BIG DATA ANALYTICS (CS-431)

Dr. Sriparna Saha
Associate Professor

Website: https://www.iitp.ac.in/~sriparna/
Google Scholar: https://scholar.google.co.in/citations?user=Fj7jA_AAAAAJ&hl=en
Research Lab: SS_Lab
Core Research AREA: NLP, GenAI, LLMs, VLMs, Multimodality, Meta-Learning, Health Care, FinTech, Conversational Agents

TAs: Sarmistha Das, Nitish Kumar, Divyanshu Singh, Aditya Bhagat, Harsh Raj

Big Data Enabling Technologies?

- Big Data is used for a collection of data sets so large and complex that it is difficult to process using traditional tools.

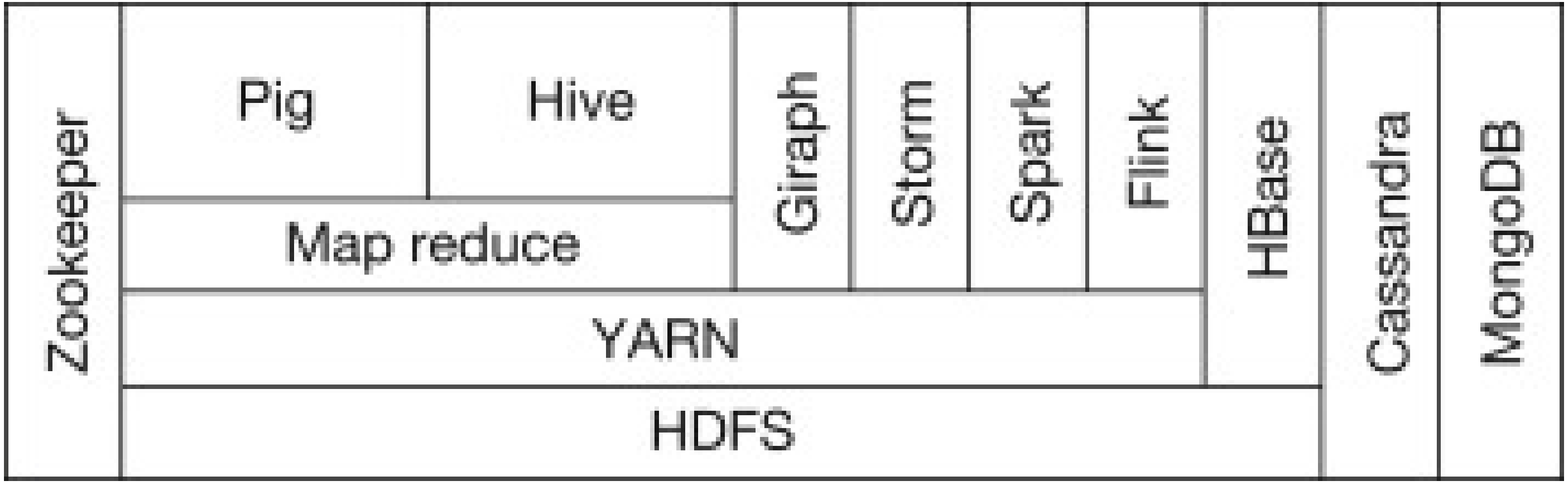
- A recent survey says that 80% of the data created in the world are unstructured.

- One challenge is how we can store and process this big amount of data. In this lecture, we will discuss the top technologies used to store and analyse Big Data.

Apache Hadoop

- ❑ **Apache Hadoop** is an open source software framework for big data.
- ❑ It has two basic parts:
  - ❑ **Hadoop Distributed File System (HDFS)** is the storage system of Hadoop which splits big data and distribut across many nodes in a cluster.
    - ❑ a. Scaling out of H/W resources
    - ❑ b. Fault Tolerant
  - ❑ **MapReduce:** Programming model that simplifies parallel programming.
    - ❑ a. Map-> apply ()
    - ❑ b. Reduce-> summarize ()
    - ❑ c. Google used MapReduce for Indexing websites.

One Layer Diagram For Hadoop System



Zookeeper

Pig

Hive

Map reduce

Giraph

Storm

Spark

Flink

HBase

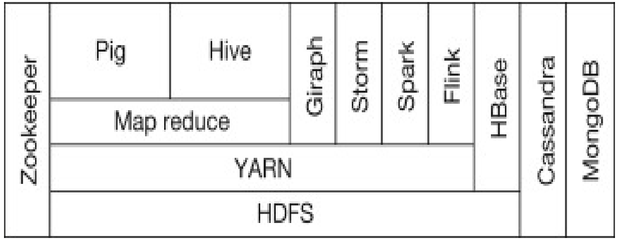
Cassandra

MongoDB

YARN

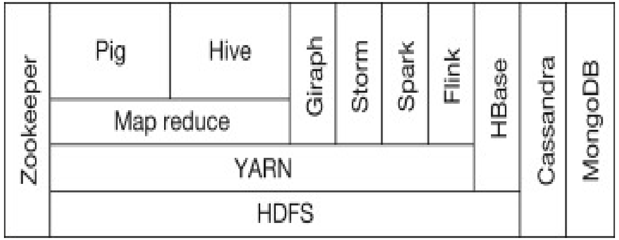
HDFS

One Layer Diagram For Hadoop System



1. **Hadoop Distributed File System: HDFS**
    a. Scaling out of H/W resources
    b. Fault Tolerant
1. **YARN: Flexible scheduling & Resource management over HDFS**
    a. Yahoo uses YARN to schedule jobs over 40,000 servers
1. **MapReduce: Programming model that simplifies parallel programming**
    a. Map -> apply()
    b. Reduce -> summarize()
    C. Google used MR for Indexing websites
1. **Pig & Hive: Augment MapReduce**
    a. Pig: Dataflow based script programming
    B. Hive: SQL like queries (Created at Facebook)

One Layer Diagram For Hadoop System



5. Giraph: Specialized models for graph processing

   a. Used by Facebook to analyze social graphs

6. Storm/Spark/Flink: Real time, in-memory processing of data on top of YARN/HDFS, 100 times faster than regular processing

