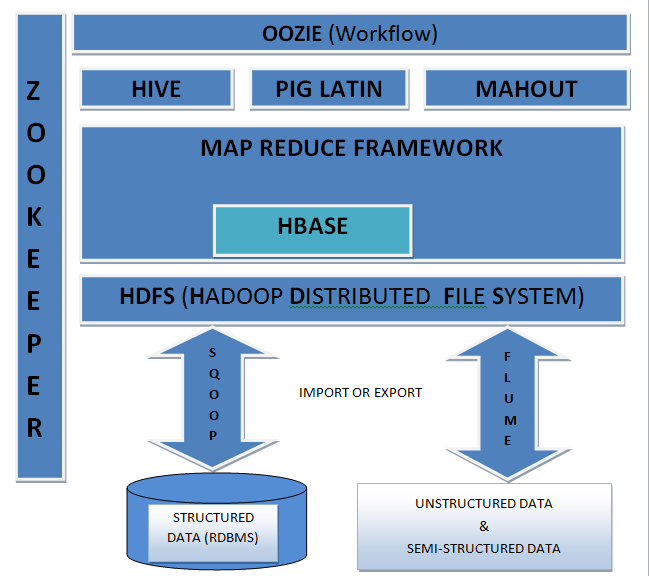
**UNDERSTANDING HADOOP ECOSYSTEM**

by [Deepesh Kumbhare](mailto:deepeshkumbhare.rcet@gmail.com)

<http://hadooptutorials.co.in/tutorials/hadoop/understanding-hadoop-ecosystem.html>

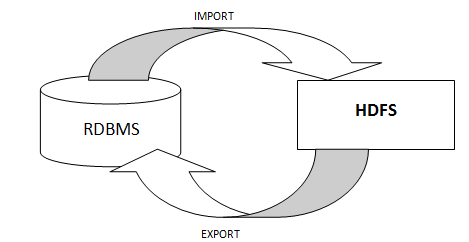
[](http://hadooptutorials.co.in/tutorials/hadoop/images/understanding-hadoop-ecosystem/Hadoop_Ecosystem.png)

We all know Hadoop is a framework which deals with Big Data but unlike any other frame work it's not a simple framework, it has its own family for processing different thing which is tied up in one umbrella called as Hadoop Ecosystem. Before jumping directly to members of ecosystem let's have a understanding of classification of data. Data is mainly categorized in 3 types under Big Data platform.

* Structured Data - Data which has proper structure and which can be easily stored in tabular form in any relational databases like Mysql, Oracle etc is known as structured data.Example- Employee data .
* Semi-Structured Data - Data which has some structure but cannot be saved in a tabular form in relational databases is known as semi structured data. Example-XML data, email messages etc.
* Unstructured Data - Data which is not having any structure and cannot be saved in tabular form of relational databases is known as unstructured data. Example- Video files, Audio files, Text file etc.

Let's try to understand each component in detail

**SQOOP : SQL + HADOOP = SQOOP**

[](http://hadooptutorials.co.in/tutorials/hadoop/images/understanding-hadoop-ecosystem/sqoop-import-export.PNG)

When we import any structured data from table (RDBMS) to HDFS a file is created in HDFS which we can process by either Map Reduce program directly or by HIVE or PIG. Similarly after processing data in HDFS we can store the processed structured data back to another table in RDBMS by exporting through Sqoop.

**HDFS (Hadoop Distributed File System)**

HDFS is a main component of Hadoop and a technique to store the data in distributed manner in order to compute fast. HDFS saves data in a block of 64MB(default) or 128 MB in size which is logical splitting of data in a Datanode (physical storage of data) in Hadoop cluster(formation of several Datanode which is a collection commodity hardware connected through single network). All information about data splits in data node known as metadata is captured in Namenode which is again a part of HDFS.

**MapReduce Framework**

It is another main component of Hadoop and a method of programming in a distributed data stored in a HDFS. We can write Map reduce program by using any language like JAVA, C++ PIPEs, PYTHON, RUBY etc. By name only Map Reduce gives its functionality Map will do mapping of logic into data (distributed in HDFS) and once computation is over reducer will collect the result of Map to generate final output result of MapReduce. MapReduce Program can be applied to any type of data whether Structured or Unstructured stored in HDFS. Example - word count using MapReduce

Thank you for reading. Happy Learning !!

**HBASE**

Hadoop Database or HBASE is a non-relational (NoSQL) database that runs on top of HDFS. HBASE was created for large table which have billions of rows and millions of columns with fault tolerance capability and horizontal scalability and based on Google Big Table. Hadoop can perform only batch processing, and data will be accessed only in a sequential manner for random access of huge data HBASE is used.

**Hive**

Many programmers and analyst are more comfortable with Structured Query Language than Java or any other programming language for which Hive is created by Facebook and later donated to Apache foundation. Hive mainly deals with structured data which is stored in HDFS with a Query Language similar to SQL and known as HQL (Hive Query Language). Hive also run Map reduce program in a backend to process data in HDFS but here programmer has not worry about that backend MapReduce job it will look similar to SQL and result will be displayed on console.

**Pig**

Similar to HIVE, PIG also deals with structured data using PIG LATIN language. PIG was originally developed at Yahoo to answer similar need to HIVE. It is an alternative provided to programmer who loves scripting and don't want to use Java/Python or SQL to process data. A Pig Latin program is made up of a series of operations, or transformations, that are applied to the input data which runs MapReduce program in backend to produce output.

**Mahout**

Mahout is an open source machine learning library from Apache written in java. The algorithms it implements fall under the broad umbrella of machine learning or collective intelligence. This can mean many things, but at the moment for Mahout it means primarily recommender engines (collaborative filtering), clustering, and classification. Mahout aims to be the machine learning tool of choice when the collection of data to be processed is very large, perhaps far too large for a single machine. In its current incarnation, these scalable machine learning implementations in Mahout are written in Java, and some portions are built upon Apache's Hadoop distributed computation project.

**Oozie**

It is a workflow scheduler system to manage hadoop jobs. It is a server-based Workflow Engine specialized in running workflow jobs with actions that run Hadoop MapReduce and Pig jobs. Oozie is implemented as a Java Web-Application that runs in a Java Servlet-Container. Hadoop basically deals with bigdata and when some programmer wants to run many job in a sequential manner like output of job A will be input to Job B and similarly output of job B is input to job C and final output will be output of job C. To automate this sequence we need a workflow and to execute same we need engine for which OOZIE is used.

**Zookeeper**

Writing distributed applications is difficult because of partial failure may occur between nodes to overcome this Apache Zookeper has been developed by maintaining an open-source server which enables highly reliable distributed coordination. ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services . In case of any partial failure clients can connect to any node and be assured that they will receive the correct, up-to-date information.