Analisis Big Data

Konsep, Arsitektur dan Algoritma

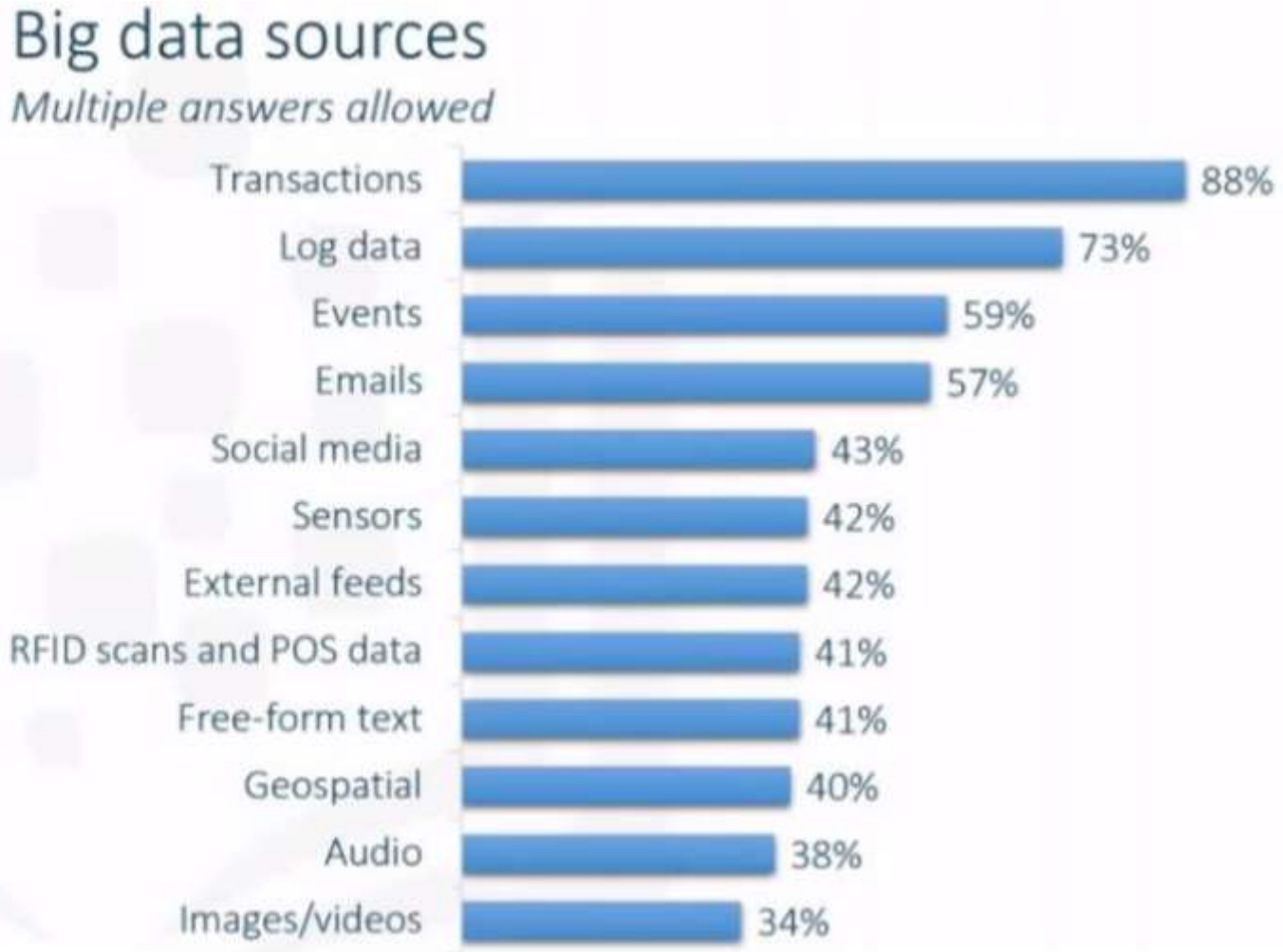
Big Data is ….

