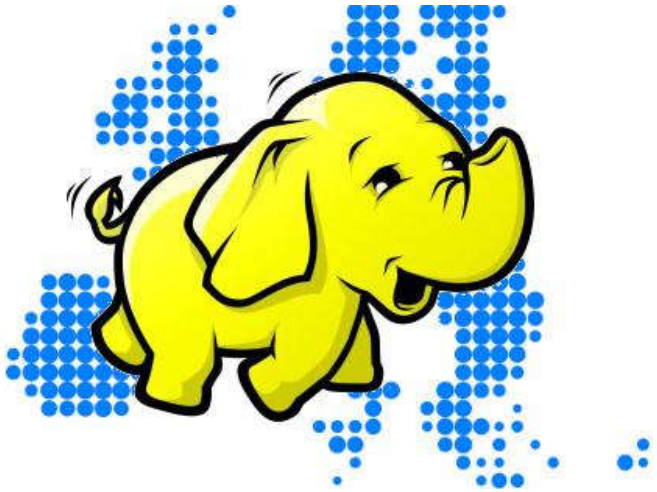
The Principle of Hadoop :

Name: อวยชัย ภิรมย์รื่น

Tel. : 086-813-5354

e-mail : p.Auoychai@gmail.com

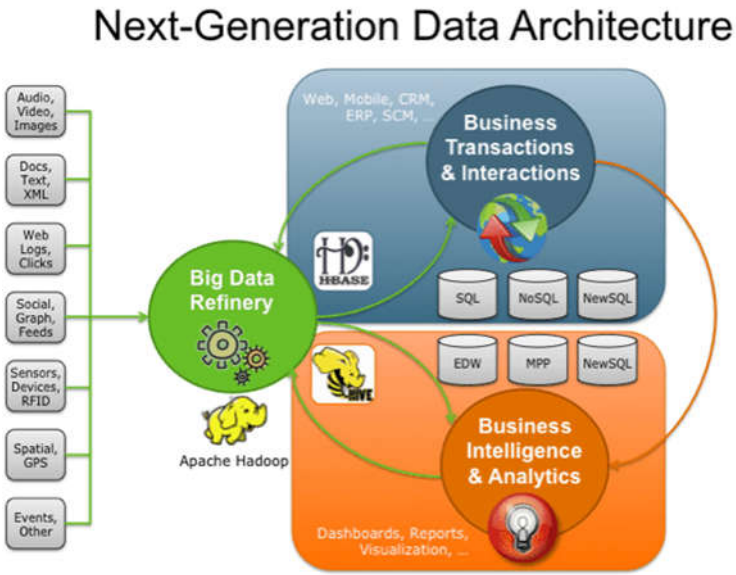
Hadoop World: Why will be Hadoop ?

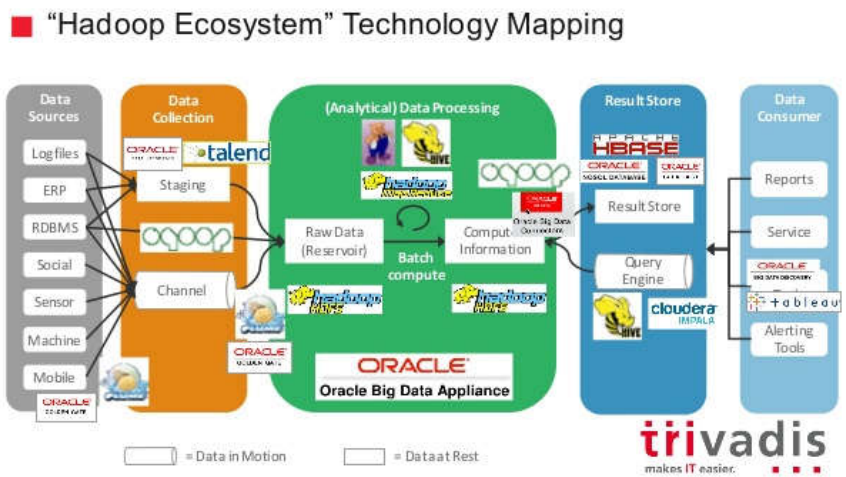


Hadoop was created by Doug Cutting and Mike Cafarella in 2005. It was originally developed to support distribution for the Nutch search engine project. Doug, who was working at Yahoo! at the time and is now Chief Architect of Cloudera, named the project after his son's toy elephant. Cutting's son was 2 years old at the time and just beginning to talk. He called his beloved stuffed yellow elephant "Hadoop"

https://opensource.com/life/14/8/intro-apache-hadoop-big-data

Hadoop World: Why will be Hadoop ?





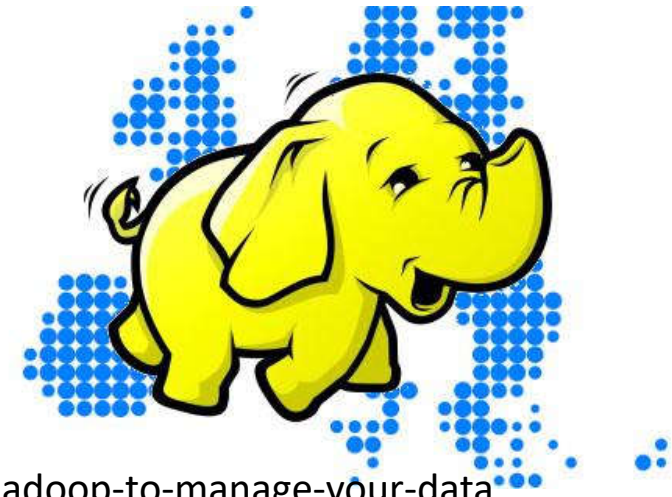
https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2012/05/nextgendataarchitecture.png

What is Hadoop :

Hadoop is software framework with optimized for distributed processing of very large datasets.

Its has two main features are the Hadoop Distributed File System(HDFS) for storing files and MapReduce for processes the stored information

Advantage: Store terabytes of data using any number of inexpensive commodity servers.

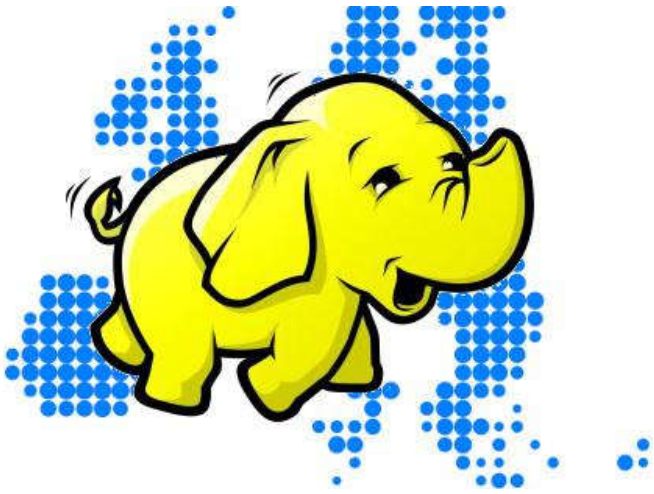


http://www.brightcomputing.com/blog/hadoop-101-do-you-need-hadoop-to-manage-your-data

What is Hadoop :

Key features that answer – Why Hadoop?

