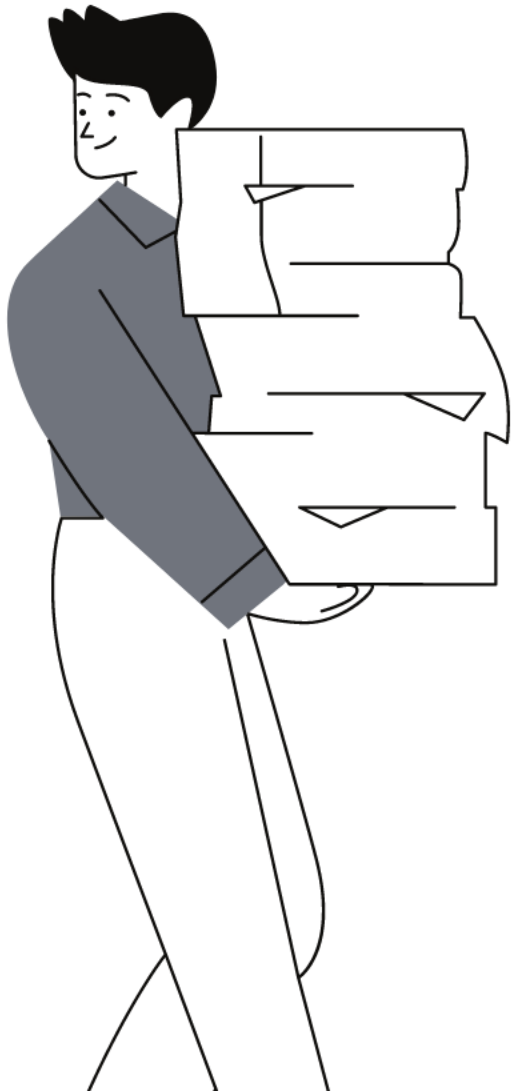
Universitas Bosowa
Teknologi Informasi

Abdillah.S.A.S.,S.Kom.,M.Pd.

Teknologi Database

Hadoop I



Universitas Bosowa
Fakultas Teknik - Prodi Teknologi Informasi

Pengertian Big Data

Big data adalah kumpulan proses yang terdiri volume data dalam jumlah besar yang terstruktur maupun tidak terstruktur dan digunakan untuk membantu kegiatan bisnis.

Big data sendiri merupakan pengembangan dari sistem database pada umumnya.

Yang membedakan disini adalah proses kecepatan, volume, dan jenis data yang tersedia lebih banyak dan bervariatif daripada DBMS (Database Management System) pada umumnya.

Definisi 3V Dari Big Data

Universitas Bosowa
Teknologi Informasi

1. **Volume**
   Ukuran data yang dimiliki oleh big data memiliki kapasitas yang besar. Anda dapat mencoba melakukan proses data dengan ukuran yang besar untuk dijalankan.
2. **Velocity**
   Kecepatan transfer data juga sangat berpengaruh dalam proses pengiriman data dengan efektif dan stabil. Big data memiliki kecepatan yang memungkinkan untuk dapat diterima secara langsung (real-time).
3. **Variety**
   Jenis variasi data yang dimiliki oleh big data lebih banyak daripada menggunakan sistem database SQL. Jenis data yang masih bersifat tradisional, lebih terstruktur daripada data yang belum terstruktur.

Fungsi Big Data

1. Dapat menentukan penyebab suatu masalah, kegagalan secara real time
2. Pengambilan sebuah keputusan yang cerdas dan tepat
3. Mendeteksi sebuah anomali atau perilaku yang menyimpang dalam struktur bisnis anda
4. Mengurangi biaya, waktu, dan meningkatkan performa produk aplikasi

Big Data Problem

Teknologi Big Data kini mulai trend di kalangan praktisi data.

Pasalnya, penggunaan Big Data bagi suatu perusahaan dapat memberikan manfaat yang positif untuk perkembangan industri tersebut.

Dengan memberikan kemudahan dalam menganalisis datanya, teknologi ini juga dapat menghasilkan keputusan yang tepat bagi perusahaan dalam menjawab segala kebutuhan, baik dari pihak customer maupun industri itu sendiri.

Big Data Problem

Tentunya, Big Data memerlukan sebuah software untuk mengelola datanya dengan baik.

Hadoop merupakan salah satu software yang mampu menghubungkan banyak komputer agar dapat bekerja sama untuk menyimpan dan mengelola data dalam satu kesatuan.

Selain itu, Hadoop didesain untuk mampu memproses data berskala besar dengan melibatkan berbagai kluster komputer.

Implementasi Hadoop dalam Big Data

Universitas Bosowa
Teknologi Informasi

Software Hadoop atau sebutan resminya adalah Apache Hadoop ini merupakan salah satu implementasi dari teknologi Big Data.

Software yang bekerja lebih dari sekedar perangkat lunak ini, dapat diakses secara terbuka atau open source.

Hadoop sendiri merupakan sekumpulan software yang mampu menyelesaikan permasalahan dari sekumpulan data dengan jumlah yang besar.

Dengan besarnya volume dan banyaknya variasi data yang diperoleh suatu perusahaan, menjadikan Hadoop solusi dalam menyelesaikan masalah tersebut.

Adapun beberapa perusahaan besar yang mengimplementasikan Hadoop dalam proses pengolahan Big Datanya, salah satunya adalah media sosial Facebook.

Universitas Bosowa
Teknologi Informasi

ANALOGI BIG DATA & PENGGUNAAN HADOOP

085242179993 Abdhi Griffindors @abdhi @abdhi_tux abdhisas.com

Analogi Big Data

Sekarang, marilah kita mencoba untuk memahami big data dan mengapa Hadoop diperlukan melalui suatu analogi yang sederhana. Bayangkan suatu skenario dimana kita mempunyai seorang *shopkeeper* (pemilik toko) bernama Tim yang menjual padi-padian. Para pelanggannya senang karena Tim sangat cepat menangani pesanan mereka.

