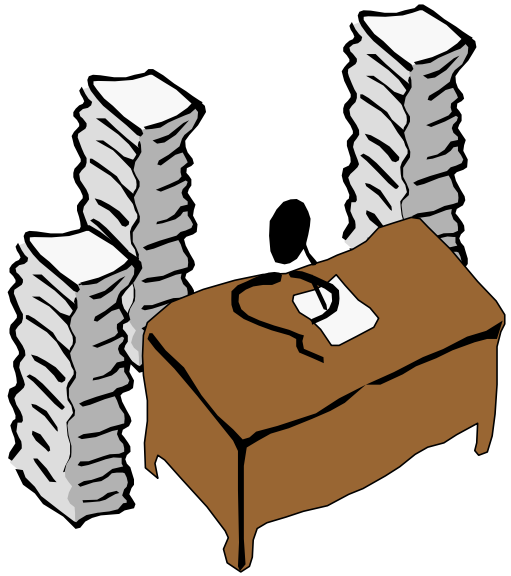
DATA LAKE

Pertemuan 1

Data is everywhere yet ...

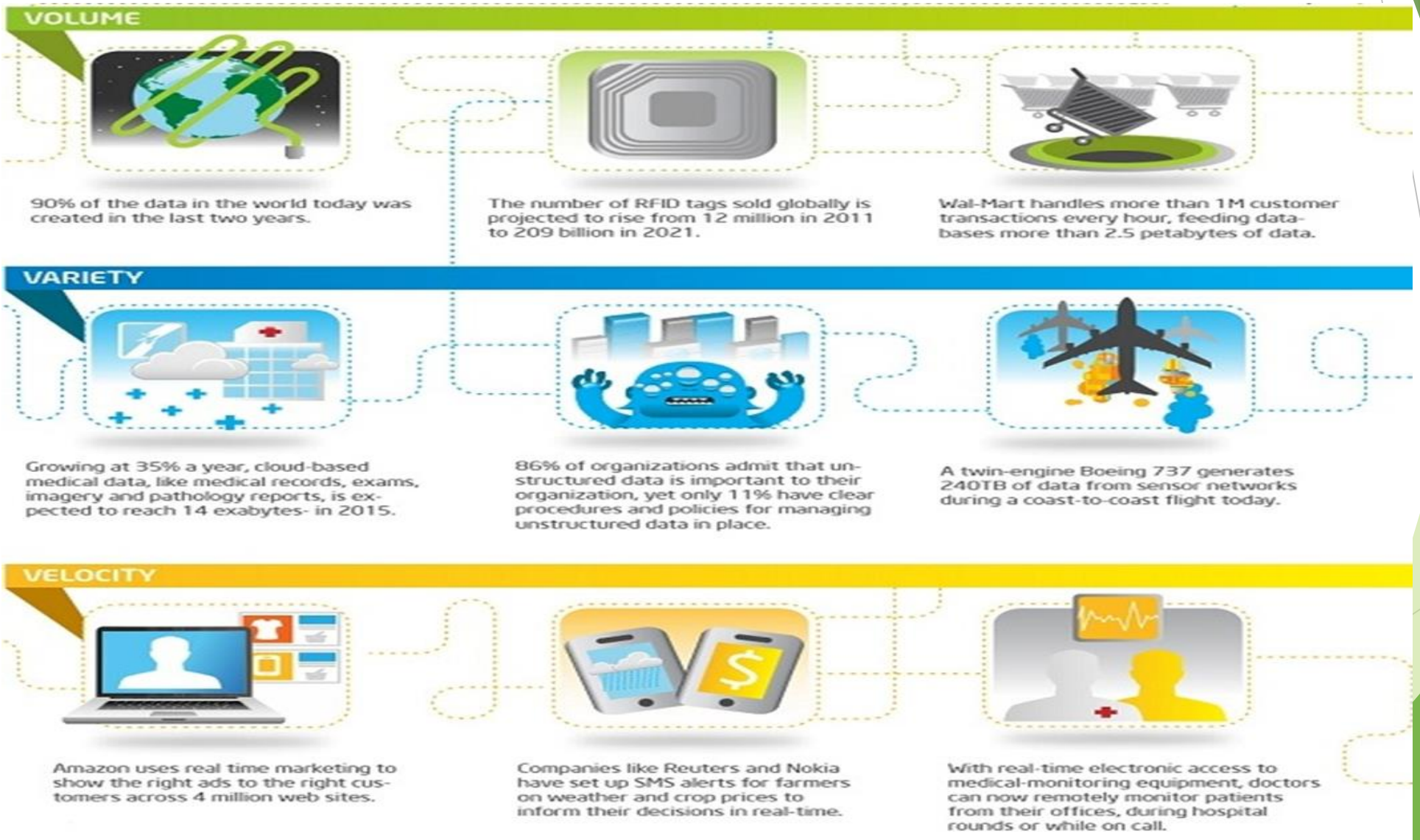


- Saya **tidak bisa menemukan** data yang saya cari
  - data tersebar dimana-mana (lintas jaringan)
  - menggunakan versi yang berbeda
- **Tidak bisa mendapatkan** data yang diperlukan
  - perlu orang yang expert untuk mendapatkan data tersebut
- Data sudah ditemukan, tapi **tidak mengerti maksud** data tersebut
  - dokumentasi data yang kacau
- Data sudah ditemukan, tapi saya **tidak bisa menggunakannya**
  - hasil data yang tidak terduga
  - data perlu ditransformasi dari bentuk satu ke bentuk yang lain

KONSEP

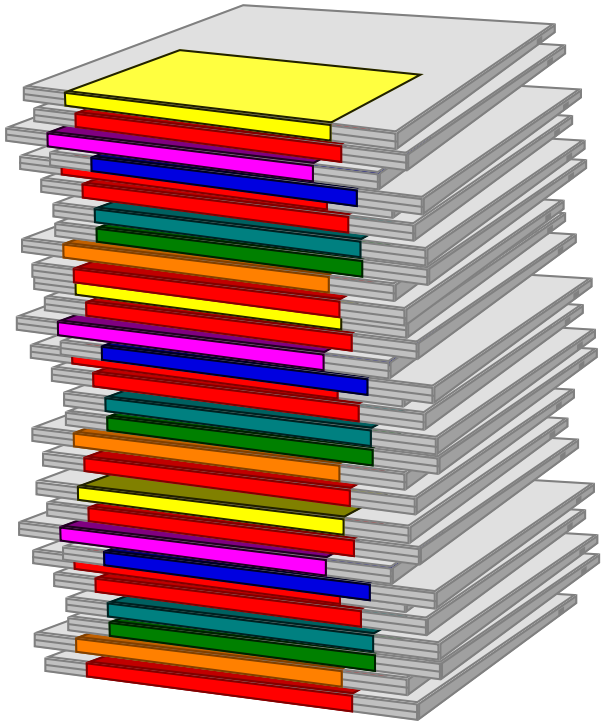
- Big Data
- Data Warehouse
- Data Lake

Big Data


VOLUME
90% of the data in the world today was created in the last two years.
The number of RFID tags sold globally is projected to rise from 12 million in 2011 to 209 billion in 2021.
Wal-Mart handles more than 1M customer transactions every hour, feeding data-bases more than 2.5 petabytes of data.
VARIETY
Growing at 35% a year, cloud-based medical data, like medical records, exams, imagery and pathology reports, is ex-pected to reach 14 exabytes- in 2015.
86% of organizations admit that un-structured data is important to their organization, yet only 11% have clear procedures and policies for managing unstructured data in place.
