1. Hadoop HDFS: Distributed Storage, DataNode
2. Hadoop YARN: Resource Scheduling and Negation (cluster resource management)

YARN enhances the power of a Hadoop computer cluster (Scalability) also improved cluster utilization and MapReduce Compatibility, supports other worldoads

1. Hadoop MapReduce: Distributed Processing, JobTracker-TaskTracker (redundant, reliable storage)

Apache Sqoop: Tool designed for efficiently transferring bulk data between Apache Hadoop and Structured datastores such as relational databases

HBASE:

1. Column-oriented database management system
2. Key-value store
3. Based on Google Big Table
4. Can hold extremely large data
5. Dynamic data model
6. Not a Relational DBMS

Pig:

1. High level programming on top of Hadoop MapReduce
2. Pig Latin
3. Data analysis problems as data flows
4. Originally developed at Yahoo 2006

Apache Hive: Data warehouse software facilitates querying and managing large datasets residing in distributed storage

Oozie:

1. Workflow scheduler system to manage Apache Hadoop jobs
2. Oozie Coordinator jobs
3. Supports: MapReduce, Pig, Apache Hive, and Sqoop, etc.

Zookeeper

1. Provides operational services for a Hadoop cluster
2. Centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing

Flume: Distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

Cloudera Hadoop Components:

Batch Processing: MapReduce, Hive, Pig

Analytic SQL: Impala (Cloudera Impala is Cloudera’s open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop), BI & Analytics applications, On-premise applications, SaaS/Cloud applications

Search Engine: Cloudera Search

Machine Learning: Spark, MapReuce, Mahout

Stream Processing: Spark

Spark: Apache SparkTM is a fast and general engine for large-scale data processing

Spark Benefits:

1. Multi-stage in-memory primitives provides performance up to 100 times faster for certain applications
2. Allows user programs to load data into a cluster’s memory and query it repeatedly:

Well-suited to machine learning