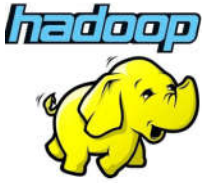
1. Flexible:
2. Scalable/fault tolerant
3. Building more efficient data economy:
4. Robust Ecosystem: MapReduce, Hive, HBase, Zookeeper, HCatalog,…
5. Hadoop is getting more “Real-Time”!
6. Cost Effective:
7. Upcoming Technologies using Hadoop:
8. Hadoop is getting cloudy!



https://www.edureka.co/blog/why-hadoop/

Hadoop Classification :

Open Source



+ Ecosystem

Distribution

cloudera MAPR TECHNOLOGIES

TERADATA

ORACLE BIG DATA

IBM

Appliance

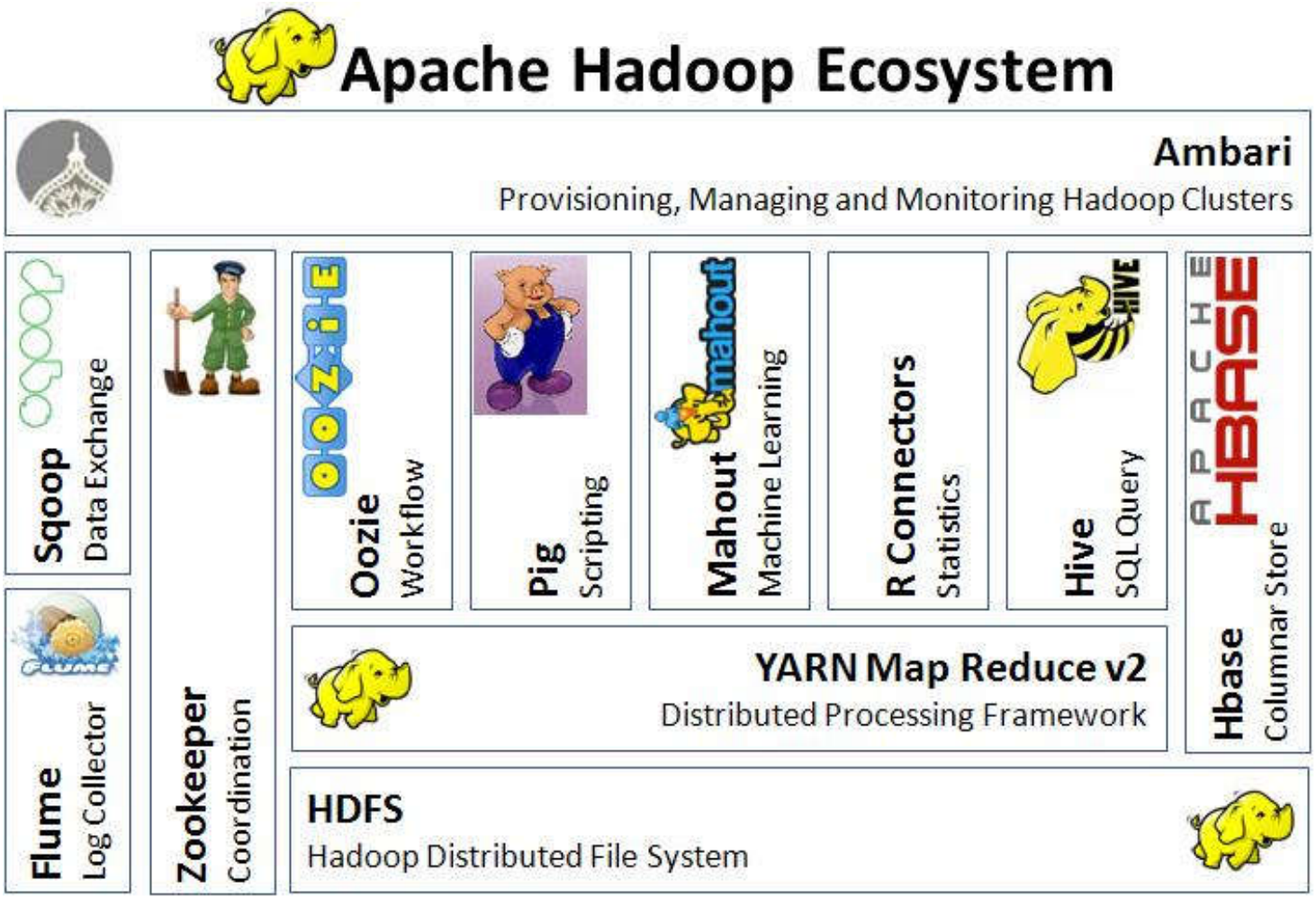
Cloud

Hadoop Component :

- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

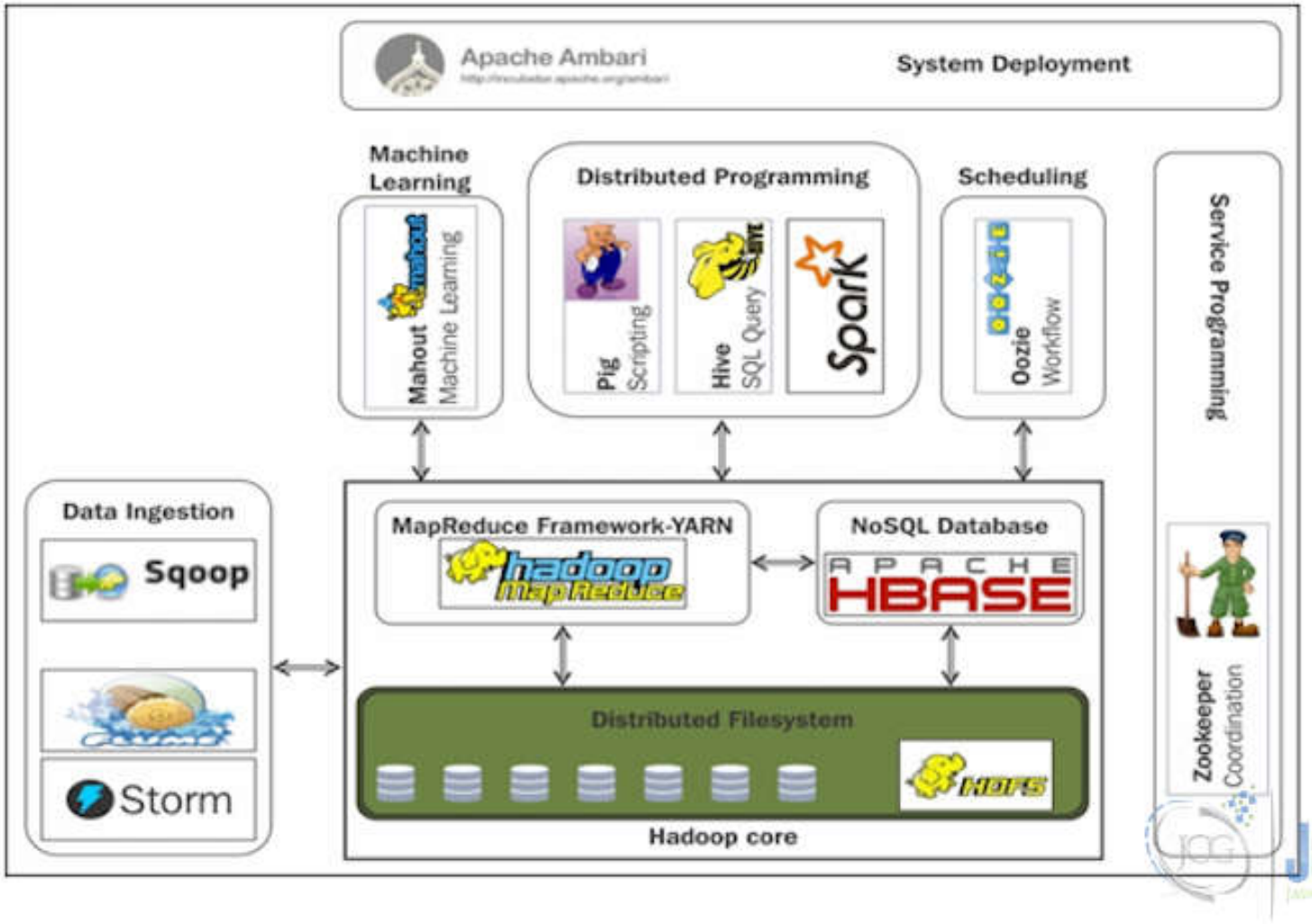
http://hadoop.apache.org , https://www.edureka.co

Hadoop Ecosystem :