Cassandra/MongoDB/HBase: NoSQL database

   a. HBase is used for Facebook's Messaging Platform

   b. Sparse tables

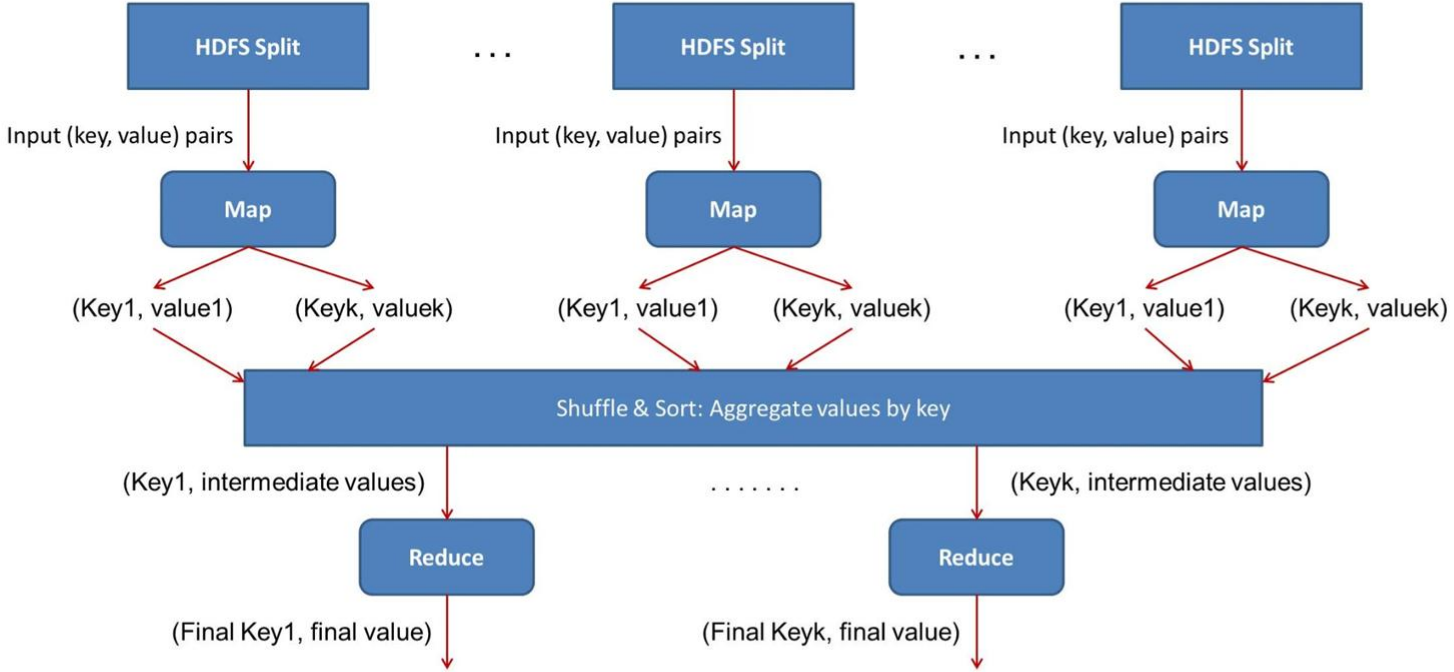
7. Zookeeper: Created by YAHOO to perform the duties of centralized management system for synchronization, configuration and to ensure high availability

Map Reduce

❑ **MapReduce** is a programming model and an associated implementation for p**rocessing and generating large data sets.**

❑ Users specify a **map** function that processes a key/value pair t generate a set of intermediate key/value pairs, and a **reduce** function that merges all intermediate values associated with the same intermediate key

Map Reduce



HDFS Split

. . .

HDFS Split

. . .

HDFS Split

Input (key, value) pairs

Input (key, value) pairs

Input (key, value) pairs

Map

Map

Map

(Key1, value1)

(Keyk, valuek)

(Key1, value1)

(Keyk, valuek)

(Key1, value1)

(Keyk, valuek)

Shuffle & Sort: Aggregate values by key

(Key1, intermediate values)

. . . .

(Keyk, intermediate values)

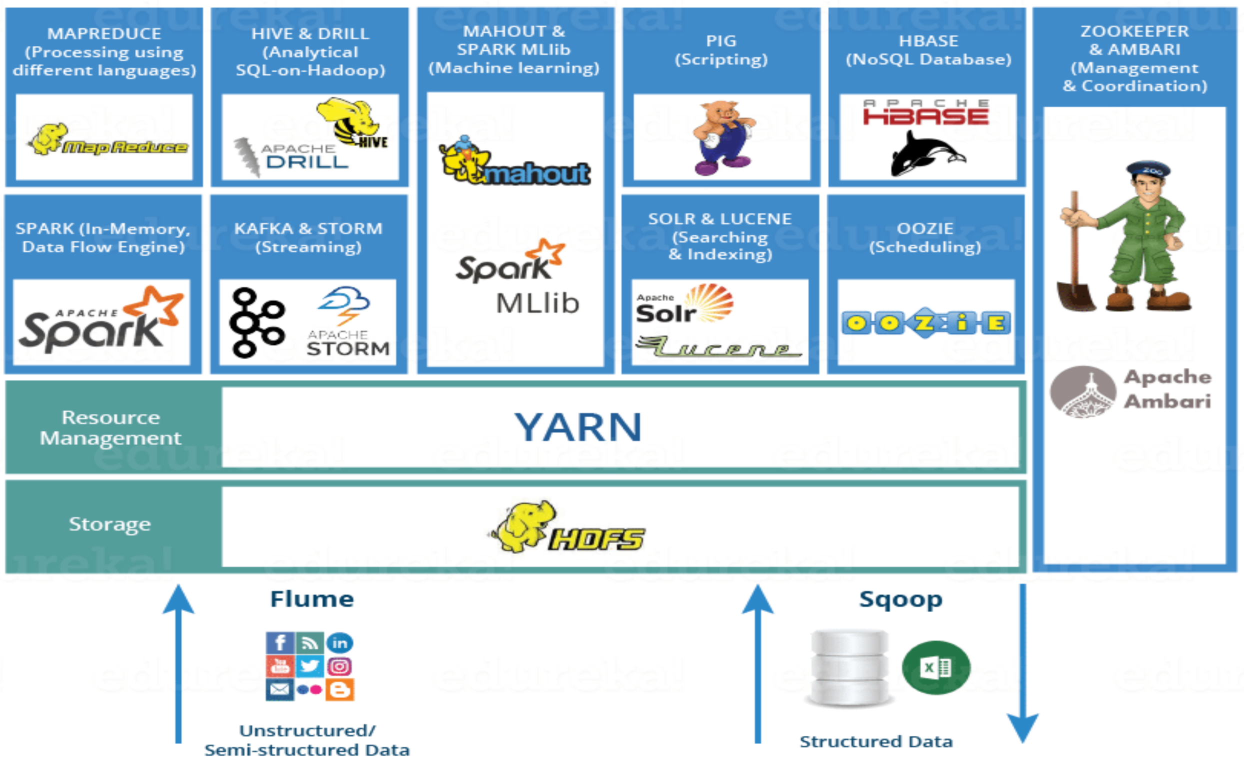
Reduce

Reduce

(Final Key1, final value)

(Final Keyk, final value)

Hadoop Ecosystem



MAPREDUCE (Processing using different languages)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

SPARK (In-Memory, Data Flow Engine)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

