YARN Architecture and Overview

HDFS is a fault-tolerant and self-healing distributed filesystem designed to turn a cluster of industry-standard servers into a massively scalable pool of storage. Developed specifically for large-scale data processing workloads where scalability, flexibility, and throughput are critical, HDFS accepts data in any format regardless of schema, optimizes for high-bandwidth streaming, and scales to proven deployments of 100PB and beyond.

Key Features: Hadoop Scalable, Flexible and Reliable

Map Reduce: One Job, multiple instances, aggregated output. Julius Caesar example for census.

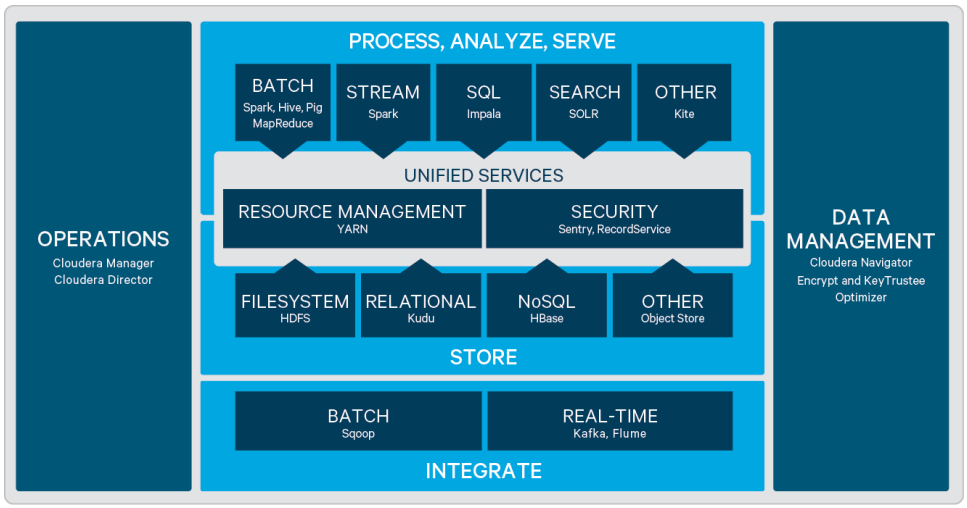
Key Features: Accessibility, Flexibility, Scalable, Reliable

While MapReduce continues to be a popular batch-processing tool, Apache Spark's flexibility and in-memory performance make it a much more powerful batch execution engine. Cloudera has been working with the community to bring the frameworks currently running on MapReduce onto Spark for faster, more robust processing.

MapReduce is designed to process unlimited amounts of data of any type that's stored in HDFS by dividing workloads into multiple tasks across servers that are run in parallel.

YARN (Yet Another Resource Negotiator) provides open-source resource management for Hadoop, so you can move beyond batch processing and open up your data to a diverse set of workloads, including interactive SQL, advanced modeling, and real-time streaming.

Key Features: Hadoop Scalable, Dynamic Mutli Tenancy, Optimal Workload Management



Core Hadoop, including HDFS, MapReduce, and YARN, is part of the foundation of Cloudera's platform. All platform components have access to the same data stored in HDFS and participate in shared resource management via YARN. Hadoop, as part of Cloudera's platform, also benefits from simple deployment and administration (through Cloudera Manager) and shared compliance-ready security and governance (through Apache Sentry and Cloudera Navigator) — all critical for running in production.