Analogi Big Data

Setelah beberapa waktu, Tim merasakan adanya permintaan mendesak terhadap produk lain, sehingga dia berpikir untuk memperluas bisnisnya. Bersama dengan padi-padian, dia mulai menjual buah-buahan, sayur-sayuran, daging, dan produk susu (*dairy*). Saat jumlah pelanggannya meningkat, Tim mengalami kesulitan untuk menjaga kualitas pelayanan (yaitu menangani pesanan yang banyak dan terus-menerus).

Analogi Big Data

Untuk mengatasi situasi ini, Tim memutuskan untuk mempekerjakan tiga orang (selain dirinya) untuk membantunya menyelesaikan pekerjaan. Mereka adalah Matt yang bertanggungjawab menangani bagian buah-buahan dan sayuran, Luke untuk menangani produk susu dan daging, dan terakhir Ann yang ditunjuk untuk menyelesaikan urusan pembayaran sebagai kasir.

Analogi Big Data

Tetapi ini tidak menyelesaikan semua permasalahan Tim. Meskipun mempunyai tenaga kerja yang diperlukan, dia mulai kehabisan ruang (*running out of space*) dalam tokonya untuk menyimpan barang-barang yang diperlukan untuk memenuhi permintaan yang terus meningkat.

Analogi Big Data

Tim menyelesaikan ini dengan mendistribusikan ruang antar lantai berbeda dari bangunan tokonya. Padi-padian dijual di lantai dasar, buah-buahan dan sayuran di lantai satu, produk susu (*dairy*) dan daging pada lantai dua, dan seterusnya.

Inilah bagaimana Tim menuntaskan masalahnya; sekarang mari kita melihat bagaimana cerita ini dapat dibandingkan dengan big data dan Hadoop.

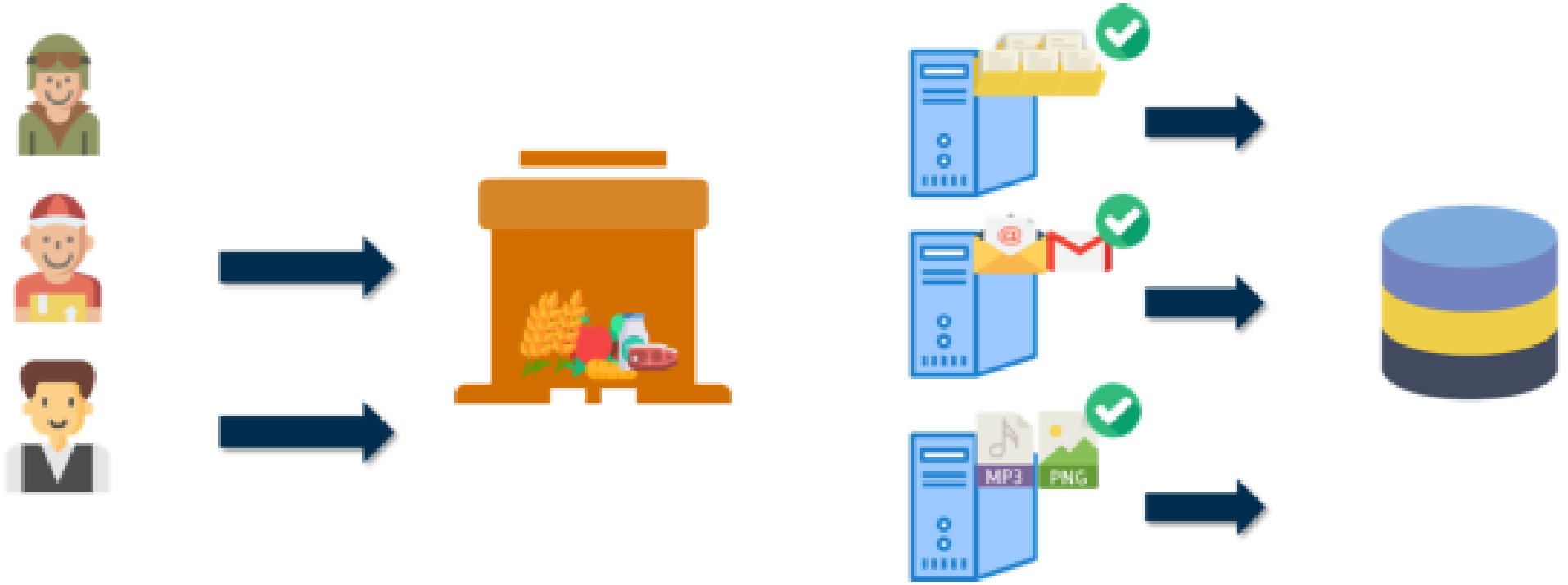
Analogi Big Data

Dulu, pembangkitan data terbatas hanya untuk satu format. Data tersebut dapat dikelola dengan hanya satu unit *storage* dan satu *processor*. Generasi data secara bertahap mulai meningkat dan variasi baru dari data telah hadir. Karena data dihasilkan dengan kecepatan tinggi maka ini mengakibatkan lebih sulit untuk ditangani oleh satu prosesor.

Ini mirip dengan bagaimana Tim menemukan kesulitan dalam mengelola sendiri bisnisnya yang baru saja diperluas.

