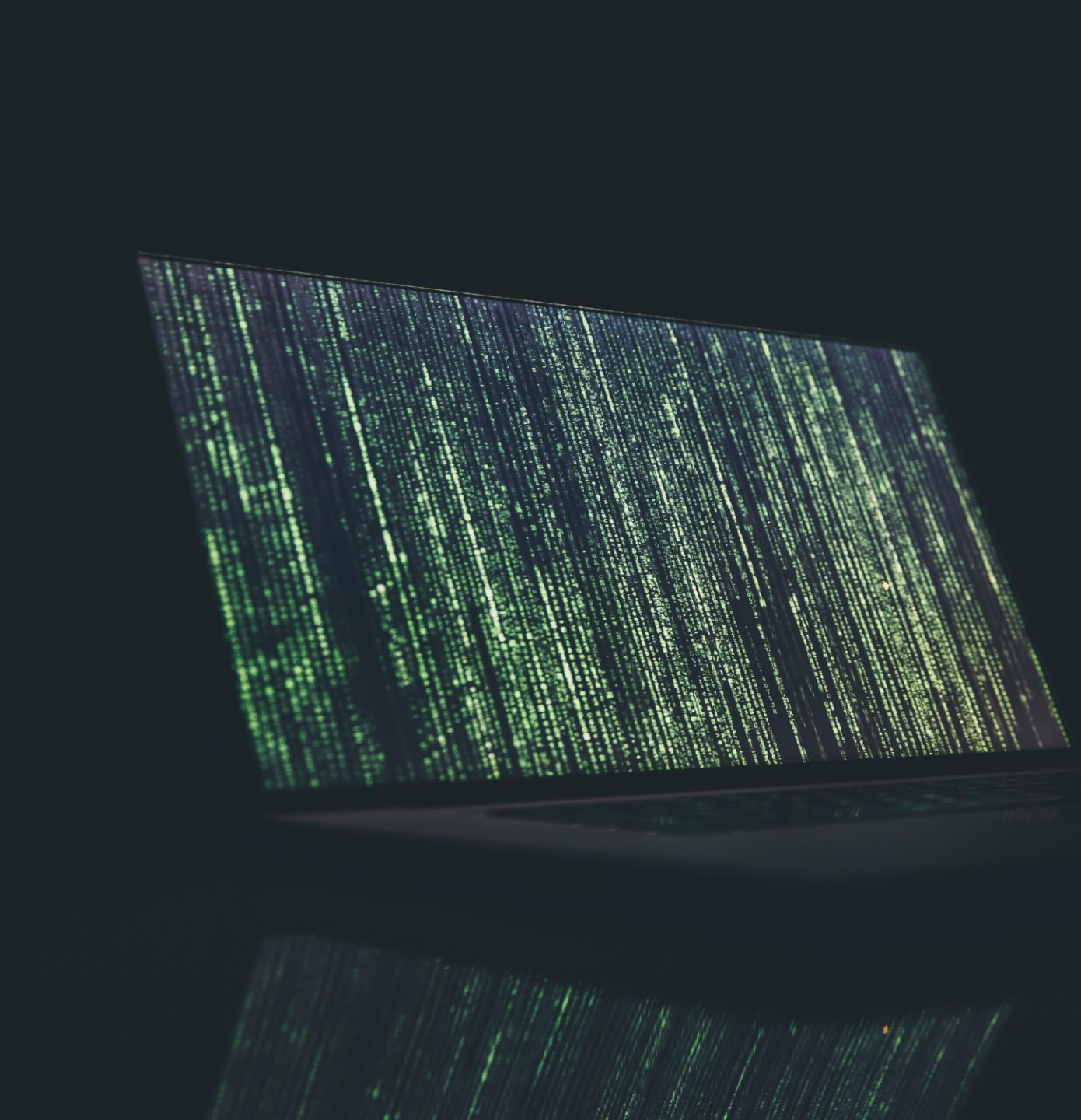


PERTEMUAN 3

PENGENALAN DAN IMPLEMENTASI INFRASTRUKTUR BIG DATA

Mata kuliah

Infrastruktur Dan Teknologi Big Data

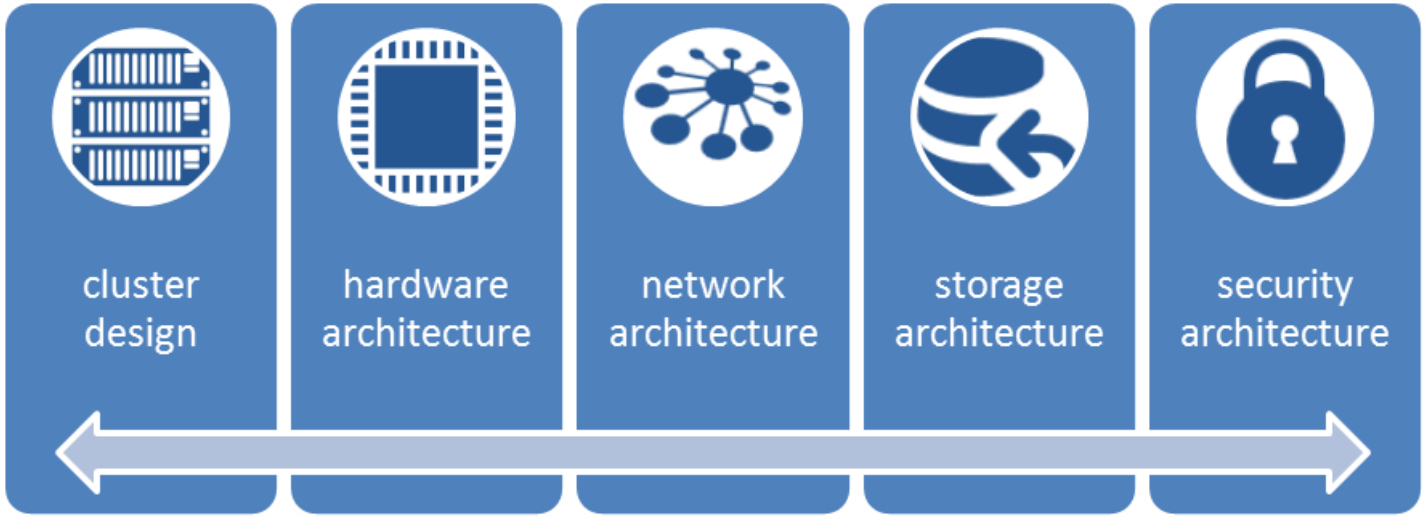
Dosen: Galih Hermawan, S.Kom., M.T.

Prodi Teknik Informatika. FTIK.

