Hadoop là một hệ sinh thái phần mềm mã nguồn mở được sử dụng để xử lý và lưu trữ dữ liệu lớn. Nó bao
gồm các thành phần như HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, và
nhiều công cụ khác. Hadoop được xây dựng trên ngôn ngữ Java và tương thích trên nhiều nền tảng khác
nhau.

Kiến trúc của Hadoop bao gồm các thành phần sau: 1. Hadoop Distributed File System (HDFS): Hệ thống
file phân tán của Hadoop, nằm ở trên cùng của hệ thống file cục bộ và giám sát quá trình. 2.
MapReduce: Mô hình xử lý dữ liệu phân tán trong Hadoop, sử dụng các công cụ để phân tách và xử lý dữ
liệu trên các nút trong cụm. 3. YARN (Yet Another Resource Negotiator): Quản lý các thành viên trong
nhóm máy chủ và phân phối tài nguyên cho các ứng dụng chạy trên Hadoop. 4. ZooKeeper: Hệ thống quản
lý và bầu cử leader trong Hadoop, đảm bảo tính nhất quán và sẵn sàng của dữ liệu. 5. Hadoop Common:
Bao gồm các thư viện và công cụ chung được sử dụng bởi các thành phần khác trong Hadoop. 6. Hadoop
Ozone: Hệ thống lưu trữ đối tượng phân tán trong Hadoop, cung cấp khả năng lưu trữ và truy xuất dữ
liệu theo kiểu đối tượng. 7. Hadoop HDFS Federation: Cung cấp khả năng mở rộng và phân chia dữ liệu
trên nhiều cụm HDFS. 8. Hadoop HDFS High Availability: Cung cấp khả năng sao lưu và khôi phục tự
động cho HDFS, đảm bảo tính sẵn sàng và tin cậy của dữ liệu. 9. Hadoop HDFS Erasure Coding: Sử dụng
mã hóa xóa để giảm bớt lưu trữ dự phòng và tăng hiệu suất lưu trữ trong HDFS. 10. Hadoop HDFS
Snapshots: Cung cấp khả năng tạo và quản lý các bản snapshot của dữ liệu trong HDFS, cho phép khôi
phục dữ liệu về trạng thái trước đó.