Indra Aulia
Program Studi Teknologi Informasi
Fakultas Ilmu Komputer dan Teknologi Informasi
Universitas Sumatera Utara

Langkah-Langkah Merancang Arsitektur Big Data

Pendahuluan

- Arsitektur big data merupakan suatu struktur logis dan/atau fisik yang menangani seberapa besar data akan disimpan, diakses dan dikelola dalam suatu big data atau lingkungan TI.
- Dalam rangka mendefininsikan seberapa besar solusi big data, kita akan bekerja berdasarkan komponen inti yang digunakan, arus informasi, keamanan dan lainnya.
- Arsitektur big data ini akan menjadi referensi dalam merancang infrastruktur big data dengan solusi-solusinya.

Pendahuluan

- Big data dapat disimpan, diperoleh, diproses dan dianalisa dengan berbagai cara dari berbagai sumber.
- Big data yang disimpan dan diproses akan juga melibatkan dimensi tambahan (tata kelola, keamanan dan kebijakan).

Tipe Big Data

- Merancang arsitektur big data merupakan aktivitas yang kompleks.
- Oleh karena itu, sebelum merancang referensi arsitektur big data, ***langkah yang paling utama adalah mengidentifikasi apakah skenario bisnis tergolong dalam masalah big data atau bukan.***

