BIG DATA

Sirojul Munir | rojulman@nurulfikri.ac.id | @rojulman

Hadoop

Sirojul Munir | rojulman@nurulfikri.ac.id | @rojulman

Apa itu Hadoop ?

- Open-Source Framework untuk memproses himpunan-data berskala besar (big data) dalam beberapa cluster hardware komputer
- Dikembangkan menggunakan bahasa Java, beberapa menggunakan C dan utilitas command line sebagai shell-scripts
- Dikembangkan oleh Apache Software foundation (apache.org) 2007 dibawah lisensi v2 Apache
- Versi Hadoop terakhir: http://hadoop.apache.org/releases.html
  - 2.6.1 : 23 Sept 2015
  - 2.7.0 : 06 July 2015
  - 2.7.3: 25 Aug 2016
  - 2.8.3: 12 Des 2017
  - 3.0.0: 13 Des 2017

Apa itu Hadoop ?

- Hadoop terinspirasi dari publikasi makalah Google MapReduce dan Google File System (GFS) oleh ilmuwan dari Google, Jeffrey Dean dan Sanjay Ghemawat pada tahun 2004.
- Hadoop diciptakan oleh Doug Cutting dan Mike Cafarella pada tahun 2005. Cutting, pada saat itu bekerja di perusahaan Yahoo!,
- Kata “Hadoop” sendiri adalah nama mainan gajah berwarna kuning milik anaknya.
- 2006 Yahoo memberikan project Hadoop ke Apache Software Foundation

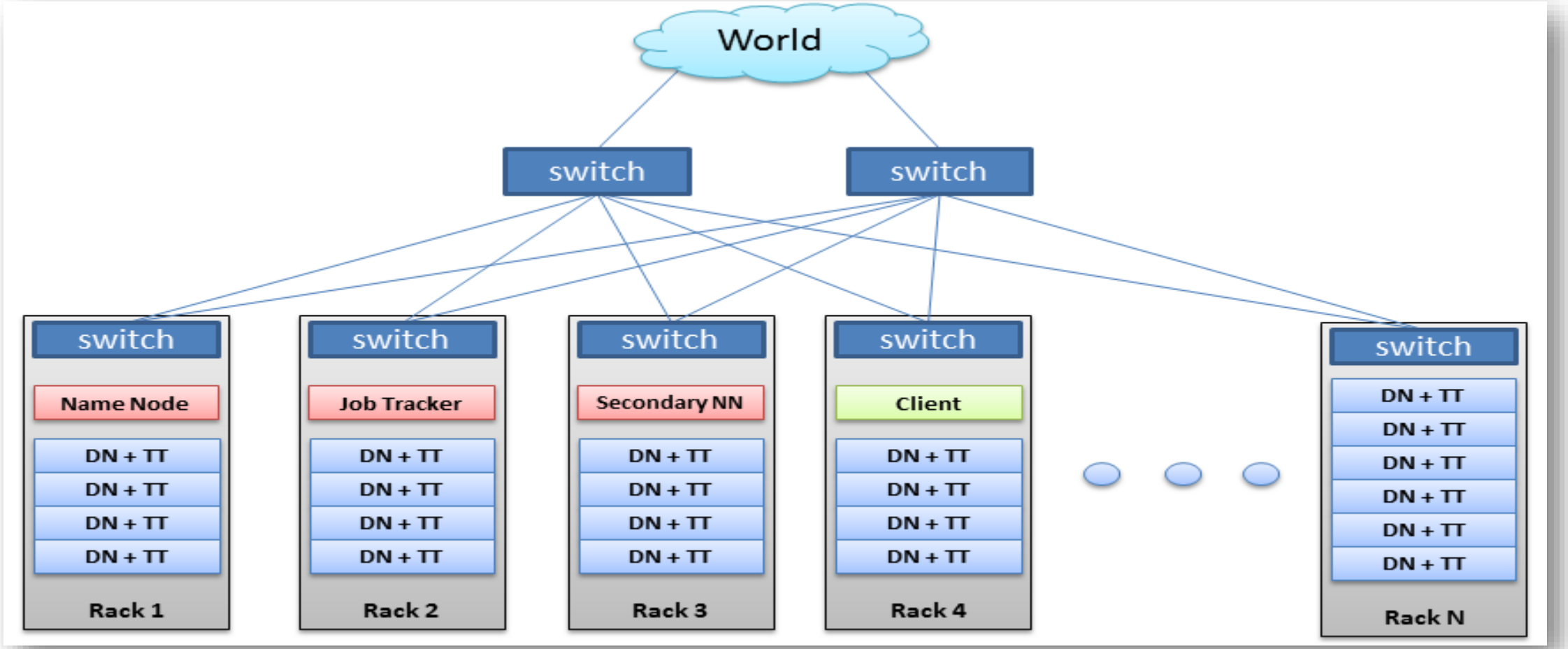
Google - Hadoop

Google calls it:	Hadoop equivalent:
MapReduce	Hadoop
GFS	HDFS
Bigtable	HBase
Chubby	Zookeeper

Arsitektur Hadoop

- Distributed, with some centralization
- Main nodes of cluster are where most of the computational power and storage of the system lies
- Main nodes run TaskTracker to accept and reply to MapReduce tasks, and also DataNode to store needed blocks closely as possible
- Central control node runs NameNode to keep track of HDFS directories & files, and JobTracker to dispatch compute tasks to TaskTracker
- Written in Java, also supports Python and Ruby

Arsitektur Hadoop



Arsitektur Hadoop

- Hadoop Distributed Filesystem
- Tailored to needs of MapReduce
- Targeted towards many reads of filestreams
- Writes are more costly
- High degree of data replication (3x by default)
- No need for RAID on normal nodes
- Large blocksize (64MB)
- Location awareness of DataNodes in network

