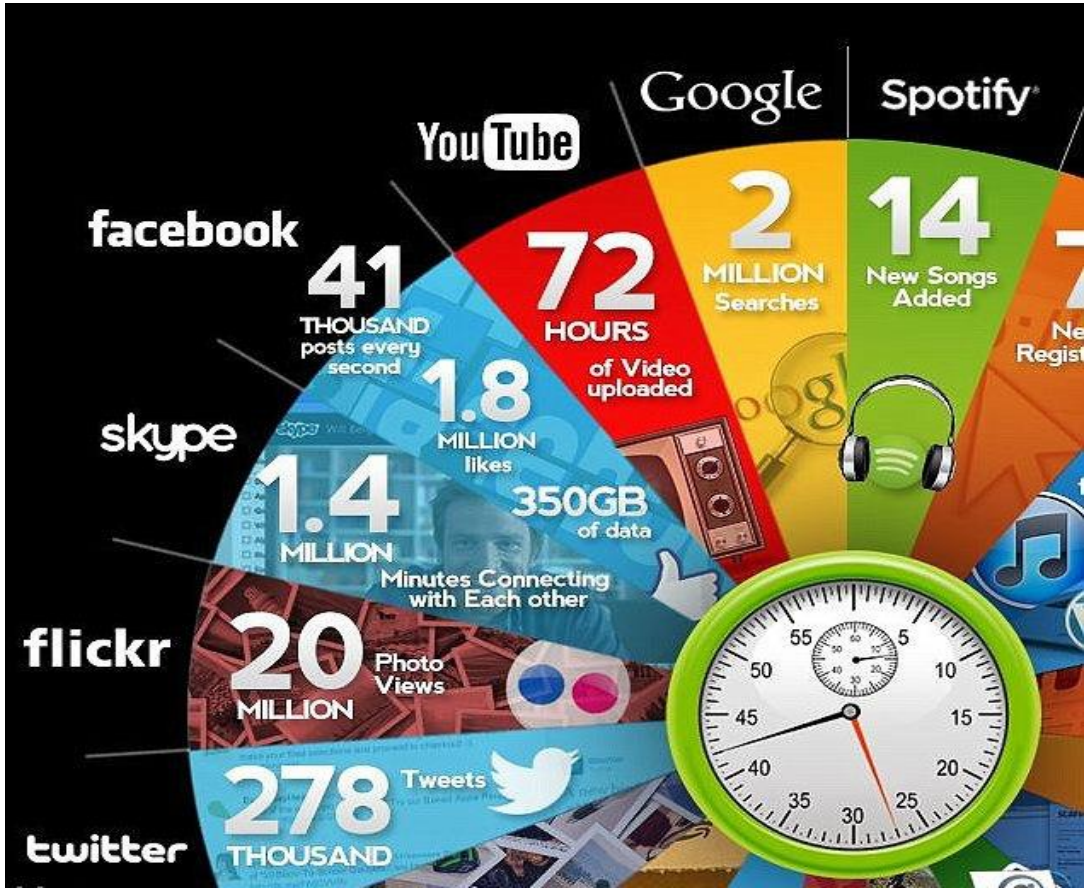
BIG DATA

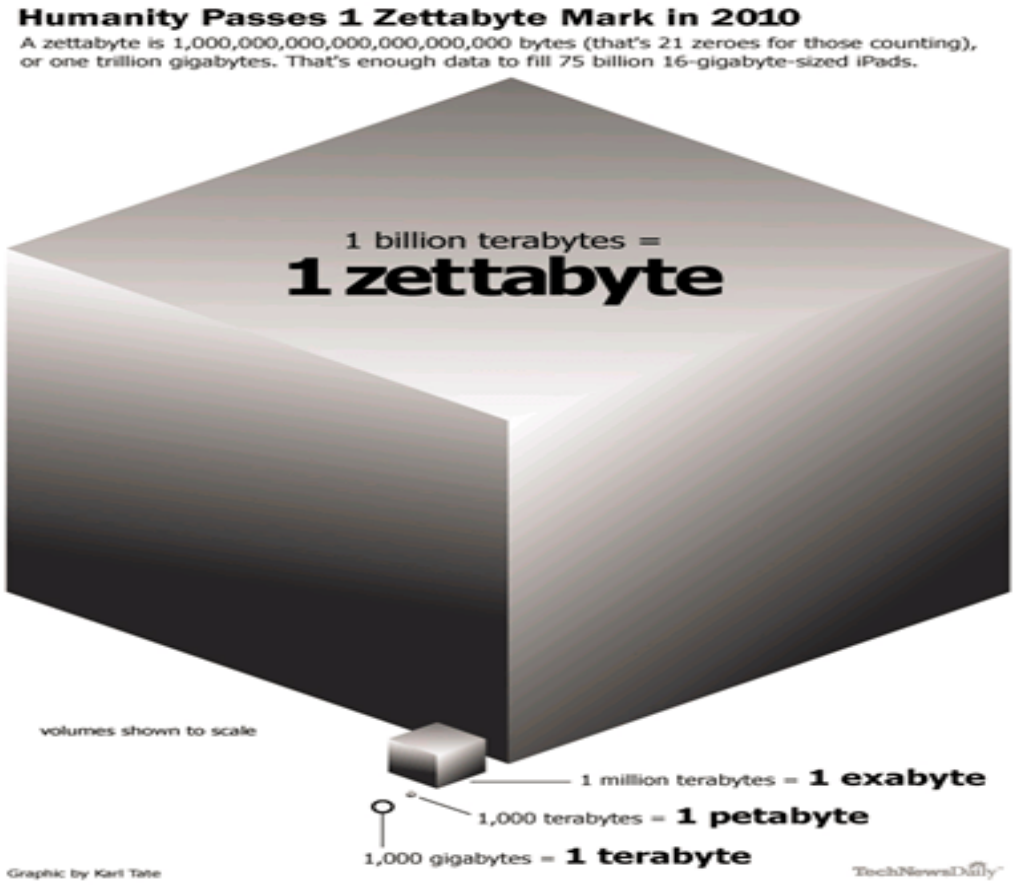
Sirojul Munir | rojulman@nurulfikri.ac.id | @rojulman

Data Processing & Data Analytics

Sirojul Munir | rojulman@nurulfikri.ac.id | @rojulman

Sumber Data

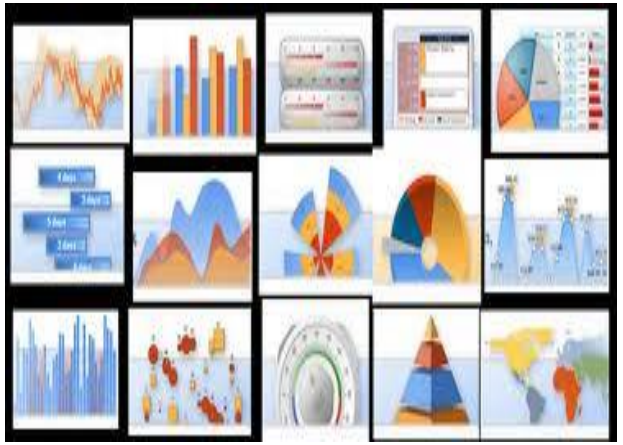


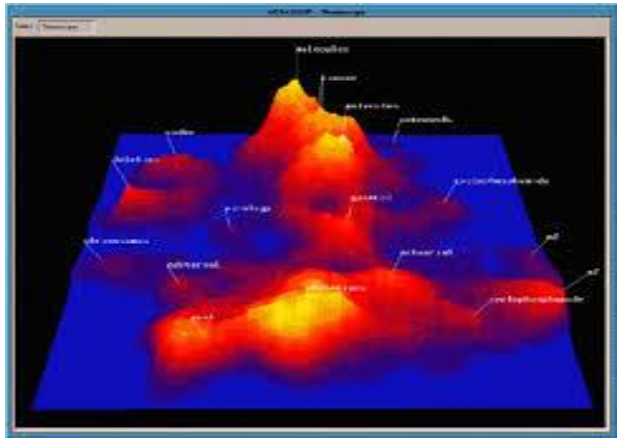


Bagaimana memprosesnya ?

Tipe Data (1)

- *Structured data*
  - *Tipe data yang dapat disimpan di database atau spreadsheet, diperlukan untuk dikelola sesuai dengan format penyimpanan standar dan ontologi, seperti : nama, alamat, telpon,*
  - *Contoh : Aplikasi sistem informasi akademik, aplikasi work flow, aplikasi SDM dll*
- *Unstructured data*
  - *text, audio, imagery, video*
  - *Contoh : email siswa, chat rooms, hasil questioner, video / audio di sistem e-elearning , RFID , barcode*