- “Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools and machines. It requires new, innovative, and scalable technology to collect, host and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management and enhanced shareholder value.” – Ernst & Young (EY)
- “Big Data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation.” – Gartner

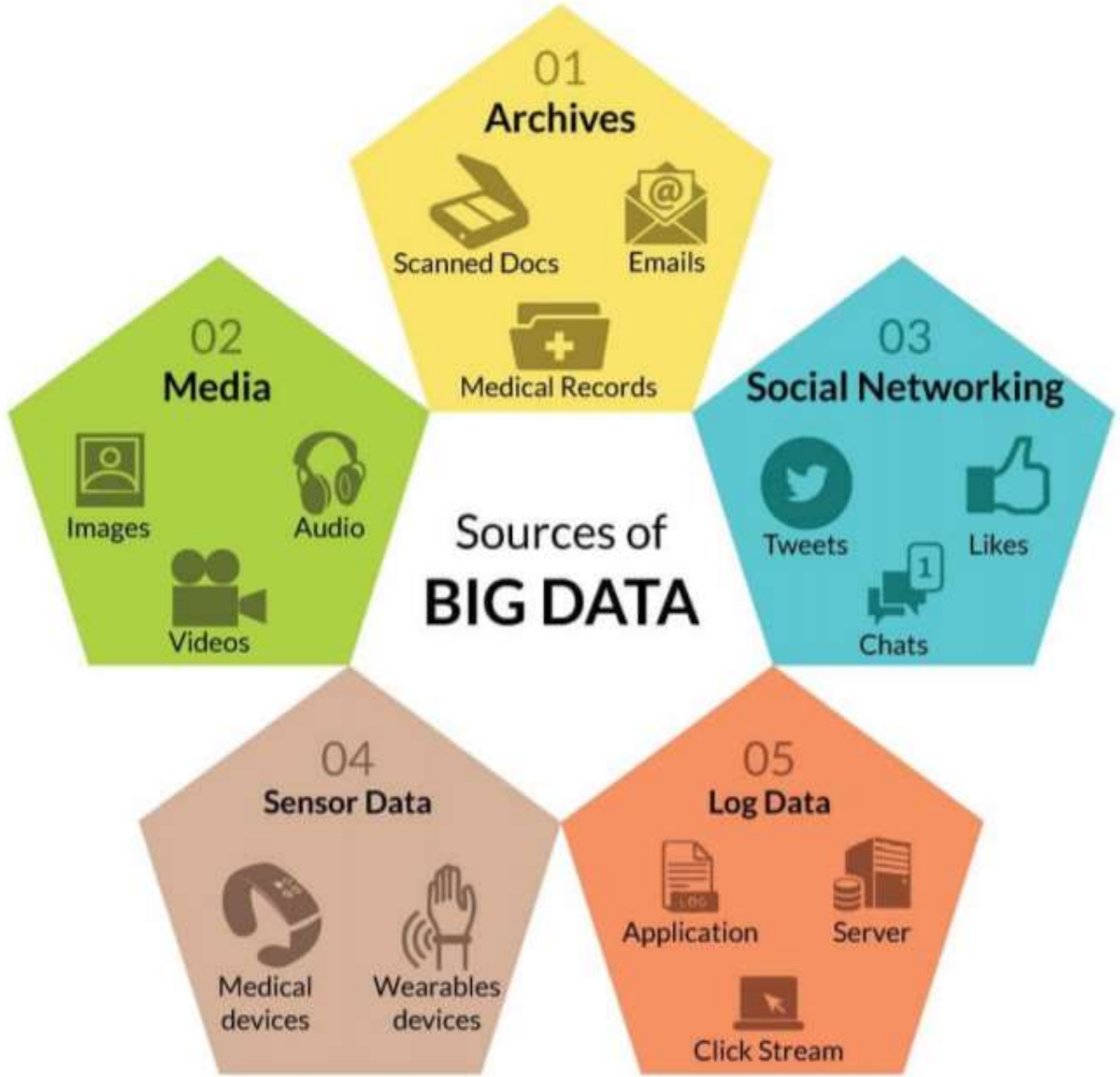
Source of Big Data

There are three major sources of Big Data:

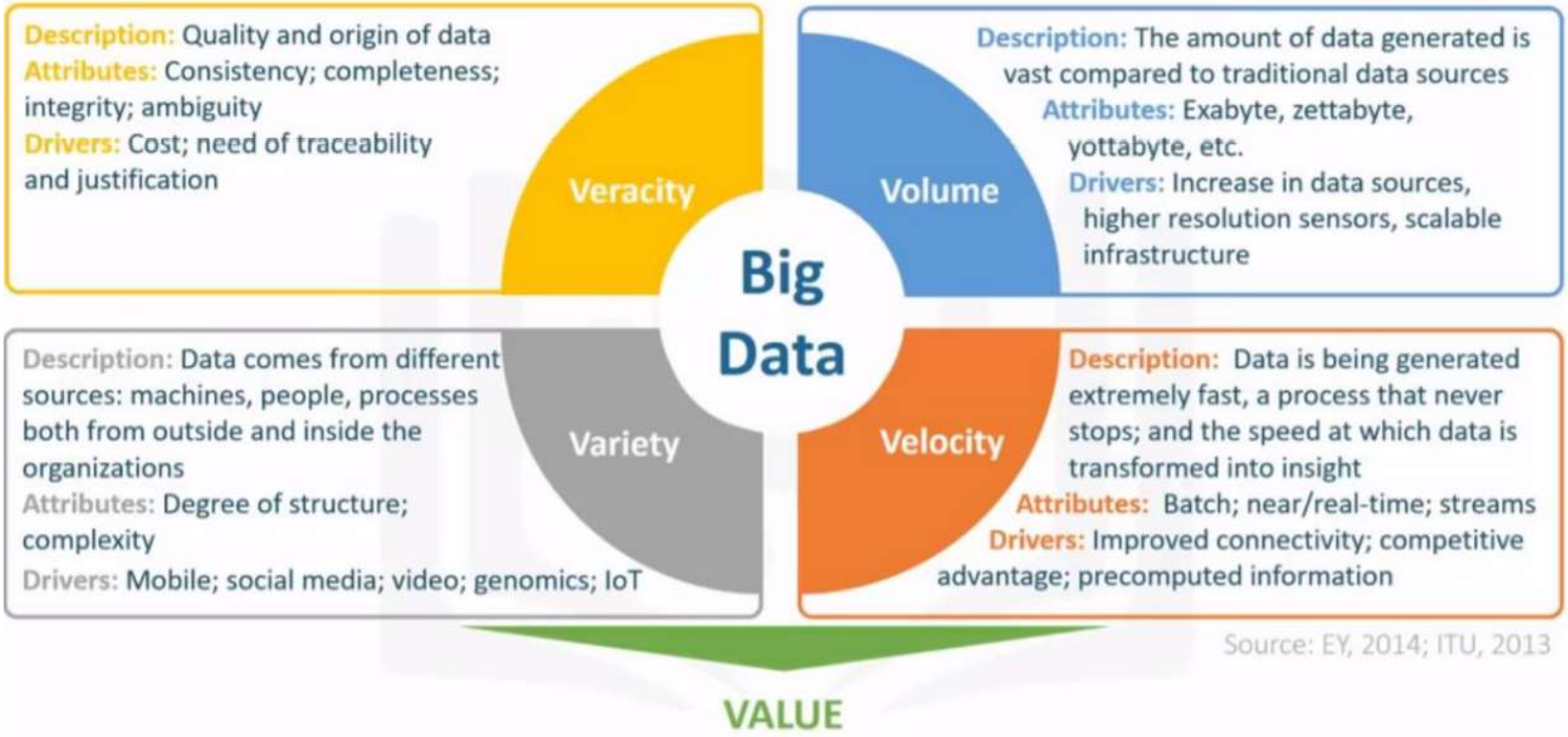
- People-generated data
- Machine-generated data
- Business-generated data