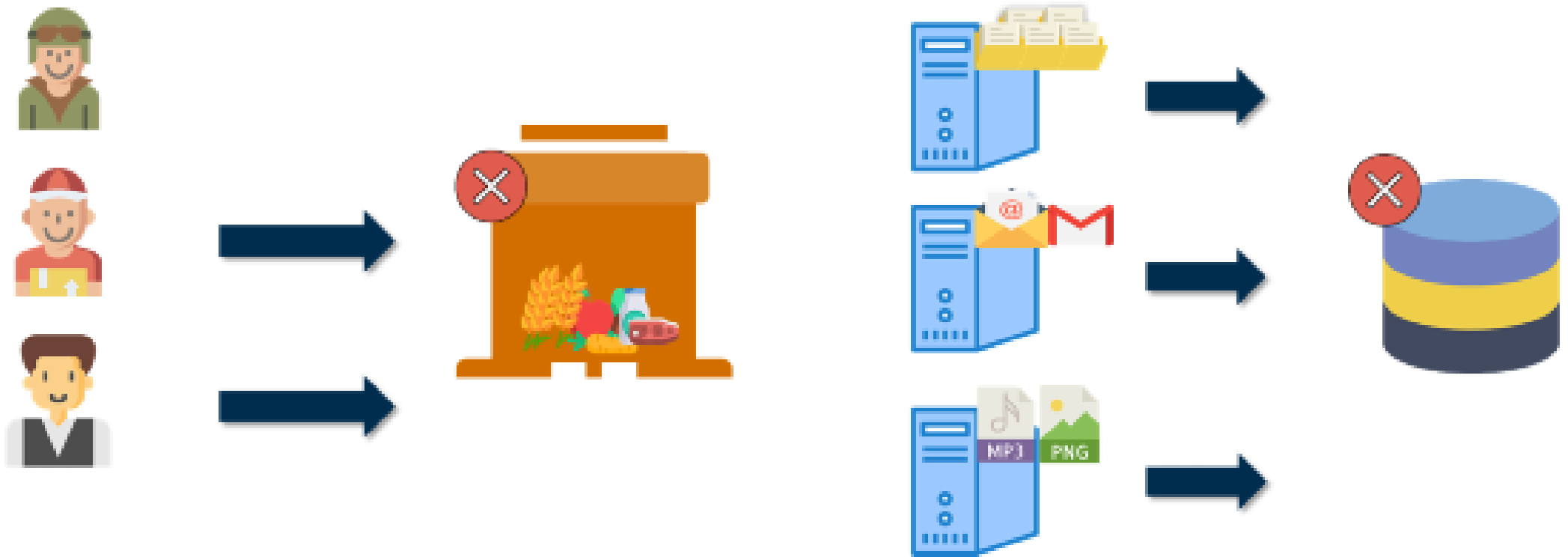
Analogi Big Data

Jadi seperti bagaimana Tim memecahkan masalah ini dengan menambah tenaga kerja, banyak prosesor dapat digunakan untuk memproses setiap jenis data.



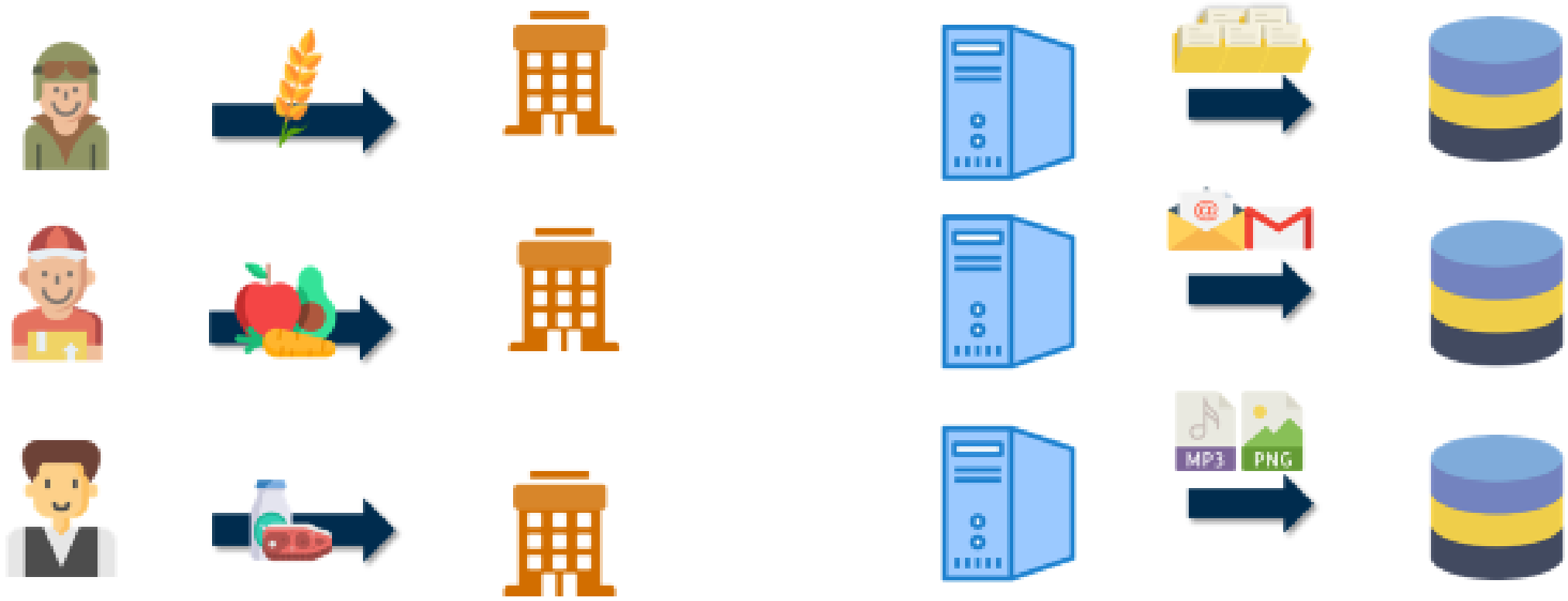
Analogi Big Data

Namun, menjadi sulit bagi banyak prosesor untuk mengakses unit penyimpanan yang sama.



