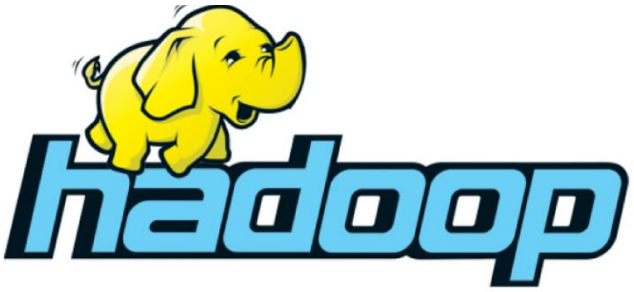
HADOOP

Mochammad Zen Samsono Hadi, ST. MSc. Ph.D

Outline

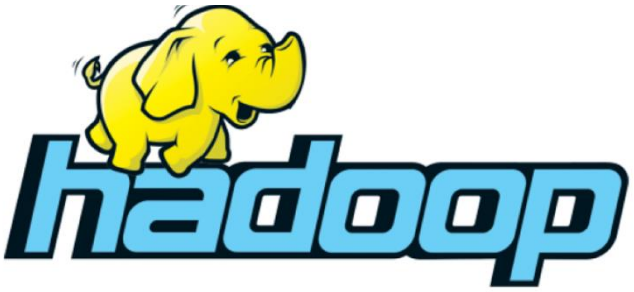
- Sejarah Hadoop
- Pengertian Hadoop
- Arsitektur Hadoop
- Cara Kerja Hadoop
- Vendor Hadoop
- Manfaat dan Penggunaan Hadoop
- Kelebihan dan Kekurangan Hadoop
- Implementasi dari Penggunaan Hadoop

Sejarah Hadoop



- Asal mula Hadoop karena terinspirasi dari makalah tentang *Google MapReduce* dan *Google File System* (GFS) yang ditulis oleh ilmuwan Google, *Jeffrey Dean* dan *Sanjay Ghemawat* pada tahun 2003.
- Proses development dimulai pada saat proyek Apache Nutch, yang kemudian menjadi sub proyek hadoop tahun 2006.
- Penamaan menjadi Hadoop oleh *Doug Cutting*, yaitu berdasarkan nama dari mainan gajah anaknya.

Sejarah Hadoop



- Hadoop sejak 2008 telah menjadi proyek tingkat atas di lingkungan Apache Software Foundation dan dikembangkan secara terbuka oleh komunitas contributor secara global.
- **Pengguna Hadoop** adalah Facebook, AOL, Baidu, IBM, ImageShack, Yahoo.
- Hadoop tersedia bebas dan menyandang lisensi Apache License 2.0.

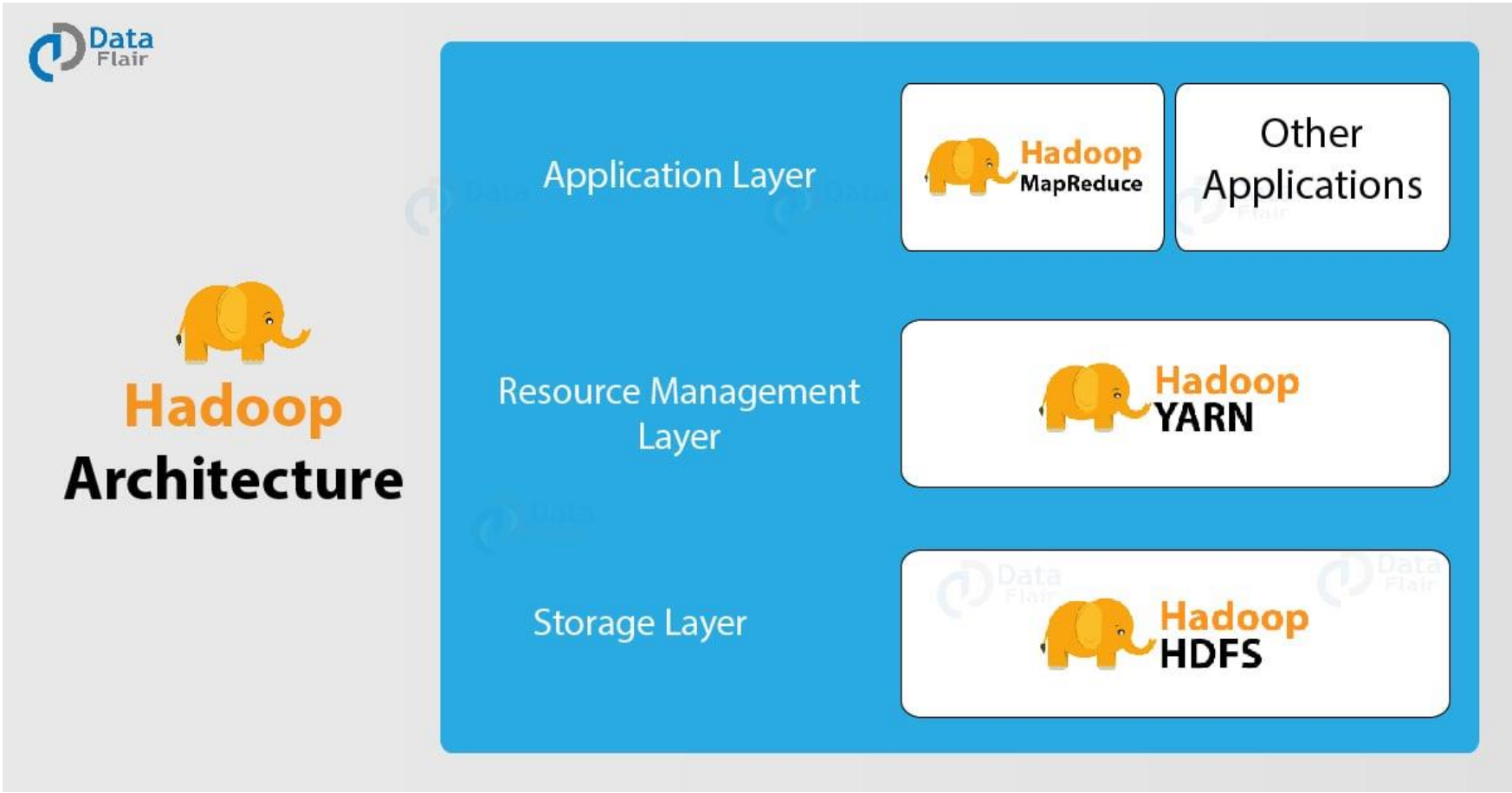
Pengertian Hadoop

- **Hadoop** atau **Apache Hadoop** adalah software bebas dan open source yang ditulis dalam Bahasa pemrograman Java untuk dijalankan secara terdistribusi dan skalable.
- Hadoop dibangun berdasarkan algoritma popular **MapReduce** dari Google Inc., system file yang disarankan GFS (Google File System), yang memungkinkan menjalankan tugas komputasi intensif dalam mengolah data jumlah besar di komputer cluster dengan hardware yang handal.