Arsitektur Hadoop : NameNode

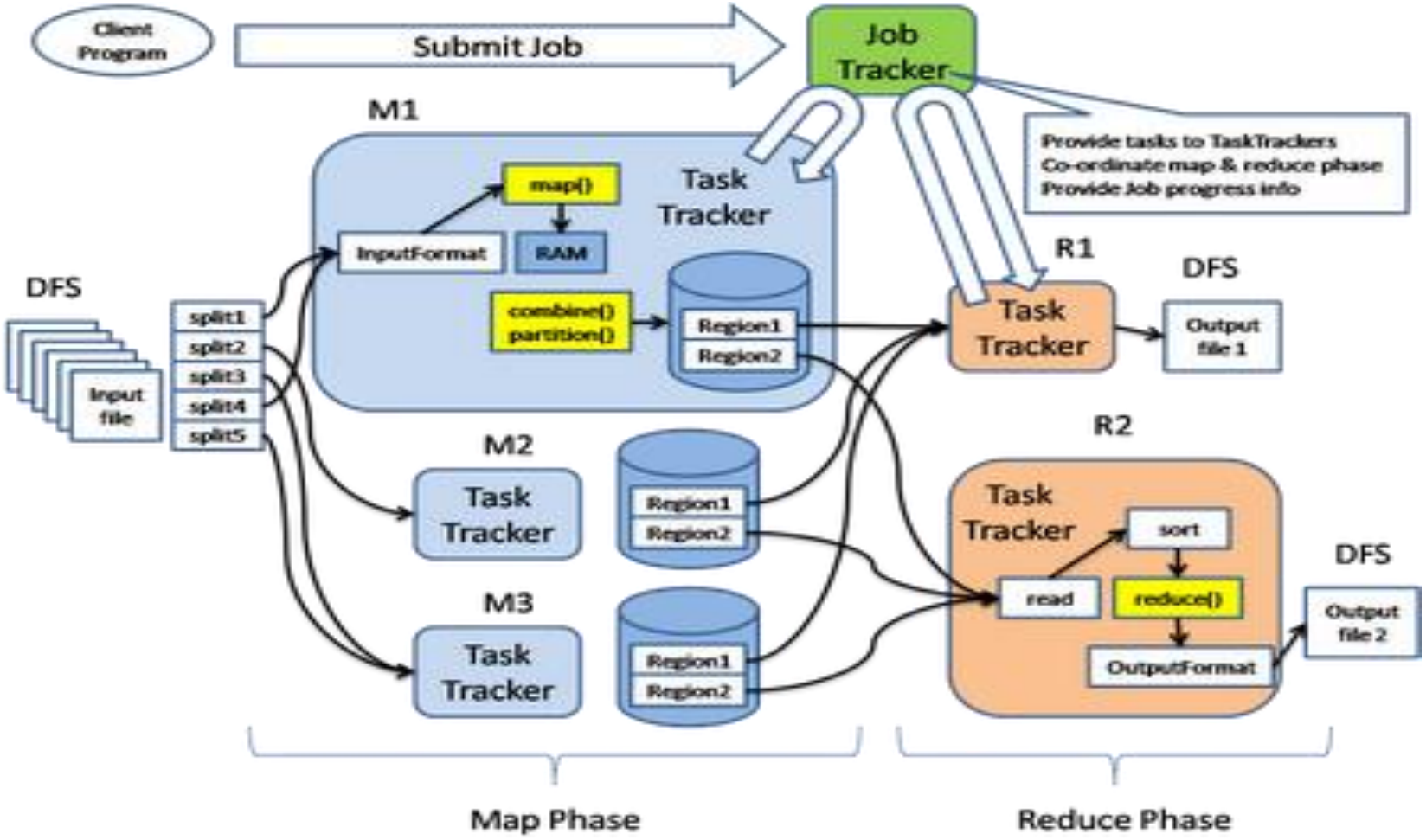
- Stores metadata for the files, like the directory structure of a typical FS.
- The server holding the NameNode instance is quite crucial, as there is only one.
- Transaction log for file deletes/adds, etc. Does not use transactions for whole blocks or file-streams, only metadata.
- Handles creation of more replica blocks when necessary after a DataNode failure

Arsitektur Hadoop : DataNode

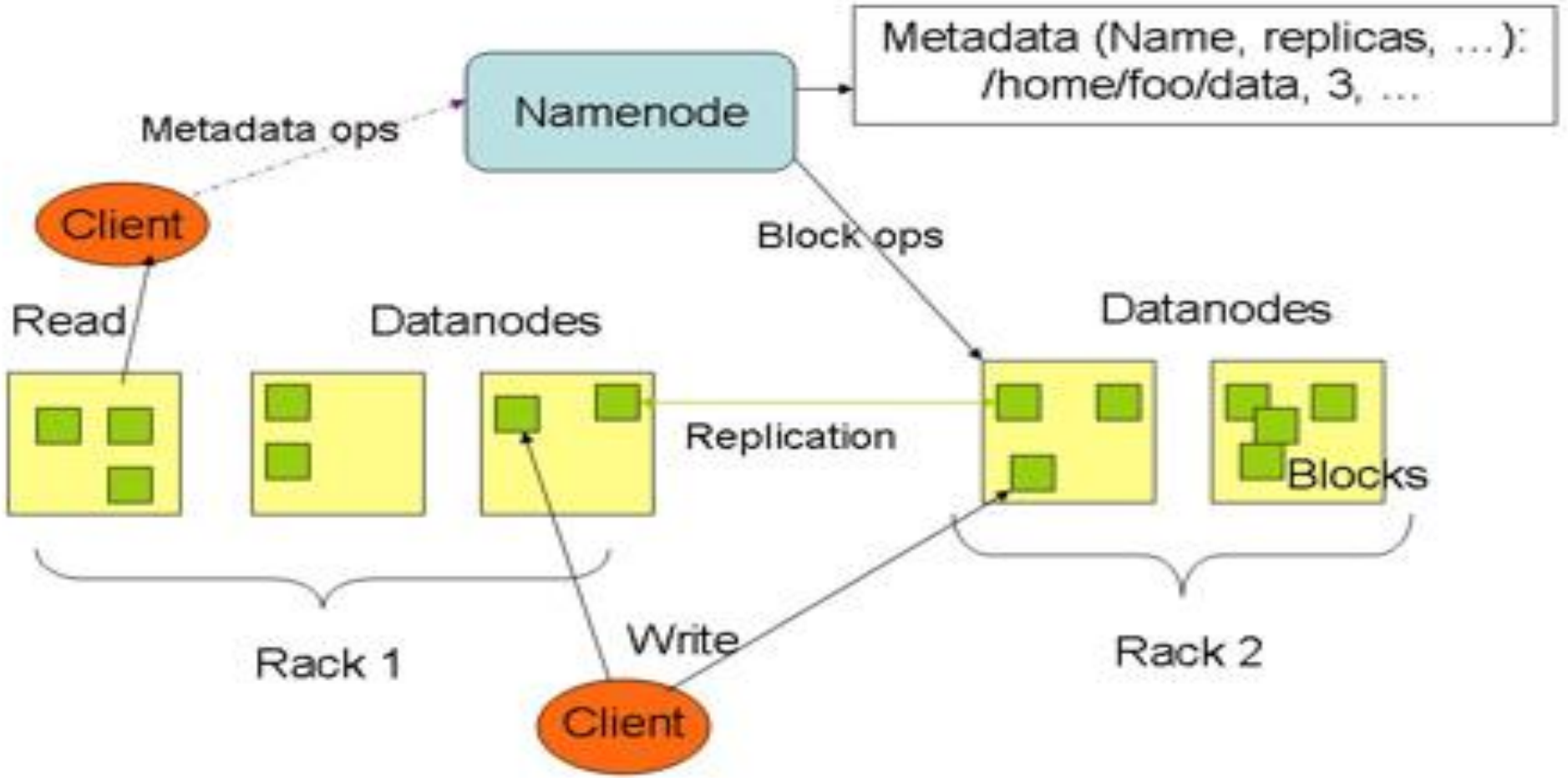
- Stores the actual data in HDFS
- Can run on any underlying filesystem (ext3/4, NTFS, etc)
- Notifies NameNode of what blocks it has
- NameNode replicates blocks 2x in local rack, 1x elsewhere
- more replica blocks when necessary after a DataNode failure

