

Hadoop是什么

Apache Hadoop is **an open source** software framework for storage and **large scale** processing of data-sets on clusters of **commodity hardware**

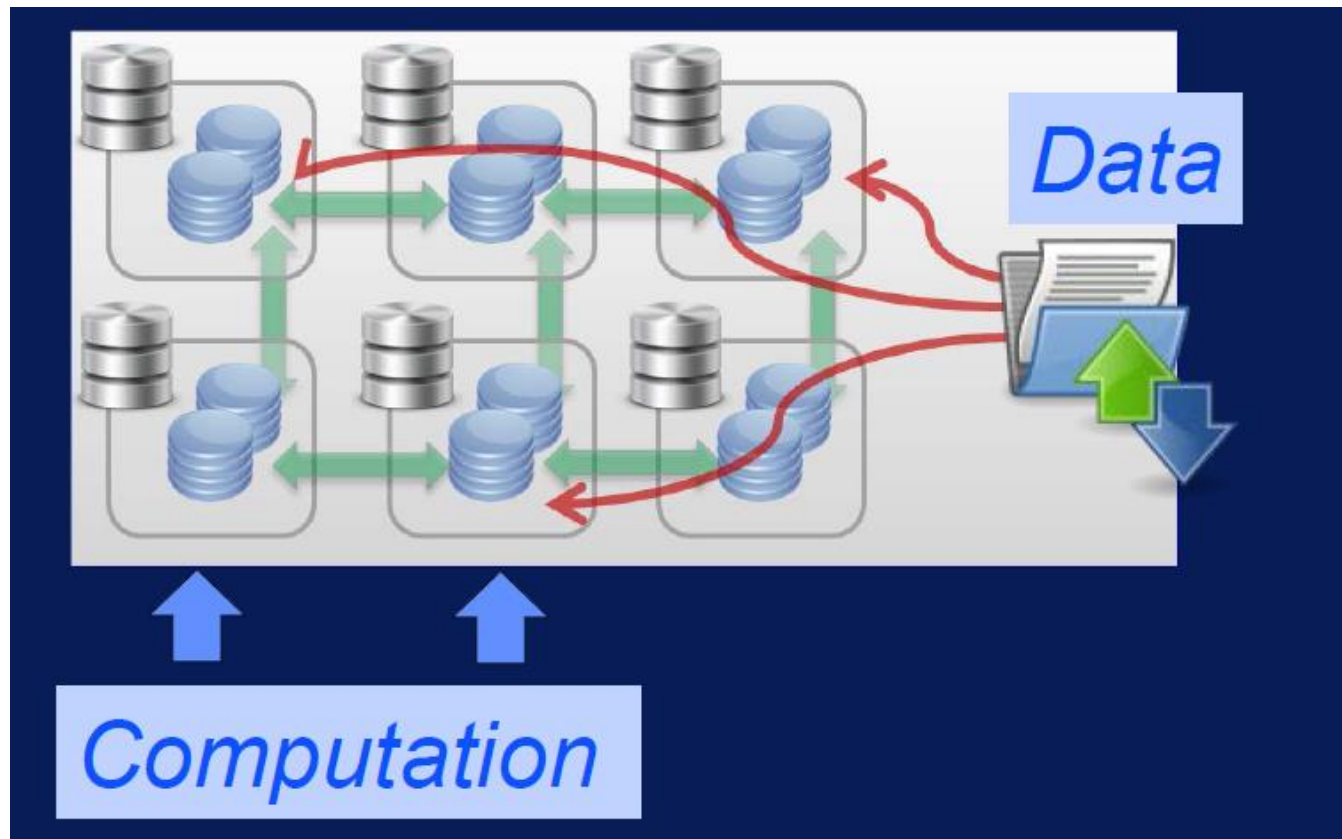
Hadoop名字的由来

