Node MongoDB Minio

A Node.js-MongoDB-Minio image upload app.

Upload an image

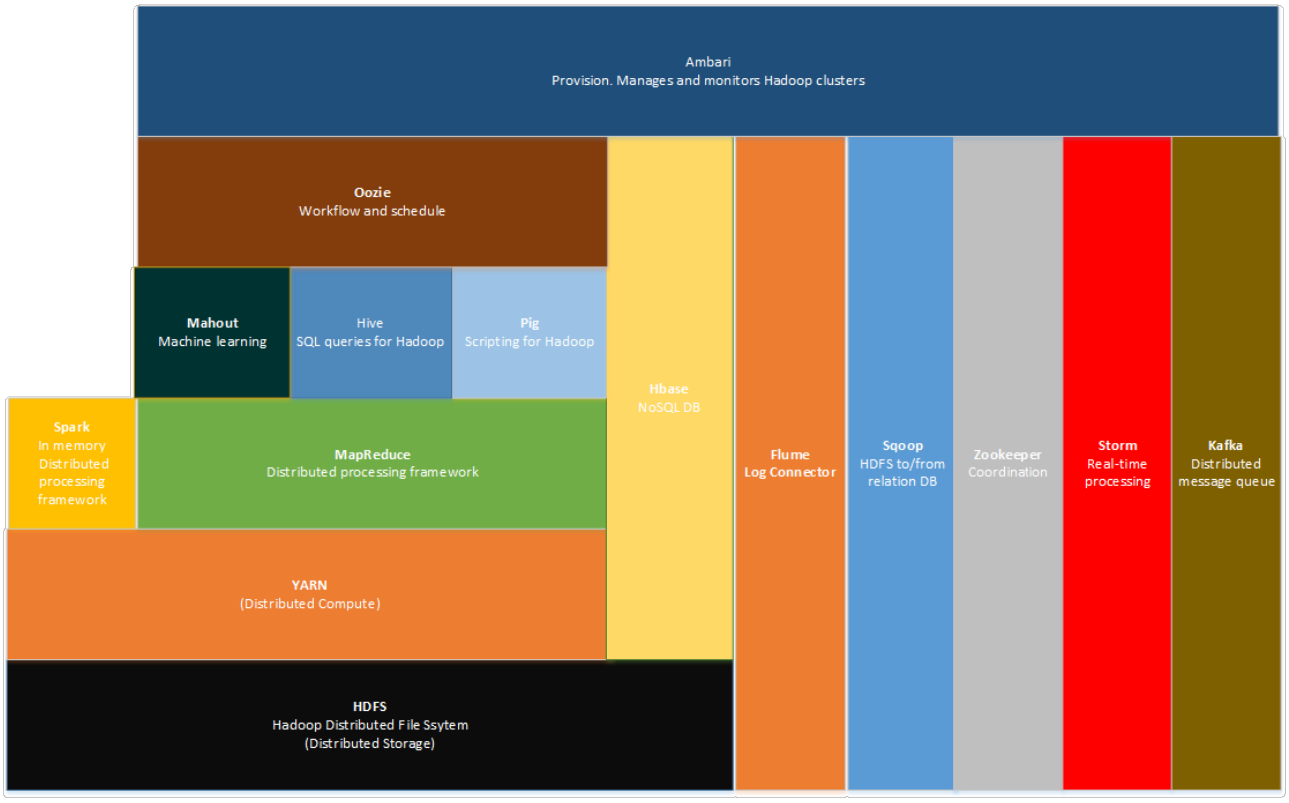
Choose File no file selected Upload

Write your content here

Publish

Notes

Apache Hadoop Stack



Apache Hadoop Ecosystem

A. Storage

1. Apache Hadoop Distributed File System (HDFS) is the primary component of the Hadoop ecosystem, it is a distributed file system in which individual Hadoop nodes operate on data that resides in their local storage. This removes network latency, providing high-throughput access to application data. In addition, administrators don't need to define schemas up front.

2. Alluxio is an open source data orchestration layer that brings data close to compute for big data and AI/ML workloads in the cloud.

Alluxio serves as a new data access layer in the big data and machine learning ecosystem. It resides between any persistent storage systems such as Amazon S3, Microsoft Azure Object Store, Apache HDFS, or OpenStack Swift, and computation frameworks such as Apache Spark, Presto, or Hadoop MapReduce.

B. Cluster Management

3. Apache Hadoop MapReduce is a programming model for large-scale data processing. In the MapReduce model, subsets of larger datasets and instructions for processing the subsets are dispatched to multiple different nodes, where each subset is processed by a node in parallel with other processing jobs. After processing the results, individual subsets are combined into a smaller, more manageable dataset.

C. Resource Management

4. Apache Hadoop Yet Another Resource Negotiator (YARN) is a resource-management platform responsible for managing compute resources in clusters and using them to schedule users' applications. It performs scheduling and resource allocation across the Hadoop system.

The fundamental idea of YARN is to split up the functionalities of resource management and job scheduling/monitoring into separate daemons. The idea is to have a global ResourceManager (RM) and per-application ApplicationMaster (AM). An application is either a single job or a DAG of jobs.

The ResourceManager and the NodeManager form the data-computation framework. The ResourceManager is the ultimate authority that arbitrates resources among all the applications in the system. The NodeManager is the per-machine framework agent who is responsible for containers, monitoring their resource usage (cpu, memory, disk, network) and reporting the same to the ResourceManager/Scheduler.

The per-application ApplicationMaster is, in effect, a framework specific library and is tasked with negotiating resources from the ResourceManager and working with the NodeManager(s) to execute and monitor the tasks.

5. Apache Mesos abstracts CPU, memory, storage, and other compute resources away from machines (physical or virtual), enabling fault-tolerant and elastic distributed systems to easily be built and run effectively.

Mesos is built using the same principles as the Linux kernel, only at a different level of abstraction. The Mesos kernel runs on every machine and provides applications (e.g., Hadoop, Spark, Kafka, Elasticsearch) with API's for resource management and scheduling across entire datacenter and cloud environments.

Mesos isn't really a part of Hadoop, but it's included in the Hadoop ecosystem as it is an alternative to YARN. It is also a resource negotiator just like YARN. Mesos and YARN solve the same problem in different ways. The main difference between Mesos and YARN is in their scheduler.

D. Batch and Interactive Data Processing

6. Apache Tez project is aimed at building an application framework, which allows for a complex directed-acyclic-graph of tasks for processing data. It is currently built atop Apache Hadoop YARN.

Apache Hadoop Ecosystem O'Reilly