Arsitektur Hadoop :: Engine MapReduce

- JobTracker & TaskTracker
- JobTracker splits up data into smaller tasks(“Map”) and sends it to the TaskTracker process in each node
- TaskTracker reports back to the JobTracker node and reports on job progress, sends data (“Reduce”) or requests new jobs



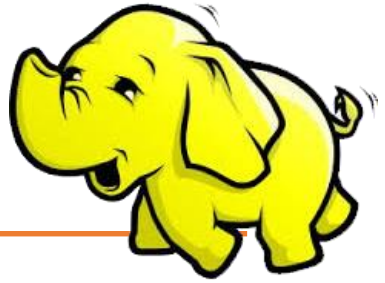
Hadoop : Namenode - Datanodes



Implementasi Aplikasi Hadoop

- Advertisement (Mining user behavior to generate recommendations)
- Searches (group related documents)
- Security (search for uncommon patterns)

Pengguna Hadoop

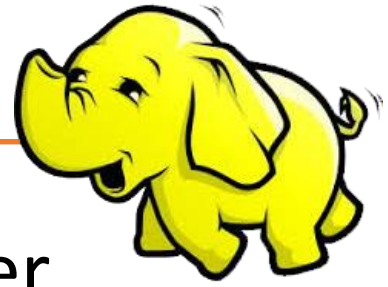


- Yahoo
- Facebook
- NY Times

  - Yahoo!’s Search Webmap runs on 10,000 core Linux cluster and powers Yahoo! Web search
  - FB’s Hadoop cluster hosts 100+ PB of data (July, 2012) & growing at ½ PB/day (Nov, 2012)
  - NY Times
    - was dynamically generating PDFs of articles from 1851-1922
    - Wanted to pre-generate & statically serve articles to improve performance
    - Using Hadoop + MapReduce running on EC2 / S3, converted 4 of TIFFs into 11 million PDF articles in 24 hrs

Big Data :: Hadoop -- Yahoo Server

Pengguna Hadoop

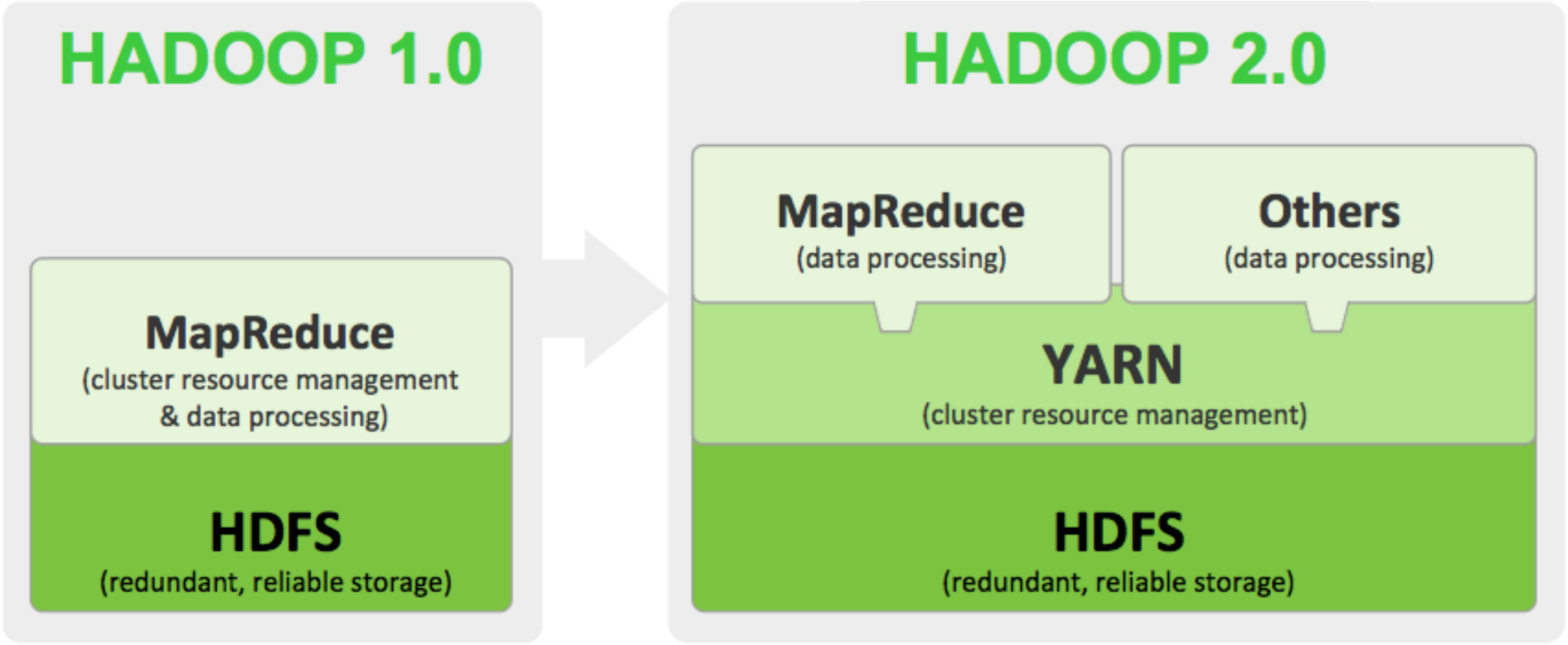


- Amazon/A9
- Facebook
- Yahoo
- Netflix
- IBM
- Joost
- Last.fm
- New York Times
- PowerSet
- Veoh

