

# Hadoop

## Sains Data IKOPIN

<https://github.com/sevinurafni/SD3204>

Thu, 7 March 2024

**What technology do we have for  
big data?**

# Big Data Landscape

## Log Data Apps



## Vertical Apps



## Business Intelligence

ORACLE | Hyperion



Microsoft | Business Intelligence



## Analytics and Visualization



## Data Providers



## Analytics Infrastructure



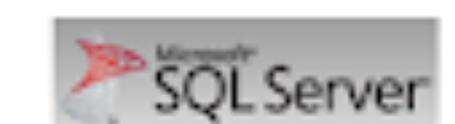
## Operational Infrastructure



## Infrastructure As A Service



## Structured Databases

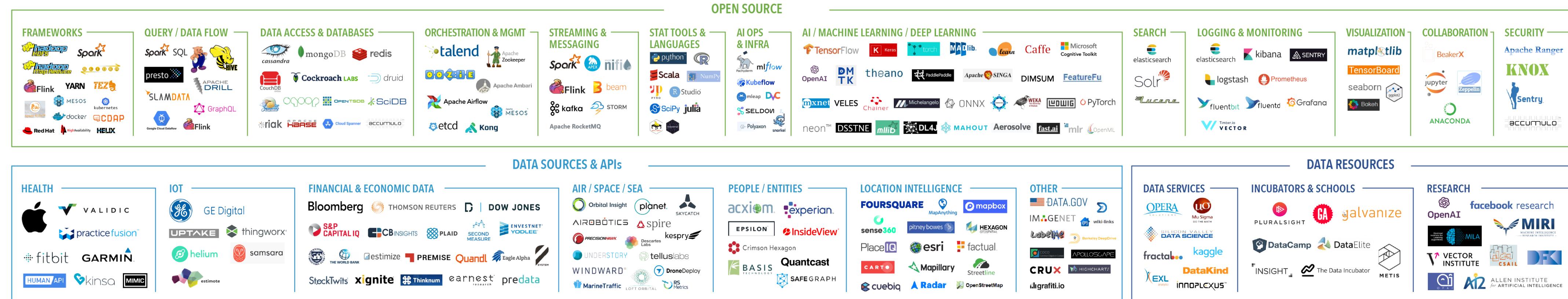
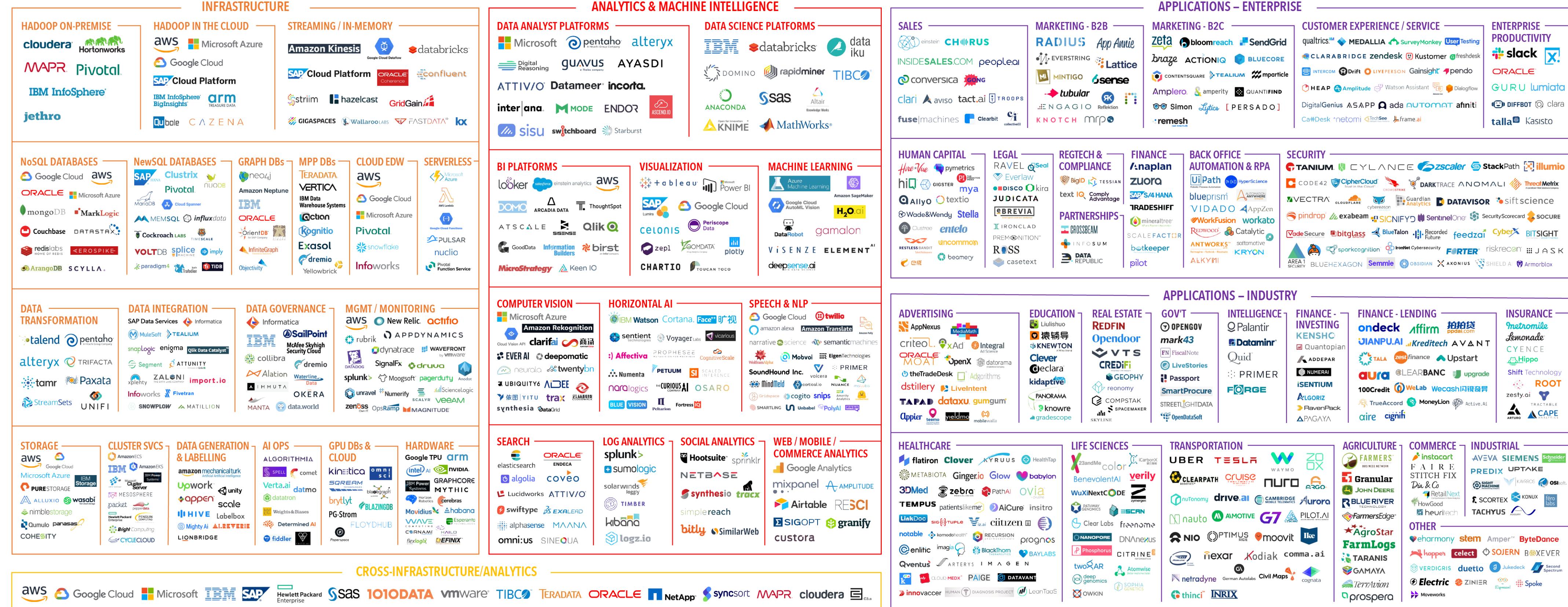