Resource Management

YARN

Storage

Flume

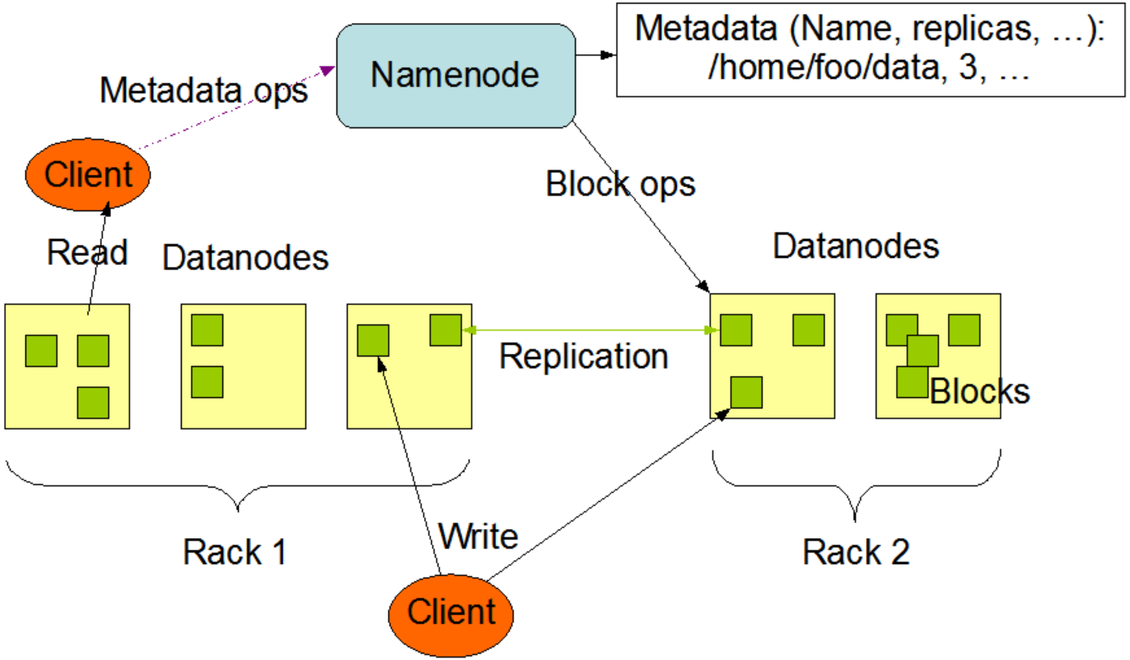
Unstructured/ Semi-structured Data

Sqoop

Structured Data

HDFS Architecture

HDFS Architecture



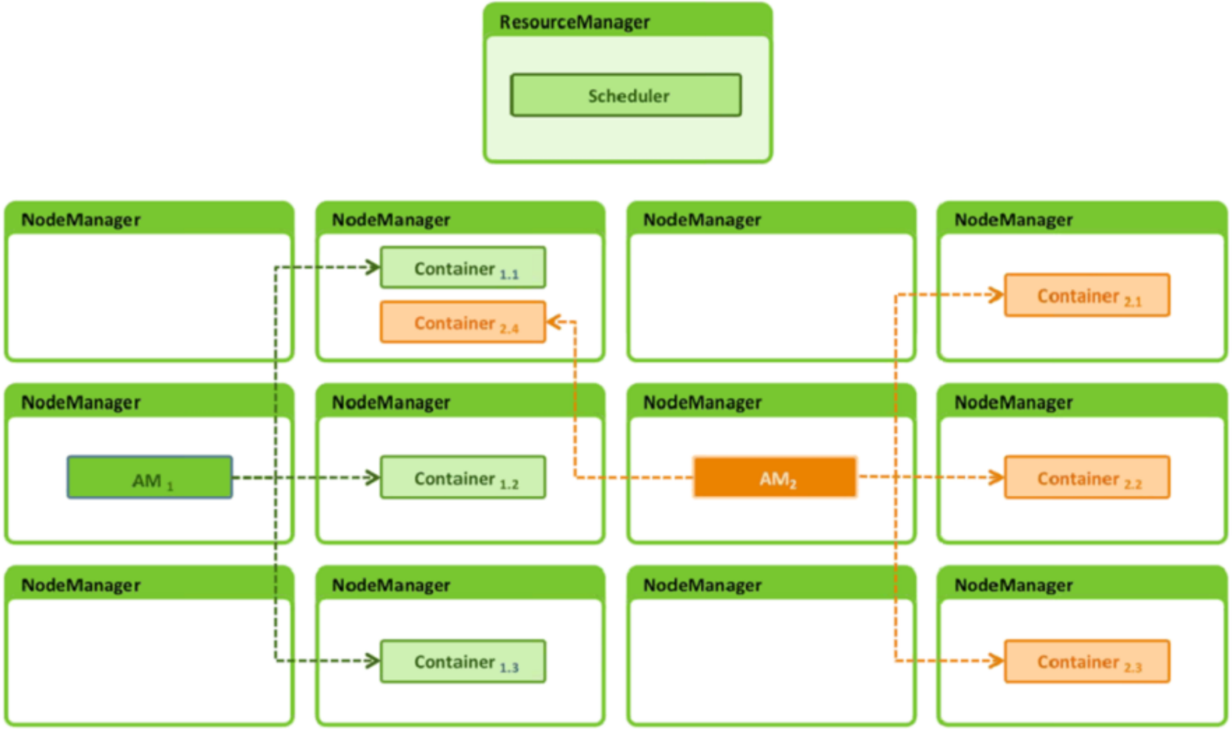
Hadoop YARN

- **YARN** – Yet Another Resource Manager.

- **Apache Hadoop YARN** is the resource management and job scheduling technology in the open source Hadoop distributed processing framework.

- **YARN** is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.

Hadoop YARN Architecture



Hive

- **Hive** is a distributed data management for Hadoop.
- It supports SQL-like query option **Hive SQL** (HSQL) to access big data.
- It can be primarily used for Data mining purpose.
- It runs on top of Hadoop.

Apache Spark

- **Apache Spark** is a big data analytics framework that was originally developed at the University of California, Berkeley's AMPLab, in 2012. Since then, it has gained a lot of attraction both in academia and in industry.

- Apache Spark is a lightning-fast cluster computing technology, designed for fast computation.

- Apache Spark is a lightning-fast cluster computing technology, designed for fast computation

ZooKeeper

- **ZooKeeper** is a highly reliable distributed coordination kernel, which can be used for distributed locking, configuration management, leadership election, work queues,....

- **Zookeeper** is a replicated service that holds the metadata of distributed applications.

- **Key attributed of such data**
  - Small size
  - Performance sensitive
  - Dynamic
  - Critical
- **In very simple words,** it is a central store of key-value using which distributed systems can coordinate. Since it needs to be able to handle the load, Zookeeper itself runs on many machines.