Source of Big Data



5 V's of Big Data



Source: EY, 2014; ITU, 2013

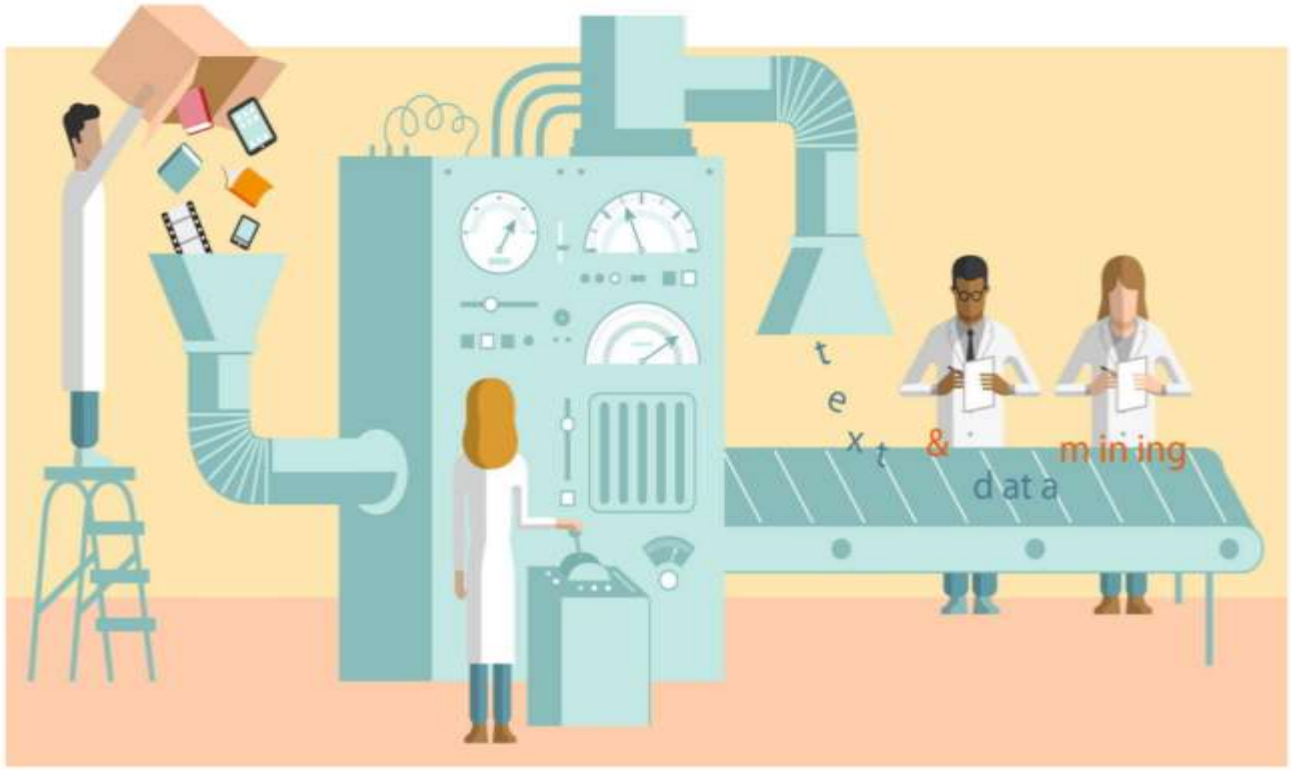
Why Big Data?