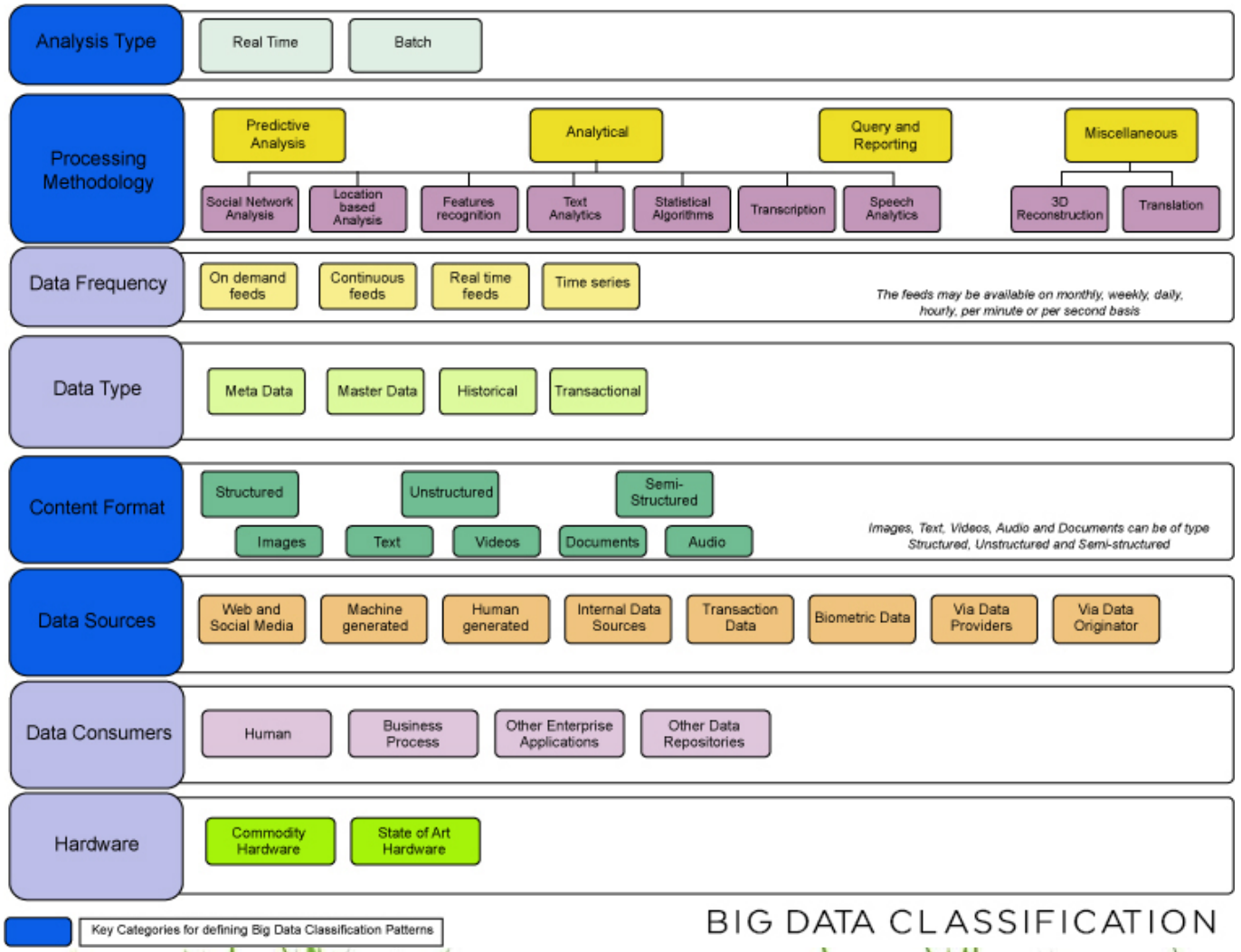
Tipe Big Data

- Merancang arsitektur big data dapat diawali dengan ***mengkategorikan masalah yang telah diidentifikasi ke dalam beberapa tipe***.
- Tujuan mengkategorikan tersebut: *untuk menentukan karakteristik masing-masing tipe big data.*

Tipe Big Data

- Tipe big data dapat dikategorikan sebagai berikut:
  - *Machine-Generated Data*
  - *Web and Social Data*
  - *Transaction Data*
  - *Human Generated*
  - *Biometrics*

Klasifikasi Karakteristik Big Data berdasarkan Tipe Big Data



Tipe Analisis

- Analisis Real Time atau Analisis Batch
- Penting karena mempengaruhi penentuan keputusan terkait produk, perangkat, sumber data dan frekuensi data yang diharapkan.

Tipe Analisis

- **Analisa Real Time** adalah Suatu proses analisis yang membutuhkan input secara kontinu, pemrosesan konstan dan output data yang jelas.
- **Spark** adalah suatu tool yang baik untuk digunakan dalam pemrosesan real time.
- **Near real-time** akan membuat kecepatan penting, namun waktu pemrosesan dalam hitungan menit dapat diterima sebagai pengganti detik.

  Contoh: Gabungan pengolahan data dan Complete Event Processing (CEP) pada Operational Intelligence.

  CEP berguna untuk mengidentifikasi peluang di kumpulan data (seperti prospek penjualan) serta ancaman (mendeteksi penyusup di jaringan).

Tipe Analisis

- **Analisa Batch** adalah Pemrosesan terkadang membutuhkan waktu berjam-jam atau berhari-hari.
- Melibatkan 3 (tiga) proses terpisah.
  - Data dikumpulkan dalam jangka waktu tertentu.
  - Data diproses oleh program yang terpisah.
  - Menghasilkan output berdasarkan pengolahan data.
- Contoh data yang dimasukkan untuk analisis dapat mencakup data operasional, data historis dan arsip, data dari media sosial, data layanan, dll.
- **MapReduce** adalah alat yang berguna untuk pemrosesan batch.
- Contoh penggunaan untuk pemrosesan batch mencakup kegiatan penggajian dan penagihan.

Metodologi Pengolahan

- Jenis teknik yang akan diterapkan untuk pengolahan data.
- Metodologi yang dipilih membantu dalam memilih Alat dan Teknik yang sesuai untuk Solusi big data.

