HADOOP

Hadoop is an open-source framework designed for distributed storage and processing of large datasets. It provides a cost-effective and scalable solution for handling big data across clusters of computers. The core components of Hadoop are the Hadoop Distributed File System (HDFS), MapReduce, Yarn and the Hadoop utilities.

HDFS is a distributed file system that allows data to be stored across multiple machines in a cluster. It provides high fault tolerance and reliability by replicating data across different nodes. Hadoop breaks large datasets into blocks and distributes them across the cluster, enabling parallel processing and efficient data access.

MapReduce is a programming model used for processing and analyzing large datasets in parallel. It divides the processing task into two phases: the Map phase and the Reduce phase. In the Map phase, data is processed and transformed into intermediate key-value pairs. In the Reduce phase, the intermediate results are aggregated and combined to produce the final output.

Yarn is a resource manager of the Hadoop which is responsible for managing resource for the job and it also does job scheduling and keeps tracking and monitoring the job.

Hadoop utilities basically contain all the contents which are necessary to run and support Hadoop cluster. It contains Java Libraries and other important files which are required to run the Hadoop.

Hadoop also offers a wide range of ecosystem tools and frameworks that extend its capabilities. These include Apache Hive, which provides a SQL-like query language for data analysis, Apache Pig for scripting data transformations, Apache HBase for real-time read/write access to large datasets, and Apache Spark for high-speed data processing.

The key advantages of Hadoop are its ability to handle massive amounts of data, fault tolerance, and scalability. It allows organizations to process and analyze large datasets that were previously challenging to handle with traditional systems. Hadoop is widely used in various industries, including finance, healthcare, e-commerce, and social media, to gain insights from big data and make informed business decisions.

In summary, Hadoop is a powerful framework that enables distributed storage and processing of big data. Its components, HDFS and MapReduce, provide scalability, fault tolerance, and parallel processing capabilities. With its ecosystem of tools, Hadoop has become a popular choice for organizations dealing with large-scale data analysis and processing.