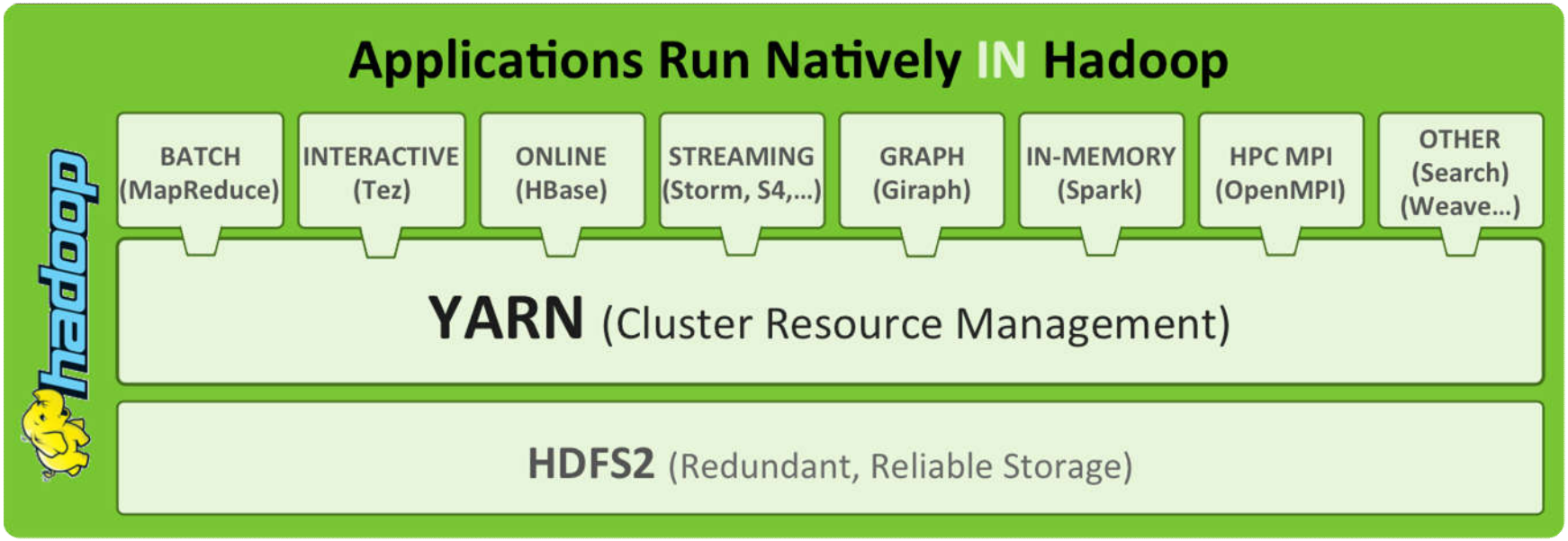
http://thebigdatablog.weebly.com/uploads/3/2/3/2/32326475/87375_orig.jpg

Hadoop Ecosystem :



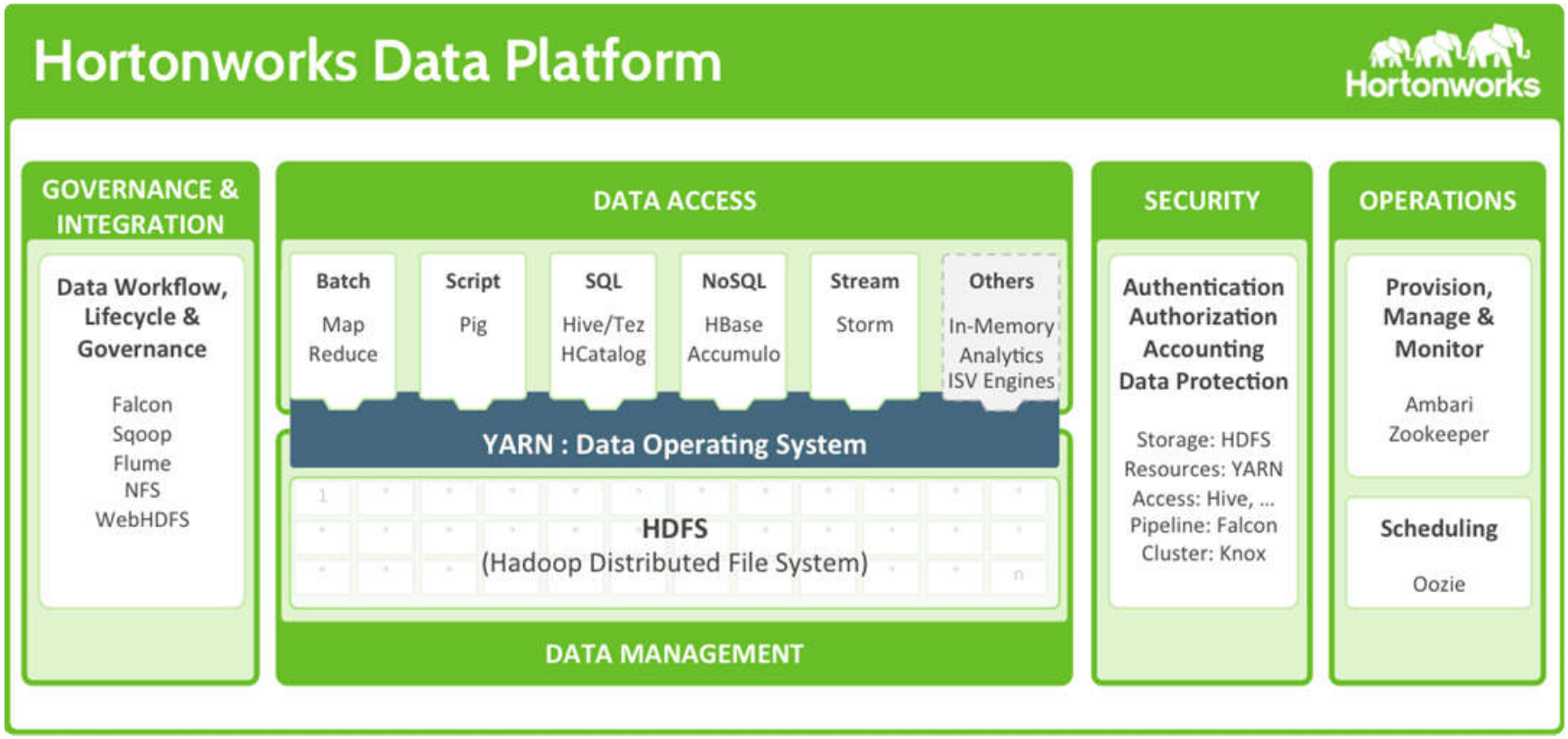
https://examples.javacodegeeks.com/wp-content/uploads/2016/08/Slide6.png

Multi-Purpose Hadoop Component Stack :