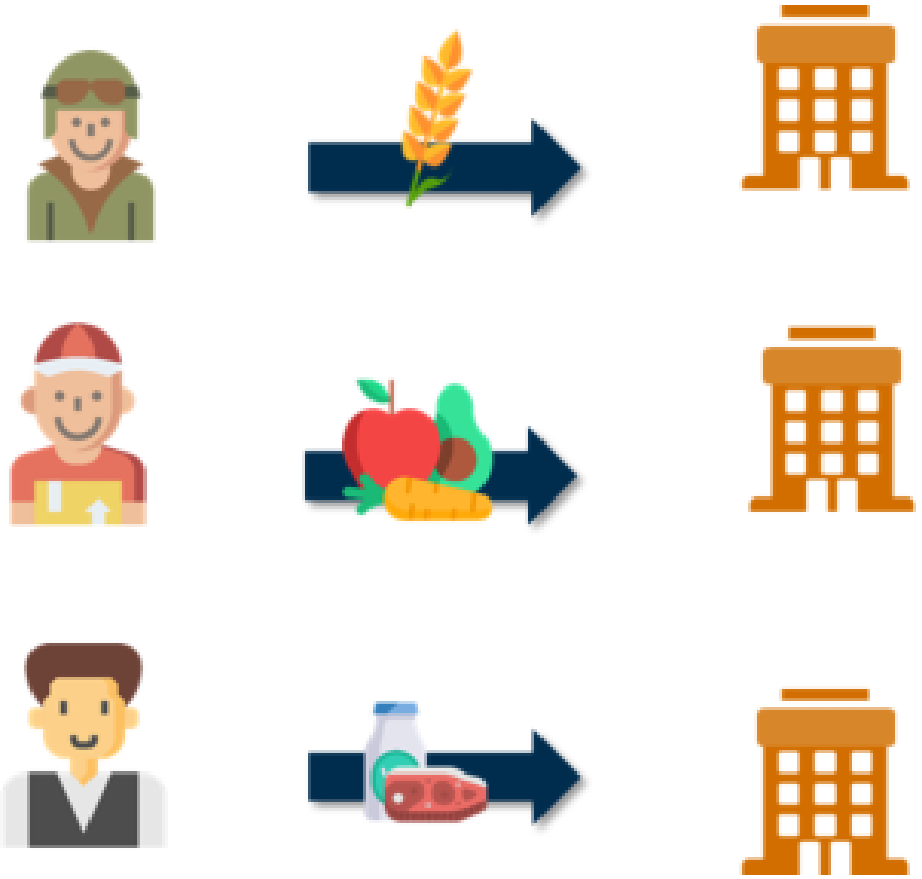
Analogi Big Data

Akhirnya, seperti cara Tim mengadopsi pendekatan penyimpanan barang terdistribusi, sistem penyimpanan data juga dapat didistribusikan, dan dengan melakukan itu, data disimpan dalam basis data individual..



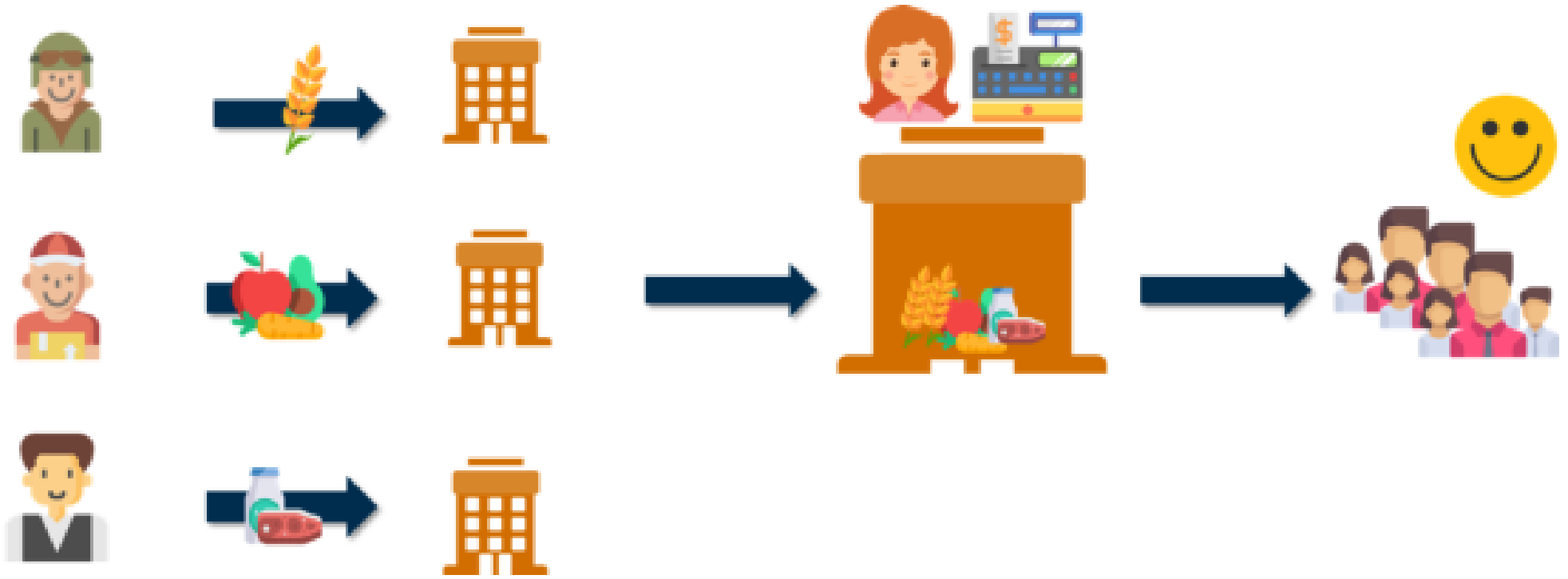
Analogi Big Data

Kisah ini membantu Kita memahami bagaimana dua komponen utama dari Hadoop: HDFS dan MapReduce. HDFS mengacu pada ruang penyimpanan terdistribusi.