NoSQL

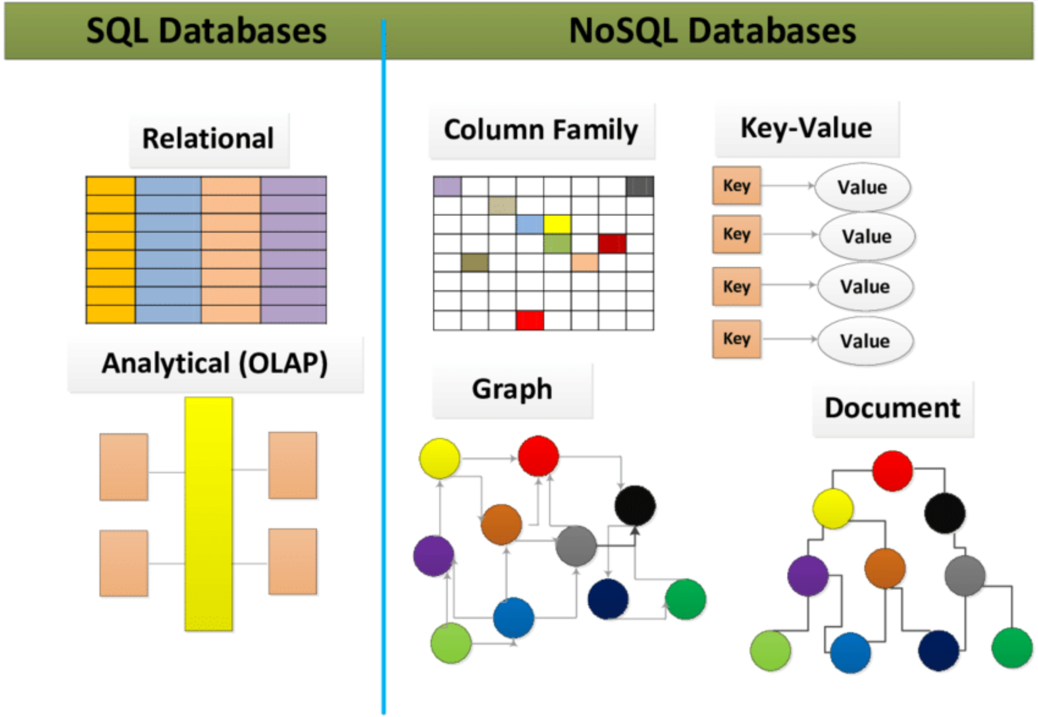
- While the traditional SQL can be effectively used to handle large amount of structured data, we need **NoSQL (Not Only SQL)** to handle unstructured data.

- NoSQL databases store unstructured data with no particular schema

- Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data.

NoSQL

Not Only SQL



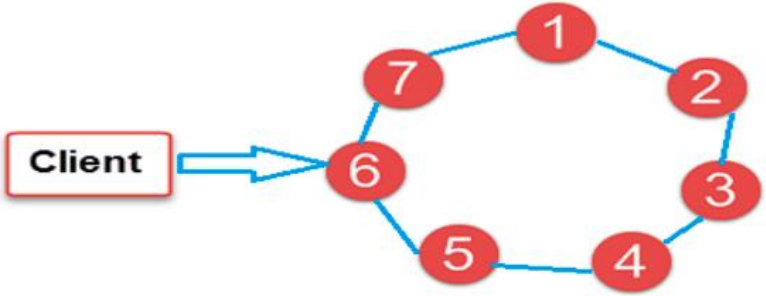
Cassandra

- **Apache Cassandra** is highly scalable, distributed and high-performance NoSQL database. Cassandra is designed to handle a huge amount of data.

- Cassandra handles the huge amount of data with its distributed architecture.

- Data is placed on different machines with more than one replication factor that provides high availability and no single point of failure.

Cassandra



In the image above, circles are Cassandra nodes and lines between the circles shows distributed architecture, while the client is sending data to the node

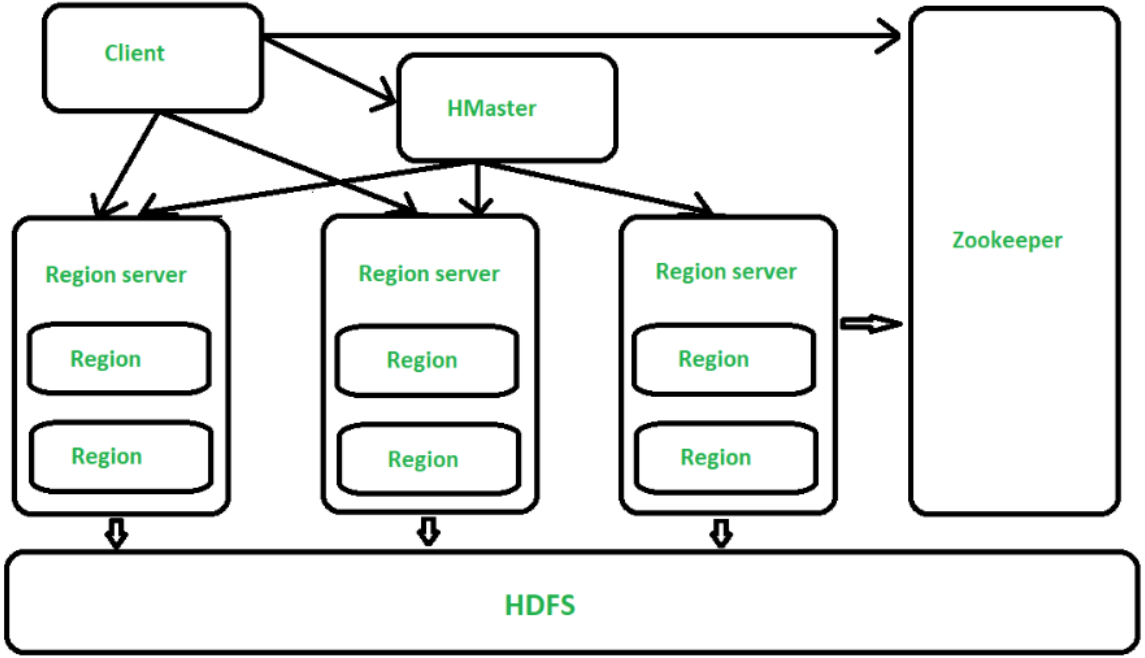
HBase

- **HBase** is an open source, distributed database, developed by Apache Software foundation.

- Initially, it was Google Big Table, afterwards it was re-named as HBase and is primarily written in Java.

- HBase can store massive amounts of data from terabytes to petabytes.