Universitas komputer indonesia

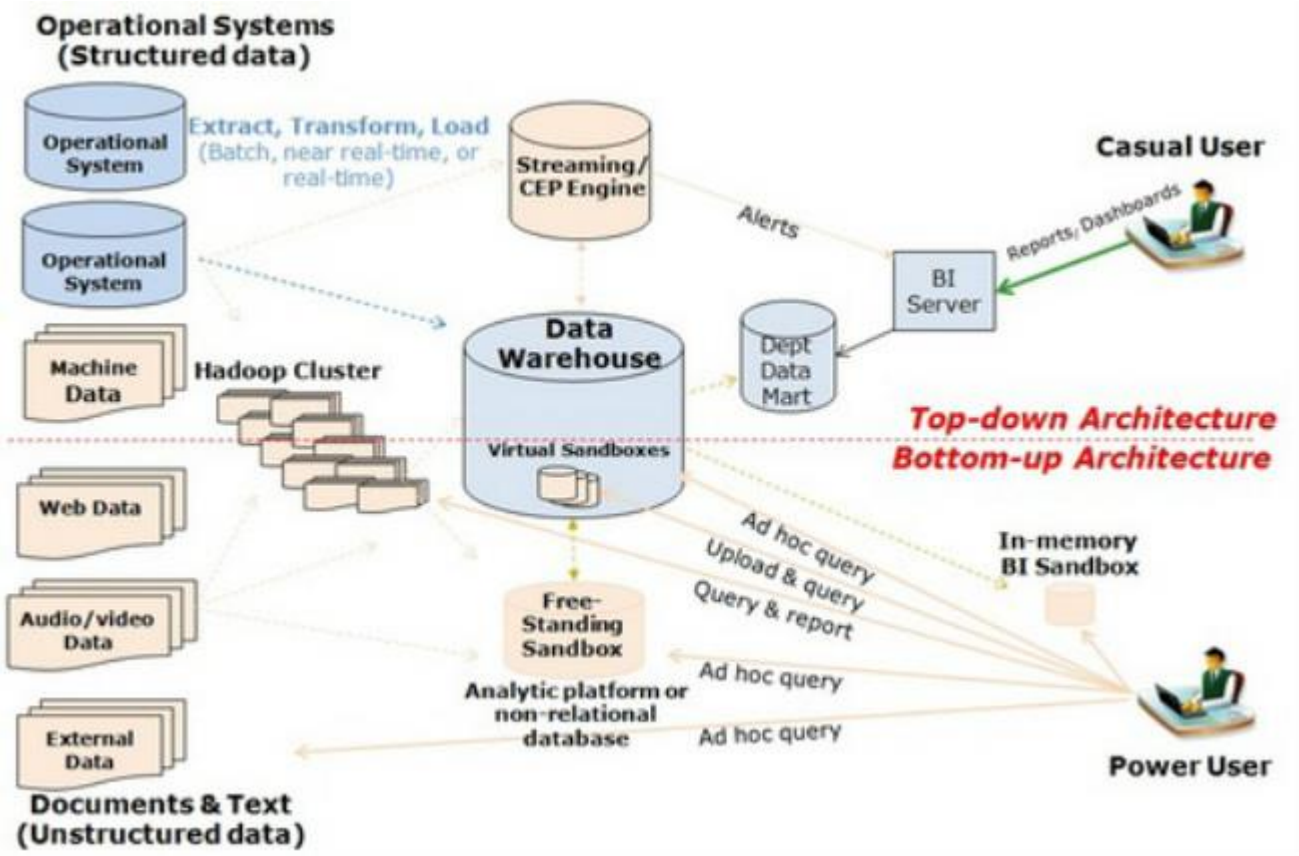
PENGANTAR INFRASTRUKTUR BIG DATA

- Infrastruktur Big Data adalah rangkaian teknologi, perangkat keras, dan perangkat lunak yang dirancang untuk mengelola, menyimpan, dan menganalisis data dalam jumlah besar.
- Tujuan → memungkinkan organisasi untuk mengambil manfaat dari data yang besar dan kompleks untuk pengambilan keputusan yang lebih baik.



Sumber gambar.
https://www.code-n.org/blog/infrastructure-building-foundations-big-data/

PERAN PENTING



- Pengelolaan dan Penyimpanan Data Masif
- Analitik Tingkat Lanjut (*Machine Learning*, Analisis *Real-Time*)
- Pengolahan Data *Real-Time* (IoT – *Internet of Things*, *Streaming*)
- Manajemen dan Orkestrasi Klaster
- Keamanan dan Kepatuhan Data
- Skalabilitas dan Fleksibilitas
- Integrasi Data dari Berbagai Sumber

Sumber gambar.
https://www.promptcloud.com/blog/big-data-for-data-driven-future/

KOMPONEN UTAMA

- Sumber Data
  - Penyimpanan data aplikasi (seperti: basis data relasional), fail statis (seperti: fail log server), sumber *real-time* (seperti: perangkat IoT), atau sumber eksternal (seperti: media sosial, *web scrapping*, open data).
- Penyimpanan Data
  - Tempat data disimpan dan didistribusikan → volume, kecepatan, varietas.
  - Contoh: HDFS – *Hadoop Distributed File Systems.*
- Pemrosesan Data