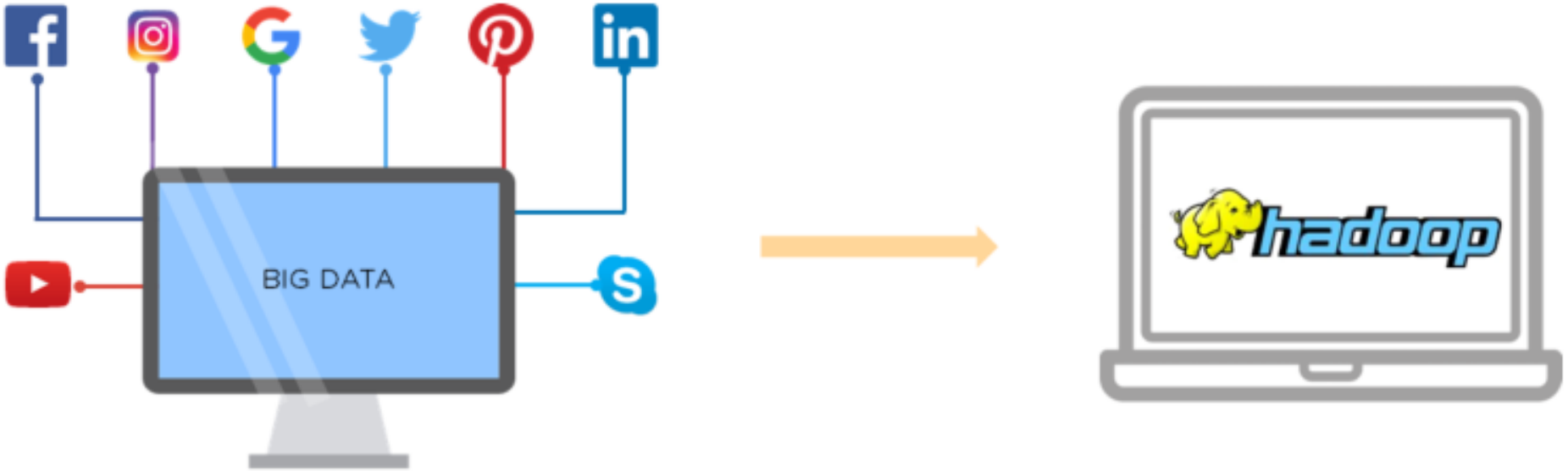
Analogi Big Data

MapReduce, di sisi lain, analog dengan bagaimana setiap orang mengurus bagian yang terpisah, dan pada akhirnya pelanggan pergi ke kasir untuk penagihan akhir. Ini mirip dengan fase pengurangan (*reduce*).



Hadoop

Hadoop adalah kerangka kerja yang menyimpan dan memproses data besar secara terdistribusi dan paralel.



Seperti yang disebutkan sebelumnya, Hadoop memiliki komponen individual untuk menyimpan dan memproses data. Mari kita pelajari lebih lanjut tentang lapisan penyimpanan Hadoop: ***Hadoop Distributed File System*** (HDFS).

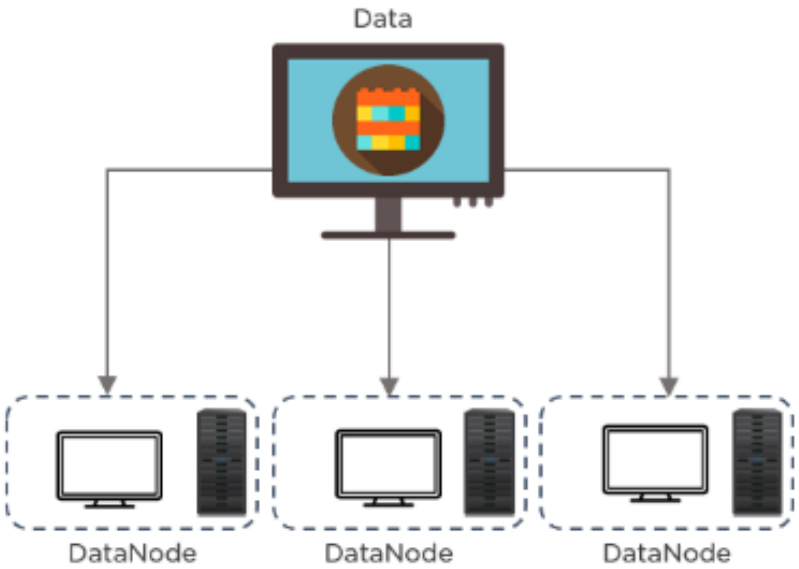
Universitas Bosowa
Teknologi Informasi

HADOOP HDFS

085242179993 Abdhi Griffindors @abdhi @abdhi_tux abdhisas.com

Hadoop HDFS

Universitas Bosowa
Teknologi Informasi



HDFS mirip dengan Google File System (GFS) karena menyimpan data di beberapa mesin.

Data direplikasi secara otomatis ke berbagai mesin untuk mencegah hilangnya data.

Dalam HDFS, data dipecah menjadi beberapa blok; masing-masing blok ini memiliki ukuran default 128 MB.

Hadoop HDFS

Jadi, bagaimana ini berbeda dari praktik penyimpanan tradisional? Perbedaannya adalah bahwa dalam sistem tradisional, semua data disimpan dalam satu basis data.

Ini dapat menjadi masalah- jika database macet (crash); semua data akan hilang. Ini juga membebani database, dan sistem demikian sangat tidak toleran terhadap kesalahan (tidak fault tolerance).

Masalah ini diatasi oleh HDFS dengan mendistribusikan data di antara beberapa mesin. Ini dirancang khusus untuk menyimpan kumpulan data besar-besaran di perangkat keras komoditas yang berarti kita dapat memiliki banyak mesin untuk digunakan (diskalakan).