HBase Architecture



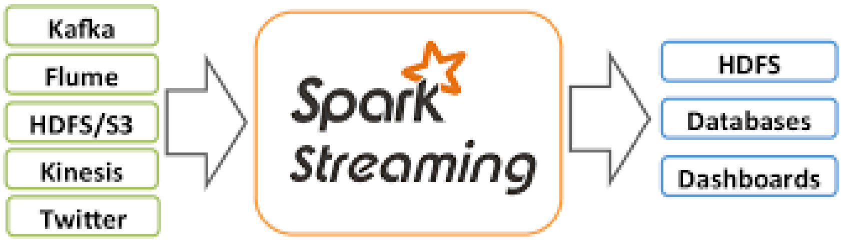
Spark Streaming

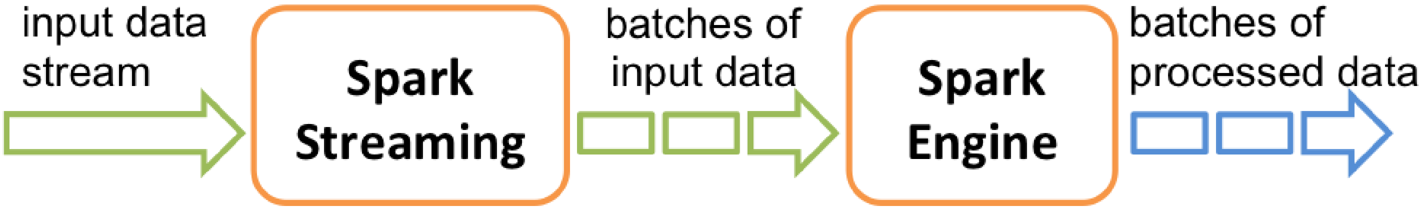
- Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams.

- Streaming data input from HDFS, Kafka, Flume, TCP sockets, Kinesis, etc.

- Spark ML (Machine Learning) functions and GraphX graph processing algorithms are fully applicable to streaming data .

Spark Streaming





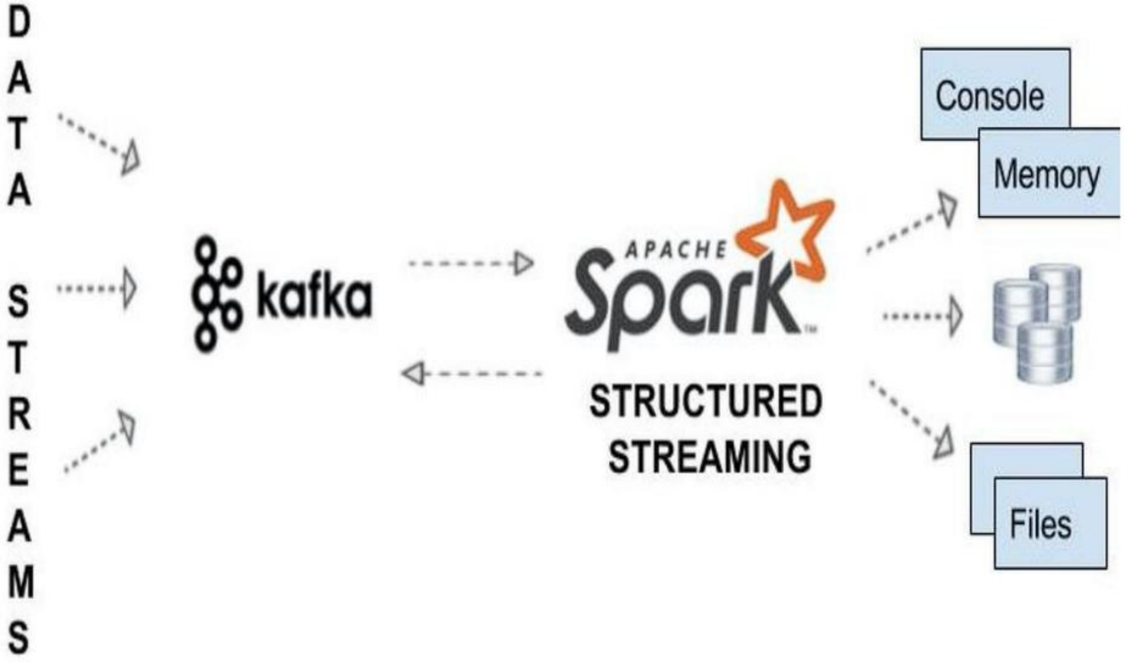
Kafka, Streaming Ecosystem

kafka Streams

- Apache Kafka is an open-source stream-processing software platform developed by the Apache Software Foundation written in Scala and Java.

- Apache Kafka is an open source distributed streaming platform capable of handling trillions of events a day, Kafka is based on an abstraction of a distributed commit log

Kafka, Streaming Ecosystem

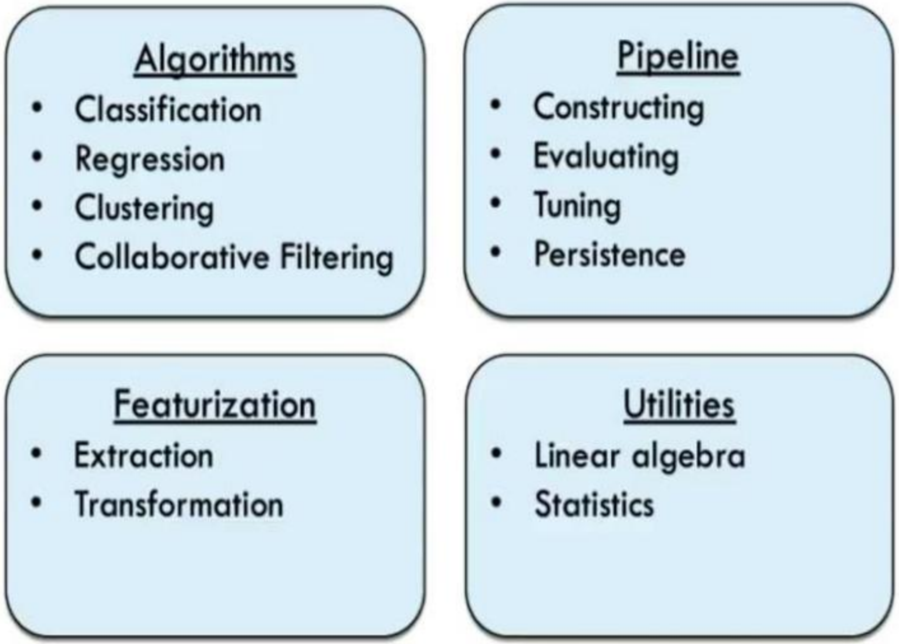


Spark MLlib

- Spark MLlib is a distributed machine-learning framework on top of Spark Core.

- MLlib is Spark's scalable machine learning library consisting of common learning algorithms and utilities,including classification, regression, clustering, collaborative filtering, dimensionality reduction.

