

# Aim : To Study of Hadoop Ecosystem in details.

The Hadoop ecosystem is an open-source framework comprising small components that help interact with big data easily. It is created and maintained by the Apache Software Foundation. The main components of the Hadoop ecosystem are:

1. **HDFS (Hadoop Distributed File System):** The primary component of the Hadoop ecosystem, responsible for storing large data sets of structured or unstructured data.
2. **YARN (Yet Another Resource Negotiator):** A core component of Hadoop that manages workloads, monitors, and implements security controls. It allocates system resources to the various applications running in a Hadoop cluster.
3. **MapReduce:** A programming model and software framework for processing large data sets in parallel across a distributed cluster.
4. **Spark:** An in-memory data processing engine suitable for use in a wide range of circumstances, featuring Java, Python, Scala, and R programming languages, and supporting SQL, streaming data, machine learning, and graph processing.
5. **Pig:** A data processing layer of Hadoop that uses a simple SQL-like scripting language called Pig Latin for performing operations and arranging the final output in the desired format.

6.

1. **Hive:** An ETL and Data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem. It has a SQL-like interface, HQL language, and features a SQL-like interface, HQL language that works similar to SQL and automatically translates queries into MapReduce jobs.
2. **HBase:** A NoSQL database that is designed after Google Big table, a distributed storage system built to handle large data sets.
3. **Mahout:** A machine learning library for Hadoop that allows invoking algorithms as per our need with the help of its own libraries.
4. **Sqoop:** A tool for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.
5. **Zookeeper:** A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.
6. **Oozie:** A workflow scheduler system to manage Hadoop jobs, which is integrated with the Hadoop stack to allow for job scheduling and monitoring.

These components collectively provide services and solutions for Big Data, including ingestion, storage, analysis, and maintenance.

The Hadoop ecosystem is a powerful platform for handling big data that provides various services such as storage, processing, analysis, and maintenance of data. It includes several components that work together to provide a comprehensive solution for big data processing.

Apart from these major elements, there are several other components in the Hadoop ecosystem that provide various services such as absorption, analysis, storage, and maintenance of data. These components include Spark, PIG, HIVE, HBase, Mahout, Spark MLLib, Solar, Lucene, Zookeeper, and Oozie. Spark is an in-memory data processing engine, PIG and HIVE are query-based processing of data services, HBase is a NoSQL database, Mahout and Spark MLLib are machine learning algorithm libraries, Solar and Lucene are searching and indexing tools, Zookeeper is used for managing clusters, and Oozie is used for job scheduling.

# Hadoop Ecosystem