- Hadoop tested on 4,000 node cluster
  - 32K cores (8 / node)
  - 16 PB raw storage (4 x 1 TB disk / node)
    (about 5 PB usable storage)

- http://developer.yahoo.com/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html

Core Hadoop System



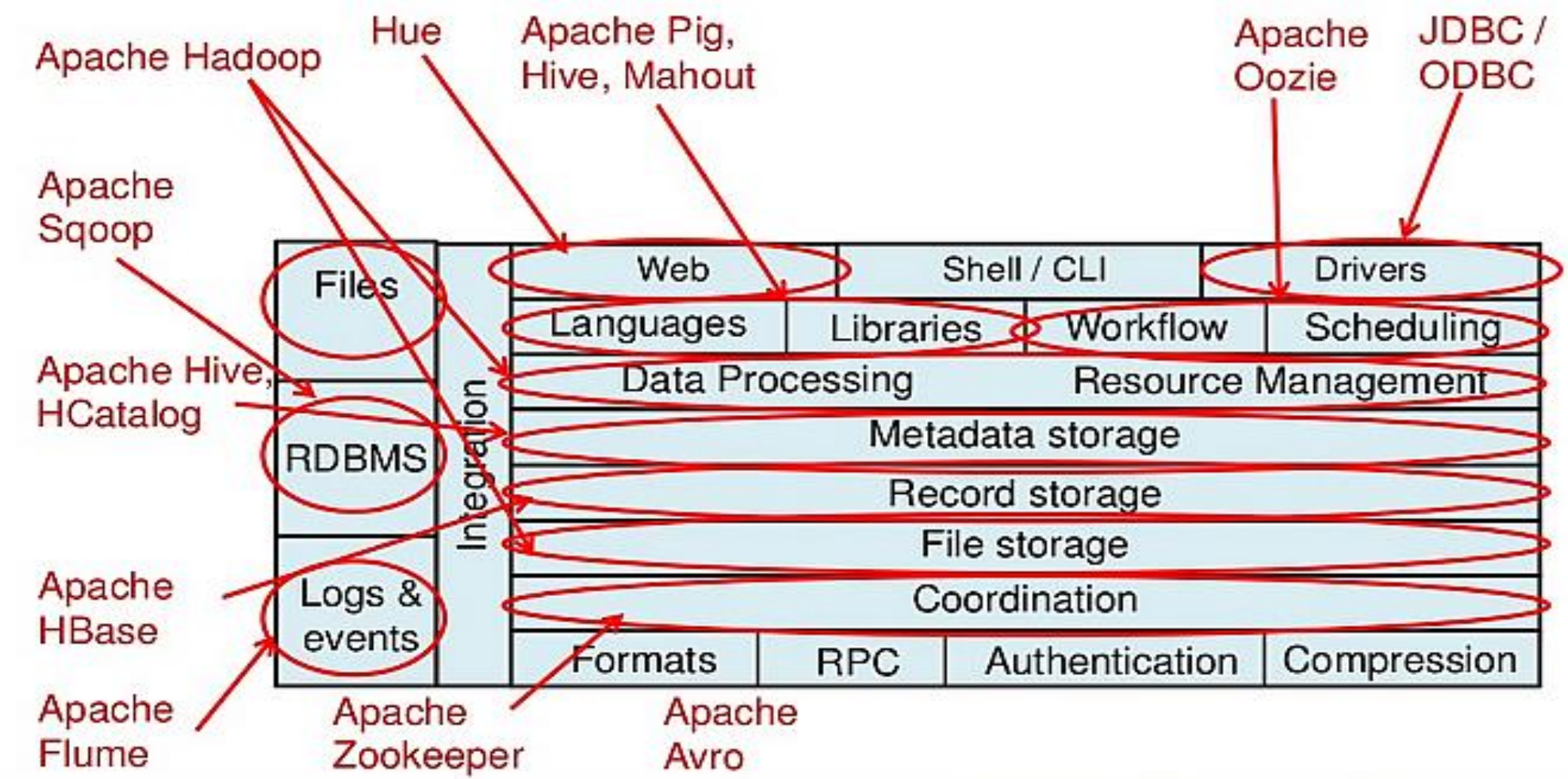
Core :: Modul Apache Hadoop

- Hadoop Common – berisi libraries dan utilities yang dibutuhkan oleh modul Hadoop lainnya.
- Hadoop Distributed File System (HDFS) – sebuah distributed file-system.
- Hadoop YARN – sebuah platform resource-management yang bertanggung jawab untuk mengelola resources dalam clusters dan scheduling.
- Hadoop MapReduce – sebuah model programming untuk pengelolaan data skala besar