  - Tempat data diproses → ekstraksi, pembersihan, integrasi, agregasi, representasi, analisis, penjelasan, dan visualisasi data.
  - Contoh: *MapReduce*.
- Aplikasi Data
  - Pemanfaatan data → membuat keputusan, memberikan rekomendasi, mendeteksi pola/anomali, membuat prediksi.
  - Contoh: *dashboard* interaktif, laporan analitik, sistem cerdas, produk/layanan berbasis data.

KARAKTERISTIK UMUM

- **Skalabilitas** → penyesuaian diri dengan pertumbuhan dan perubahan data secara dinamis.
- **Fleksibilitas** → penanganan berbagai jenis dan sumber data dengan mudah.
- **Keandalan** → jaminan kualitas dan keamanan data serta kinerja sistem.
- **Keterbukaan** → interaksi dengan berbagai *platform* dan alat analitik lainnya.

KOMPONEN INFRASTRUKTUR BIG DATA

- *Storage* (Penyimpanan)
  - Hadoop Distributed File System (HDFS): Sistem file terdistribusi untuk menyimpan data di lingkungan Hadoop.
  - NoSQL Databases: Basis data yang dirancang untuk menangani data yang tidak terstruktur dan semi-terstruktur.
  - Data Warehouses: Sistem penyimpanan data terstruktur untuk analisis bisnis.
- *Processing* (Pemrosesan)
  - Hadoop MapReduce: Model pemrograman dan sistem eksekusi untuk memproses data besar di lingkungan Hadoop.
  - Apache Spark: Platform pemrosesan data cepat dan umum yang dapat digunakan untuk berbagai keperluan analitik.
  - Apache Flink: Sistem pemrosesan data real-time dan batch yang kuat.