## Technologies



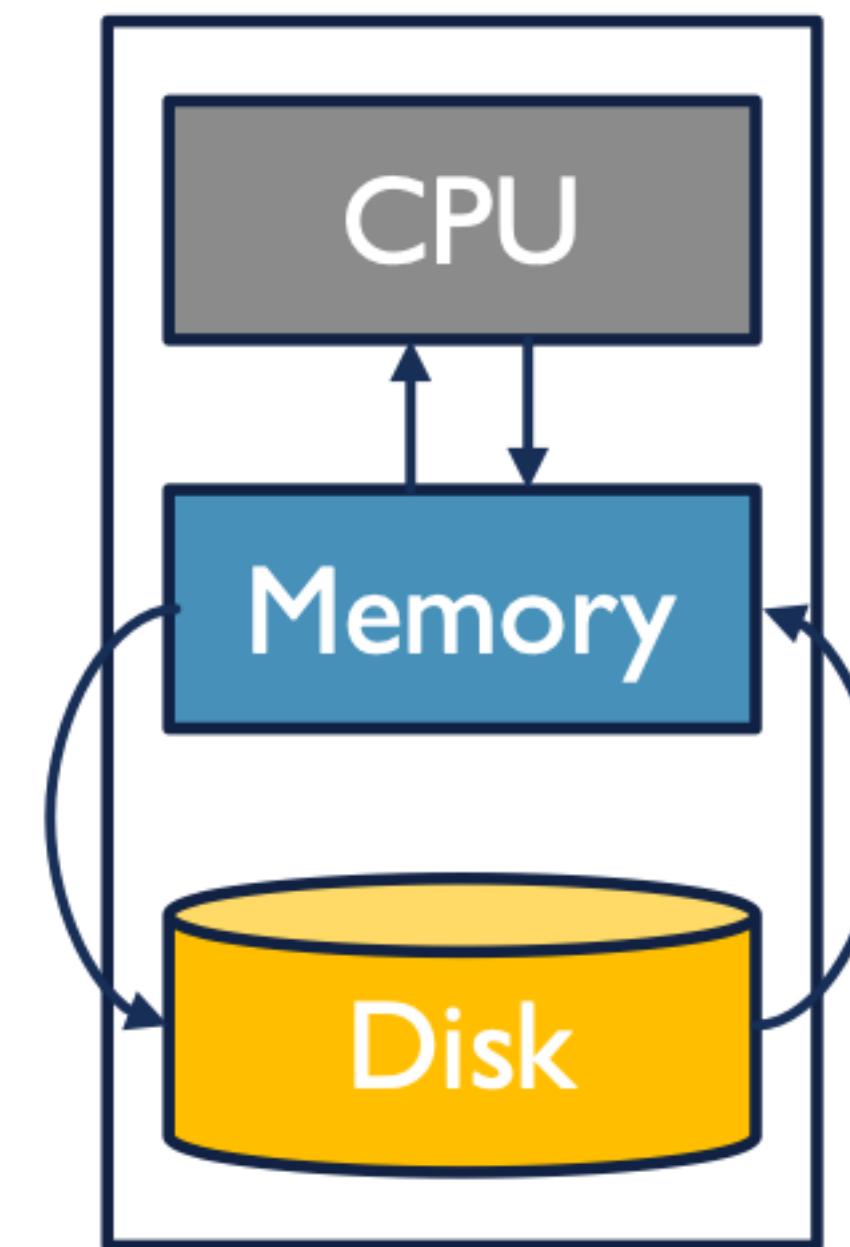


DATA & AI LANDSCAPE 2019



# Traditional Data Processing

- Membaca kata dari disk vs. memori utama: 105 lebih lambat. Disk normal hanya mencapai 150MB/detik untuk pembacaan berurutan.
- IO Bounded: hambatan kinerja terbesar adalah membaca/menulis file besar ke disk.
- 100 GB ~ 10 menit.
- 200 TB ~ 20.000 menit = 13 hari.



# What is Hadoop?

Hadoop is named after a toy elephant belonging to Doug Cutting's son.

Cutting is 'father' of Hadoop.

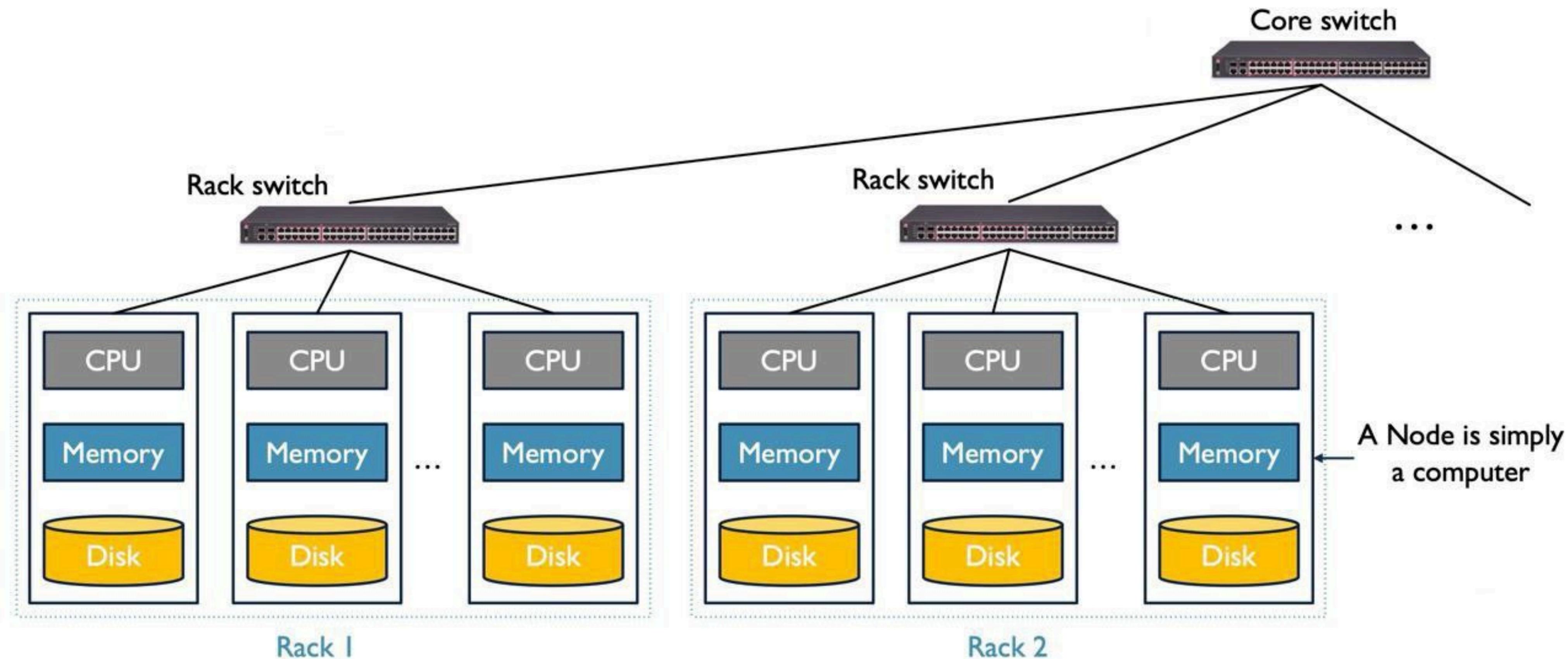
- Hadoop adalah sebuah kerangka kerja untuk menyimpan data pada kelompok besar perangkat keras komoditas dan menjalankan aplikasi terhadap data tersebut.
- Tidak perlu server yang besar dan mahal.
- Banyak komputer yang terjangkau dan mudah didapat dengan CPU tunggal yang disatukan.
- Cluster adalah sekelompok komputer yang saling terhubung (komputer dalam cluster sebagai node) yang dapat bekerja bersama pada masalah yang sama.
- Menggunakan jaringan sumber daya komputasi yang terjangkau untuk mendapatkan wawasan bisnis adalah proposisi nilai utama Hadoop.