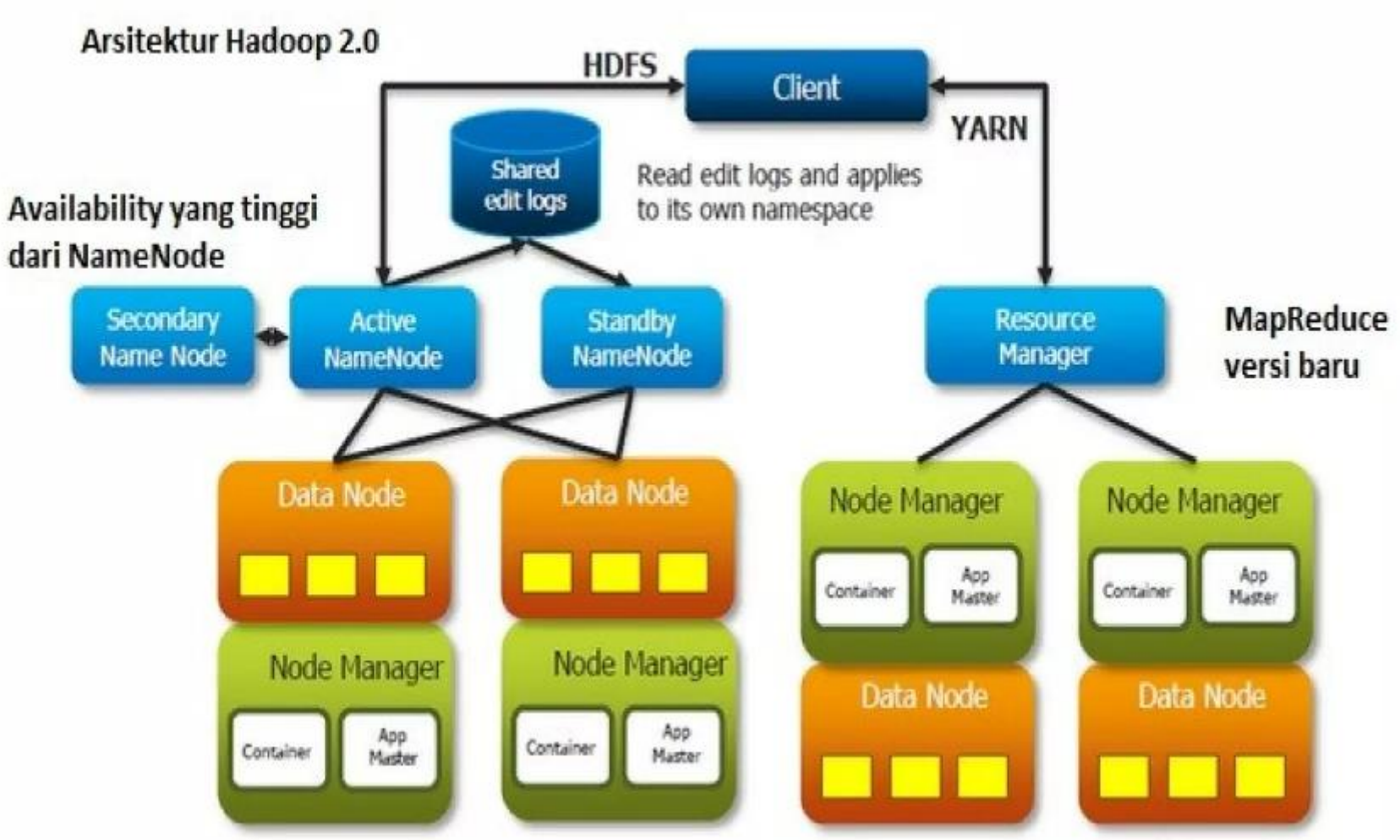
Pengertian Hadoop

- **Hadoop** bisa dijalankan di satu komputer saja (single node) ataupun dalam cluster yang berisi banyak komputer (multi node).
- **Single node** biasanya untuk development atau training saja.
- Hadoop memerlukan **Java** untuk bisa berjalan.