Hadoop was created by Doug Cutting and Mike Cafarella in 2005

Named the project after son's toy elephant



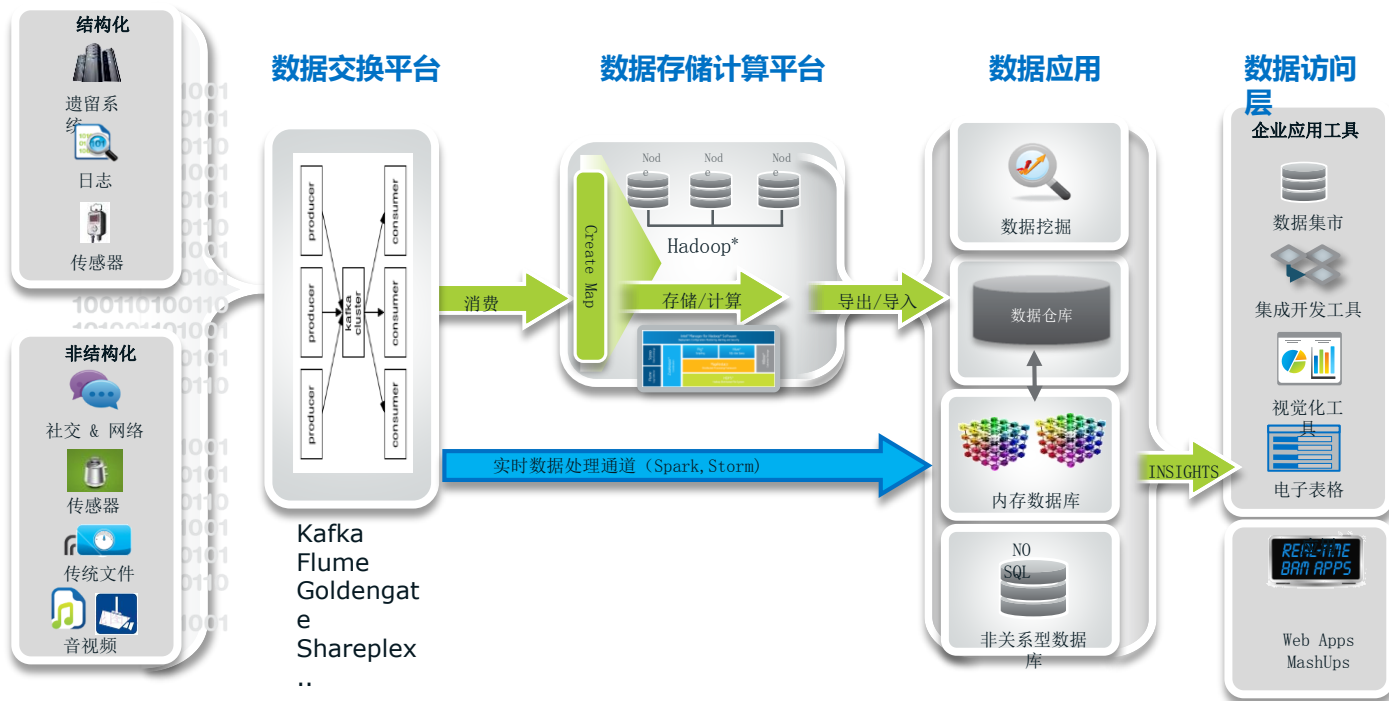
从移动数据到移动算法



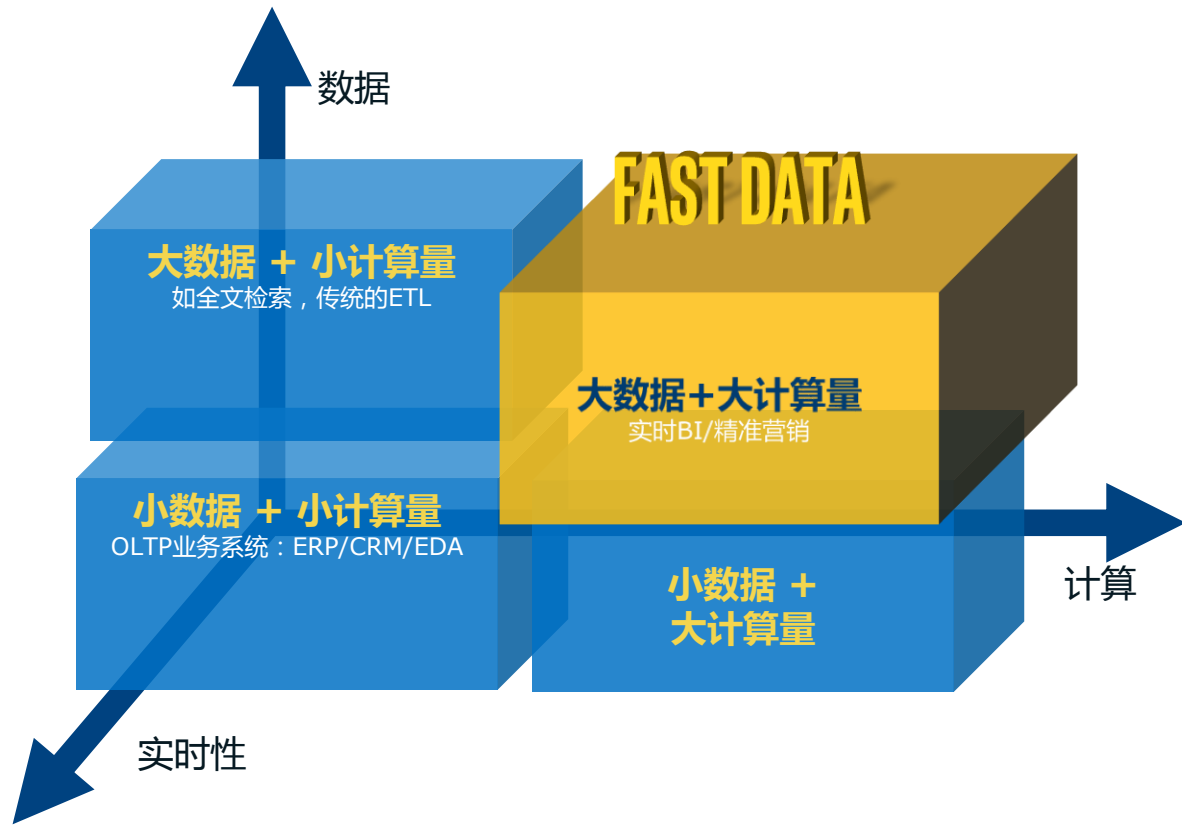
Hadoop的核心理念

- 可扩展性
- 可靠性