- Rack adalah kumpulan 30 atau 40 node yang secara fisik disimpan berdekatan dan semuanya terhubung ke sakelar yang sama.
- Bandwidth jaringan antara dua node di dalam rak lebih besar daripada bandwidth antara dua node di rak yang berbeda.
- Cluster Hadoop adalah kumpulan rack.



Navigation icons: back, forward, search, etc.

Switch: a networking hardware that connects devices on a computer network. (not Nintendo switch) 😊



# Challenges for IO Cluster Computing

Hadoop bekerja dengan dua komponen utama: HDFS dan MapReduce

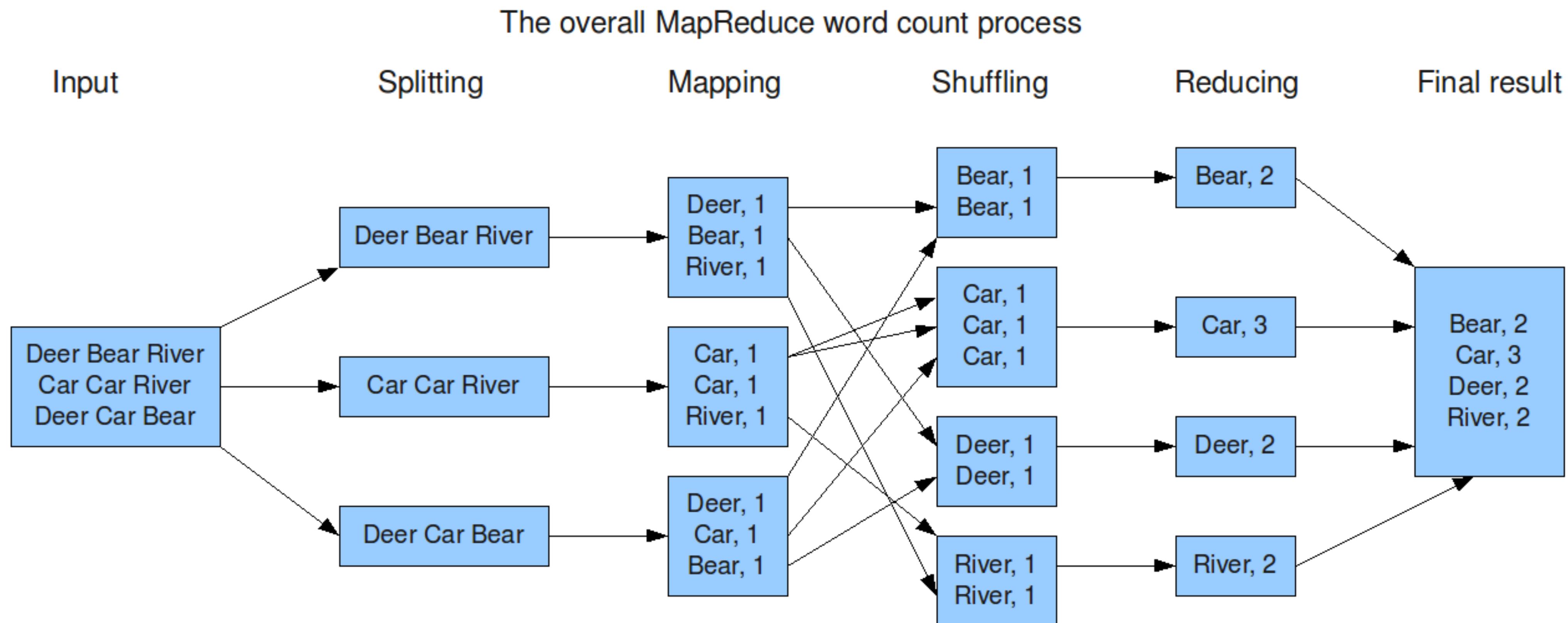
- Node gagal: rata-rata 1 dari 1000 node gagal dalam sehari.
  - Solusi: duplikasi data.
- Kemacetan jaringan: biasanya throughput 1-10 GB/s.
  - Solusi: membawa komputasi ke node, bukan data ke node.
- Pemrograman terdistribusi tradisional sering kali bersifat ad-hoc dan rumit.
  - Solusi: tetapkan sistem pemrograman yang dapat dengan mudah didistribusikan.

# MapReduce

MapReduce adalah sebuah kerangka kerja untuk komputasi paralel. Biasanya terdiri dari tiga operasi:

1. Map: setiap node menerapkan fungsi map ke data lokal, dan menulis output ke penyimpanan sementara.
2. Shuffle: node mendistribusikan ulang data berdasarkan kunci keluaran (yang dihasilkan oleh fungsi map), sehingga semua data yang termasuk dalam satu kunci berada pada node yang sama.
3. Reduce: node sekarang memproses setiap kelompok data keluaran, per kunci, secara paralel.