Arsitektur Hadoop



Arsitektur Hadoop



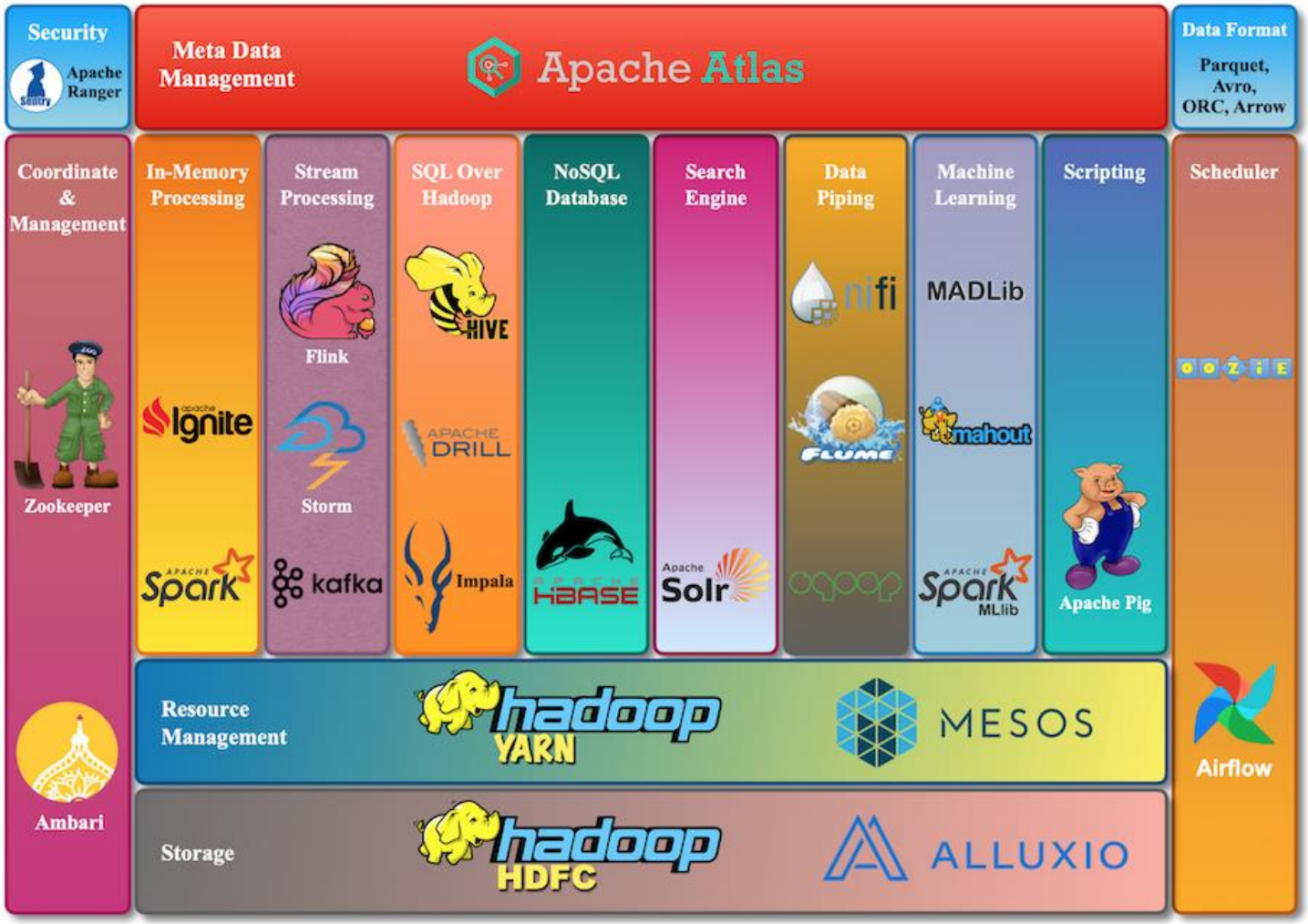
Arsitektur Hadoop

Framework Hadoop terdiri atas empat modul atau komponen utama, yaitu:

- **Hadoop Distributed File System (HDFS)**, yaitu sebuah system file yang terdistribusi
- **Hadoop MapReduce**, yaitu sebuah model programming / algoritma untuk pengolahan data skala besar dengan komputasi secara terdistribusi.
- **Hadoop YARN**, yaitu sebuah platform resource management yang bertanggung jawab untuk mengelola resources dalam clusters dan scheduling.
- **Hadoop Common**, yaitu berisi libraries dan utilities yang dibutuhkan oleh modul Hadoop lainnya.