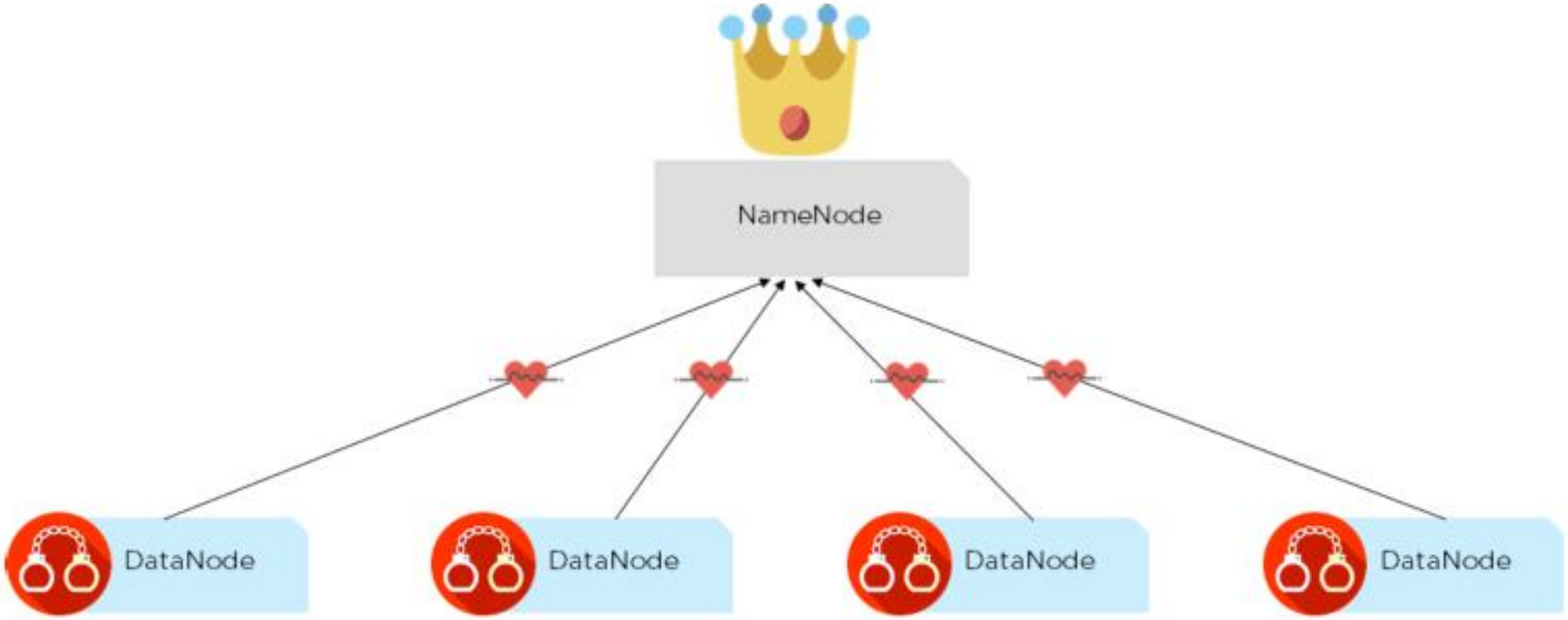
Hadoop HDFS

HDFS memiliki dua komponen yang berjalan di banyak mesin. yaitu:

1. NameNode. NameNode adalah master dari lapisan penyimpanan HDFS. Mesin ini menyimpan semua metadata. Jika mesin tempat proses NameNode lumpuh, gugus (cluster) tidak akan tersedia.

2. DataNode. DataNodes dikenal juga sebagai slave node. Mesin-mesin ini menyimpan data aktual, dan mereka melakukan operasi baca / tulis. FS

Pada dasarnya, NameNode mengelola semua DataNodes. Sinyal yang dikenal sebagai detak jantung (heartbeats) dikirim secara periodik oleh DataNodes ke NameNode untuk memberikan pembaruan status.

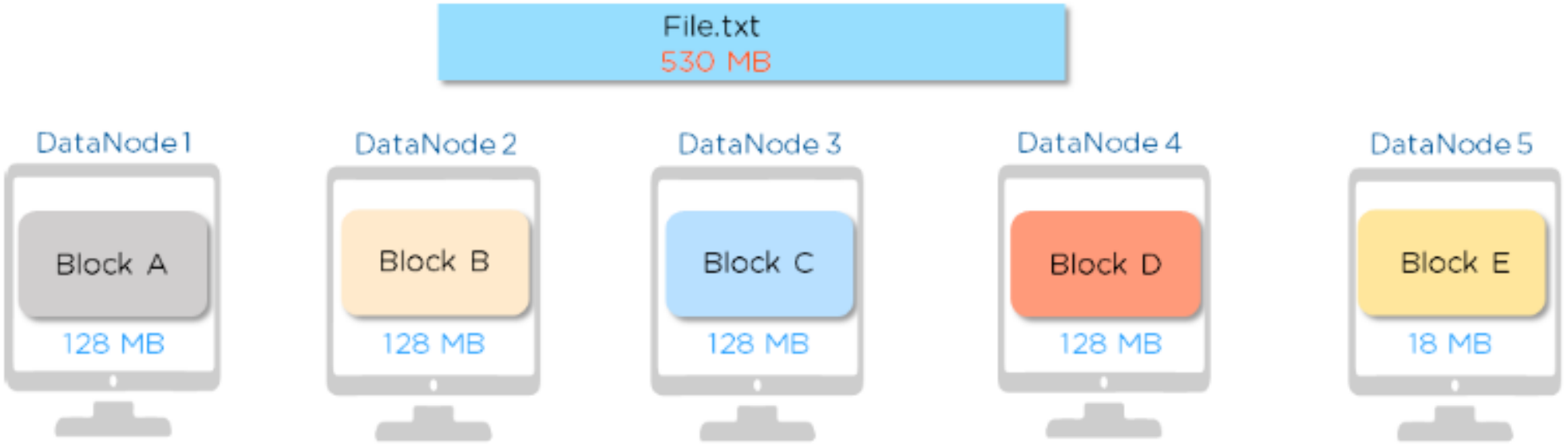

NameNode
DataNode
DataNode
DataNode
DataNode

Hadoop HDFS

Sekarang, mari kita lihat bagaimana data dipecah (*split*) dalam HDFS.

Seperti yang dapat terlihat dari contoh di atas, kita memiliki file berukuran 530 MB. File ini tidak akan disimpan apa adanya; tetapi akan dipecah menjadi lima blok yang berbeda (ukuran blok default 128MB). Blok terakhir hanya akan menggunakan ruang yang tersisa untuk penyimpanan. Blok data disimpan dalam beberapa DataNodes yang pada dasarnya hanya komputer.