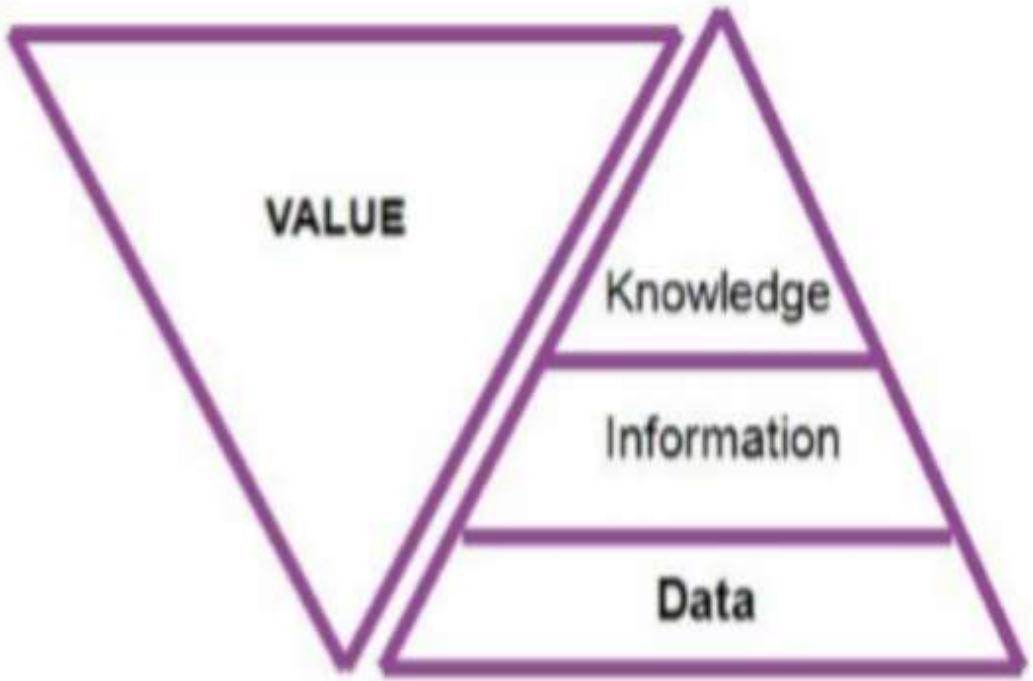
We are drowning in data, but starving for knowledge!

(John Naisbett, Megatrends, 1988)