Ekosistem Hadoop

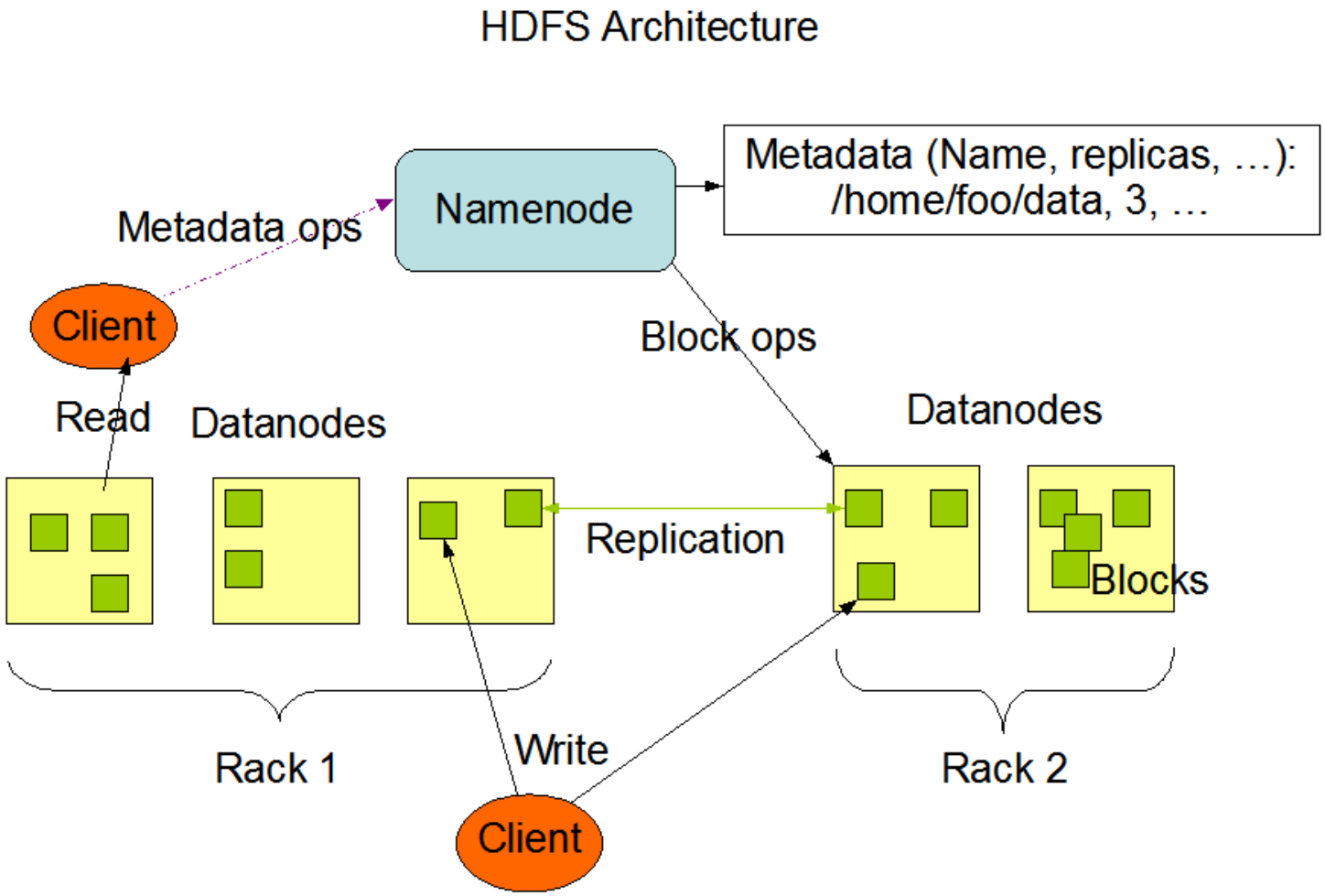
Ekosistem Hadoop



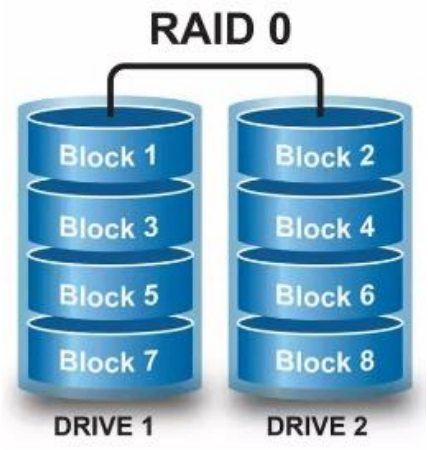
Ekosistem Hadoop

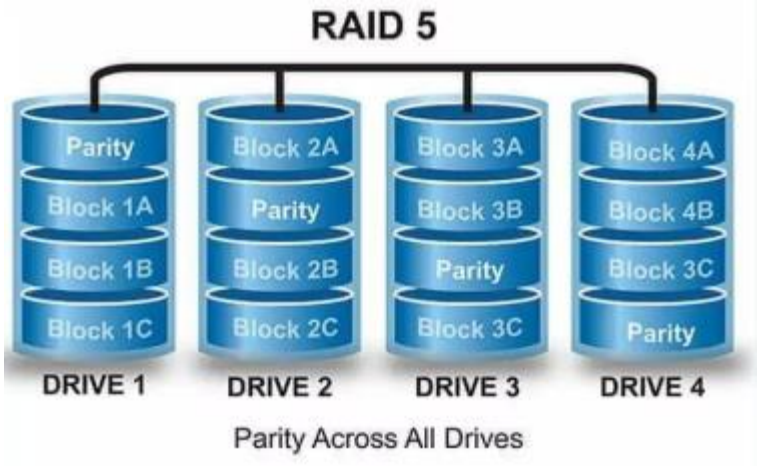
- **Framework Hadoop** bukan hanya empat modul utama namun merupakan kumpulan modul open source seperti Hive, Pig, Oozie, Zookeeper, Mahout, dsb.
- **Hadoop Hive**: dilengkapi dengan fungsi Data-Warehouse, yaitu Bahasa query HiveQL dan indeks. HiveQL adalah Bahasa query berbasis SQL dan memungkinkan pengembang untuk menggunakan sintaks seperti SQL.
- **Hadoop Pig**: dapat digunakan sebagai Bahasa pemrograman high-level untuk menulis program pada Hadoop MapReduce.
- **Hadoop Base**: database sederhana dan skalabel untuk mengelola data dengan jumlah yang sangat besar dalam cluster Haddop. Menggunakan Hbase dapat mengelola jutaan baris data secara efisien.

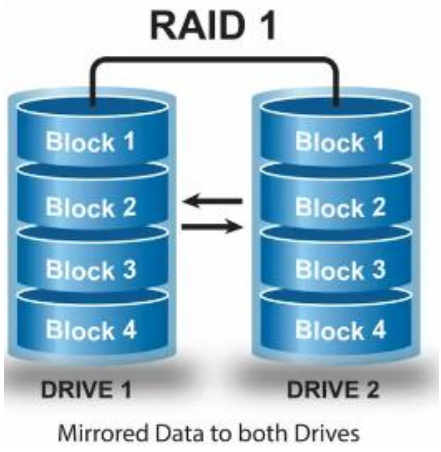
Cara Kerja Hadoop

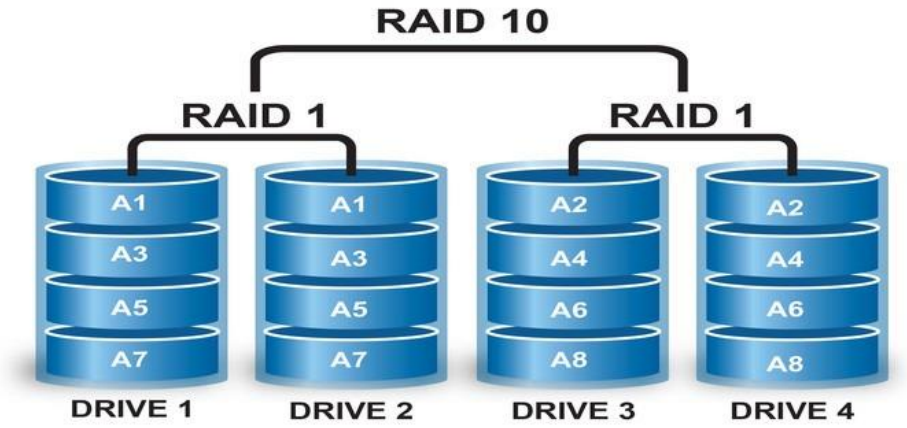


Sistem RAID (Redundant Array of Independent Disks)