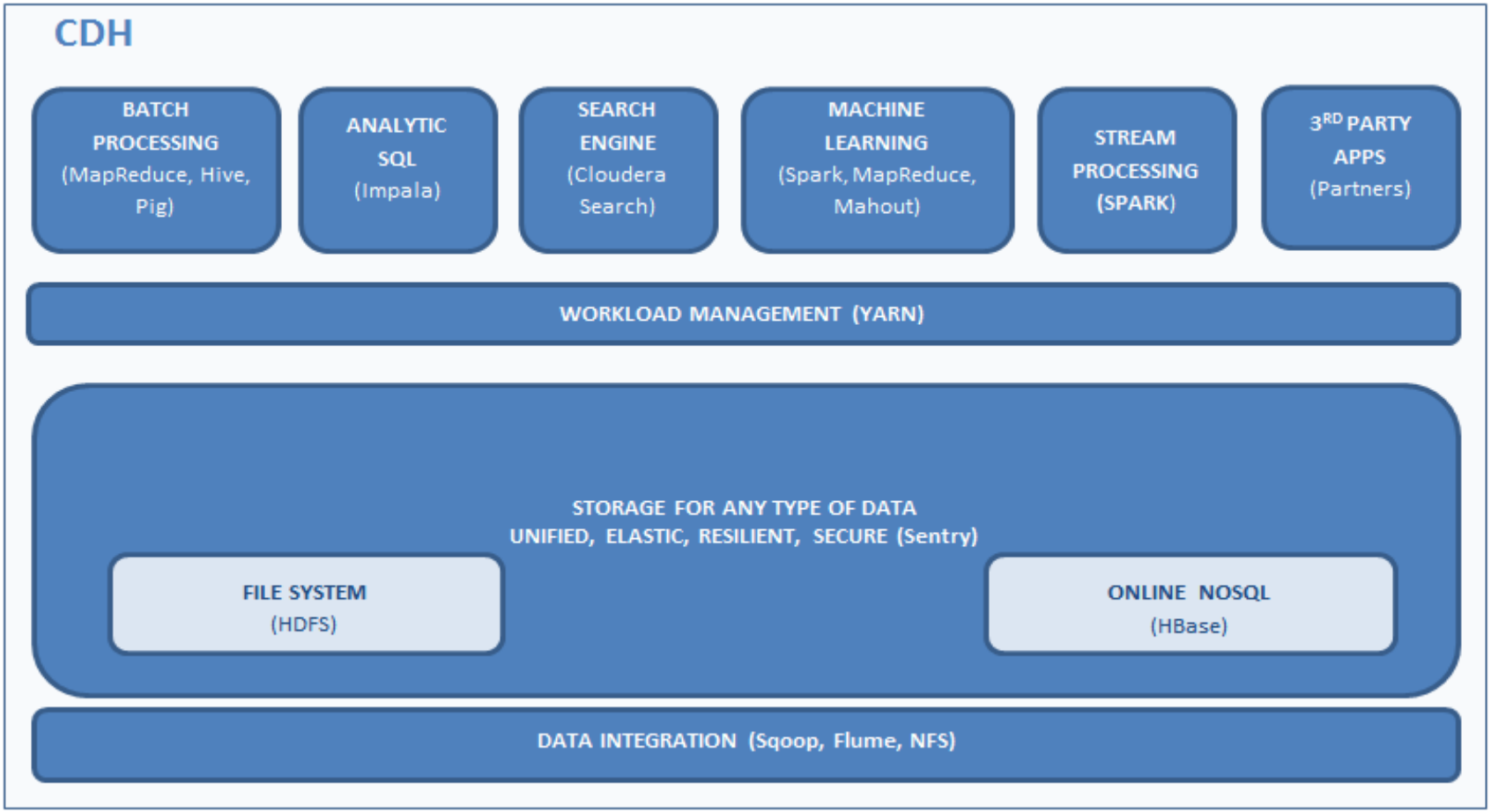
http://tm.durusau.net/wp-content/uploads/2013/06/YARN2.png

Enterprise Hadoop Component Stack :



Apache Hadoop :

Enterprise Hadoop Component Stack:



Hadoop Component & Terminology :

Hadoop Architecture :

Master/Slave Architecture:

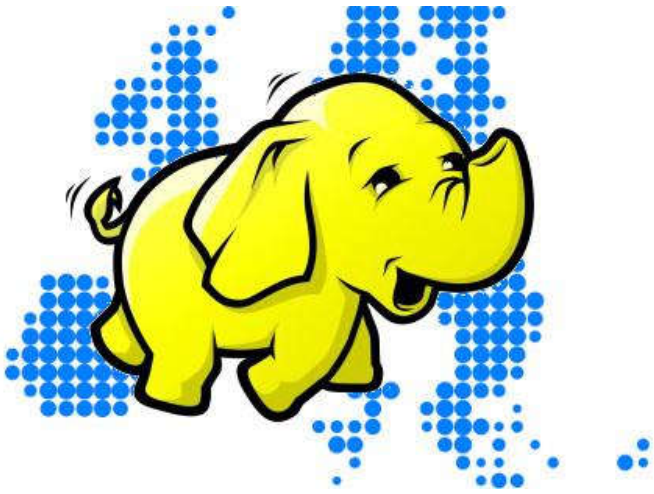
- Master / Secondary NameNode
- DataNodes

Data Management:

- HDFS
- Data Bock
- Replication Management
- Rack Awareness
- HDFS Read/Write – Behind the scenes

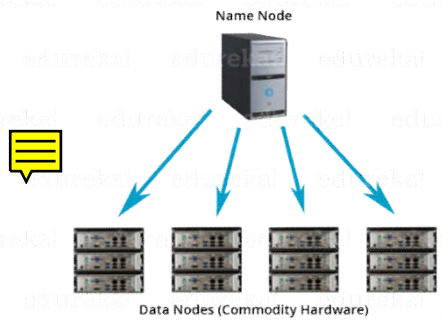
Process & Resource Management :

- YARN

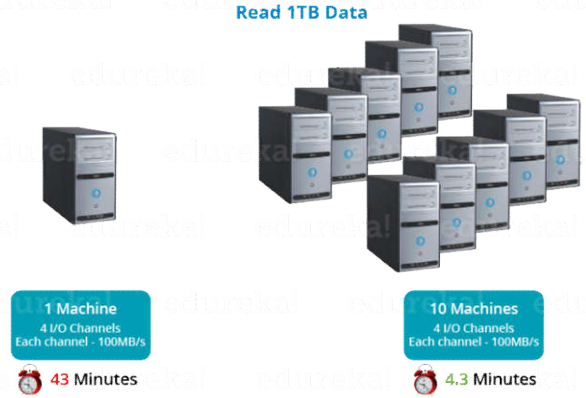


Hadoop Component & Terminology :

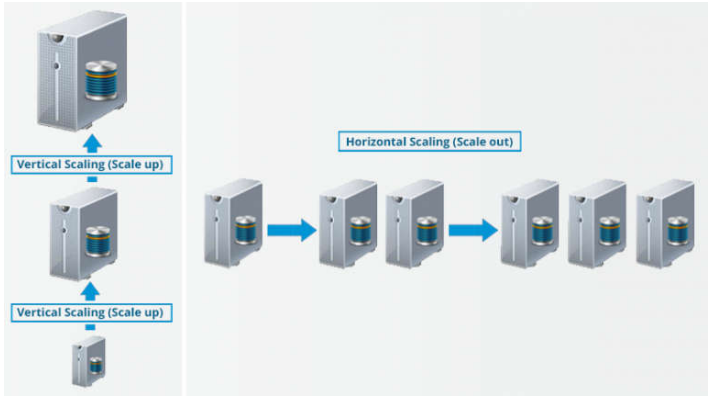
Hadoop is software framework with optimized for distributed processing of very large datasets. The main features are the Hadoop Distributed File System(HDFS) and MapReduce.



1. Distributed Storage:



2. Distributed & Parallel Computation:

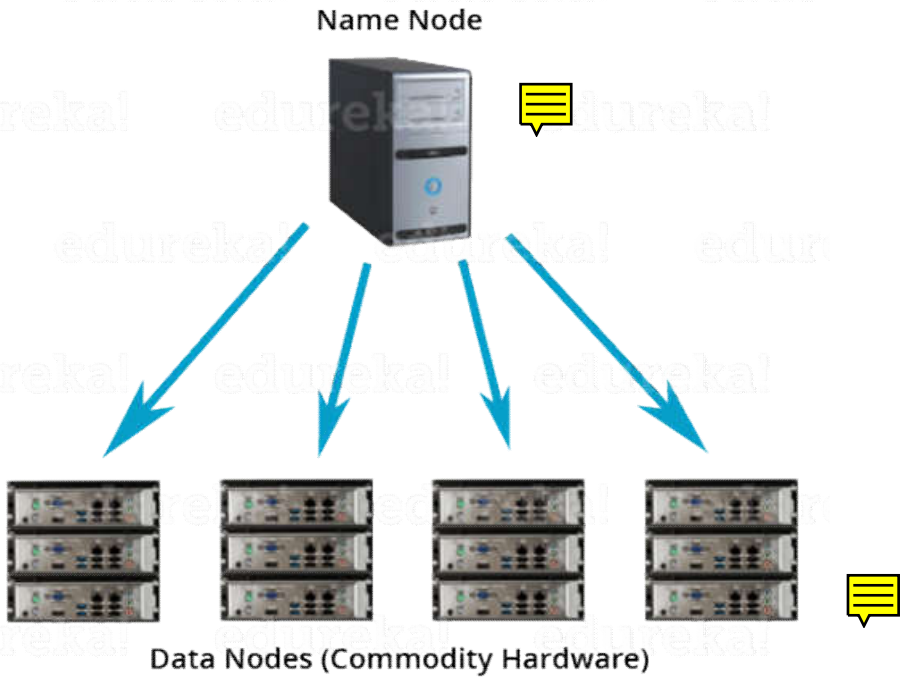


3. Horizontal Scalability:

Hadoop Component & Cluster :

