**Main Components**

* **Hadoop Common-** A set of common libraries and utilities used by other Hadoop modules.
* **HDFS-** The default storage layer for Hadoop.
* **MapReduce-** Executes a wide range of analytic functions by analyzing datasets in parallel before ‘reducing’ the results. The “Map” job distributes a query to different nodes, and the “Reduce” gathers the results and resolves them into a single value.
* **YARN-** Present in version 2.0 onwards, YARN is the cluster management layer of Hadoop. Prior to 2.0, MapReduce was responsible for cluster management as well as processing. The inclusion of YARN means you can run multiple applications in Hadoop (so you’re no longer limited to MapReduce), which all share common cluster management.

**Some the more well-known components include:**

* **Spark-** Used on top of HDFS, Spark promises speeds up to 100 times faster than the two-step MapReduce function in certain applications. Allows data to loaded in-memory and queried repeatedly, making it particularly apt for machine learning algorithms
* **Hive-** Originally developed by Facebook, Hive is a data warehouse infrastructure built on top of Hadoop. Hive provides a simple, SQL-like language called HiveQL, whilst maintaining full support for MapReduce. This means SQL programmers with little former experience with Hadoop can use the system easier, and provides better integration with certain analytics packages like Tableau. Hive also provides indexes, making querying faster.
* **HBase-** Is a NoSQL columnar database which is designed to run on top of HDFS. It is modelled after Google’s BigTable and written in Java. It was designed to provide BigTable-like capabilities to Hadoop, such as the columnar data storage model and storage for sparse data.
* **Flume-** Flume collects (typically log) data from ‘agents’ which it then aggregates and moves into Hadoop. In essence, Flume is what takes the data from the source (say a server or mobile device) and delivers it to Hadoop.
* **Mahout-** Mahout is a machine learning library. It collects key algorithms for clustering, classification and collaborative filtering and implements them on top of distributed data systems, like MapReduce. Mahout primarily set out to collect algorithms for implementation on the MapReduce model, but has begun implementing on other systems which were more efficient for data mining, such as Spark.
* **Sqoop-** Sqoop is a tool which aids in transitioning data from other database systems (such as relational databases) into Hadoop.