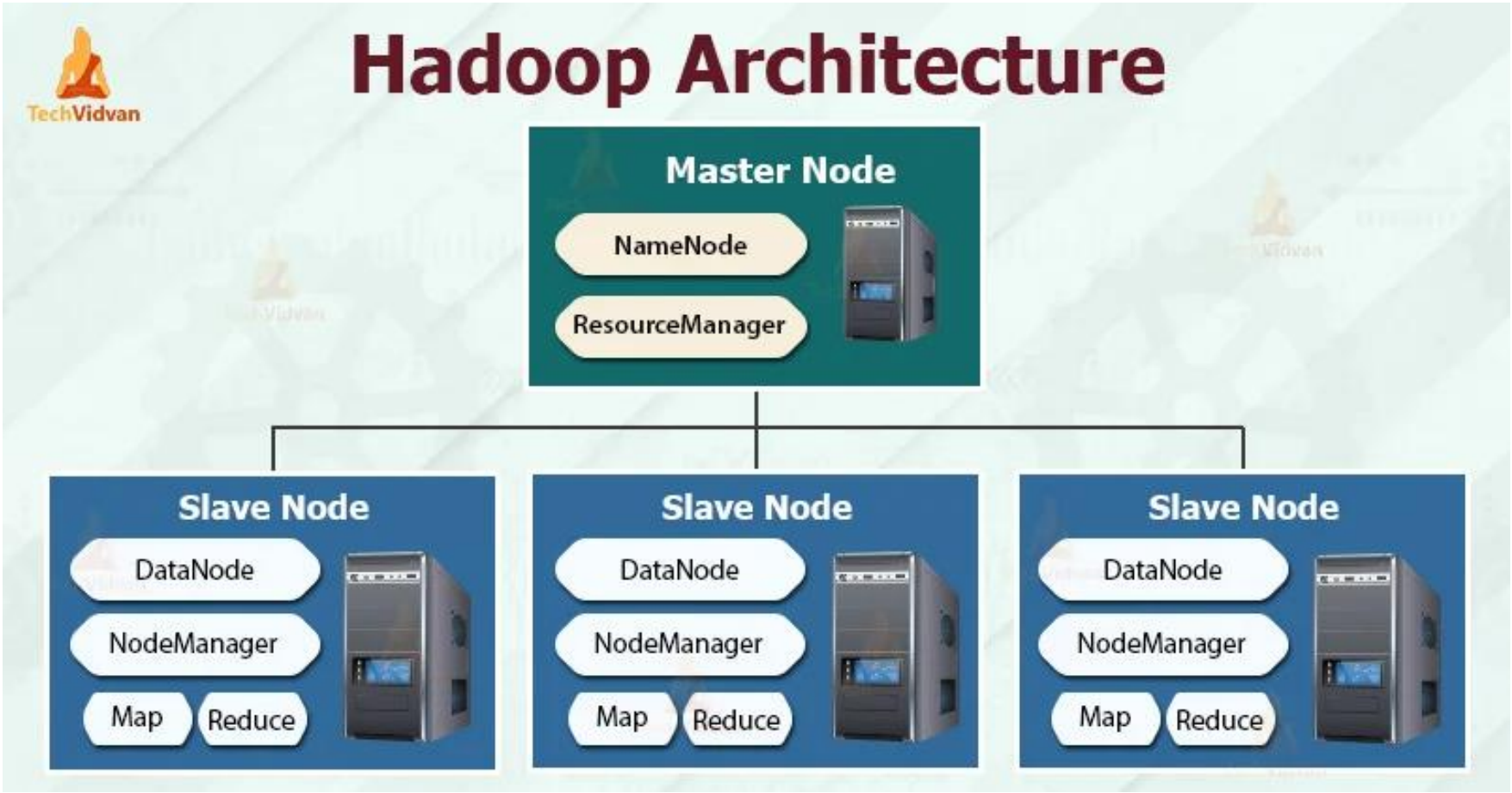
Cara Kerja Hadoop

- **Hadoop Distributed File System** adalah sebuah file system terdistribusi dengan high-availability yang dapat menyimpan data pada mesin komoditas, digunakan untuk menyediakan bandwidth sangat tinggi yang diagregasi ke semua cluster (node).
- File dibagi menjadi **blok data** dengan panjang yang baku dan didistribusikan secara redundan pada simpul (node) yang berpartisipasi.
- Sebuah cluster HDFS terdiri dari **NameNode**, yang mengelola metadata dari klister, dan **DataNode** yang menyimpan data/file
- File dan direktori diwakili pada NameNode oleh ***inode***. Inode menyimpan atribut seperti *permission*, modifikasi dan waktu akses, atau kuota *namespace* dan *diskspace*.

Cara Kerja Hadoop

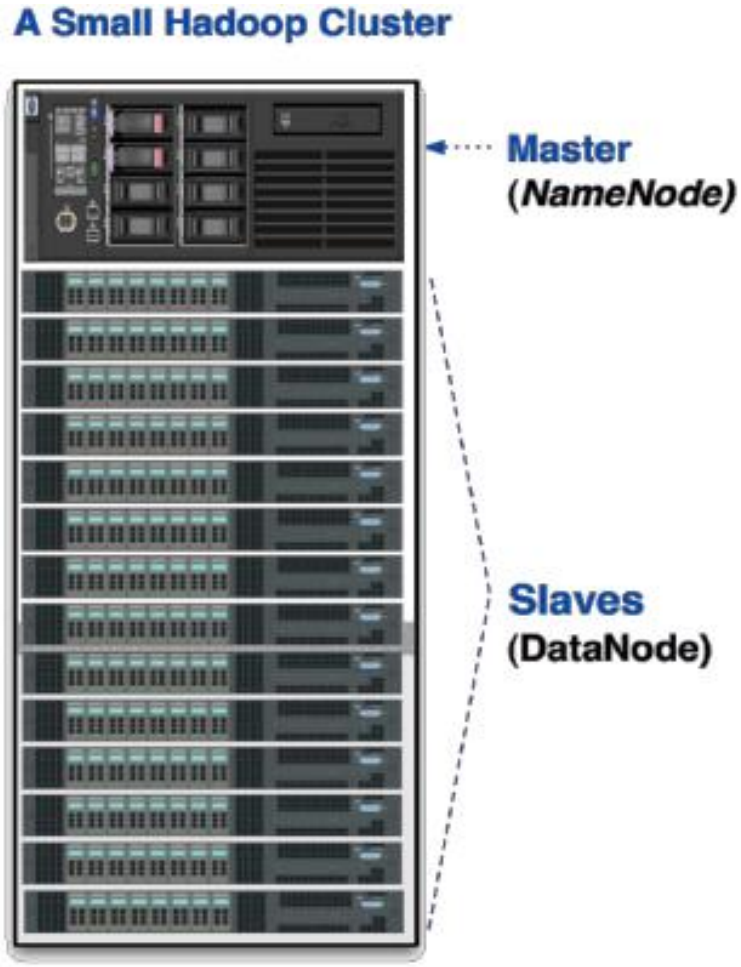
- Isi file dibagi menjadi **blok-blok file** (biasanya 128 MB), dan setiap blok file tersebut di replikasi di beberapa DataNodes
- Blok file disimpan pada system file local dari DataNode
- **NameNode** yang aktif memonitor jumlah salinan/replica blok file. Ketika ada salinan blok file yang hilang karena kerusakan pada DataNode, NameNode akan mereplikasi kembali blok file tersebut ke DataNode lainnya yang berjalan baik.
- **NameNode** mengelola struktur namespace dan memetakan blok file pada DataNode.

Cara Kerja Hadoop



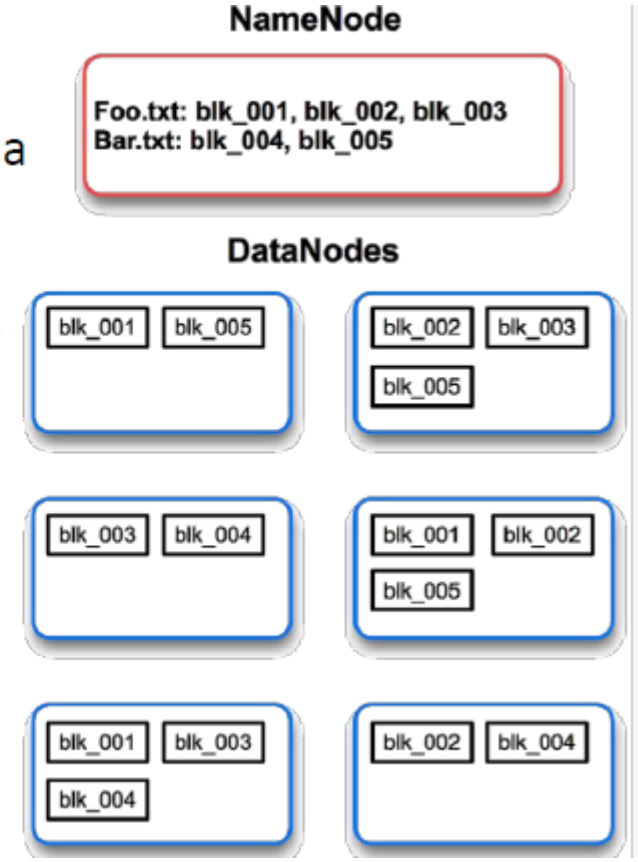
Arsitektur Hadoop

- Hadoop memiliki arsitektur master/slave
- HDFS master daemon: Name Node
  - Mengatur namespace (pemetaan file ke block) dan metada (pemetaan block ke mesin/komputer)
- HDFS slave daemon: Data Node
  - Membaca dan menulis data sebenarnya
  - Dapat berjalan di atas filesystem sebenarnya (ext3/4, NTFS, dll)