Master/Slave Architecture:

- Master / Secondary NameNode
- DataNodes



NameNode (Master Node)

- Manage data block on DataNodes with Metadata , block location , file size , permission , hierarchy ,etc.
- Maintain DataNodes on cluster , ensure DNs are live

DataNode (Slave Node)

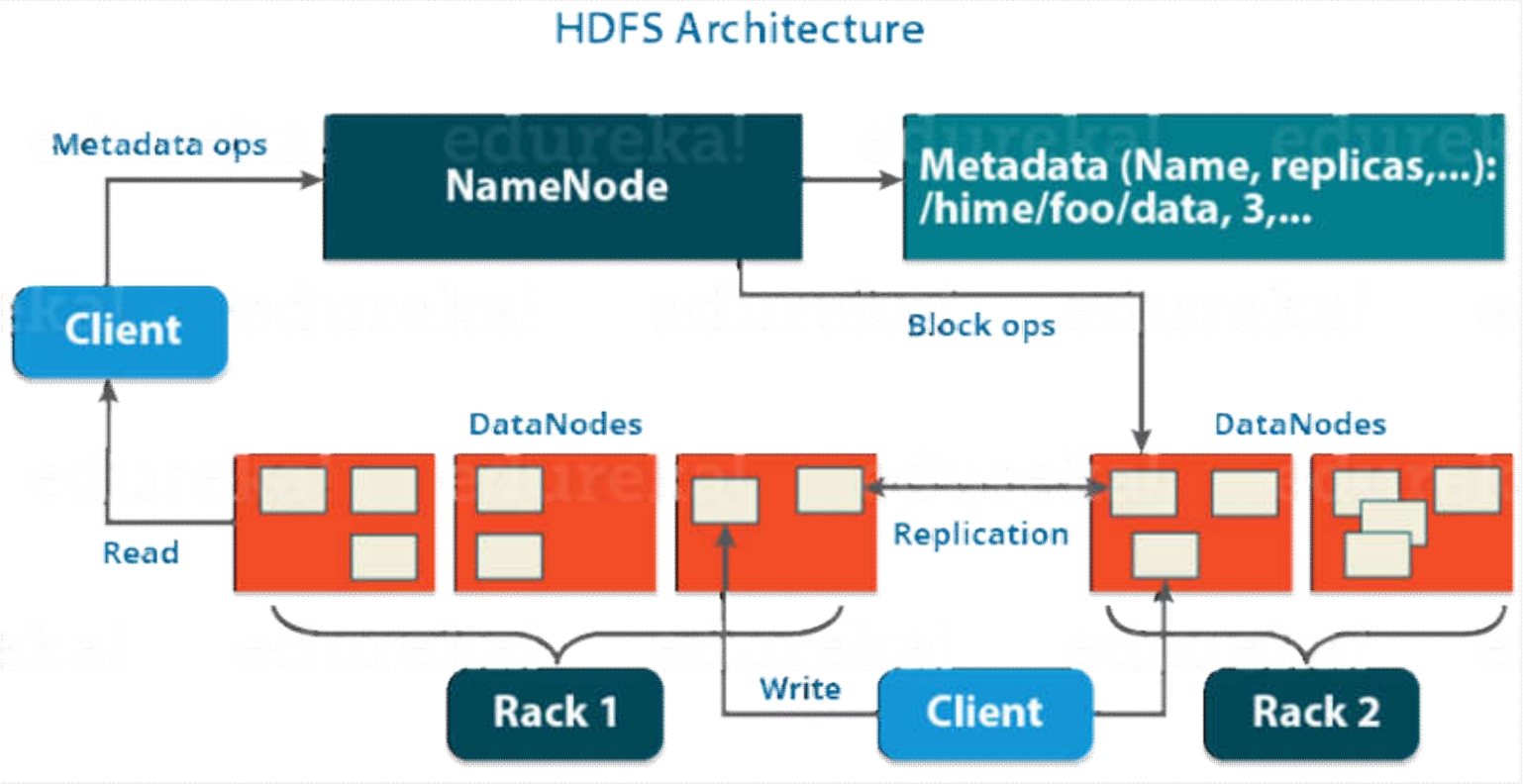
- Actual stored data
- Servicer read and write to Clients.
- Sent Heartbeats to NameNode

HDFS Architecture :

Data Management:

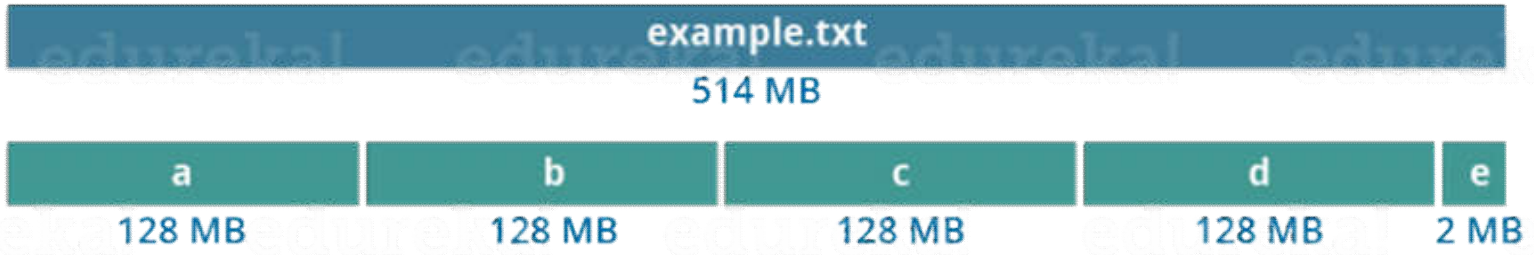
HDFS | Data Bock

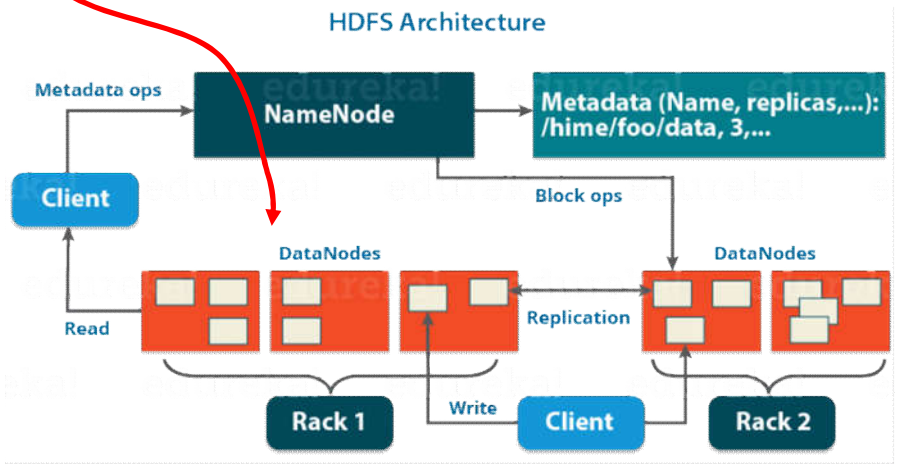
Replication | Rack Awareness



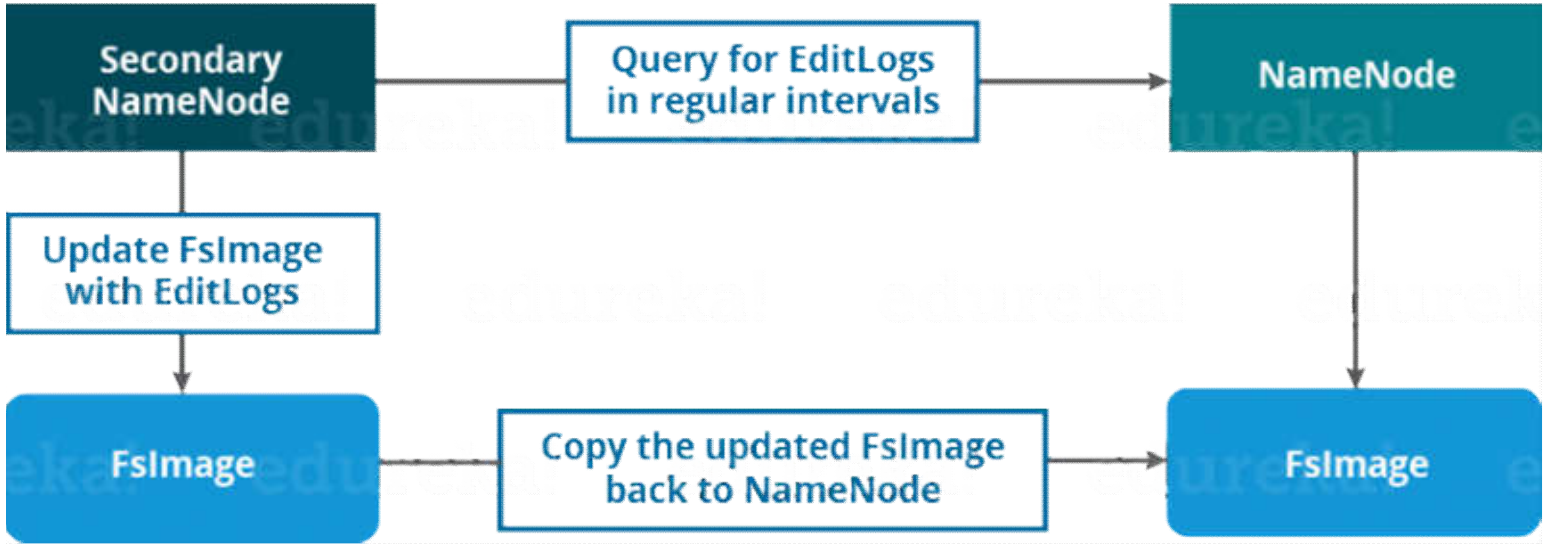
Data Block :

64MB , 128MB / Block





Secondary NaeNode :

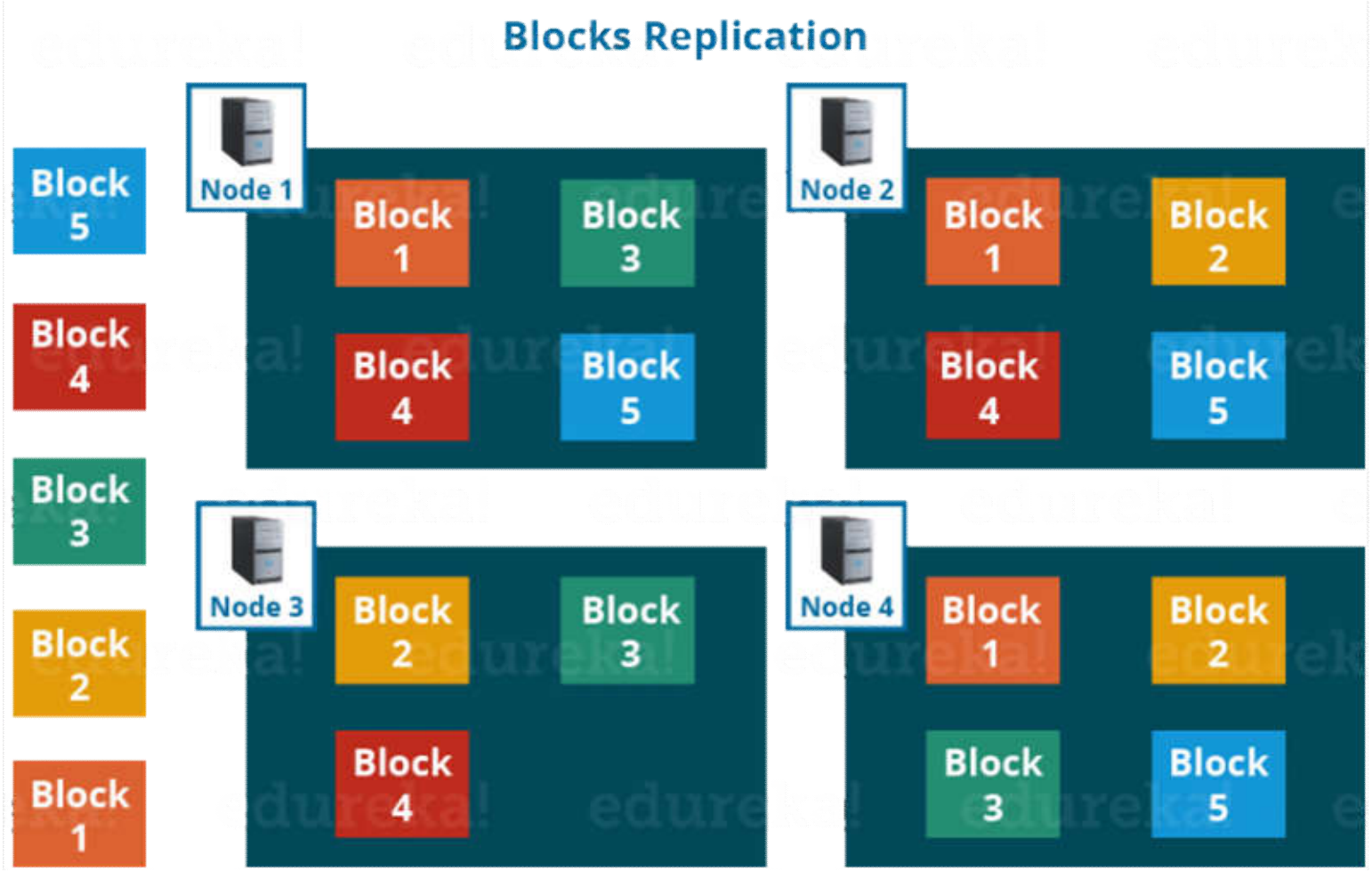


don’t be confused about the Secondary NameNode being a **backup NameNode because it is not.**

Automatic backup Metadata of Master NameNode , Metadata , FsImage , EditLogs

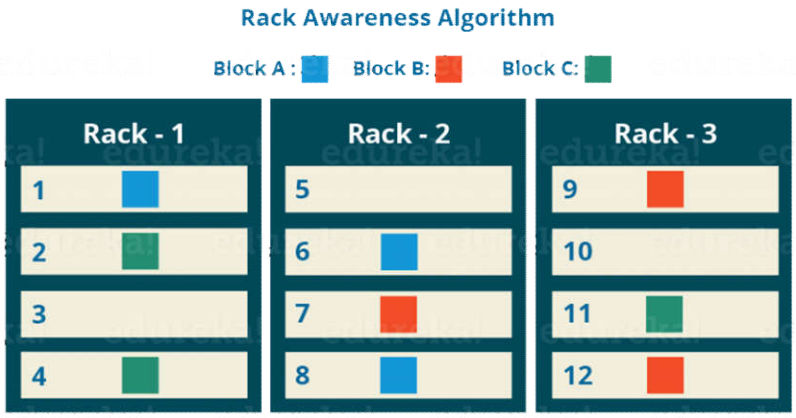
Secondary NameNode act as Checkpoint Node