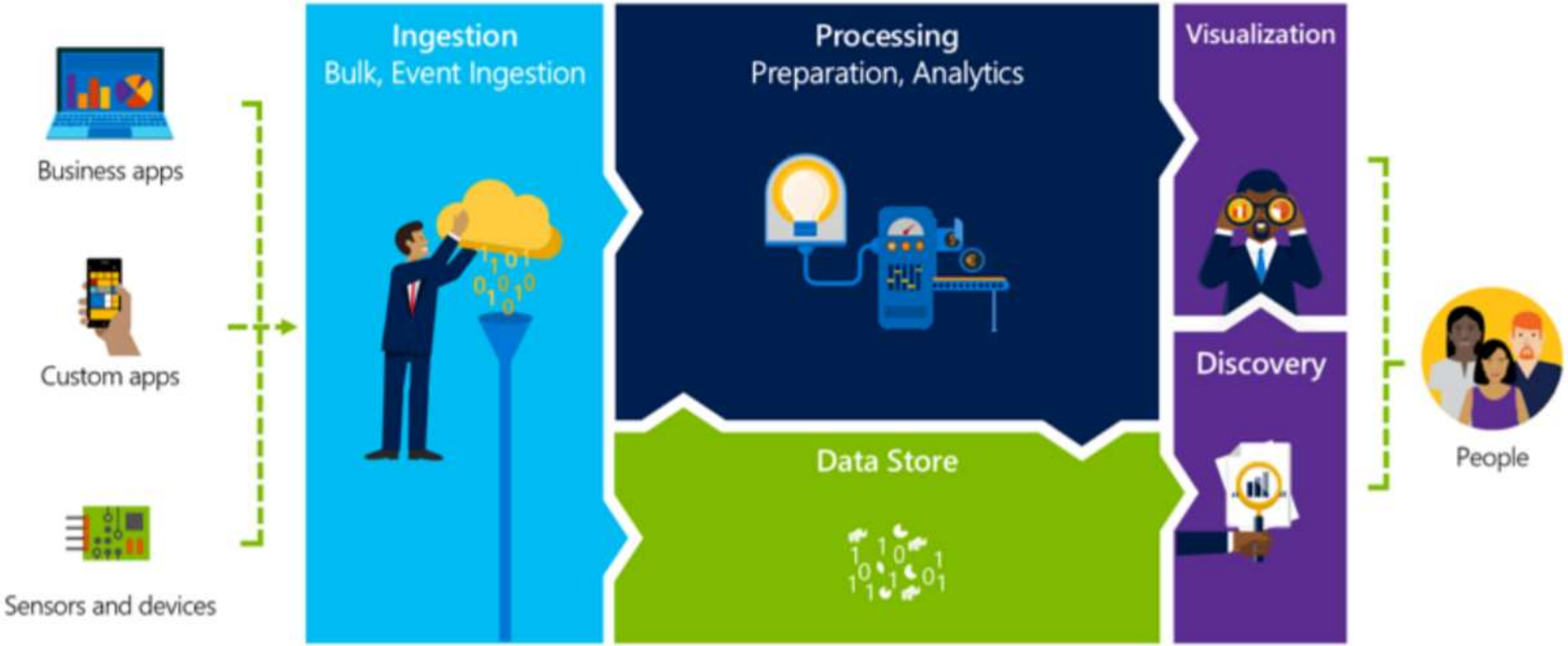
Mengubah Data Menjadi Pengetahuan

- Data harus diolah menjadi pengetahuan melalui proses data mining supaya bisa bermanfaat bagi manusia.