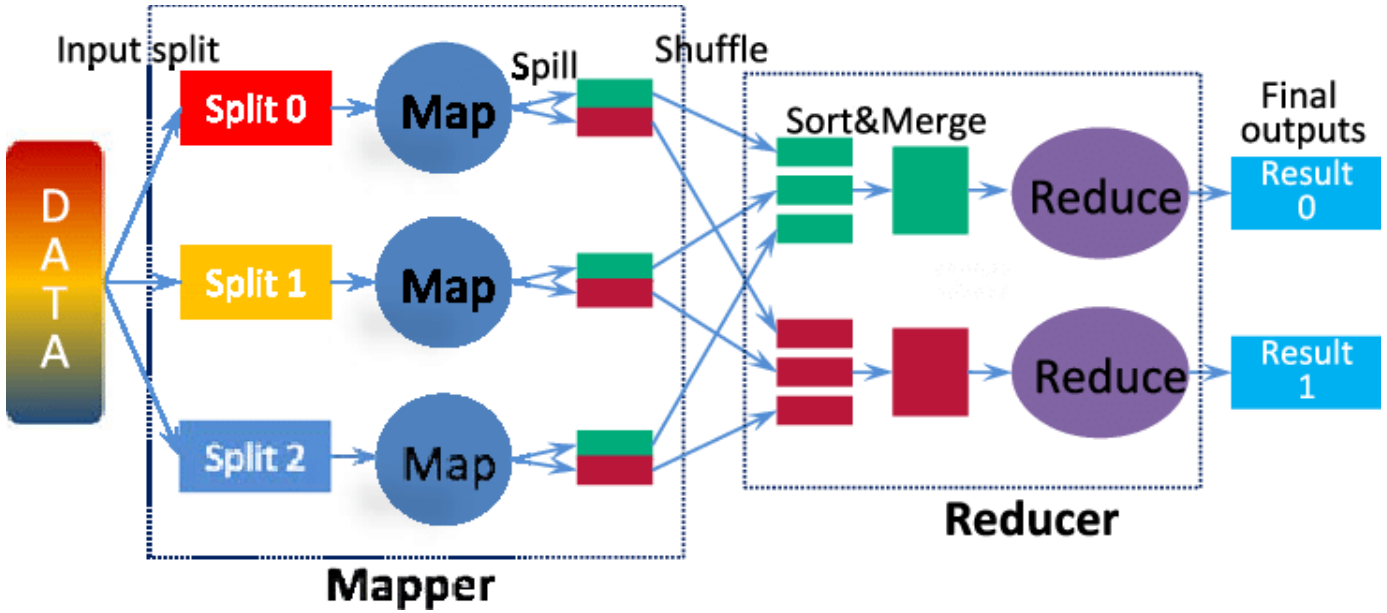


Arsitektur Hadoop

- Contoh:
- NameNode berisi metadata untuk dua file
  - Foo.txt (300MB) dan Bar.txt (200MB)
  - Dimisalkan HDFS dikonfigurasikan dengan block sebesar 128MB
- DataNodes berisi block data sebenarnya
  - Tiap block berukuran 128MB
  - Tiap block direplikasi 3x pada cluster
  - Laporan block dikirimkan ke NameNode secara periodik