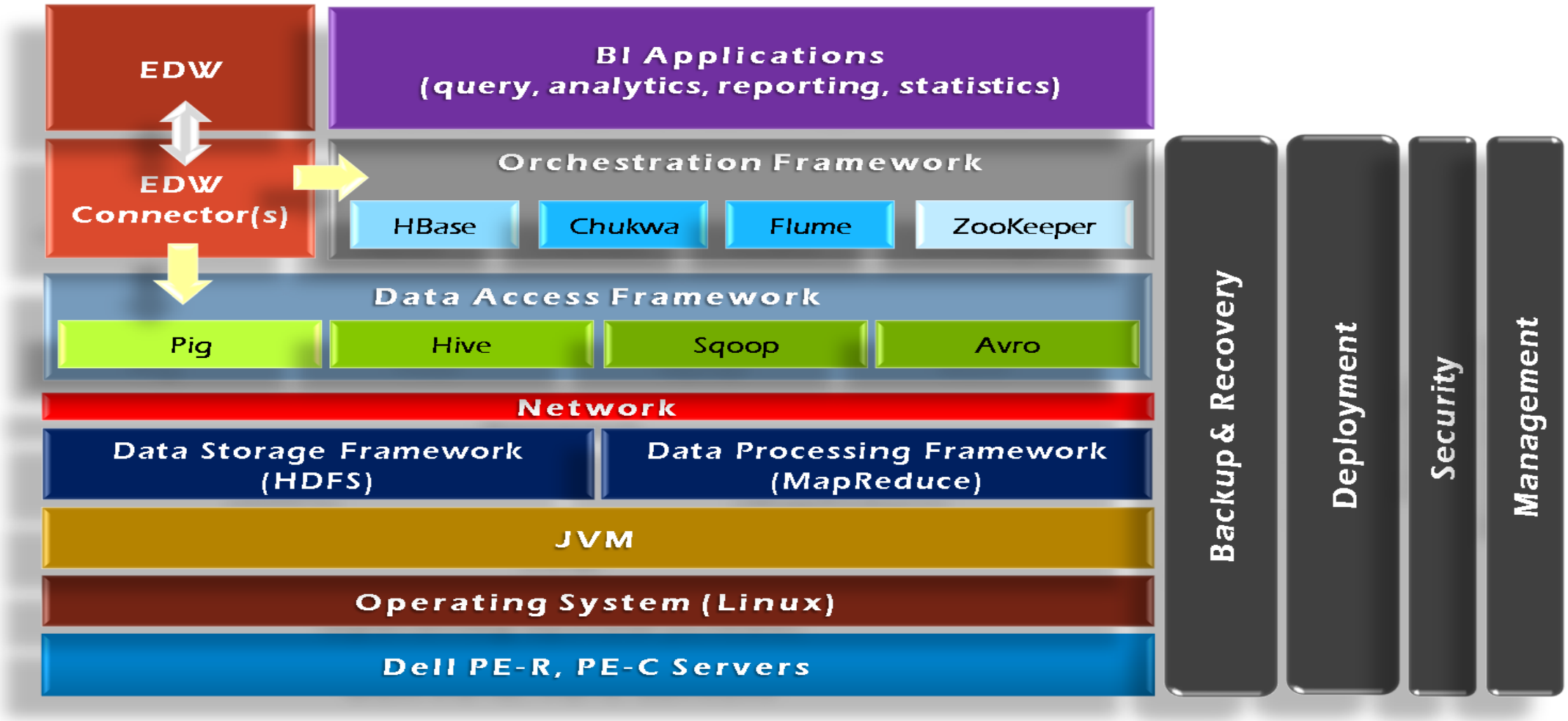
Extends :: Modul Apache Hadoop

- Ambari, Zookeeper (managing & monitoring)
- HBase, Cassandra (database)
- Hive, Pig (data warehouse and query language)
- Mahout (machine learning)
- Chukwa, Avro, Oozie, Giraph, and many more