相对于传统的BI架构转变



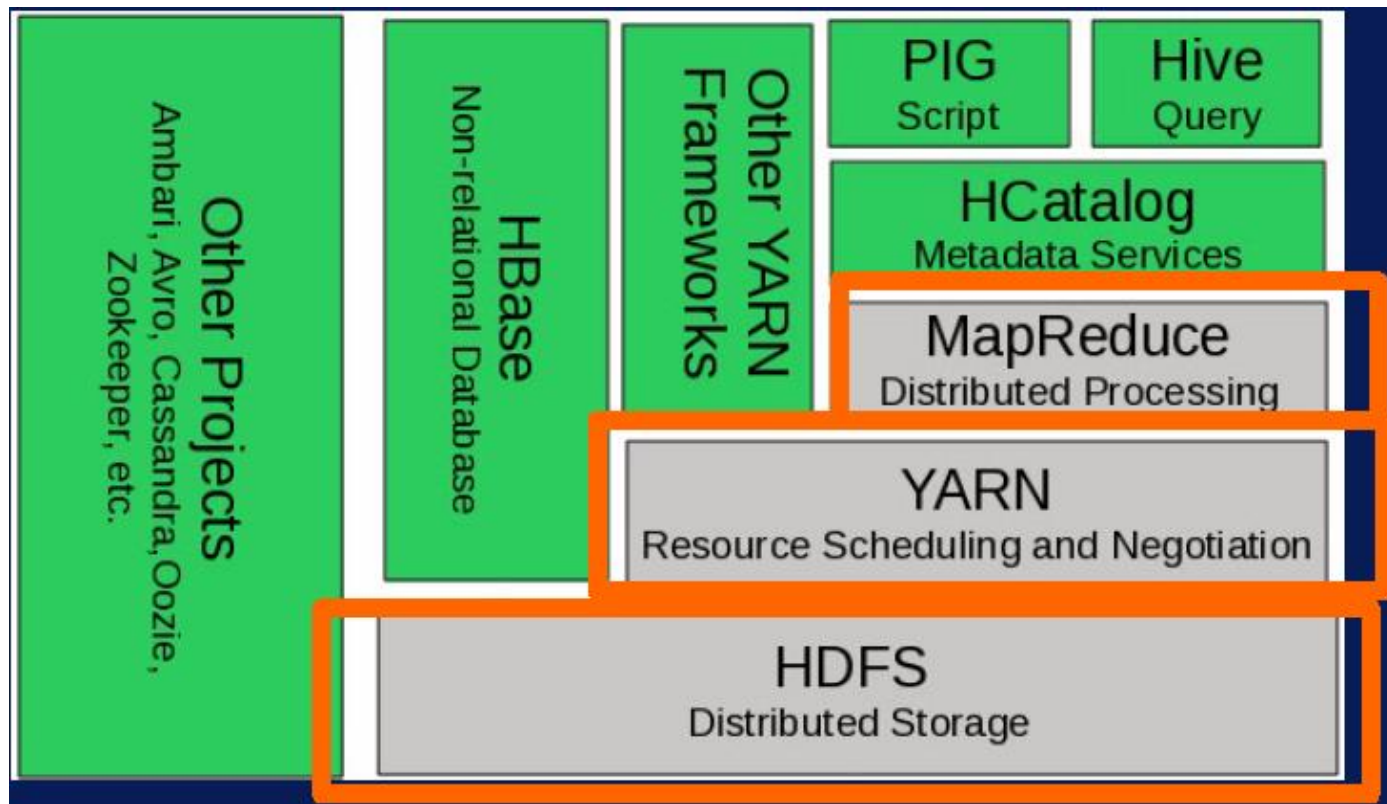
hadoop的适用场景

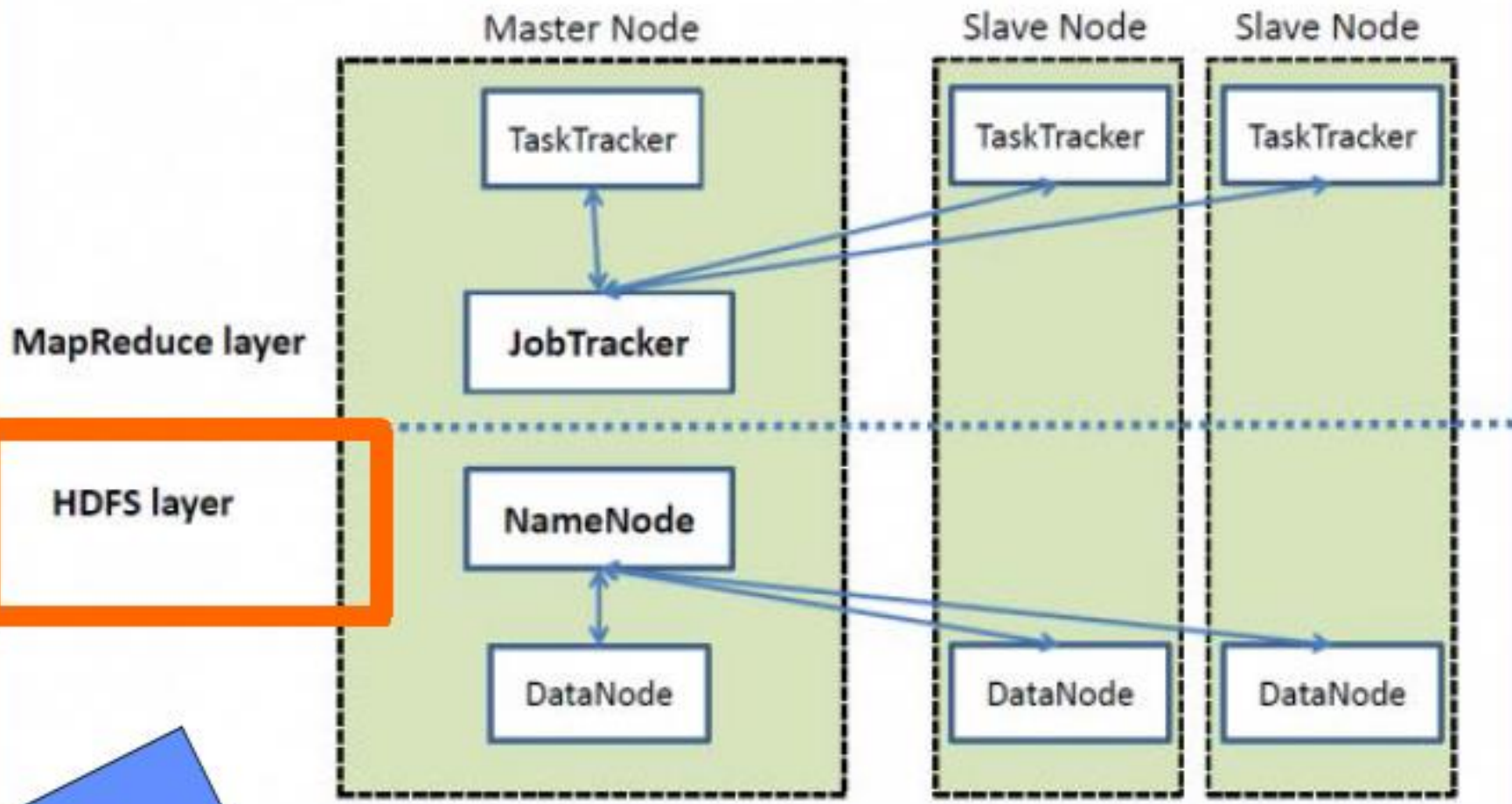


Hadoop基础组件

- **Hadoop Common**
- **Hadoop Distributed File System (HDFS)**
- **Hadoop YARN**
- **Hadoop MapReduce**