Spark MLlib Component



Algorithms

- Classification
- Regression
- Clustering
- Collaborative Filtering

Pipeline

- Constructing
- Evaluating
- Tuning
- Persistence

Featurization

- Extraction
- Transformation

Utilities

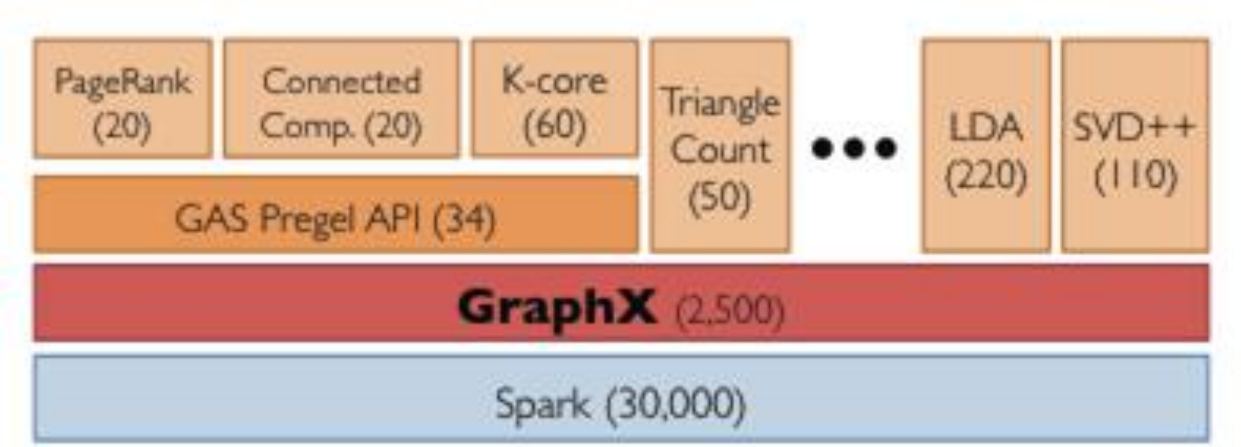
- Linear algebra
- Statistics

Spark MLlib

- GraphX is a new component in Spark for graphs and graph-parallel computation. At a high level, GraphX extends the Spark RDD by introducing a new graph abstraction.

- GraphX reuses Spark RDD concept, simplifies graph analytics tasks, provides the ability to make operations on a directed multigraph with properties attached to each vertex and edge.

Spark MLlib



GraphX is a thin layer on top of the Spark general-purpose dataflow framework (lines of code).

Conclusion

In this lecture, we given a brief overview of following Big Data Enabling Technologies:

- Apache Hadoop
- Hadoop Ecosystem
- HDFS Architecture
- YARN
- NoSQL
- Hive
- Map Reduce
- Apache Spark
- Zookeeper
- Cassandra
- Hbase
- Spark Streaming
- Kafka
- Spark MLlib
- GraphX

Hadoop Stack for Big Data

What is Hadoop ?

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

Hadoop was created by Doug Cutting and Mike Cafarella in 2005

It was originally developed to support distribution of the Nutch Search Engine Project.

Doug, who was working at Yahoo at the time, who is now actually a chief architect at Cloudera, has named this project after his son’s toy elephant, Hadoop.

Scalability & Reliability

Scalability:

*Scalability's at it's core of a Hadoop system.
*We have cheap computing storage.
*We can distribute and scale across very easily
*in a very cost effective manner.

Reliability:

*Hardware Failures Handles Automatically!

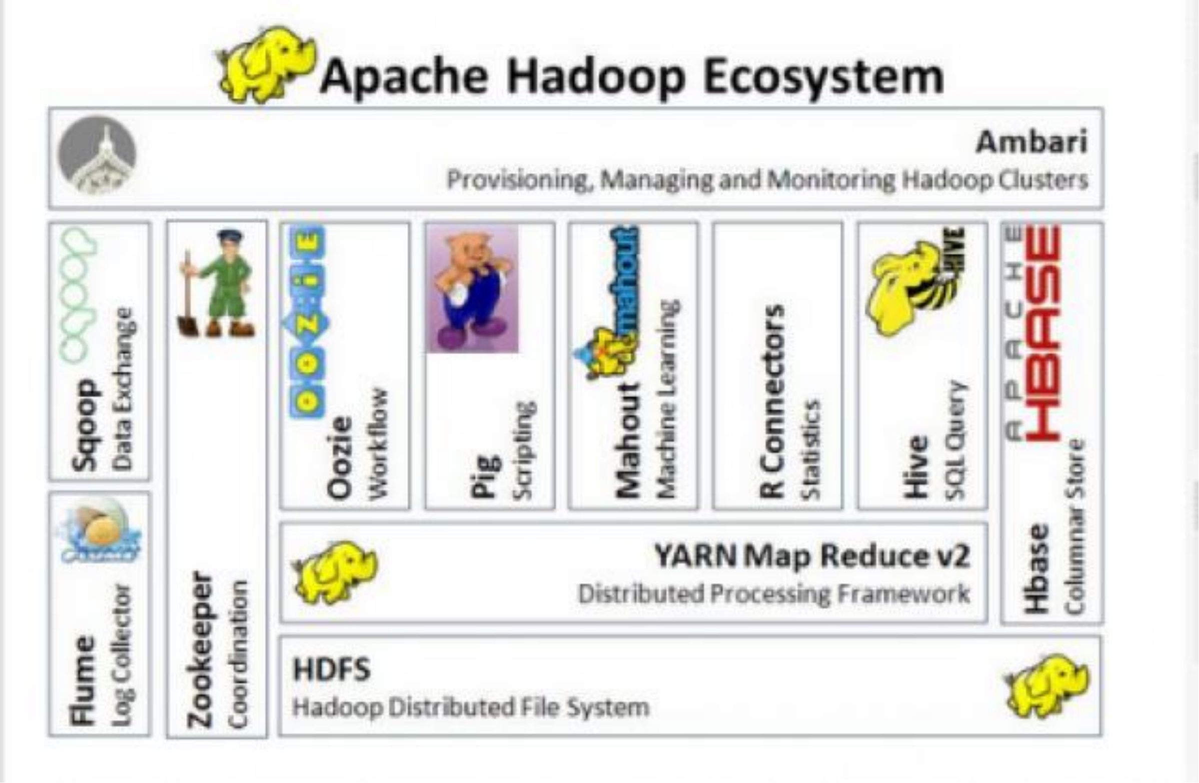
New Approach to Data: Keep all data

- A new approach is, we can keep all the data that we have, and we can take that data and analyze it in new interesting ways. We can do something that's called schema and read style.
- And we can actually allow new analysis. We can bring more data into simple algorithms, which has shown that with more granularity, you can actually achieve often better results in taking a small amount of data and then some really complex analytics on it.

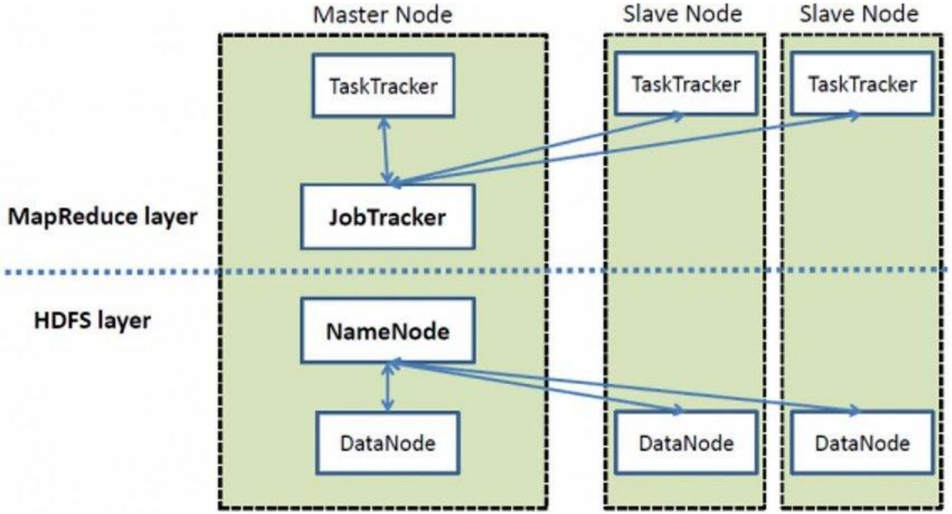
Apache Framework Basic Modules

- **Hadoop Common**: It contains libraries and utilities needed by other Hadoop modules.
- **Hadoop Distributed File System (HDFS)**: It is a distributed file system that stores data on a commodity machine. Providing very high aggregate bandwidth across the entire cluster.
- **Hadoop YARN**: It is a resource management platform responsible for managing compute resources in the cluster and using them in order to schedule users and applications.
- **Hadoop MapReduce**: It is a programming model that scales data across a lot of different processes.