Eco-System Hadoop



Hadoop Framework Tools

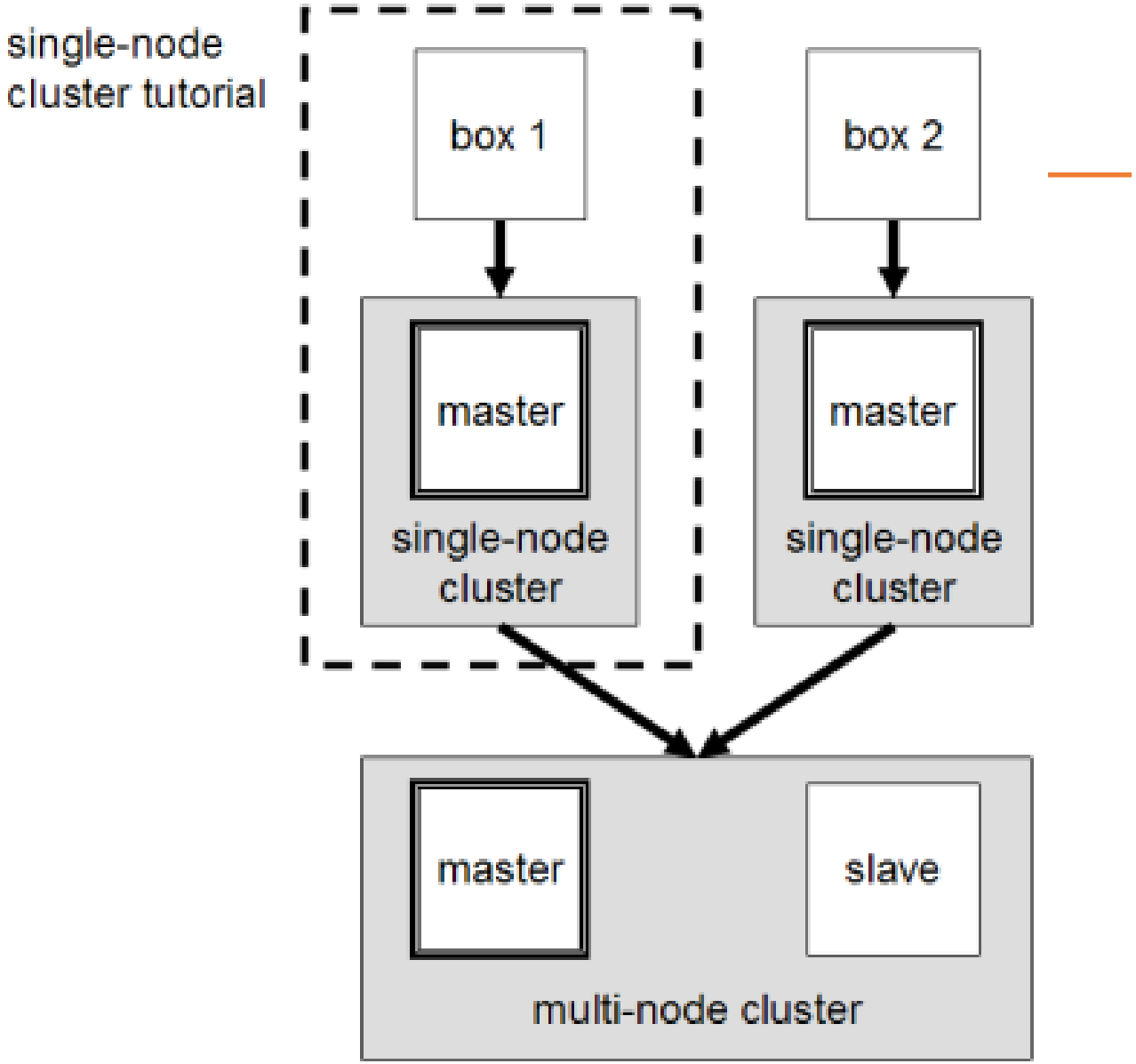


Fungsi – Manfaat :: Hadoop

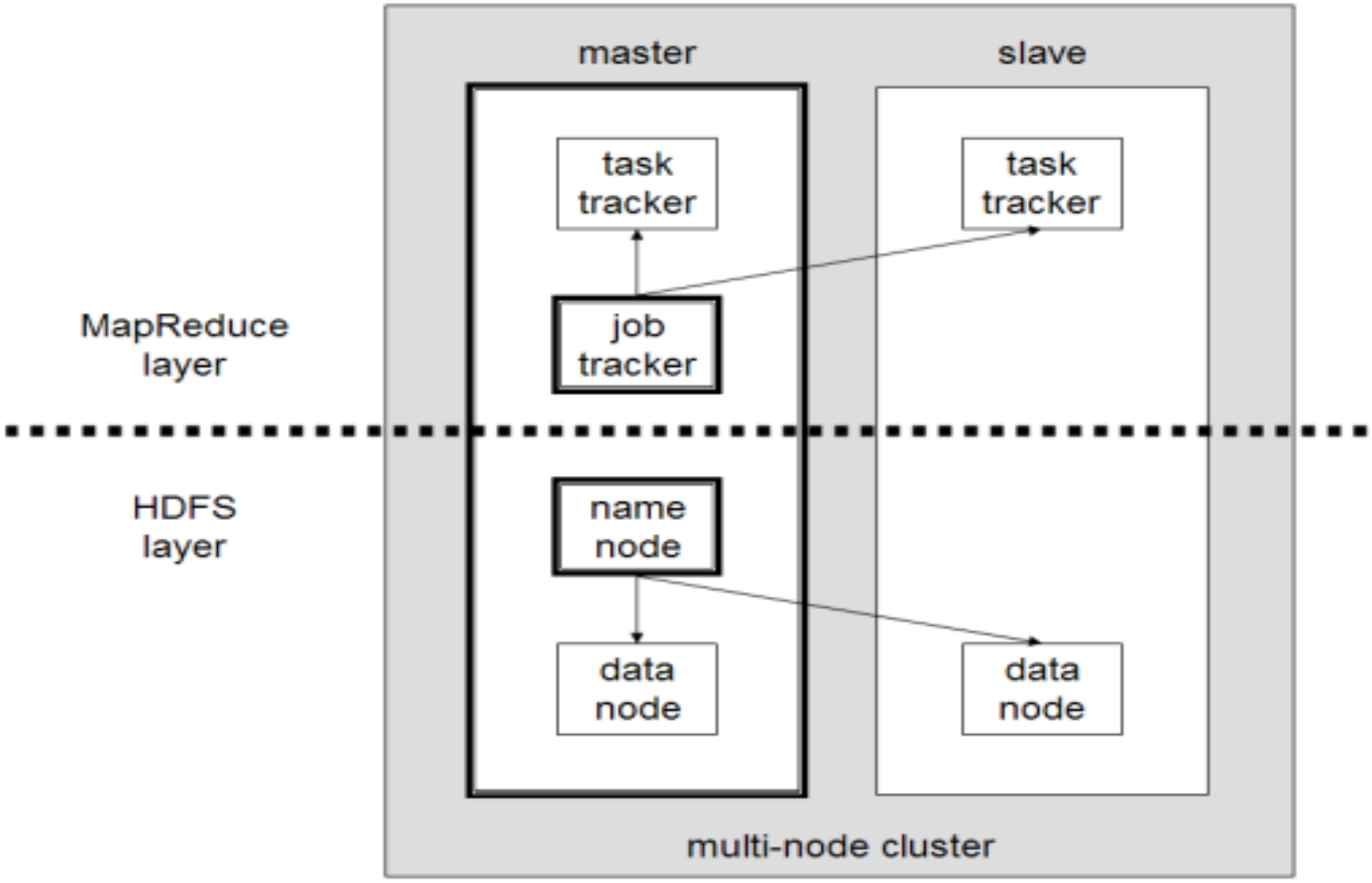
- Abstract and facilitate the storage and processing of large and/or rapidly growing data sets
    - Structured and non-structured data
    - Simple programming models
- High scalability and availability
- Use commodity (cheap!) hardware with little redundancy
- Fault-tolerance
- Move computation rather than data