Hadoop基础组件



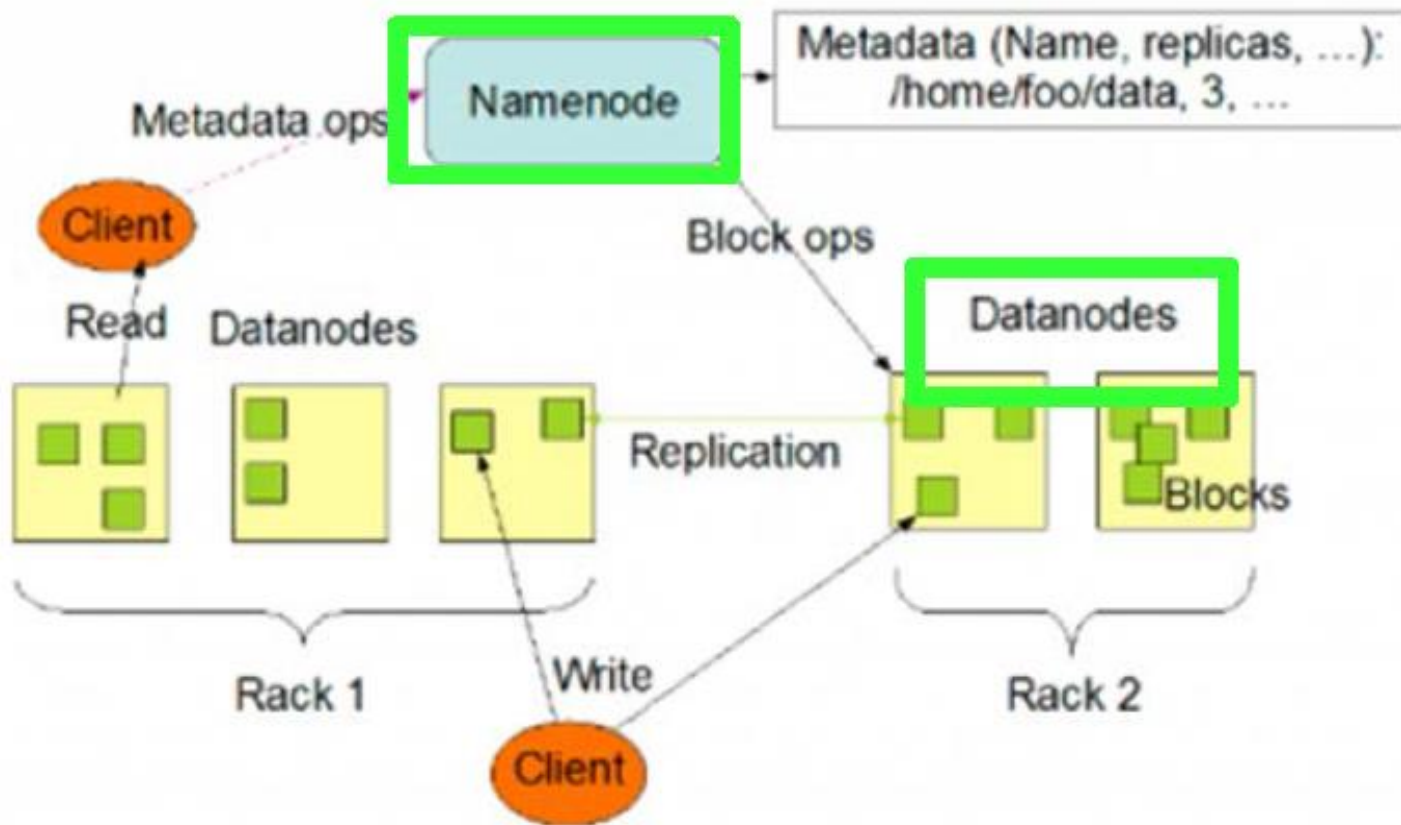


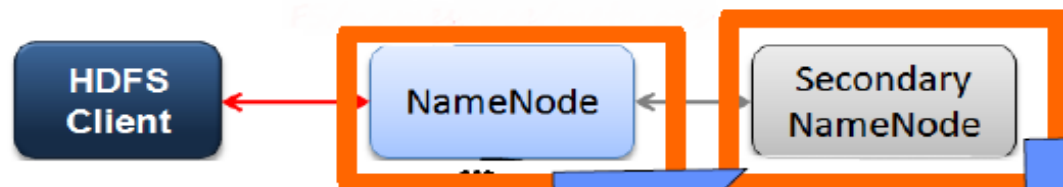
HDFS

Hadoop Distributed File System

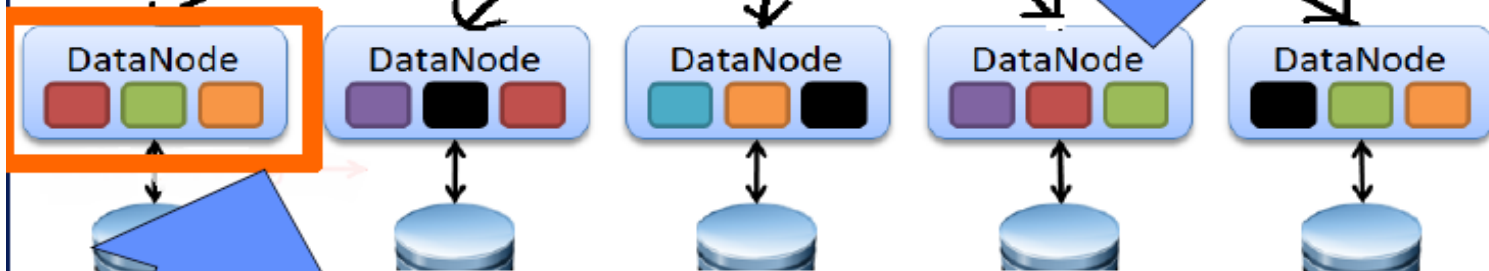
*Distributed, scalable, and portable file-system
written in Java for the Hadoop framework*

