The Hadoop Ecosystem

Oozie

A sequence of steps is called a Workflow. For example, to build a recommender system for sales

1. Copy logs from production system
2. Dump the database
3. Apply Big Data algorithms for machine learning over the dump and logs
4. Pre-compute recommendations
5. Store in production database

The Oozie architecture allows users to define workflows and run them periodically, with error handling.

1. User submits workflow as XML
2. Oozie Tomcat parses this workflow and synchronizes this with the database
3. Allows workflow usage to begin with supported applications such as Hadoop, Pig, etc

The oozie workflow usually looks like a DAG with a set of action nodes and control nodes.

Action nodes do something, for example, run MR, and have a normal exit and an error exit.

Control nodes are usually START, END and KILL. Usually are decisions to be taken based on action nodes.

Ambari

The architecture of Ambari is used to set up cluster environments autonomously given the specification. The architecture comprises of

1. Stacks
    a. These describe the applications to be installed
        i. Stack : Set of services, where to get the software
        ii. Service : Components that make up the service, such as HDFS
        iii. Component : Building blocks of the service - namenode, datanode
        iv. Category : Master, slave, client
2. Blueprints
    a. Creation of the cluster
3. Views
    a. User Interface

It also allows basic monitoring with visualisations on its UI.

Pig

Some of the disadvantages of MR are

1. Very low level for data analysis
2. Need a SQL like scripting language

Pig allows dataflows to be executed as a set of MR jobs on a cluster. It uses a DAG structure to construct these dataflows, and each component of the dataflow is provisioned as a MR job.

SQOOP

This is mainly used to use data periodically for analysis and write results back to the SQL database. SQL on Hadoop or SQOOP allows this.

It is a bulk transfer tool which

1. Allows import/export of data from databases
2. Integrates with Oozie as an action
3. Allows support plugins via a connector base architecture

Import in SQOOP works as

1. Gather metadata about the data to be imported
2. Transfer using map only jobs to HDFS with newline separator

Flume

Flume is a tool/service/data ingestion mechanism for collecting, aggregating and transporting large amounts of streaming data such as log files, events (etc..) from various sources to a centralised source.

It is a highly reliable, distributed and configurable tool. It is designed to mainly copy streaming data from various servers to HDFS.

Advantages of flume

1. Can store data into any centralised store
2. Flume provides flow control for streaming data
3. Contextual routing is provided
4. Channel based transactions for reliable delivery
5. Fault tolerant, scalable, manageable and customizable.