Frekuensi dan Ukuran Data

- Jumlah data dan kecepatan perolehannya.
- Karakteristik data ini membantu dalam menentukan mekanisme penyimpanan, format dan alat pra-pemrosesan.
- Ukuran dan Frekuensi bervariasi untuk sumber data yang berbeda:
  - On Demand - Data Media Sosial
  - Continuous Feed / Real Time - Data Cuaca, Data Transaksional
  - Time Series - Data Berbasis Waktu

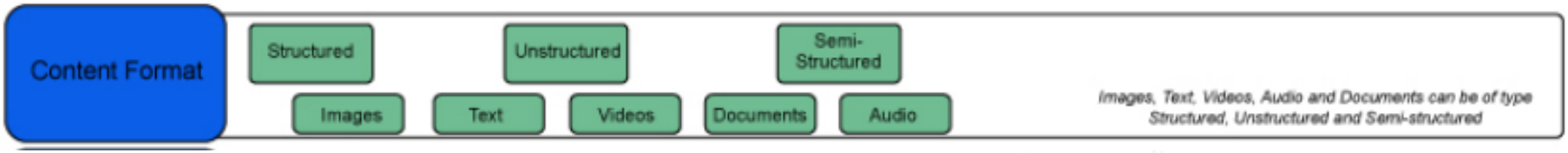
Tipe Data

- Jenis Data yang akan diolah.

  Mengetahui tipe data membantu dalam pemisahan data dalam penyimpanan.

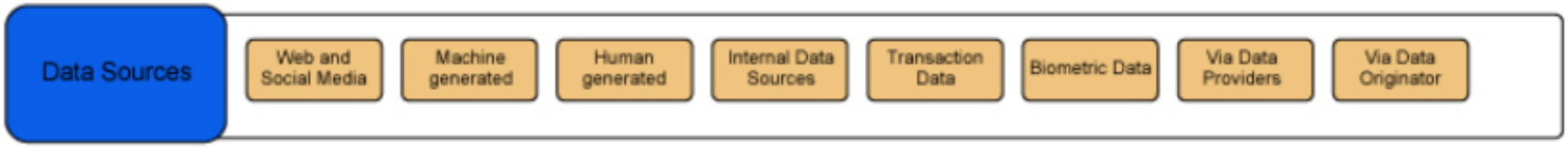
Format Konten

- Format memberitahu kita tentang bagaimana data yang masuk perlu diproses dan alat dan teknik apa yang harus digunakan.
- Format bisa berupa Structured (RDBMS) atau Un-Structured (Audio, Video, Images) atau Semi-Structured.



Sumber Data

- Mengidentifikasi Sumber Data sangat penting dalam menentukan lingkup dari perspektif bisnis. Misalnya. Media Web dan Sosial, mesin yang dihasilkan, buatan manusia dll.



Data Konsumen

- Daftar kemungkinan konsumen data yang diproses
  - Proses bisnis
  - Pengguna Bisnis
  - Aplikasi Enterprise
  - Individu dalam Berbagai Peran Bisnis
  - Bagian dari proses mengalir
  - Repositori data atau aplikasi enterprise lainnya