HDFS

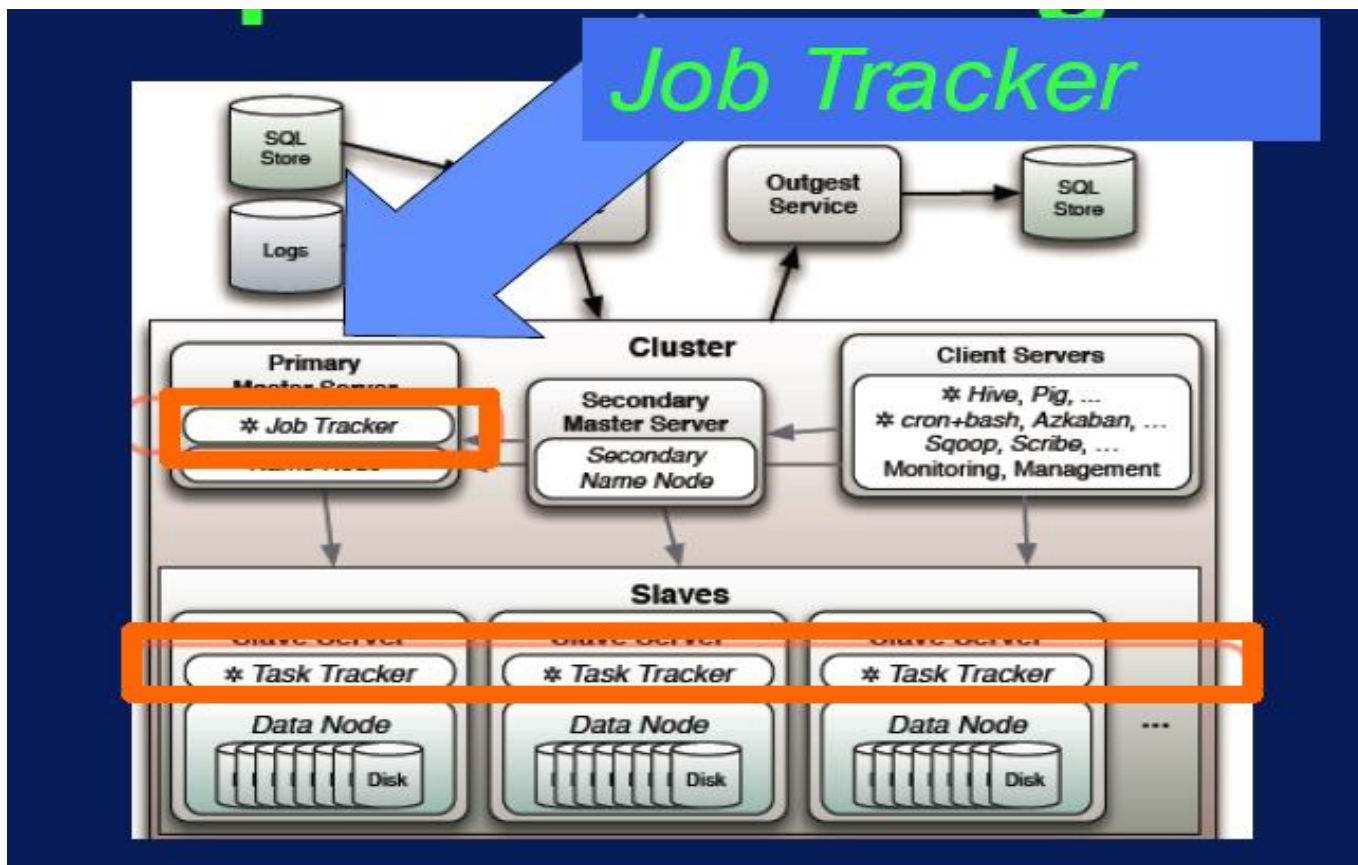




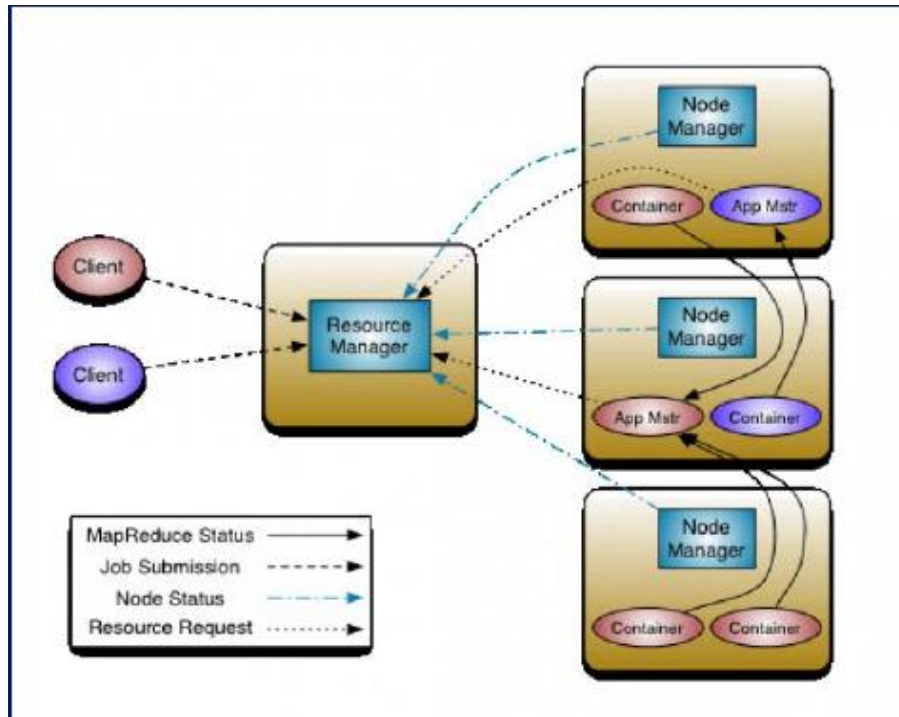
Heartbeats, Replication, and Erasure Coding



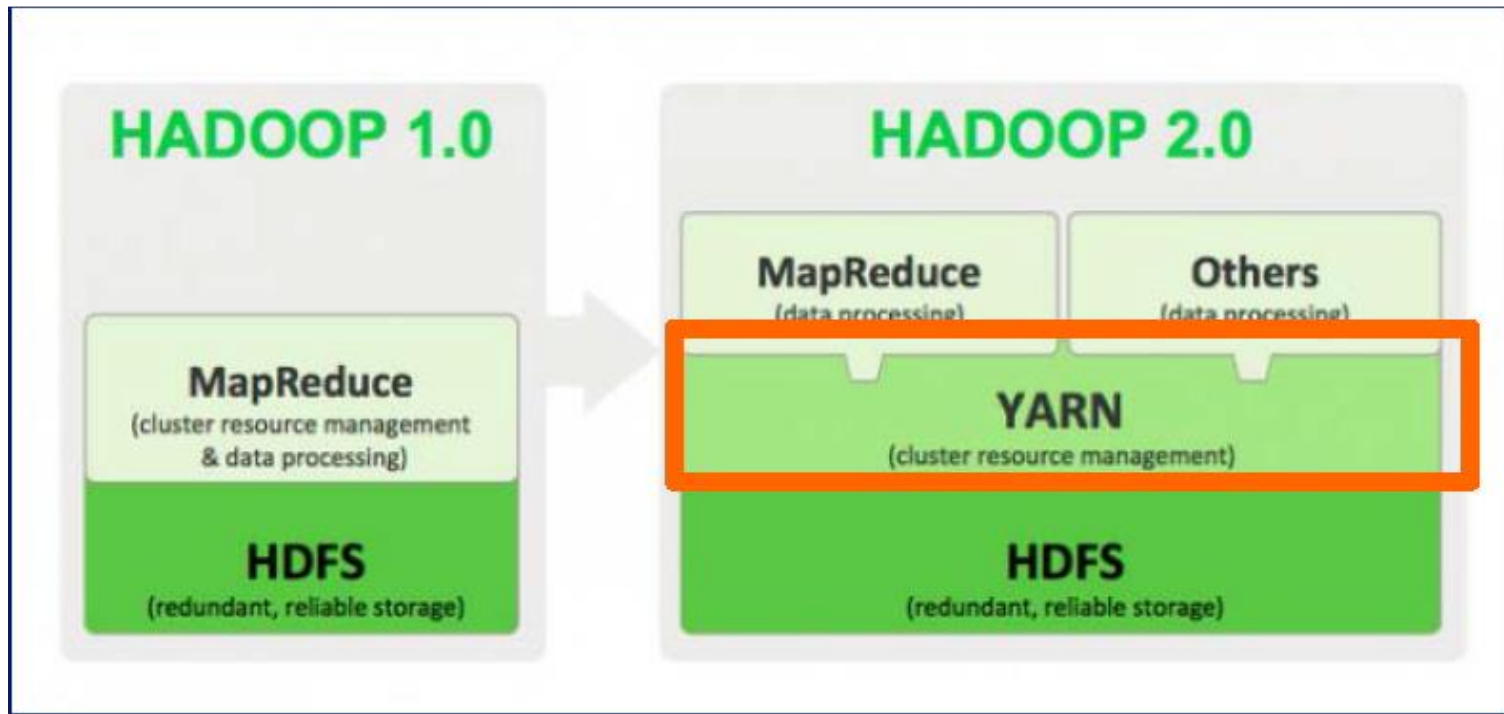
MapReduce



YARN



Hadoop 1.0和2.0MR的主要区别



YARN

资源管理器,可以高效管理集群内的计算资源,除了Hadoop,Yarn也可以和其它框架结合使用,目前市场上除了Yarn,还有Mesos.