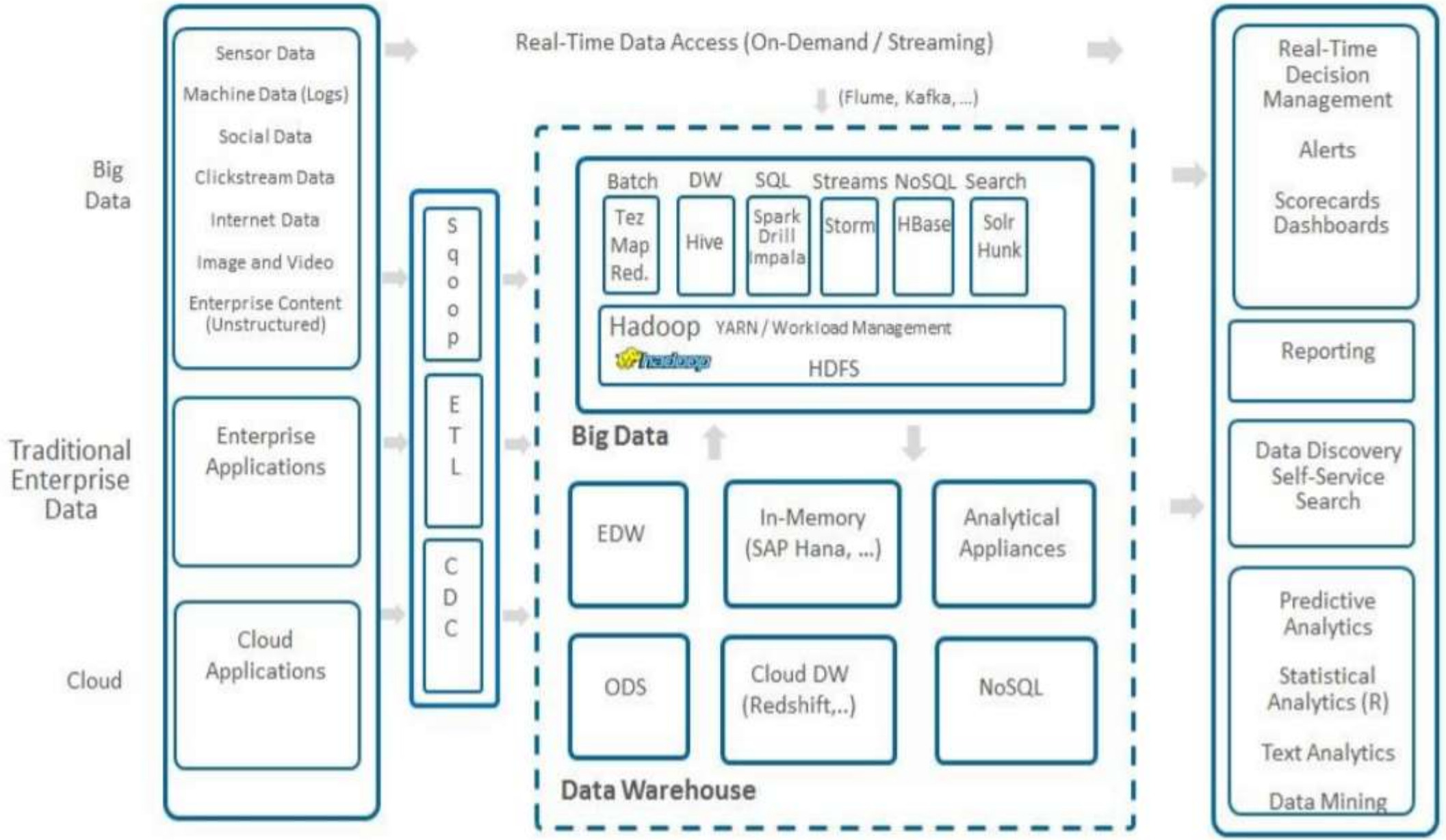




Big Data Flow



Big Data Architecture



Algoritma

1. Estimation (Estimasi):
   - Linear Regression, Neural Network, Support Vector Machine, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
   - Linear Regression, Neural Network, Support Vector Machine, etc
3. Classification (Klasifikasi):
   - Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression, etc
4. Clustering (Klastering):
   - K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc
5. Association (Asosiasi):
   - FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

Big Data Use Cases

- **Marketing**: product recommendation, market basket analysis, product targeting, customer retention.
- **Finance:** investment support, portfolio management, price forecasting.
- **Banking and Insurance**: credit and policy approval, money laundry detection.
- **Security**: fraud detection, access control, intrusion detection, virus detection.
- **Manufacturing**: process modeling, quality control, resource allocation
- **Web and Internet**: smart search engine, web marketing
- **Software Engineering**: effort estimation, fault prediction
- **Telecommunication**: network monitoring, user behavior analysis

Rencana Perkualiahan

- Bagian I: Pengolahan Data
- Bagian II: Infrastuktur Big Data

Rencana Perkuliahan Bagian Pertama

- Metode pembelajaran: Problem-based Learning
- Pertemuan 1-6: studi kasus
- Pertemuan 7-8: presentasi usulan tugas individu oleh mahasiswa
- Pertemuan 15-16: presentasi hasil tugas individu oleh mahasiswa

Presentasi **usulan** tugas individu oleh mahasiswa

- Memilih usecase
- Dapat menggunakan dataset yang dikumpulkan sendiri atau dataset dari Internet (Kaggle dataset, UCI Machine Learning Repository)
- Mengusulkan salah satu algoritma pengolahan data

Presentasi **hasil** tugas individu oleh mahasiswa

- Presentasi hasil tugas menggunakan:
  - Library
  - Program yang ditulis dari scratch (optional)

Penilaian

- Tugas Individu: 60%
- Tugas-tugas studi kasus: 40%

Tools:

- Python
- Anaconda merupakan distribution open source yang memberikan kemudahan dalam penggunaan Python. Dapat dijalankan di OS: Linux, Windows, and Mac OS X
- Link download: https://www.anaconda.com/distribution/