Apache Framework Basic Modules



High Level Architecture of Hadoop



- Two major pieces of Hadoop are: Hadoop Distribute the File System and the MapReduce, a parallel processing framework that will map and reduce data. These are both open source and inspired by the technologies developed at Google.
- If we talk about this high level infrastructure, we start talking about things like TaskTrackers and JobTrackers, the NameNodes and DataNodes.

HDFS: Hadoop distributed file system

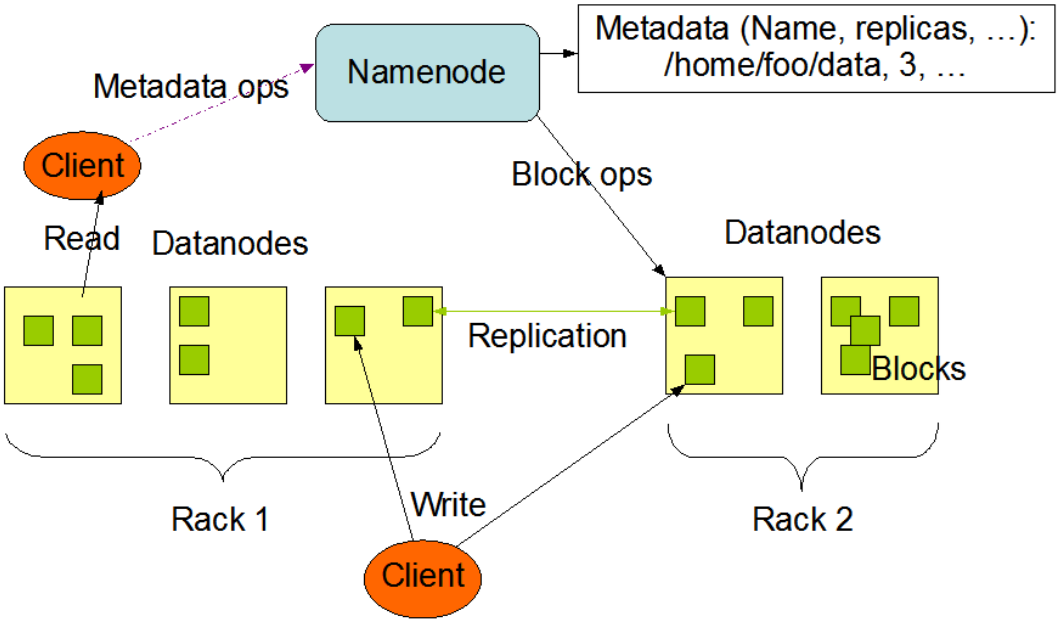
- Distributed, scalable, and portable file-system written in Java for the Hadoop framework.

- Each node in Hadoop instance typically has a single namenode, and a cluster of data nodes that formed this HDFS cluster.

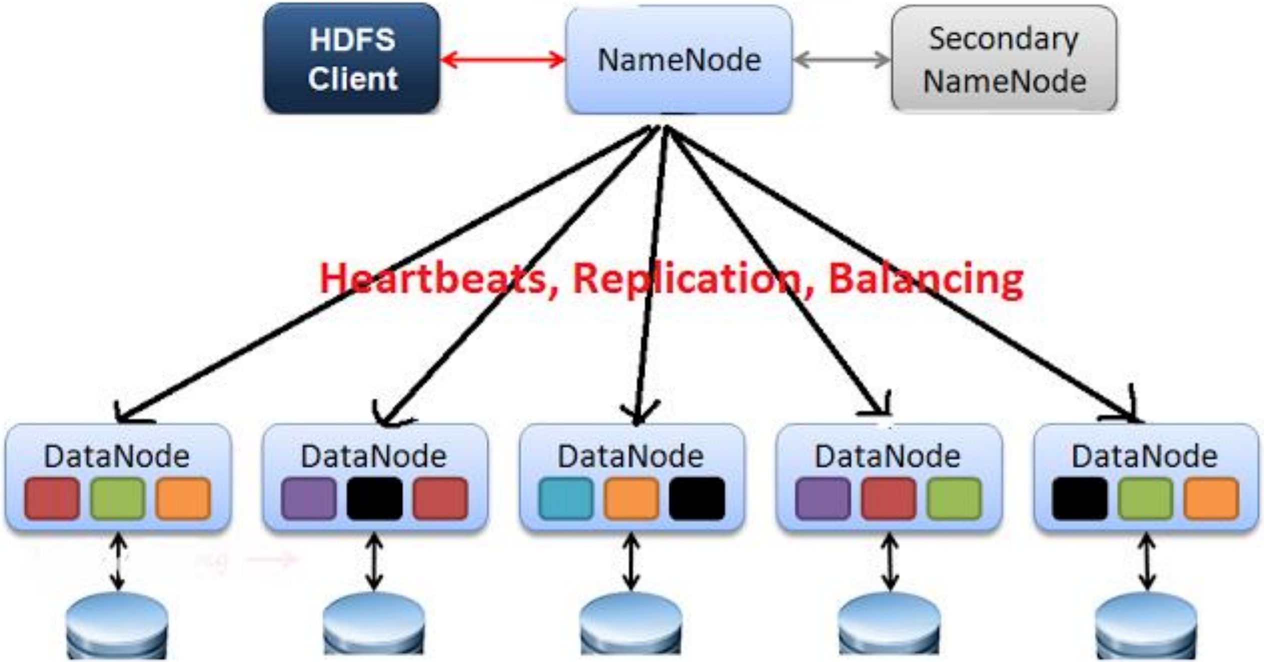
- Each HDFS stores large files, typically in ranges of gigabytes to terabytes, and now petabytes, across multiple machines. And it can achieve reliability by replicating the cross multiple hosts, and therefore does not require any range storage on hosts.