Hadoop Zoo



Apache Hadoop Ecosystem



Ambari

Provisioning, Managing and Monitoring Hadoop Clusters



Scoop
Data Exchange



Zookeeper
Coordination



Oozie
Workflow



Pig
Scripting



Mahout
Machine Learning

R Connectors
Statistics




Hive
SQL Query

Apache HBase
Columnar Store



Flume
Log Collector

Zookeeper
Coordination

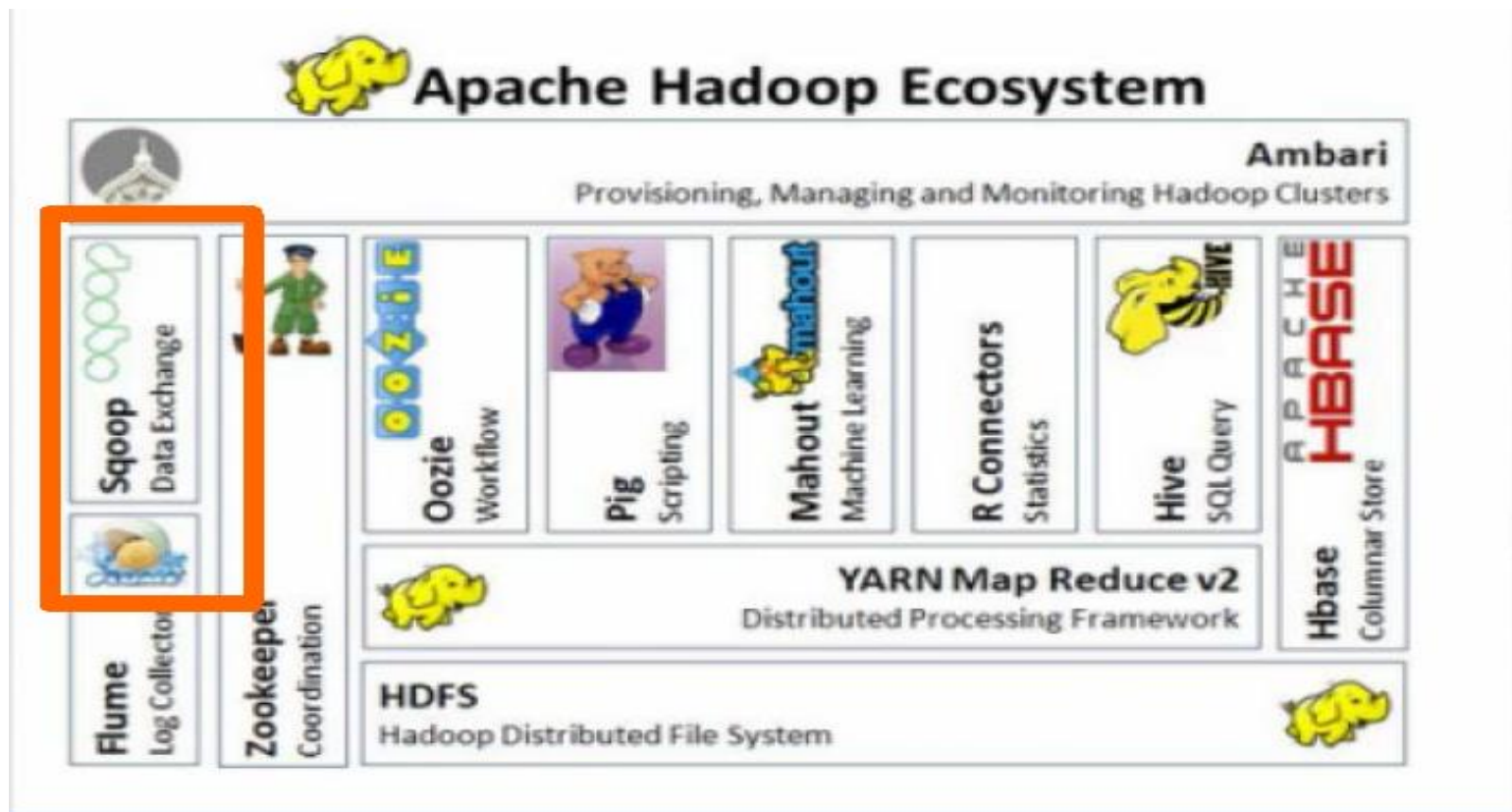


YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System



动物园成员1:sqoop



Apache Sqoop

- Tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases

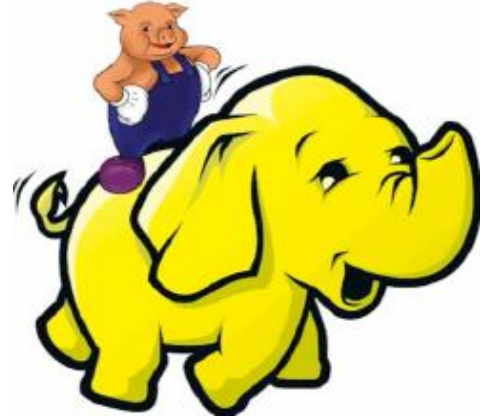


HBASE

- **Column-oriented database management system**
- **Key-value store**
- **Based on Google Big Table**
- **Can hold extremely large data**
- **Dynamic data model**
- **Not a Relational DBMS**

PIG

- **Originally developed at Yahoo 2006**
- **High level programming on top of Hadoop MapReduce**
- **The language: Pig Latin**
- **Data analysis problems as data flows**



Apache Hive

- **Data warehouse software facilitates querying and managing large datasets residing in distributed storage**
- **SQL Like Language**
- **Facilitates querying and managing large datasets in HDFS**
- **Mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL**