Hadoop HDFS

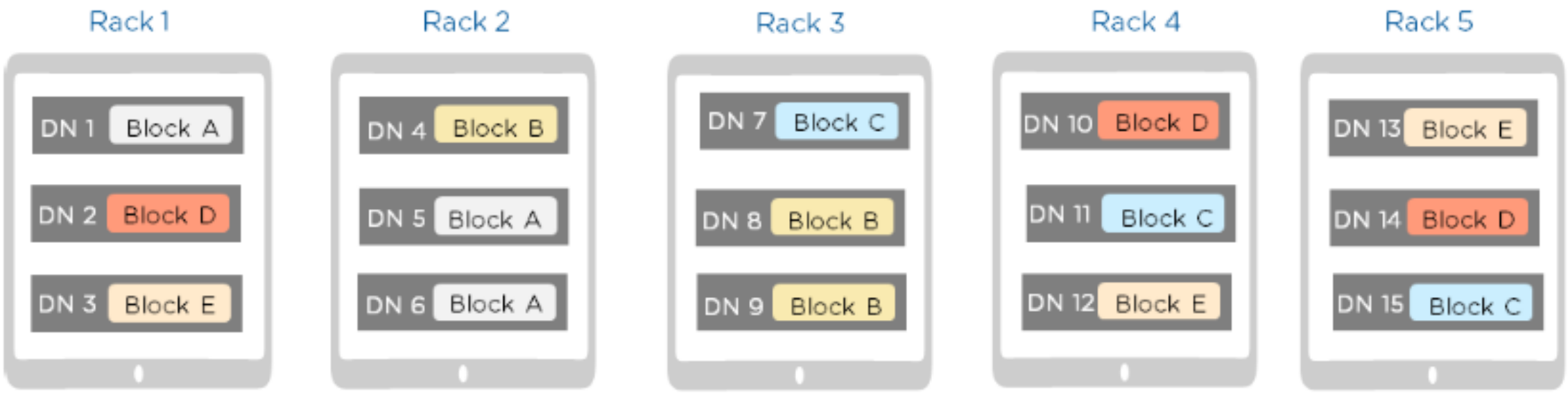


Jadi, apa yang terjadi jika komputer yang berisi Blok A lumpuh? Apakah kita akan kehilangan data kita? TIDAK: itulah keindahan HDFS; salah satu fitur utama HDFS adalah replikasi data.

HDFS membuat salinan data di beberapa mesin, dan dengan cara ini jika komputer yang memegang blok A *crash*, data kita akan aman pada komputer yang lain. Faktor replikasi default dalam HDFS adalah tiga. Ini berarti secara keseluruhan kita memiliki tiga salinan dari setiap blok data.

Hadoop HDFS

Mari kita lihat lebih dekat konsep dari replikasi. Konsep Kesadaran Rak (*Rack Awareness*) membantu untuk memutuskan di mana replika blok data harus disimpan. Di sini, rak mengacu pada koleksi 30-40 DataNodes. Menurut aturan replikasi, blok data dan salinannya tidak dapat disimpan pada DataNode yang sama.



Dari gambar di atas, dapat dilihat bahwa kita memiliki Blok A di rak 1 dan rak 2.

Hadoop HDFS

Sudah menjadi aturan, kita tidak dapat memiliki blok dan replikanya semua berada di rak yang sama. Namun, sebagai contoh dari Blok D, juga tidak ideal untuk memiliki blok yang tersebar di semua rak, karena akan meningkatkan kebutuhan bandwidth.

Oleh karena itu, jika cluster yang dibangun bersifat sadar rak, sebaran yang bagus adalah seperti terlihat pada Blok A dan Blok B.

Di sini, blok data tidak semua disimpan pada rak yang sama, karena itu jika satu rak crash maka kita tidak kehilangan data karena kita memiliki salinan di rak lain.

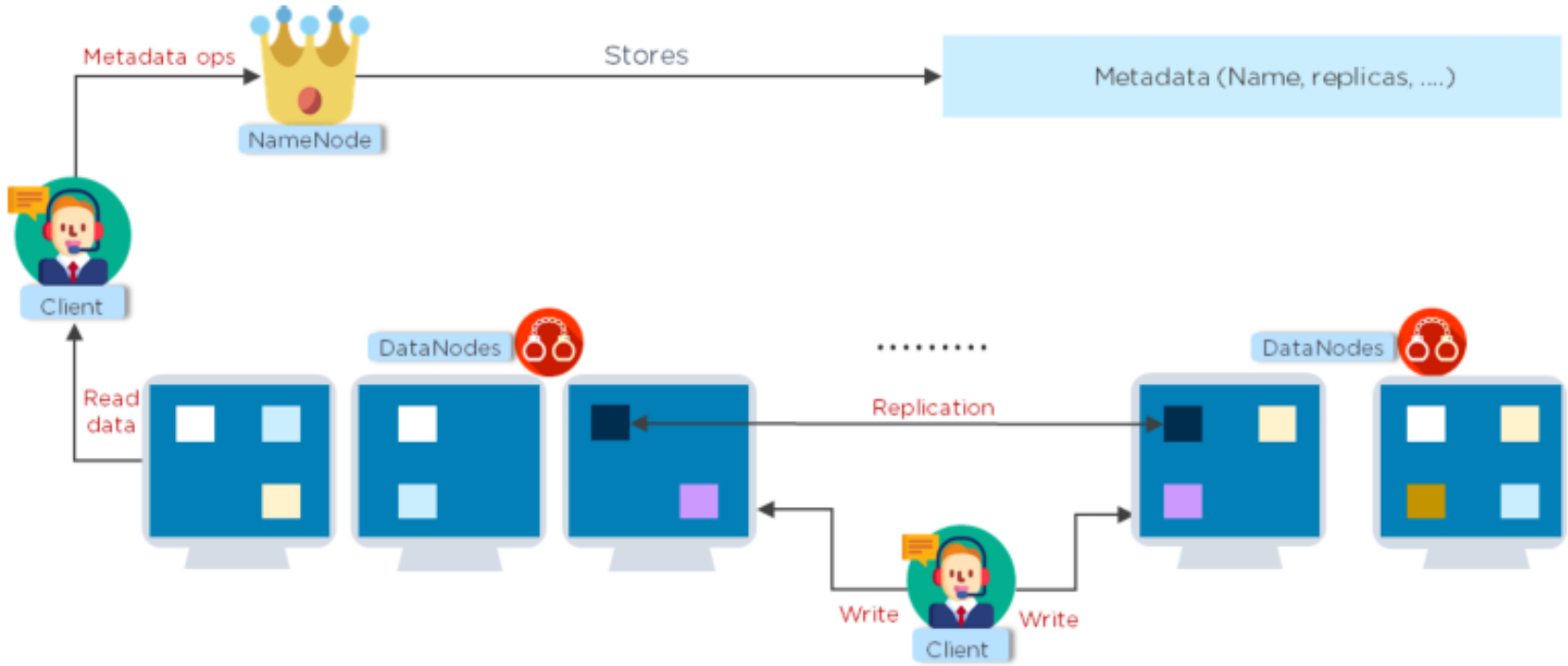
Inilah cara HDFS menyediakan toleransi kesalahan.

