

Hadoop Eco System and Related Components

HDFS and MapReduce

Hadoop Common, HDFS, YARN and MapReduce are kernel components of Hadoop

HDFS provides storage. MapReduce provides processing.

- CLI, JAVA Library API, GUI
- Hadoop Streaming for other programming platform binaries to interact
- Integration support to various file system sources i.e. Local FS, WebHDFS, S3 FS, and others

Various Hadoop Distributors

Apache Hadoop, Cloudera, Hortonworks, MAPR

Approaches to do computing with Hadoop

- HDFS for storage baseline
- Native Map Reduce for java programmers and others
- Flume, Sqoop like data ingestors
- PIG like tools for scripting developers
- HIVE like tools for query language experts
- HBase like Distributed Databases
- Zookeeper like service coordinator
- Oozie like workflow scheduler
- Mahout like Machine Learning
- Ambari like monitoring and provisioning tools

SQL-NoSQL-NewSQL

Data Stores of various types and functionalities support handling of Big Data

For example, mongoDB provides NoSQL document store and also has inbuilt support to write map and reduce function with support of java script.

P.s. No that some tools let you write SQL/Script and behind the scene automatically converts SQL/Script to run as Map-Reduce program logic to achieve the task.

In Memory computation

Apache Spark like engine for large scale data processing and in-memory computation

Spark Streaming

For real time stream processing applications

Spark GraphX

For Graphs and graph-parallel computation

Complete Reference

- <https://hadoopecosystemtable.github.io>