Oozie

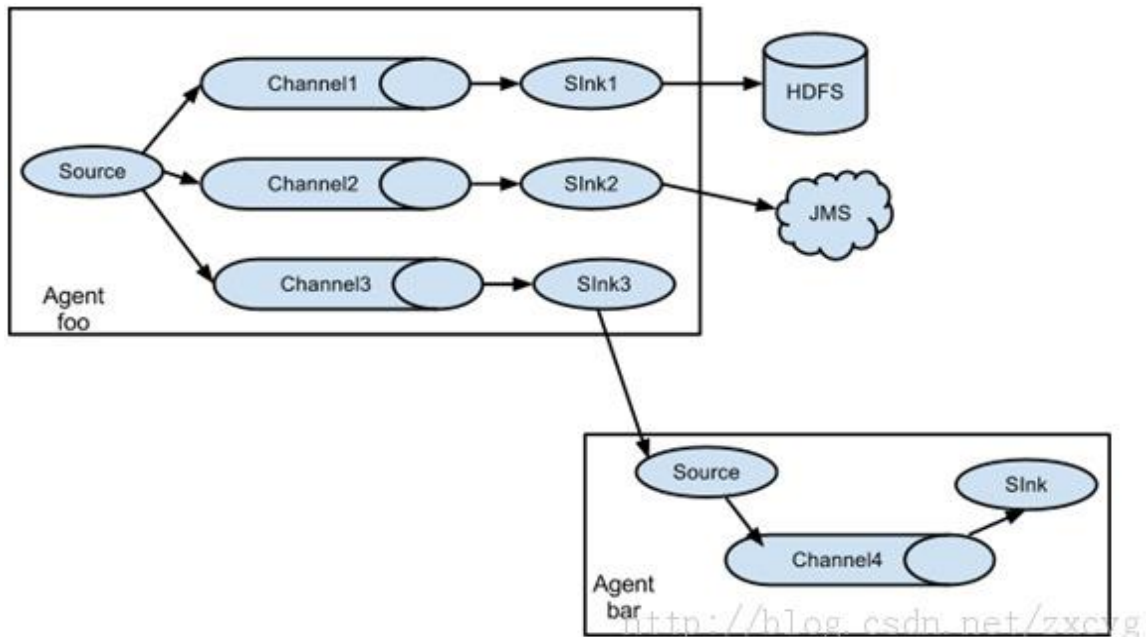
- **Workflow scheduler system to manage Apache Hadoop jobs**
- **Oozie Coordinator jobs!**
- **Supports MapReduce, Pig, Apache Hive, and Sqoop, etc.**

Zookeeper

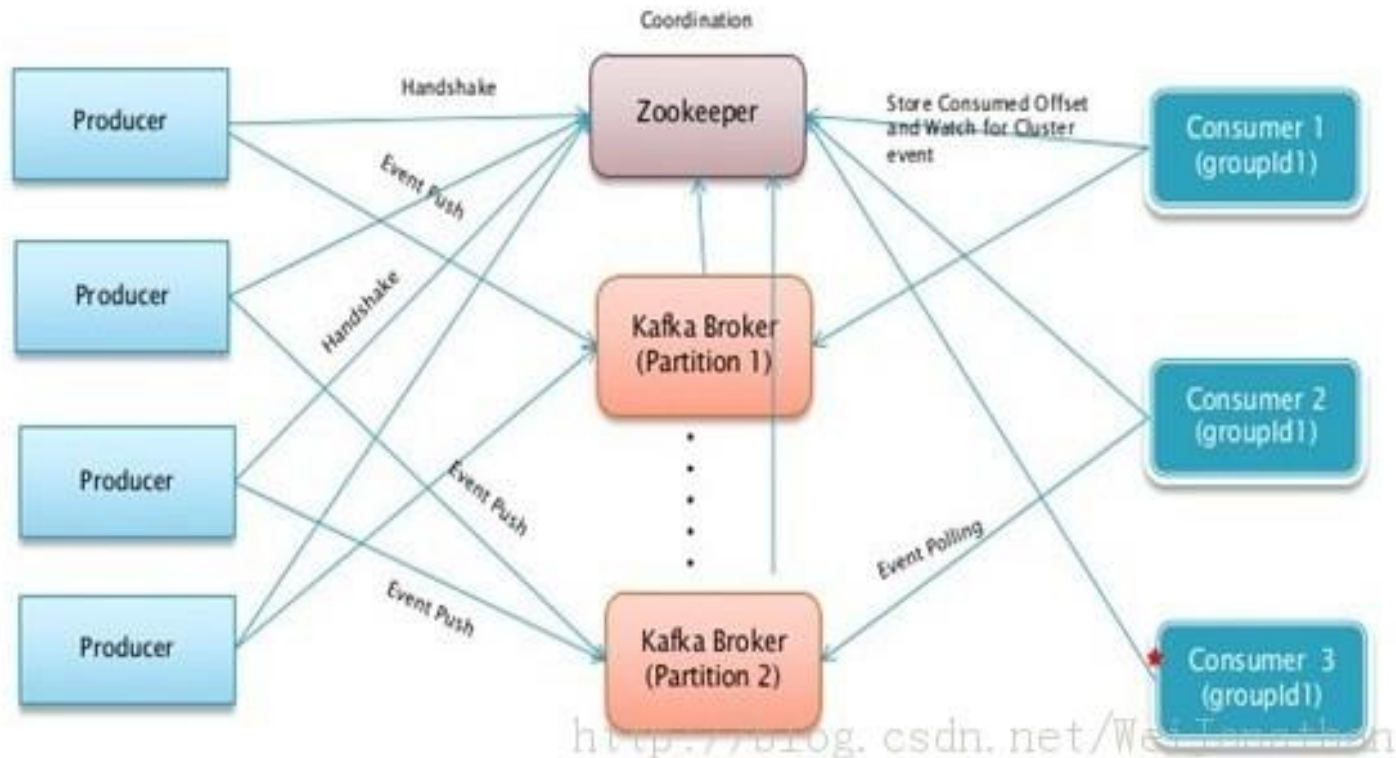
- **Provides operational services for a Hadoop cluster group services**
- **Centralized service for:**
 - **maintaining configuration information**
 - **naming services**
 - **providing distributed synchronization**
 - **and providing group services**