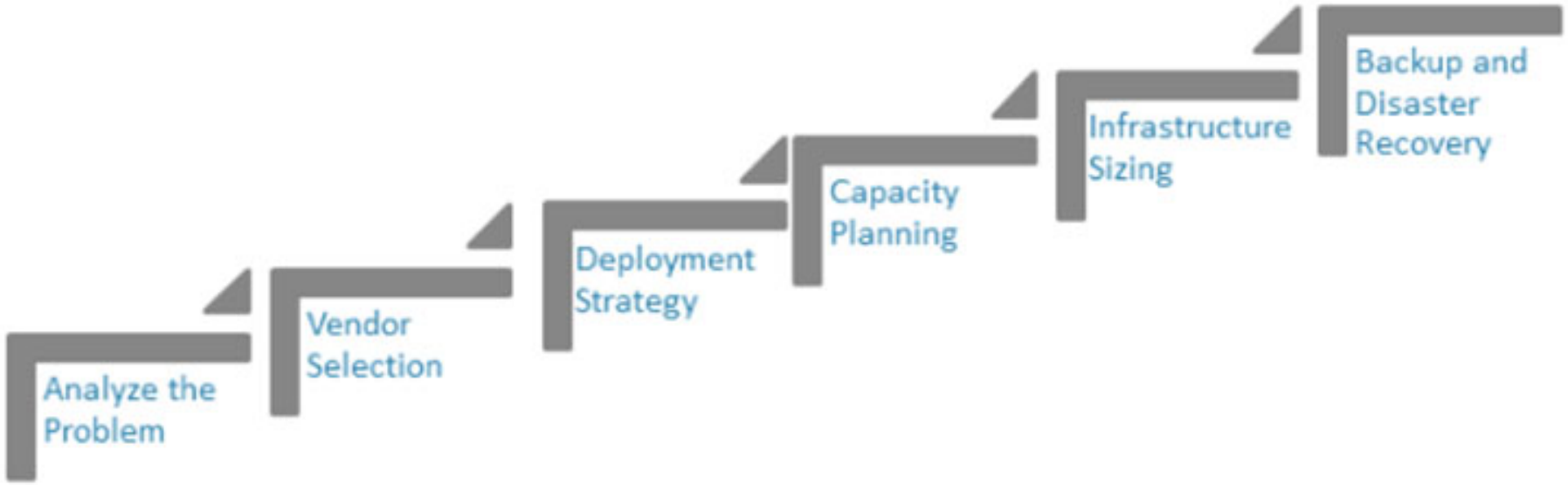
Hardware

- Hardware dimana Big Data Solution akan diimplementasikan
- Memahami keterbatasan perangkat keras membantu menginformasikan pilihan Big Data Solution

Langkah-Langkah Dasar Merancang Arsitektur Big Data

- Setalah menganalisa dan mengetahui skenario big data, karakternya dan pola big data di suatu perusahaan, maka kita dapat memulai dengan merancang Referensi Arsitektur Big Data.
- Langkah-langkahnya terdiri atas 6 dasar langkah yakni:
  - *Analyze the Problem*
  - *Vendor Selection*
  - *Deployment Strategy*
  - *Capacity Planning*
  - *Infrastructure Sizing*
  - *Backup and Disaster Recovery*

Langkah-Langkah Dasar Merancang Arsitektur Big Data



Analyze the Problem

- Menganalisa apakah kita memerlukan Big Data Solution atau tidak, karakteristik Data dan tipe Big Data Pattern.

Vendor Selection

- Berdasarkan jenis fungsionalitas yang harus kita capai melalui alat bantu.
- Ada banyak vendor di pasar dengan berbagai alat yang handal untuk berbagai tugas.
- Penentuannya diserahkan kepada organisasi dalam menentukan jenis alat yang ingin mereka pilih.

Deployment Strategy

- Menentukan apakah akan didasarkan pada premis, berbasis awan atau campuran keduanya.
- **Solusi premis** cenderung lebih aman, namun pemeliharaan perangkat keras akan menghabiskan banyak uang, usaha dan waktu.
- **Solusi berbasis cloud** lebih hemat biaya dalam hal skalabilitas, pengadaan dan perawatan.

Capacity Planning

- Mengevaluasi ukuran perangkat keras dan infrastruktur berdasarkan faktor-faktor:
  - *Volume Data untuk One-Historical Load*
  - *Volume konsumsi data harian*
  - *Periode retensi data*
  - *Replikasi Data untuk Data Penting*
  - *Periode waktu dimana cluster berukuran, setelah itu cluster diberi skala horizontal*
  - *Penyebaran Multi Datacenter*

Infrastructure Sizing

- Kesimpulan dari langkah sebelumnya membantu dalam perencanaan infrastruktur seperti jenis perangkat keras yang dibutuhkan.
- Faktor Penting yang harus dipertimbangkan:
  - *Jenis pengolahan Memori atau I / O intensif*
  - *Jenis Disk*
  - *Tidak ada disk per mesin*
  - *Ukuran Memori Ukuran HDD*
  - *Tidak ada CPU dan core*
  - *Data disimpan dan disimpan di setiap lingkungan*

Backup and Disaster Recovery

- Bagian perencanaan yang sangat penting
- Pertimbangan:
  - *Kekritisan data yang tersimpan*
  - *Persyaratan RPO (Recovery Point Objective) dan RTO (Recovery Time Objective)*
  - *Pemulihan Bencana Aktif-Aktif atau Aktif-Pasif*
  - *Multi datacenter deployment*
  - *Interval Cadangan (dapat berbeda untuk berbagai jenis data)*