Hadoop Federation, Architecture and Overview

Technologies for Big Data Analytics based on Distributed Computing:

1. Hadoop – which provides a distributed file system.
2. YARN – a resource manager through which multiple applications can perform computations simultaneously on data.
3. Spark – an open-source framework for the analysis of data that can be run on Hadoop, its architecture and its mode of operation in comparison to MapReduce.

Hadoop is an ecosystem of open-source components that fundamentally changes the way enterprises store, process, and analyze data. Unlike traditional systems, Hadoop enables multiple types of analytic workloads to run on the same data, at the same time, at massive scale on industry-standard hardware. CDH, Cloudera's open-source platform, is the most popular distribution of Hadoop and related projects in the world.

Hadoop's infinitely scalable flexible architecture (based on the HDFS filesystem) allows organizations to store and analyze unlimited amounts and types of data—all in a single, open-source platform on industry-standard hardware.

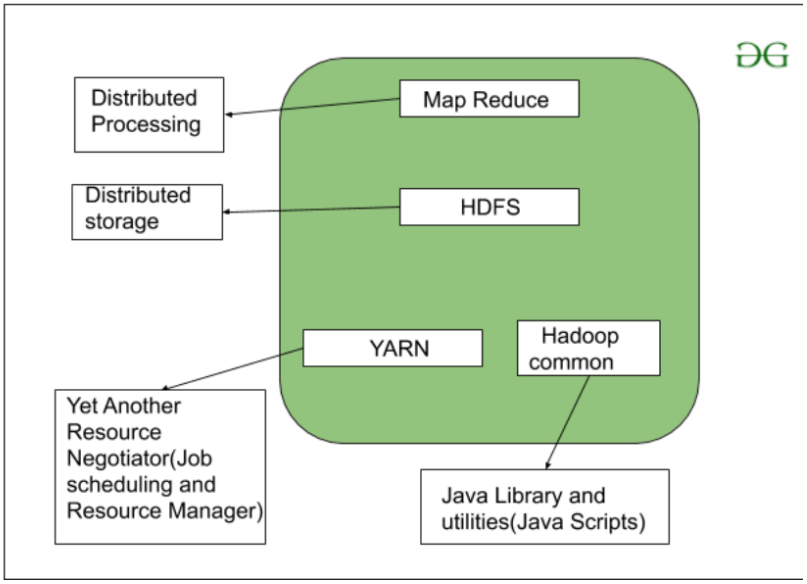
Quickly integrate with existing systems or applications to move data into and out of Hadoop through bulk load processing (Apache Sqoop) or streaming (Apache Flume, Apache Kafka).

Transform complex data, at scale, using multiple data access options (Apache Hive, Apache Pig) for batch (MR2) or fast in-memory (Apache Spark™) processing. Process streaming data as it arrives in your cluster via Spark Streaming.

Analysts interact with full-fidelity data on the fly with Apache Impala, the data warehouse for Hadoop. With Impala, analysts experience BI-quality SQL performance and functionality plus compatibility with all the leading BI tools.

Using Cloudera Search, an integration of Hadoop and Apache Solr, analysts can accelerate the process of discovering patterns in data in all amounts and formats, especially when combined with Impala.

With Hadoop, analysts and data scientists have the flexibility to develop and iterate on advanced statistical models using a mix of partner technologies as well as open-source frameworks like Apache Spark™.



The distributed data store for Hadoop, Apache HBase, supports the fast, random reads/writes (“fast data”) required for online applications.

CDH, the world's most popular Hadoop distribution, is Cloudera’s 100% open-source platform. It includes all the leading Hadoop ecosystem components to store, process, discover, model, and serve unlimited data, and it's engineered to meet the highest enterprise standards for stability and reliability.

CDH is based entirely on open standards for long-term architecture. And as the main curator of open standards in Hadoop, Cloudera has a track record of bringing new open-source solutions into its platform (such as Apache Spark™, Apache HBase, and Apache Parquet) that are eventually adopted by the entire ecosystem.