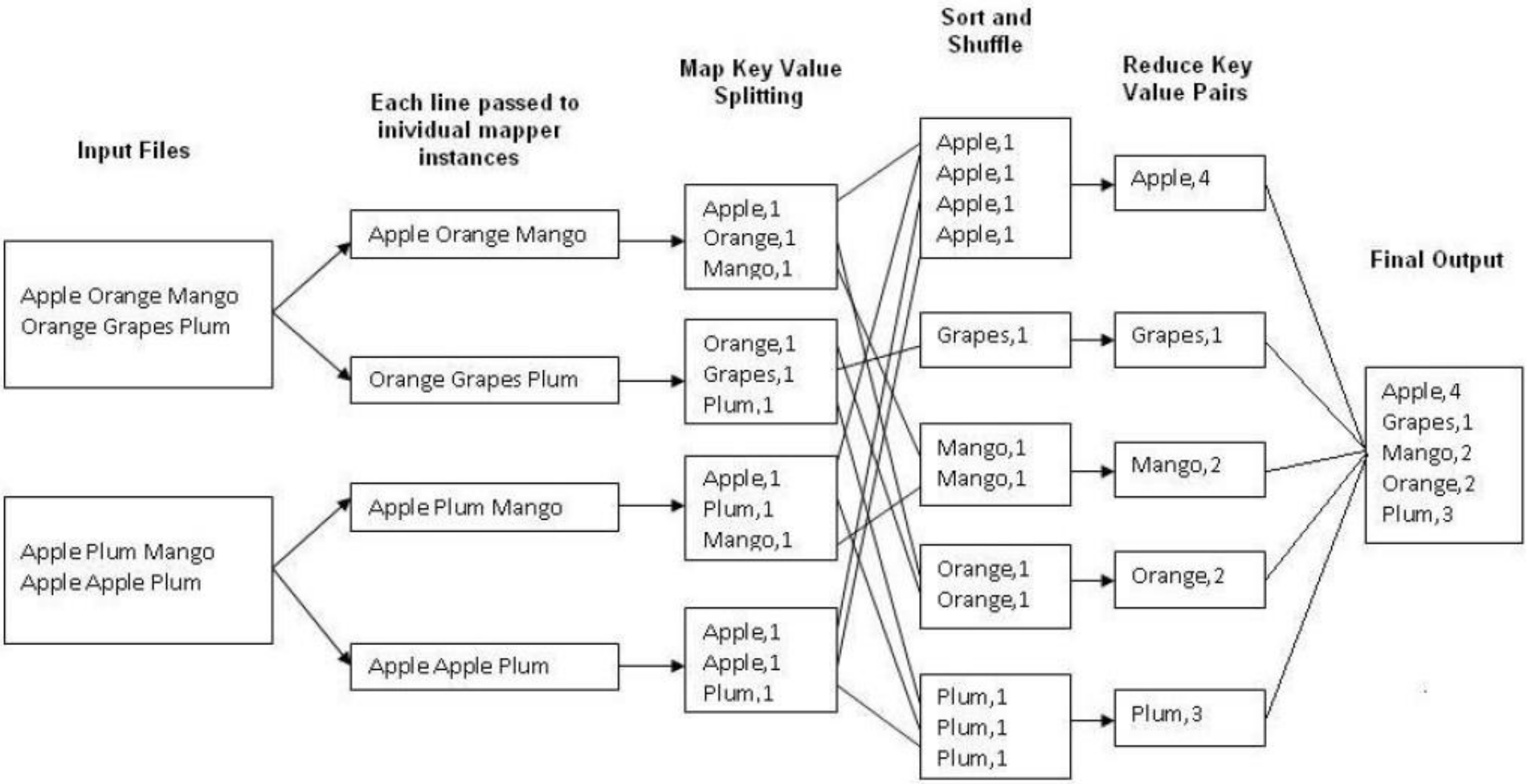


Cara Kerja Hadoop



- MapReduce bertugas membagi data yang besar ke dalam potongan lebih kecil dan mengatur mereka ke dalam bentuk tupel untuk pemrosesan parallel. Tupel adalah kombinasi antara key dan value-nya, dapat disimbolkan dengan notasi "***(k1, v1)***".
- Dengan pemrosesan bersifat parallel tersebut, tentunya akan meningkatkan kecepatan dan keandalan komputasi pada system klastering.

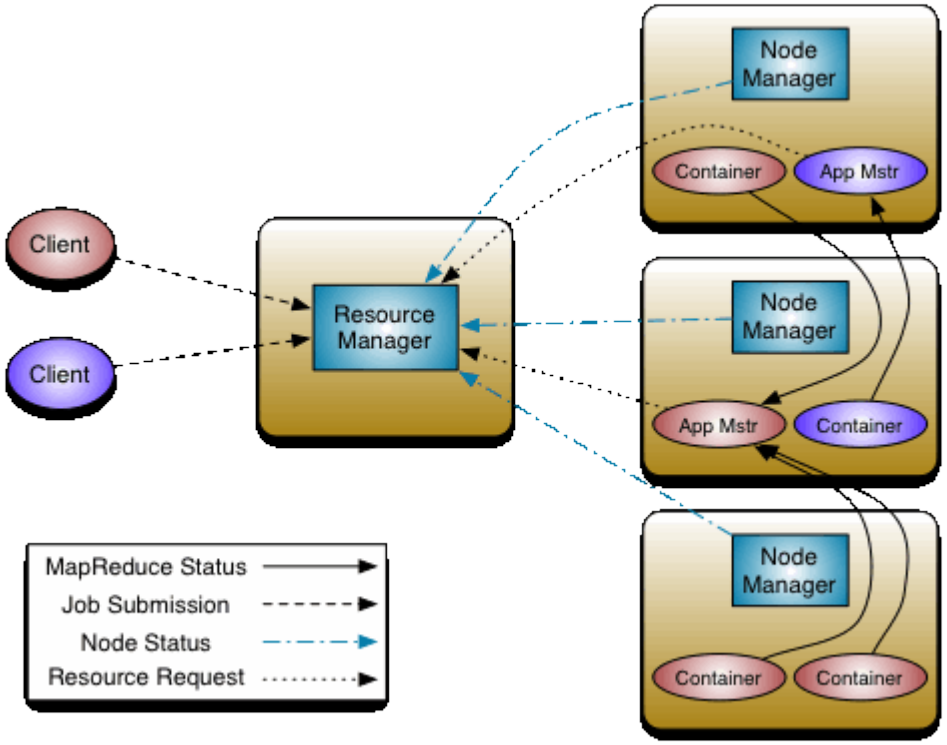
MapReduce


Input Files
Each line passed to inividual mapper instances
Map Key Value Splitting
Sort and Shuffle
Reduce Key Value Pairs
Final Output
Apple Orange Mango Orange Grapes Plum
Apple Plum Mango Apple Apple Plum
Apple Orange Mango
Orange Grapes Plum
Apple Plum Mango
Apple Apple Plum
Apple,1 Orange,1 Mango,1
Orange,1 Grapes,1 Plum,1
Apple,1 Plum,1 Mango,1
Apple,1 Apple,1 Plum,1
Apple,1 Apple,1 Apple,1 Apple,1
Grapes,1
Mango,1 Mango,1
Orange,1 Orange,1
Plum,1 Plum,1 Plum,1
Apple,4
Grapes,1
Mango,2
Orange,2
Plum,3
Apple,4 Grapes,1 Mango,2 Orange,2 Plum,3

Cara Kerja Hadoop

- **MapReduce** terdiri atas 3 tahap yaitu tahap *map, shuffle*, dan *reduce*. Shuffle dan reduce digabungkan ke dalam satu tahap saja yaitu tahap ***reduce***.
- **Map** berfungsi memproses data input yang umumnya berupa file yang tersimpan dalam HDFS, input tersebut kemudian diubah menjadi tuple yaitu pasangan antara key dan value-nya.
- Tahap **reduce**, memproses data input dari hasil proses map, yang kemudian dilakukan tahap shuffle dan reduce yang hasil data set baru-nya disimpan di HDFS kembali.

Cara Kerja Hadoop (YARN)