HDFS: Hadoop distributed file system

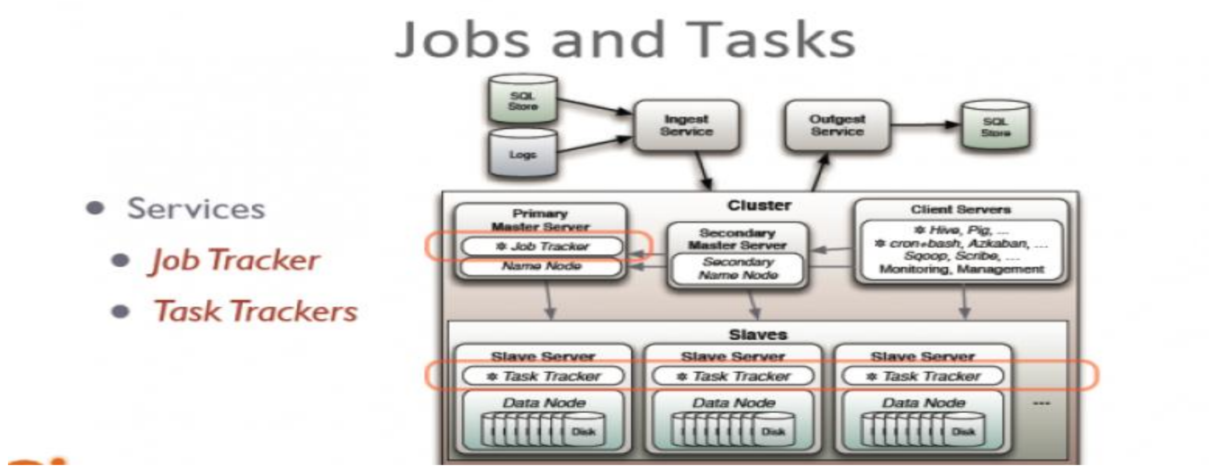
HDFS Architecture



HDFS: Hadoop distributed file system



MapReduce Engine



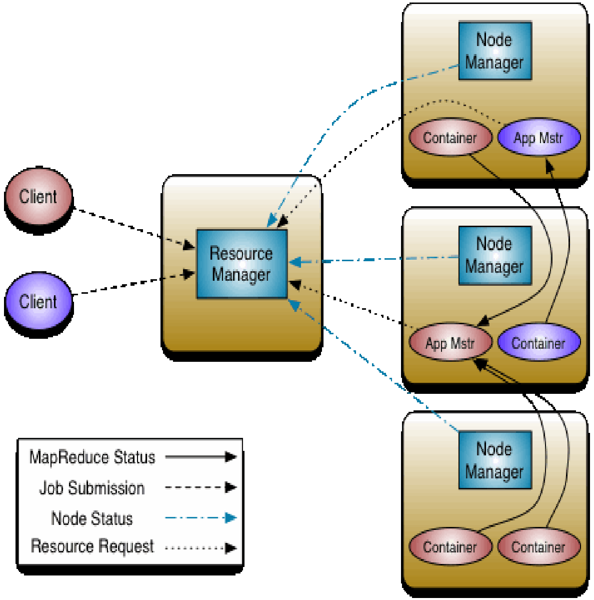
Jobs and Tasks

- Services
  - *Job Tracker*
  - *Task Trackers*

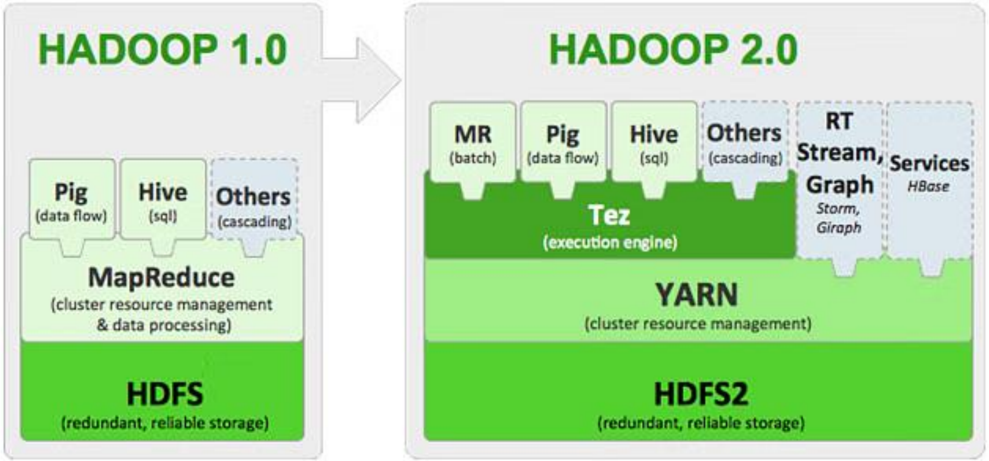
- The typical MapReduce engine will consist of a job tracker, to which client applications can submit MapReduce jobs, and this job tracker typically pushes work out to all the available task trackers, now it's in the cluster. Struggling to keep the word as close to the data as possible, as balanced as possible.

Apache Hadoop NextGen MapReduce (YARN)

- Yarn enhances the power of the Hadoop compute cluster, without being limited by the map produce kind of framework.
- It's scalability's great. The processing power and data centers continue to grow quickly, because the YARN research manager focuses exclusively on scheduling. It can manage those very large clusters quite quickly and easily.
- YARN is completely compatible with the MapReduce. Existing MapReduce application end users can run on top of the Yarn without disrupting any of their existing Processes.



Hadoop 1.0 vs. Hadoop 2.0



- Hadoop 2.0 provides a more general processing platform, that is not constraining to this map and reduce kinds of processes.
- The fundamental idea behind the MapReduce 2.0 is to split up two major functionalities of the job tracker, resource management, and the job scheduling and monitoring, and to do two separate units. The idea is to have a global resource manager, and per application master manager.

What is Yarn ?

- Yarn enhances the power of the Hadoop compute cluster, without being limited by the map produce kind of framework.
- It's scalability's great. The processing power and data centers continue to grow quickly, because the YARN research manager focuses exclusively on scheduling. It can manage those very large clusters quite quickly and easily.
- YARN is completely compatible with the MapReduce. Existing MapReduce application end users can run on top of the Yarn without disrupting any of their existing processes.
- It does have a Improved cluster utilization as well. The resource manager is a pure schedule or they just optimize this cluster utilization according to the criteria such as capacity, guarantees, fairness, how to be fair, maybe different SLA's or service level agreements.

Scalability

MapReduce Compatibility

Improved cluster utilization

What is Yarn ?

- It supports other work flows other than just map reduce.
- Now we can bring in additional programming models, such as graph process or iterative modeling, and now it's possible to process the data in your base. This is especially useful when we talk about machine learning applications.
- Yarn allows multiple access engines, either open source or proprietary, to use Hadoop as a common standard for either batch or interactive processing, and even real time engines that can simultaneous acts as a lot of different data, so you can put streaming kind of applications on top of YARN inside a Hadoop architecture, and seamlessly work and communicate between these environments.

Fairness

Iterative Modeling

Machine Learning

Multiple Access Engines

Supports Other Workloads