A twin-engine Boeing 737 generates 240TB of data from sensor networks during a coast-to-coast flight today.
VELOCITY
Amazon uses real time marketing to show the right ads to the right cus-tomers across 4 million web sites.
Companies like Reuters and Nokia have set up SMS alerts for farmers on weather and crop prices to inform their decisions in real-time.
With real-time electronic access to medical-monitoring equipment, doctors can now remotely monitor patients from their offices, during hospital rounds or while on call.

Data Warehouse

basis data yang menyimpan data **sekarang** dan data **masa lalu** yang berasal dari **berbagai sistem operasional dan sumber yang lain** (sumber eksternal) yang menjadi perhatian penting bagi manajemen dalam organisasi dan ditujukan untuk keperluan **analisis** dan **pelaporan manajemen** dalam rangka **pengambilan keputusan**



Empat karakteristik Data Warehouse

- Subject Oriented
- Integrated
- Time – Variant
- Non volatile

Data Lake

- Pusat berkumpulnya data-data dalam format dan skala aslinya.
- Data dapat disimpan tanpa perlu menyusunnya dalam struktur, pengelompokan, atau hierarki tertentu.
- Data yang terdapat dalam data lake adalah data mentah yang belum diproses atau dianalisis.
- Menyimpan data dari beragam sumber.
- Data-data di dalamnya pun terdiri dari berbagai tipe dan skema.
- Berbagai macam pengguna dari mana saja dapat mengakses *data lake* dan mengambil sampel data dari dalamnya.