# How MapReduce Works?



# Hadoop Key Features

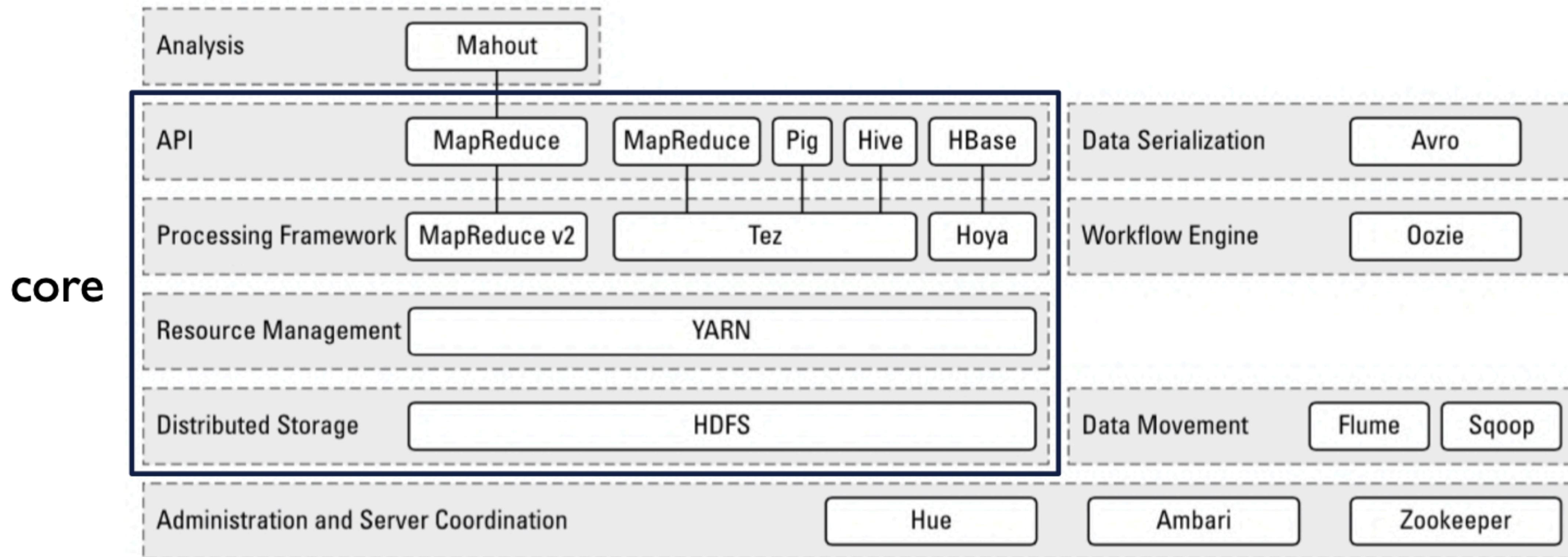
- Arsitektur berbagi tidak ada: Hadoop adalah sebuah cluster dengan mesin-mesin independen (kluster dengan node), di mana setiap node menjalankan tugasnya dengan menggunakan sumber daya sendiri.
- Fault Tolerance: Ketika data dikirim ke sebuah node individual, data tersebut juga direplikasi (biasanya x3) ke node-node lain dalam kluster. Jika terjadi kegagalan, ada salinan lain yang tersedia untuk digunakan.
- Commodity hardware: Hadoop tidak memerlukan server yang sangat canggih dengan memori besar dan daya pemrosesan tinggi. Hadoop berjalan pada JBOD (sekumpulan disk saja), sehingga setiap node independen dalam Hadoop.
- Skalabilitas horizontal: Kita tidak perlu meningkatkan daya node. Saat data terus bertambah, kita hanya perlu menambahkan node-node baru.

# Hadoop Workflow

Hadoop menjalankan kode di seluruh kluster komputer. Proses ini mencakup tugas inti berikut yang dilakukan oleh Hadoop:

- \* Data awalnya dibagi ke dalam direktori dan file-file. File-file ini dibagi menjadi blok-blok berukuran seragam sebesar 128M.
- \* File-file ini kemudian didistribusikan ke berbagai node kluster untuk pemrosesan lebih lanjut.
- \* HDFS, yang berada di atas sistem file lokal, mengawasi pemrosesan.
- \* Blok-blok direplikasi untuk menangani kegagalan perangkat keras.
- \* Memeriksa bahwa kode dieksekusi dengan sukses.
- \* Melakukan pengurutan yang terjadi antara tahap map dan reduce.
- \* Mengirimkan data yang sudah diurutkan ke komputer tertentu.
- \* Menulis log debugging untuk setiap pekerjaan.

# Apache Hadoop Ecosystem



# Tugas 2

- Cari tahu cara menginstall Hadoop dan prerequisite apa saja yang diperlukan

# Thank You!

- Reference (recommend for further reading):  
**The official website:** <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>
- **The paper:** K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, Incline Village, NV, 2010, pp. 1-10.
- **The book:** Chapter 4, DeRoos, Dirk. *Hadoop for dummies*. John Wiley & Sons, 2014.