing loads to other systems |

YARN coordinates data ingest from Apache Sqoop and other services that deliver data into the Enterprise Hadoop cluster.

The following image describes the workflow of Sqoop.



## 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.