Hadoop :: Node

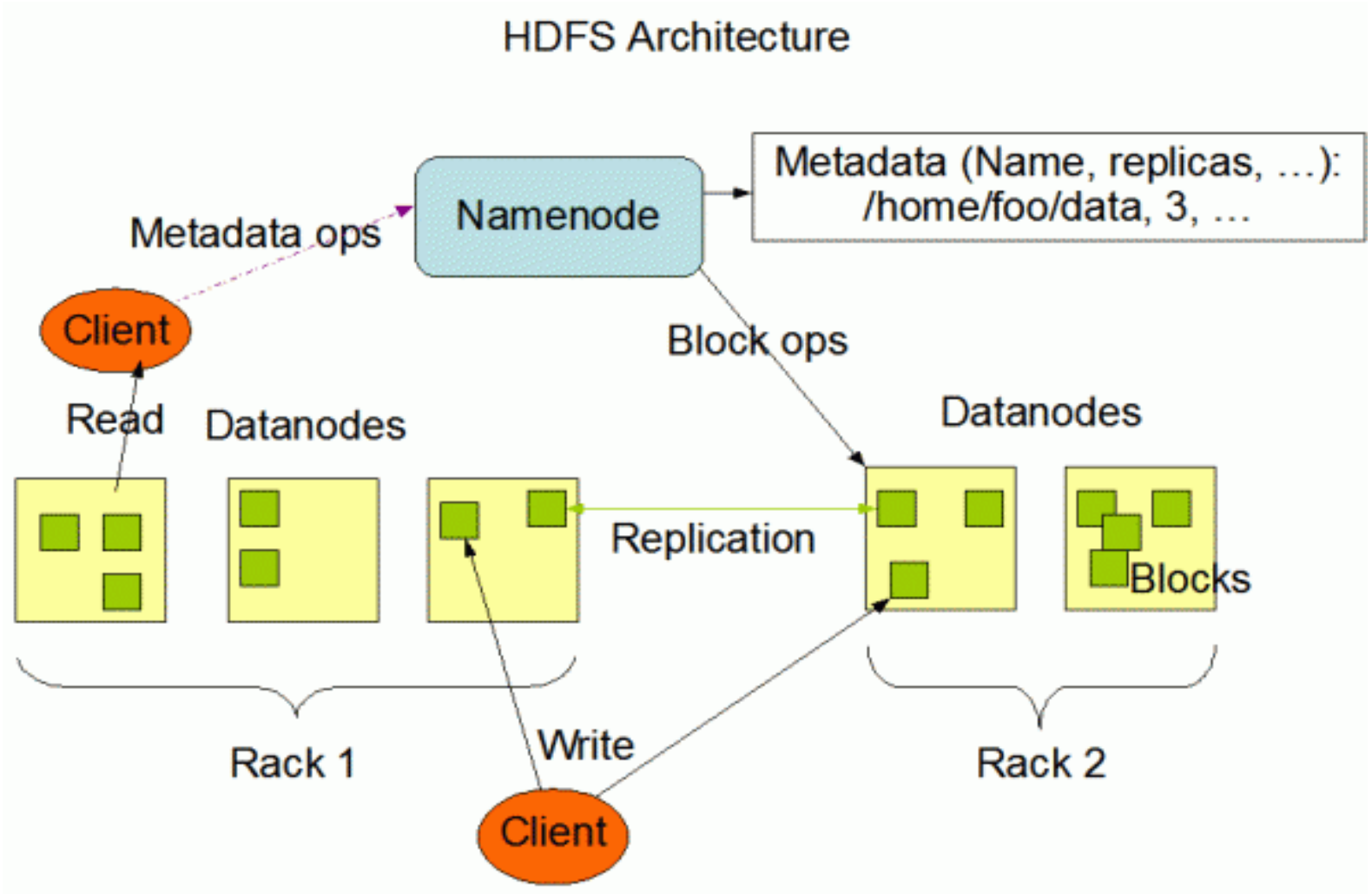
- Single Node
- Multiple Node



Multi-node Cluster



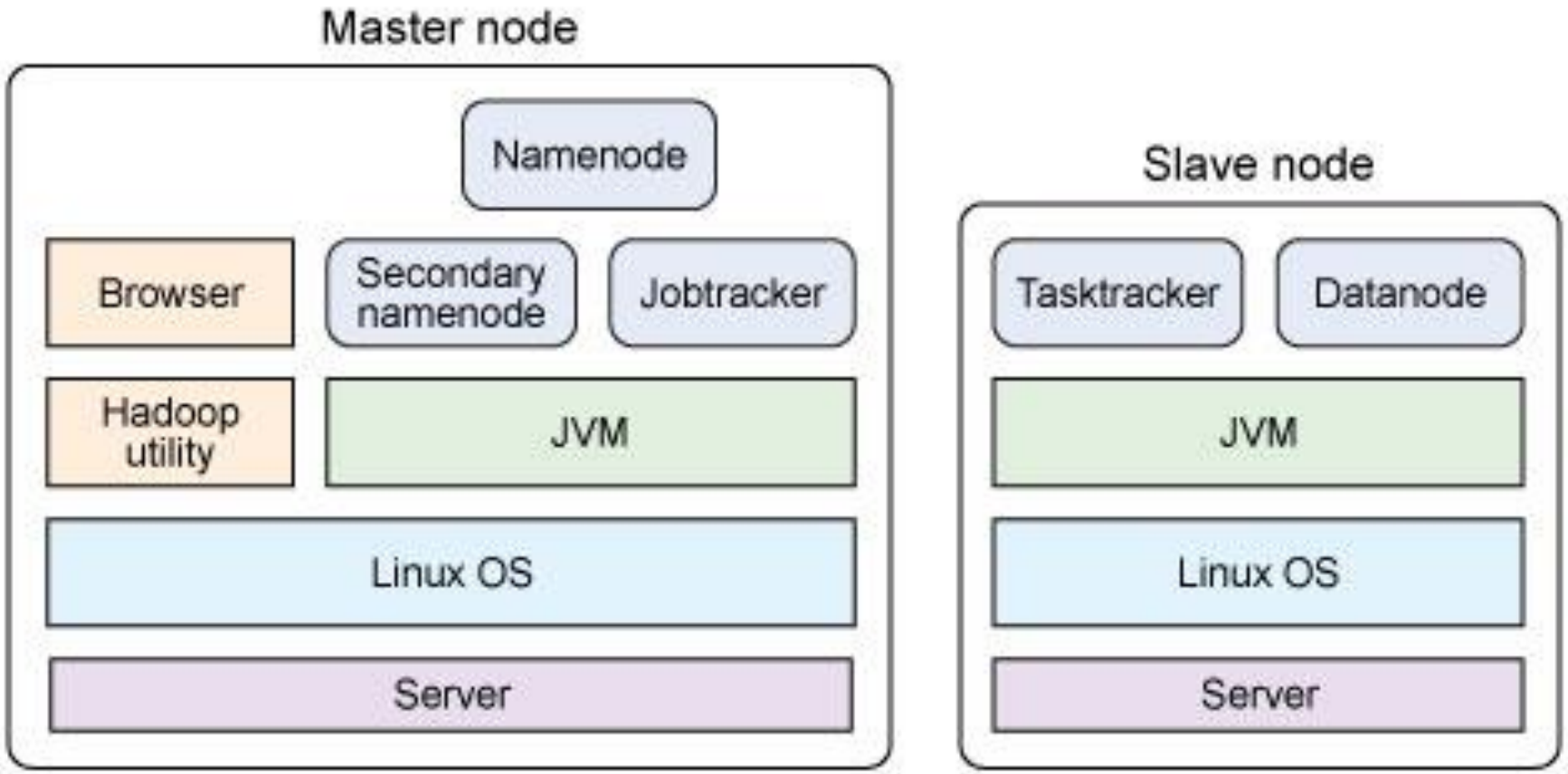
HDFS Architecture



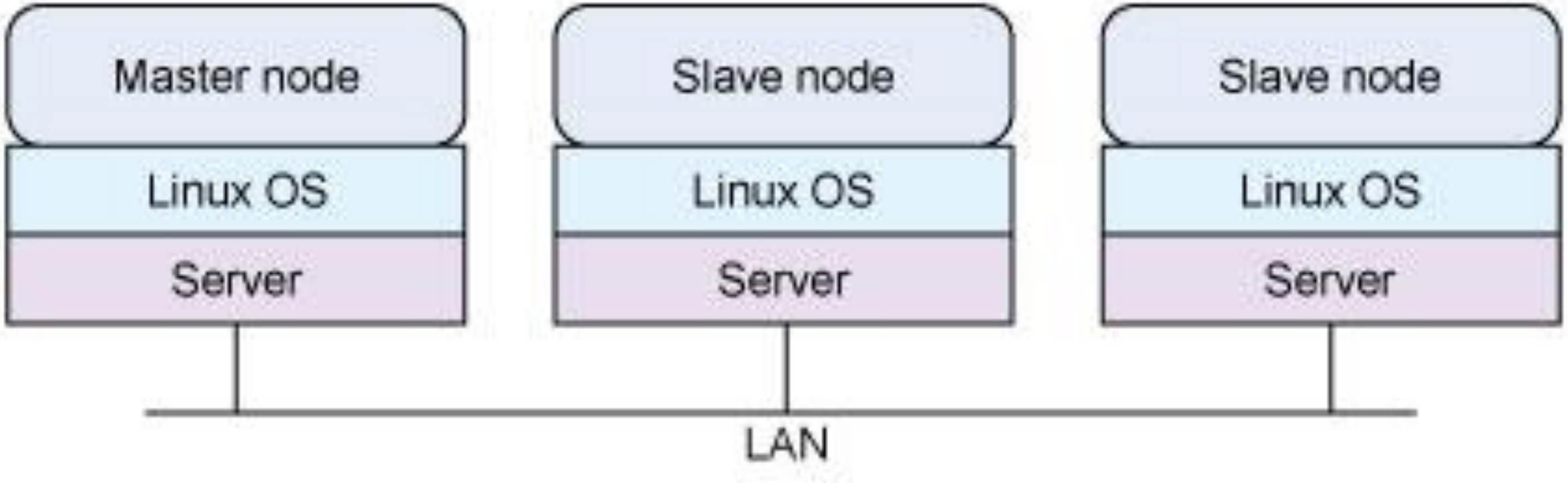
Pre-requisites Instalasi

- Sistem Operasi :
  - Ubuntu
  - Centos
- Java
  - Instalasi JDK , versi 1.6 keatas
  - Environment User / Superuser
- OpenSSL
  - SSH
  - RSYNC
- Hadoop
- Referensi instalasi:
  - https://hadoop.apache.org/docs/r1.2.1/single_node_setup.html
  - Single : http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