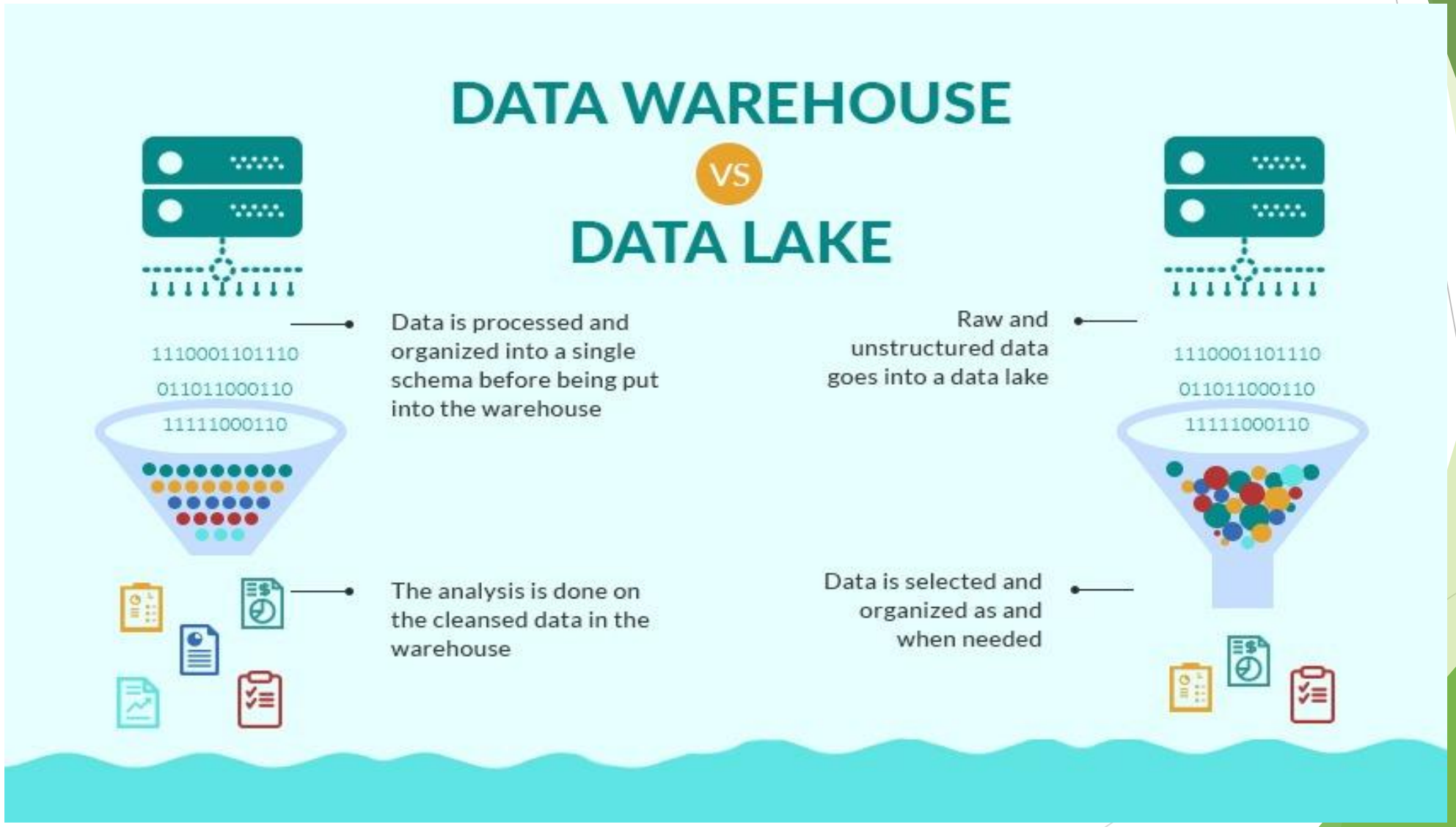
Komponen Penyusun Data Lake

- ***Data Ingestion and Storage***
- ***Data Processing***
- ***Data Analysis***
- ***Data Integration***

Mengapa Data Lake penting?

- 1. Mengindeks data
- 2. *Machine learning*
- *3*. Mengembangkan interaksi dengan konsumen
- 4. Analisis

DATA WAREHOUSE VS DATA LAKE



Key Differences Between the Data Lake and Data Warehouse

DATA WAREHOUSE	vs.	DATA LAKE
Structured, processed	DATA	Structured/semi-structured/unstructured/raw
Schema-on-write	PROCESSING	Schema-on-read
Expensive for large data volumes	STORAGE	Designed for low-cost storage
Less agile, fixed configuration	AGILITY	Highly agile, configure and reconfigure as needed
Mature	SECURITY	Maturing
Business pros	USERS	Data scientists et al.

Analysis Source: "A Big Data Cheat Sheet: What Marketers Want to Know" by Tamara Dull