KOMPONEN INFRASTRUKTUR BIG DATA (2)

- *Management* (Manajemen)
  - Apache Ambari: Alat manajemen klaster untuk proyek-proyek Apache Hadoop.
  - Cloudera Manager: Platform manajemen lengkap untuk ekosistem Hadoop.
  - Kubernetes: Platform orkestrasi kontainer untuk mengelola aplikasi di lingkungan kontainer.
- *Tools* (Alat)
  - Hive: Data warehouse yang memungkinkan kueri SQL terhadap data yang disimpan di HDFS.
  - Pig: Bahasa skrip untuk memproses dan menganalisis data di lingkungan Hadoop.
  - Kafka: Platform pengiriman pesan distribusi untuk mengelola aliran data.
  - HBase: Basis data non-relasional untuk menyimpan dan mengelola data terstruktur.

PERANGKAT KERAS (HARDWARE)

- Pemilihan Server: Kapasitas, Memori, Prosesor
- Storage: SSD vs. HDD, Kapasitas Penyimpanan
- Jaringan: *Bandwidth*, Latensi

ARSITEKTUR BIG DATA

- *Single Node vs. Cluster*
  - *Single Node* → semua komponen infrastruktur dijalankan pada satu server
  - *Cluster* → semua komponen infrastruktur dijalankan pada beberapa server yang saling terhubung
- *Master-Slave Architecture*
  - *Master Node* → mengelola dan mengkoordinasikan semua node dalam cluster, serta menyimpan metadata data.
  - *Slave Node* → menyimpan dan memproses data, serta menjalankan aplikasi analisis data.
- *Shared vs. Distributed Storage*
  - *Shared storage* → penyimpanan data yang diakses oleh beberapa server atau *node* secara bersamaan.
  - *Distributed storage* → penyimpanan data yang tersebar di beberapa server atau *node* yang berbeda.

PENERAPAN INFRASTRUKTUR BIG DATA

- Instalasi dan Konfigurasi: Hadoop, Spark, dll.
- Manajemen Sumber Daya:
  - *Cluster Scaling* → proses menambah atau mengurangi node dalam cluster Big Data
  - *Monitoring* → proses mengumpulkan data tentang kinerja infrastruktur Big Data
- *Data Ingestion* (Penyerapan Data)*:*
  - *Data Loading* → proses pemuatan data dari sumber data ke sistem penyimpanan data terpusat.
  - ETL (*Extract, Transform, Load*) → proses mengekstrak data dari sumber data, mentransformasi data ke format yang sesuai dengan sistem penyimpanan data terpusat, dan memuat data ke sistem penyimpanan data terpusat.
  - *Data Warehousing* → proses penyimpanan data yang telah diproses dan ditransformasikan ke dalam data warehouse untuk keperluan analisis data.

SKALABILITAS DAN PERFORMA

- *Horizontal vs. Vertical Scaling*
  - *Vertical scaling*: Proses menambah atau mengurangi sumber daya pada node yang ada.
  - *Horizontal scaling*: Proses menambah atau mengurangi node baru ke dalam cluster.
- Pembagian Tugas dan Paralelisasi
- Penyusunan Data untuk Performa Optimal

KEAMANAN DAN KEPATUHAN

- Proteksi Data: *Encryption, Access Control*
  - Proses melindungi data dari akses, penggunaan, pengungkapan, gangguan, modifikasi, atau penghancuran yang tidak sah.
  - Penting dilakukan untuk melindungi data dari serangan siber dan pencurian data.
- Kepatuhan Hukum: GDPR, HIPAA, dll.
  - ***General Data Protection Regulation* (GDPR)**: GDPR adalah peraturan Uni Eropa yang mengatur perlindungan data pribadi. GDPR berlaku untuk semua organisasi yang memproses data pribadi warga negara Uni Eropa, terlepas dari lokasi organisasi tersebut.
  - ***Health Insurance Portability and Accountability Act* (HIPAA):** HIPAA adalah undang-undang federal Amerika Serikat yang mengatur perlindungan data kesehatan. HIPAA berlaku untuk semua organisasi yang memproses data kesehatan pasien, termasuk organisasi medis, asuransi kesehatan, dan penyedia layanan kesehatan.
  - ***California Consumer Privacy Act* (CCPA)**: CCPA adalah undang-undang negara bagian California yang mengatur perlindungan data pribadi konsumen. CCPA berlaku untuk semua organisasi yang mengumpulkan atau menjual data pribadi konsumen California.
  - ***Personal Information Protection and Electronic Documents Act* (PIPEDA)**: PIPEDA adalah undang-undang federal Kanada yang mengatur perlindungan data pribadi. PIPEDA berlaku untuk semua organisasi yang memproses data pribadi warga negara Kanada.
  - ***Personal Information Protection Act* (PIPA)**: PIPA adalah undang-undang negara bagian New York yang mengatur perlindungan data pribadi. PIPA berlaku untuk semua organisasi yang memproses data pribadi warga negara New York.
  - Indonesia → Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi (UU PDP).

TANYA JAWAB

Terima Kasih