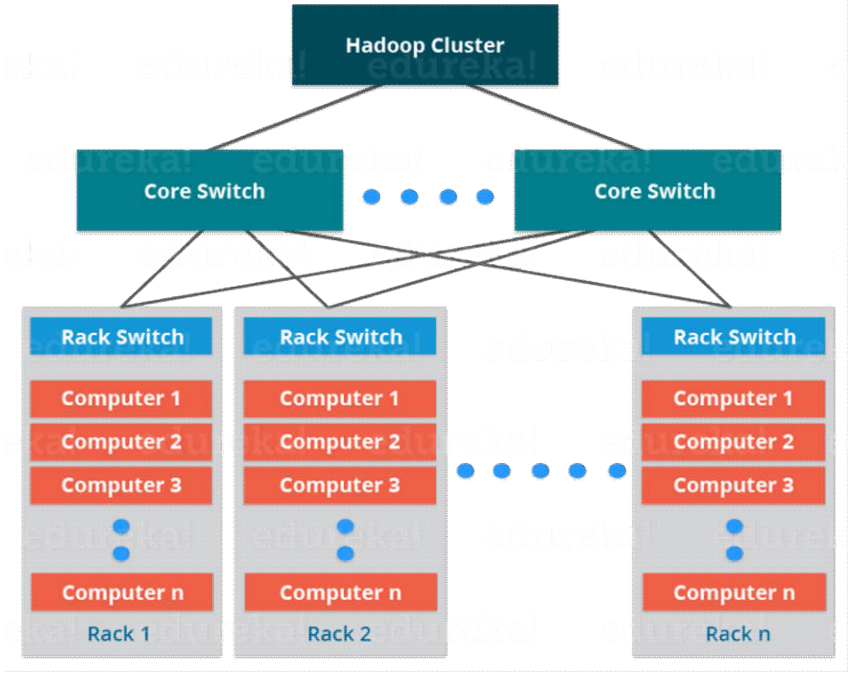
Replication Management :



Hadoop Topology :

Rack Awareness :



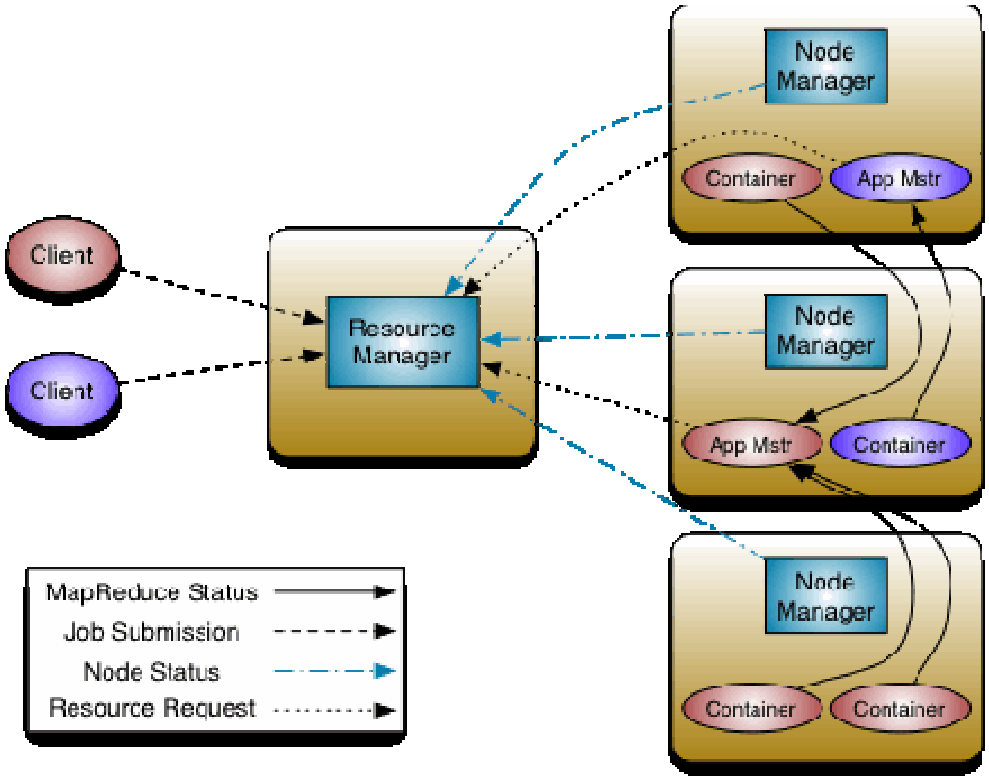


Balance the network traffic for reduce latency

YARN Architecture :

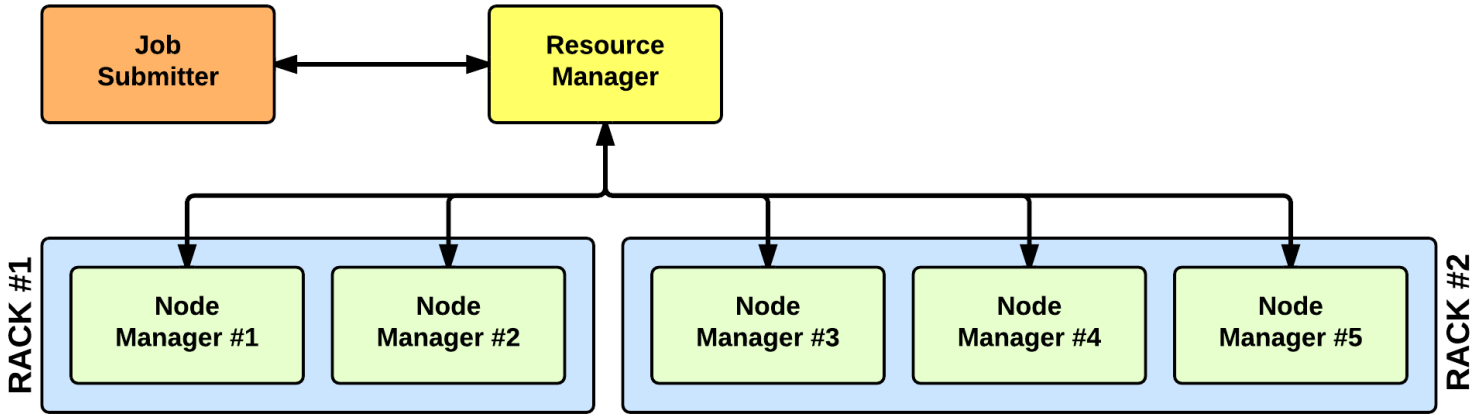
Process & Resource Management :