Harus diingat bahwa ukuran blok dan faktor replikasi dapat disesuaikan.

Hadoop HDFS

Universitas Bosowa
Teknologi Informasi

Sekarang mari kita beralih ke arsitektur HDFS. Gambar di bawah ini menunjukkan bagaimana HDFS beroperasi. Kita memiliki NameNode, DataNodes, dan permintaan klien (request).



Pada dasarnya sistem file HDFS menyediakan dua operasi, yaitu operasi membaca (*read*) dan menulis (*write*).

Fitur Unggulan Hadoop HDFS

1. HDFS toleran terhadap kesalahan (fault tolerance) karena ada banyak salinan data yang dibuat
2. HDFS menyediakan enkripsi ujung ke ujung (end-to-end encryption) untuk melindungi data pengguna
3. Dalam HDFS, beberapa node dapat ditambahkan ke cluster tergantung pada kebutuhan
4. Hadoop HDFS fleksibel dalam menyimpan semua jenis data, seperti data terstruktur, semi-terstruktur atau tidak terstruktur

Kesimpulan

Sekarang semua data disimpan dalam HDFS, langkah selanjutnya adalah mengolahnya untuk mendapatkan informasi yang bermakna. Untuk menyelesaikan pemrosesan, kita menggunakan Hadoop MapReduce.

Ada Pertanyaan ?

Ayo Diskusi, dan
Belajar Bersama.

085242179993 Abdhi Griffindors @abdhi @abdhi_tux abdhisas.com

Universitas Bosowa
Teknologi Informasi

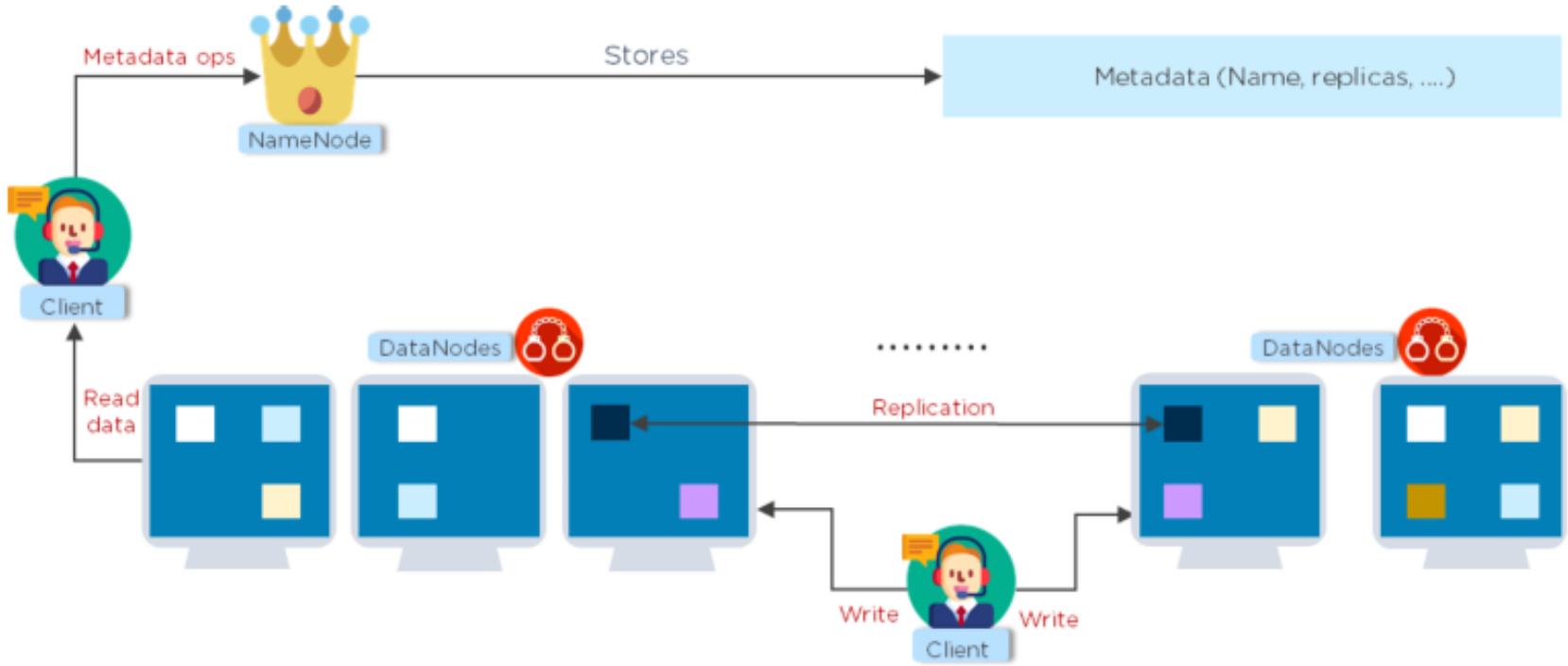
Terima Kasih

085242179993 Abdhi Griffindors @abdhi @abdhi_tux abdhisas.com

Universitas Bosowa
Fakultas Teknik - Prodi Teknologi Informasi

Tugas

Berdasarkan pemahaman anda tentang Hadoop HDFS, jelaskan cara kerja gambar dibawah ini menurut pendapat anda masing-masing



Pada dasarnya sistem file HDFS menyediakan dua operasi, yaitu operasi membaca (*read*) dan menulis (*write*).