Flume

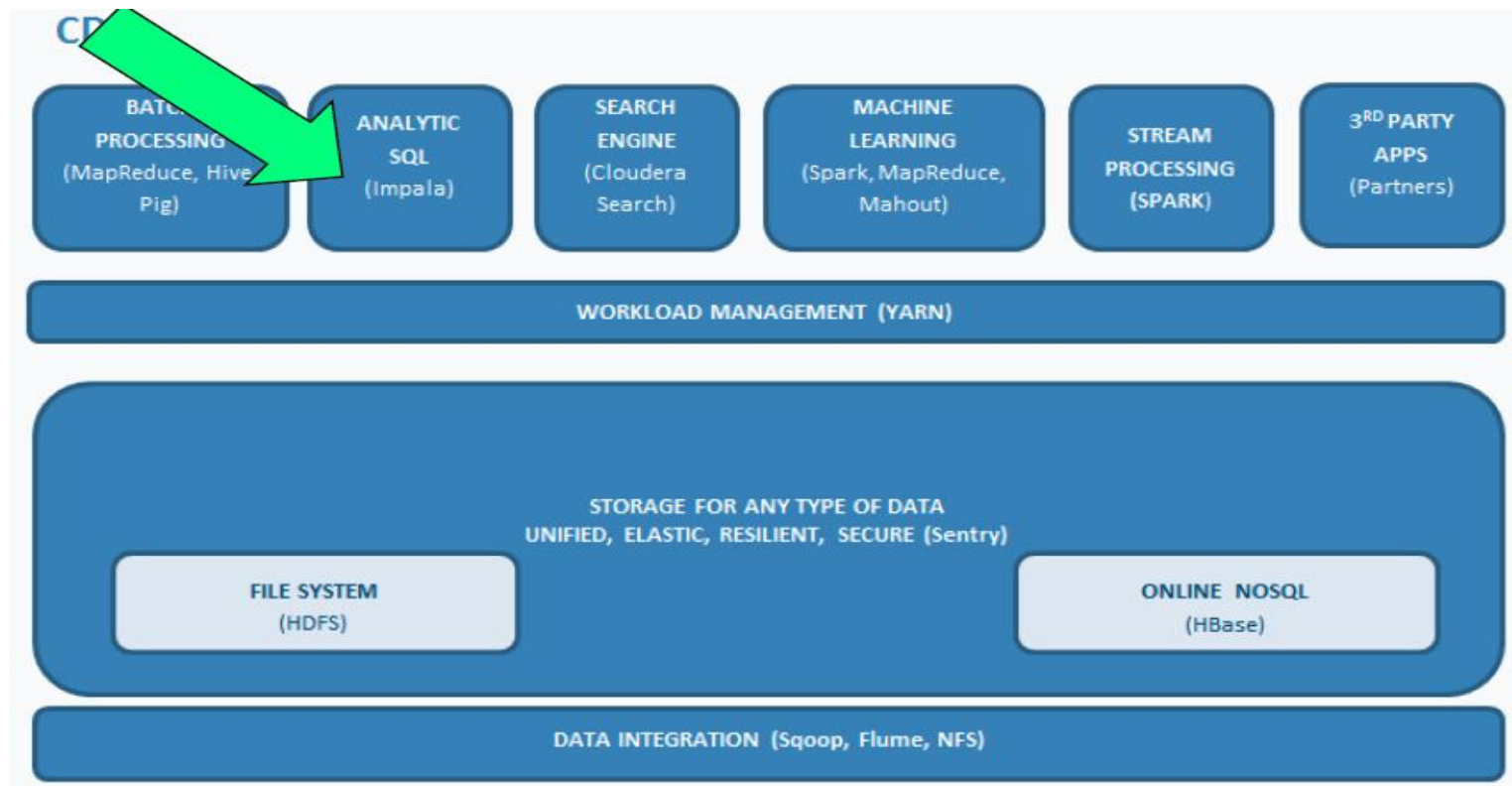
- **Distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data**



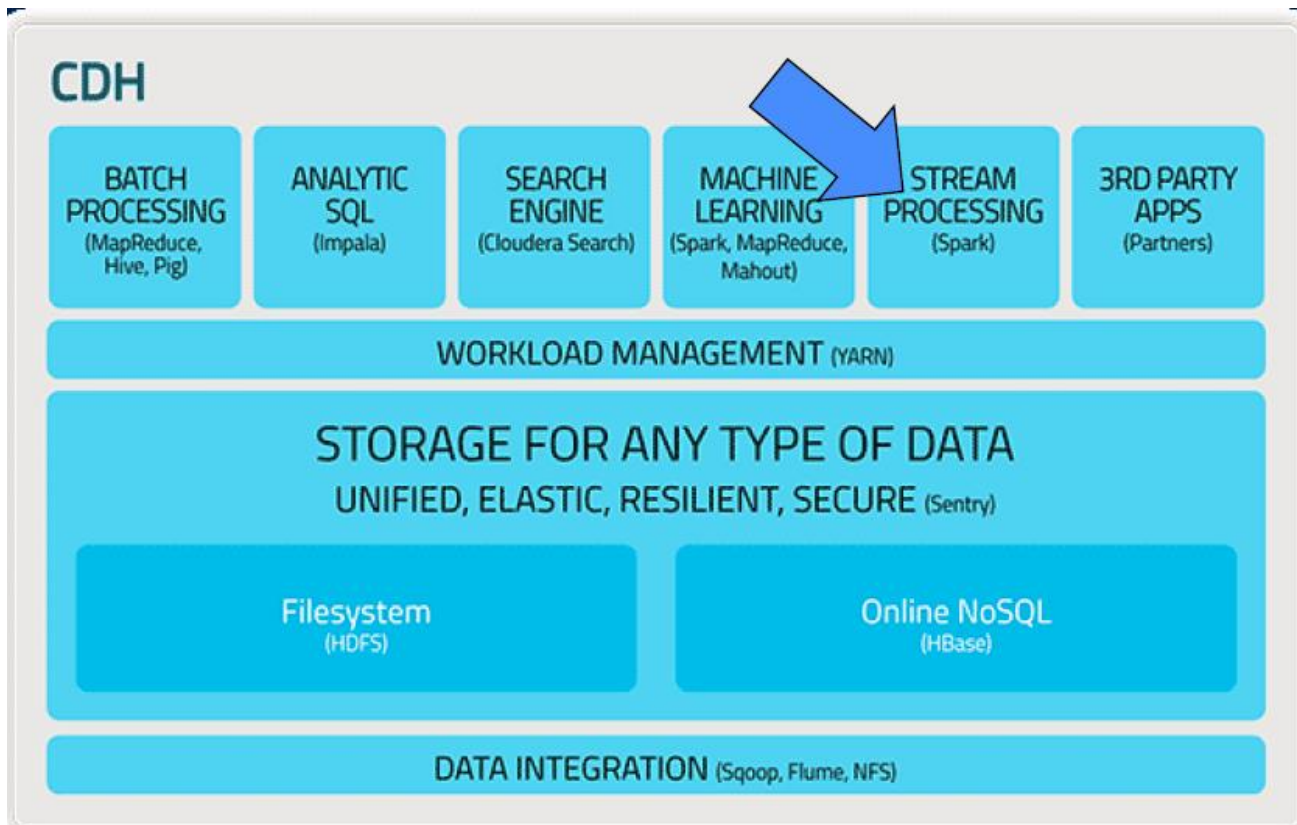
Kafka



Impala



Spark



Storm