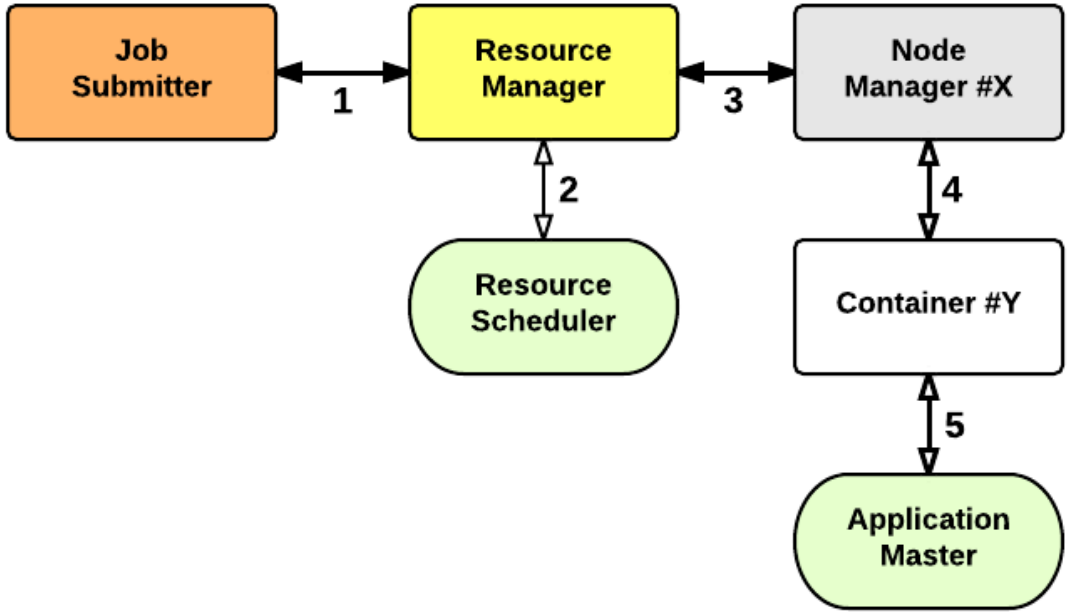
YARN



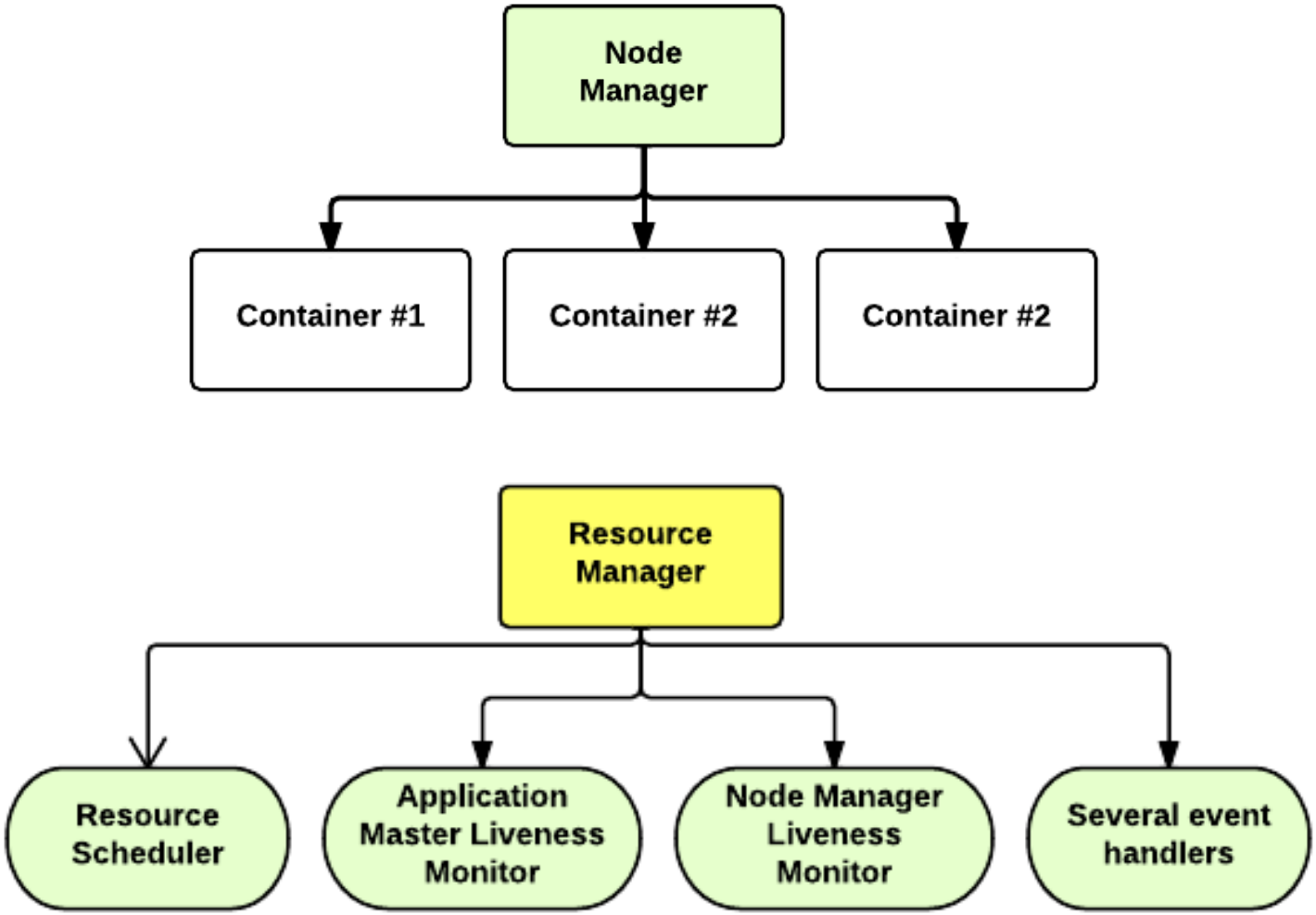
https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/YARN.html

YARN Architecture :



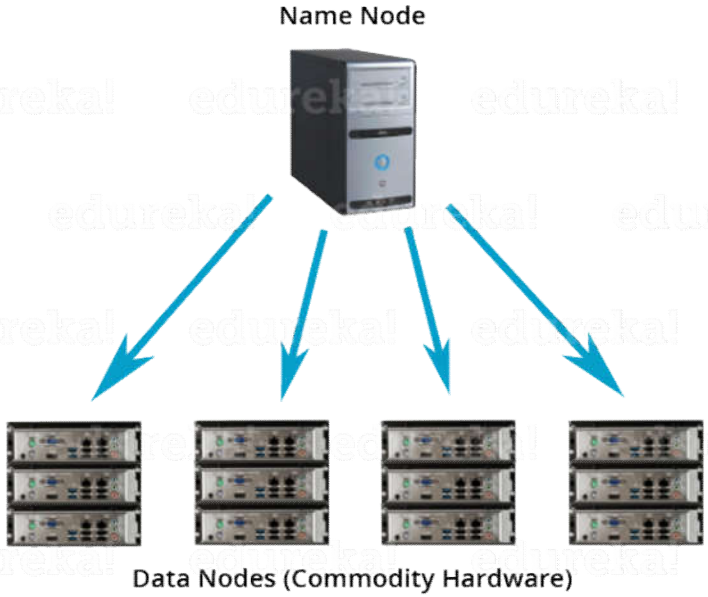


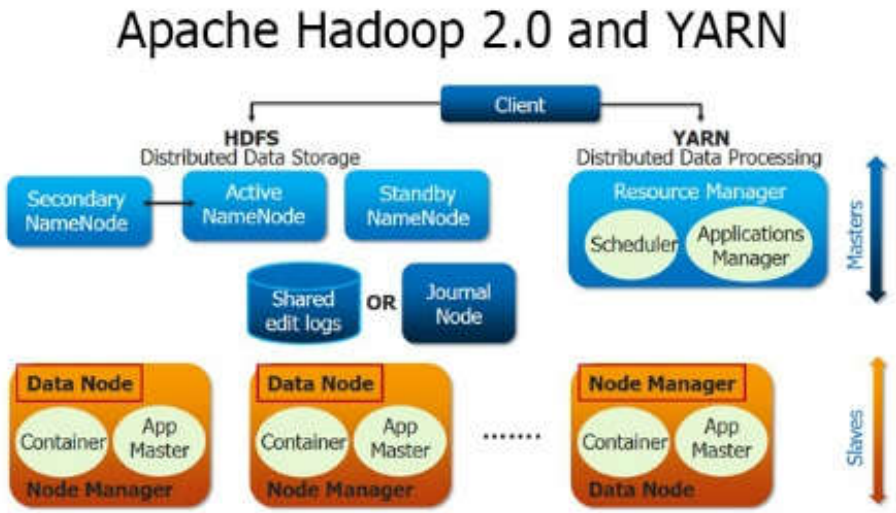
YARN Architecture :



https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/YARN.html

HDFS & YARN Daemon Process :





MasterNode:

NameNode , ResourceManager

SlaveNode:

DataNode , NodeManager


The End
Big
data
Shift