  - Multi : http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/

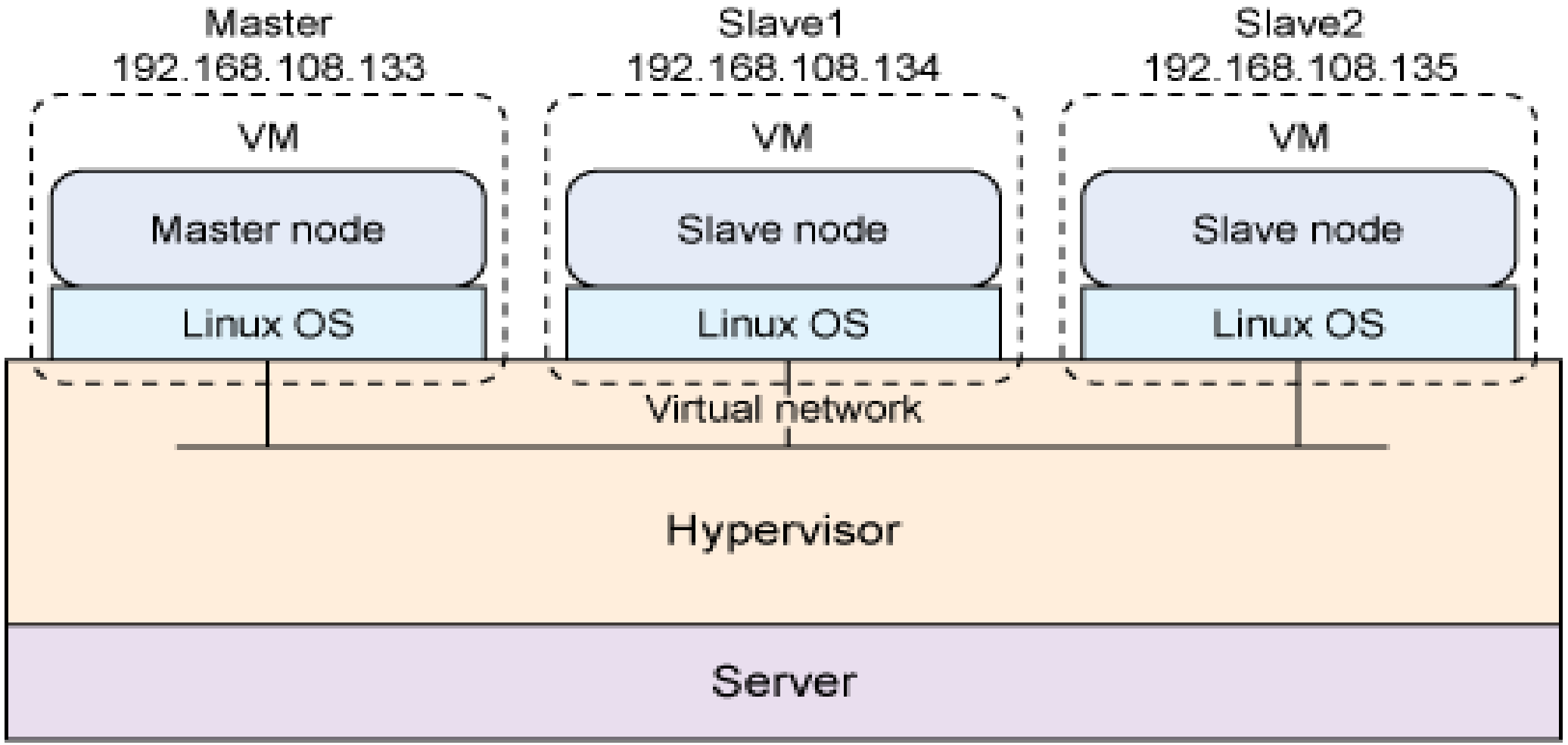
Master – Slave Server



LAN - Hadoop


Master node
Slave node
Slave node
Linux OS
Linux OS
Linux OS
Server
Server
Server
LAN

Master – Slave Virtual Network



Tugas:

- Instalasi Hadoop, berdasarkan tutorial dari website berikut ini:
  - https://www.digitalocean.com/community/tutorials/how-to-install-hadoop-in-stand-alone-mode-on-ubuntu-16-04
  - http://hadoop.apache.org/docs/r2.7.3/hadoop-project-dist/hadoop-common/SingleCluster.html
  - http://doctuts.readthedocs.io/en/latest/hadoop.html