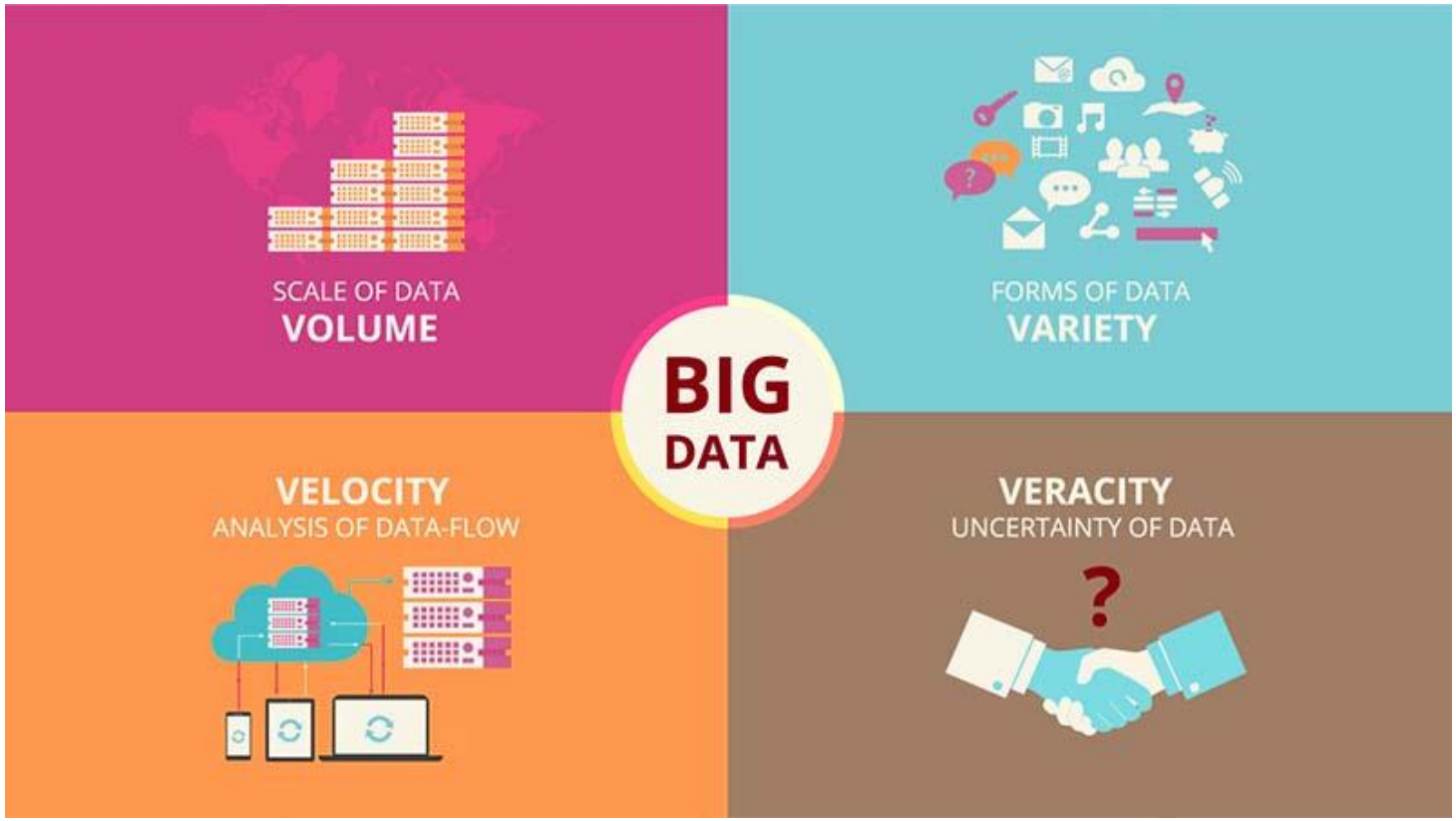


- **Hadoop YARN** adalah sebuah platform manajemen sumber daya yang bertanggung jawab atas pengelolaan sumber daya komputasi dalam sebuah klaster dan digunakan untuk penjadwalan aplikasi pengguna.

Cara Kerja Hadoop

- Tujuan awal YARN adalah untuk memisahkan dua tanggung jawab dari Job Tracker atau Task Tracker menjadi beberapa entitas yang terpisah.
- **Global Resource Manager** di node master, yang berfungsi mengatur semua resource yang digunakan aplikasi dalam system.
- **Aplication Master** di setiap aplikasi, yang berfungsi untuk negosiasi resource dengan Resouce Manager dan kemudian bekerja sama dengan Node Manager untuk mengeksekusi dan memonitor *tasks*.
- **Node Manager** di Agen Framework setiap node slave, yang bertanggung jawab terhadap Container, dengan memantau penggunaan *resource* / sumber daya dari container (cpu, memori, disk, jaringan) dan melaporkannya pada Resource Manager.
- **Container** di setiap aplikasi yang jalan di Node Manager, sebagai wadah penyimpanan data/file.

Hadoop dalam Big Data



Hadoop dalam Big Data

- Big data memiliki 4 tantangan utama sehingga Hadoop sangat dibutuhkan, yaitu:
- **Volume**, keperluan menyimpan dan mengelola data dalam jumlah yang sangat besar, dan data tersebut selalu tambah besar setiap saat.
- **Velocity**, begitu cepat data yang muncul dan keperluan untuk bisa mengakses data besar tersebut dengan cepat.
- **Variety**, semakin bervariasinya data saat ini sehingga dengan teknologi relational database (RDBMS) saat ini sudah tidak bisa ditangani lagi.

Hadoop optimal digunakan untuk menangani data dalam jumlah besar baik data *Structured*, *Semi-structured* maupun *Unstructured*.

Hadoop mereplikasi data di beberapa komputer (*clustering*) sehingga jika salah satu komputer mati/bermasalah maka data dapat diproses dari salah satu komputer lainnya yang masih hidup.

Kelebihan dan Kekurangan Hadoop

Kelebihan Hadoop

- Hadoop merupakan software open source
- Hadoop dapat menampung data dengan jumlah yang sangat besar

Kekurangan Hadoop

- Map reduce hanya bisa berjalan secara serial untuk mengolah data. Artinya tidak bisa dilakukan pemrosesan data secara parallel
- Map reduce hanya bisa berjalan dalam batch atau secara periodic dan tidak bisa terus menerus secara realtime. Hal ini membuat Map Reduce tidak bisa mengolah data dalam bentuk streaming misalnya tweet dari twitter.

Implementasi Hadoop

Penggunaan Hadoop saat ini sudah semakin luas, diantaranya adalah:

- **Yahoo**

  Terdiri dari 24.000 server di 17 cluster. Lebih dari 10 petabytes data user. Mengerjakan ratusan ribu jobs setiap bulannya. Dan digunakan untuk news, search dan mail.

- **New York Times**

  Menggunakan Hadoop untuk mengkonversi artikel NYTimes menjadi pdf dari tahun 1851 sampai dengan 1922. Berjalan diatas 100 server Amazon EC2 selama 24 jam dengan input data sebesar 4TB dan outuput 1,5TB.

- **Facebook**

  Digunakan untuk data mining dan data warehousing, user data analysis. Dan dijalankan di 600 server.

Kesimpulan

- **Hadoop** merupakan framework yang digunakan sebagai solusi untuk Big Data dan bersifat open source.
- **Hadoop HDFS** adalah system file terdistribusi yang bersifat fault-tolerant dan mendukung untuk mengolah data set yang besar (Big Data).
- **Hadoop MapReduce** adalah model komputasi berbasis Java pada Sistem Terdistribusi dalam rangka mendukung aplikasi Big Data.
- **Hadoop YARN** adalah platform untuk resource-management yang muncul untuk mengatasi limitasi MapReduce pada arsitektur Hadoop 1.0.