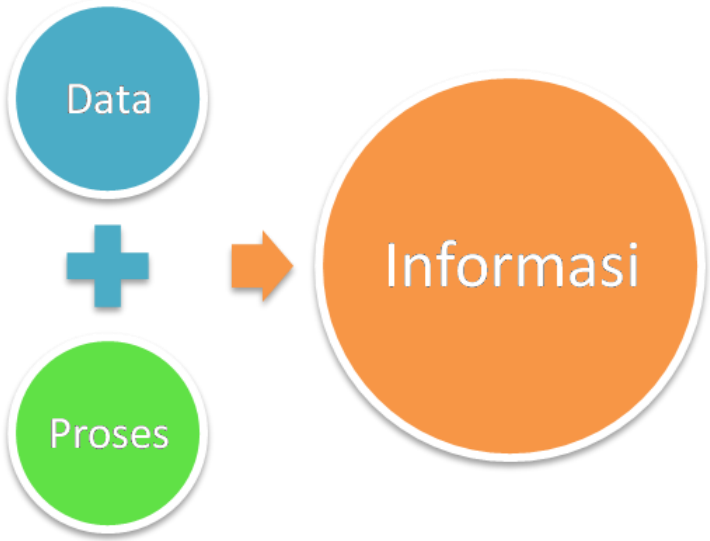


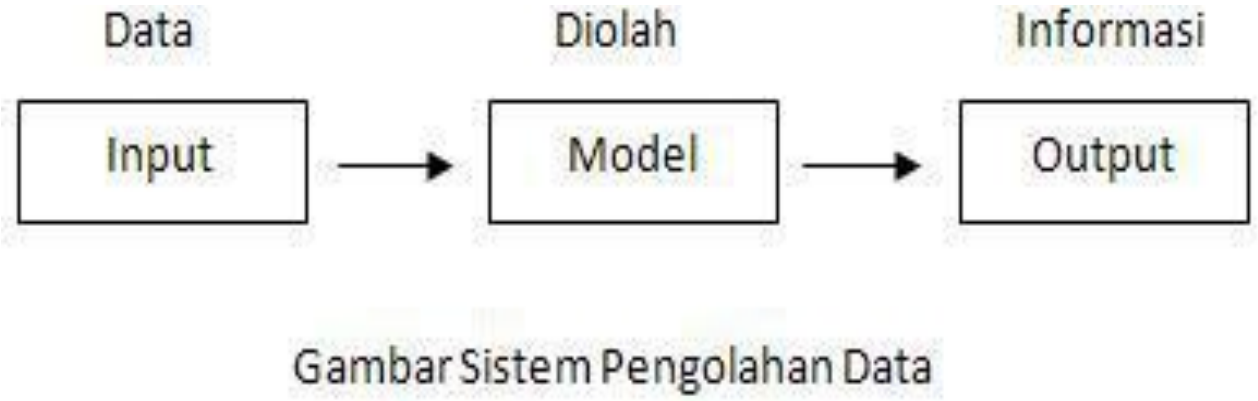
Tipe Data (2)

- *Jenis data yang berbeda memerlukan analisa teknis yang berbeda pula, Data tidak terstruktur terkadang memerlukan pre- processing lebih utama agar struktur data dapat di analisis*
- *Unstructured data analysis*
  - *Text : document clustering , topic detection, entity extraction (people, places, locations, dates, times etc., sentiment analysis (+,-)*
  - *Audio : speaker identification, language identification, speech to text, keyword spotting*
  - *Video analysis : face recognition, object recognition, target tracking*

Data Processing : Data -> Informasi -> Knowledge

Definisi Informasi:



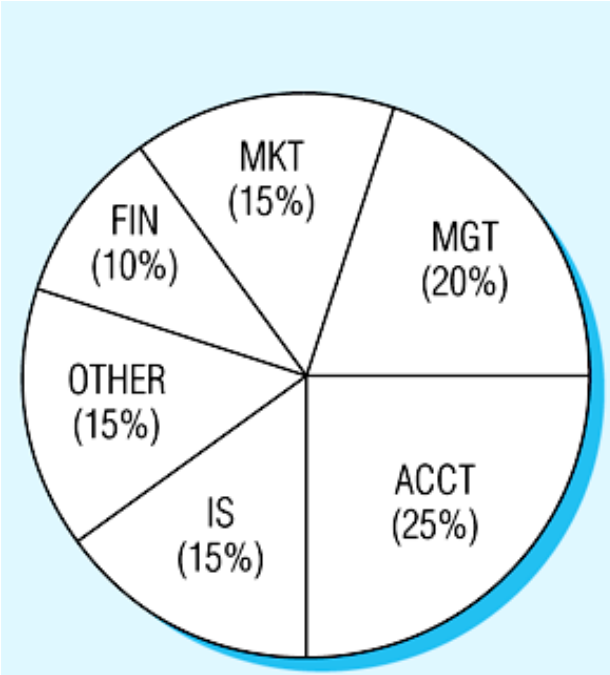


Gambar Sistem Pengolahan Data

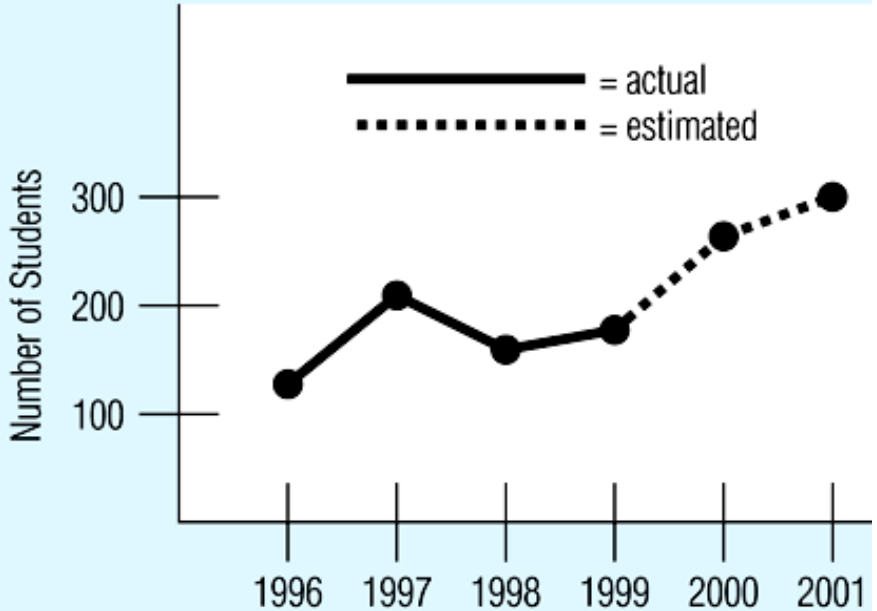
- **Informasi**: data yang telah diproses sebagai bahan dalam proses pengambilan keputusan.

Data Processing : Data -> Informasi -> Knowledge

Informasi - dapat dimanfaatkan sebagai dasar untuk pengambilan keputusan dan memahami permasalahan/situasi



Percent Enrollment by Major (199X)



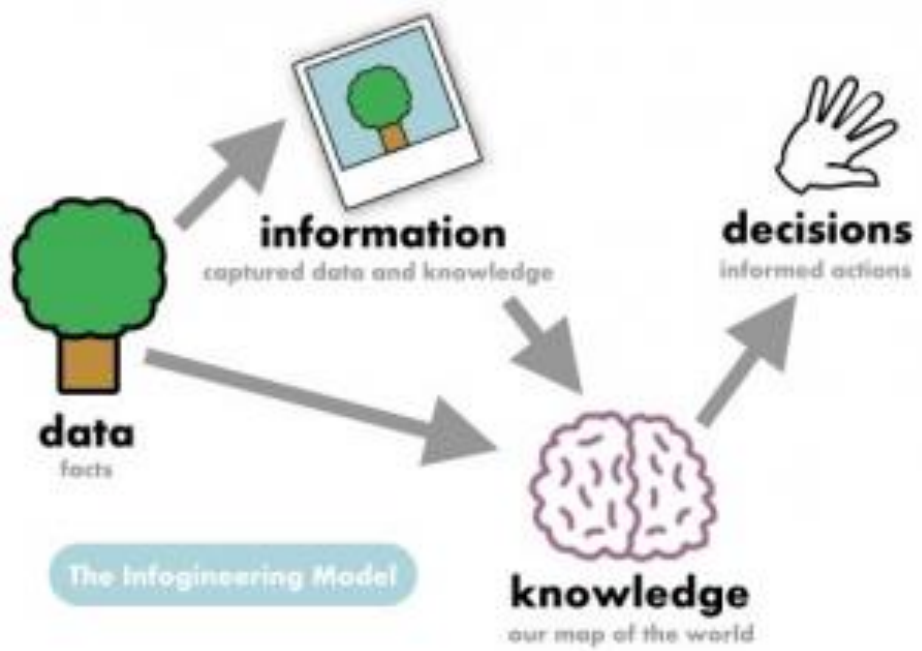
Enrollment Projections

Data Processing: Data -> Informasi -> Knowledge

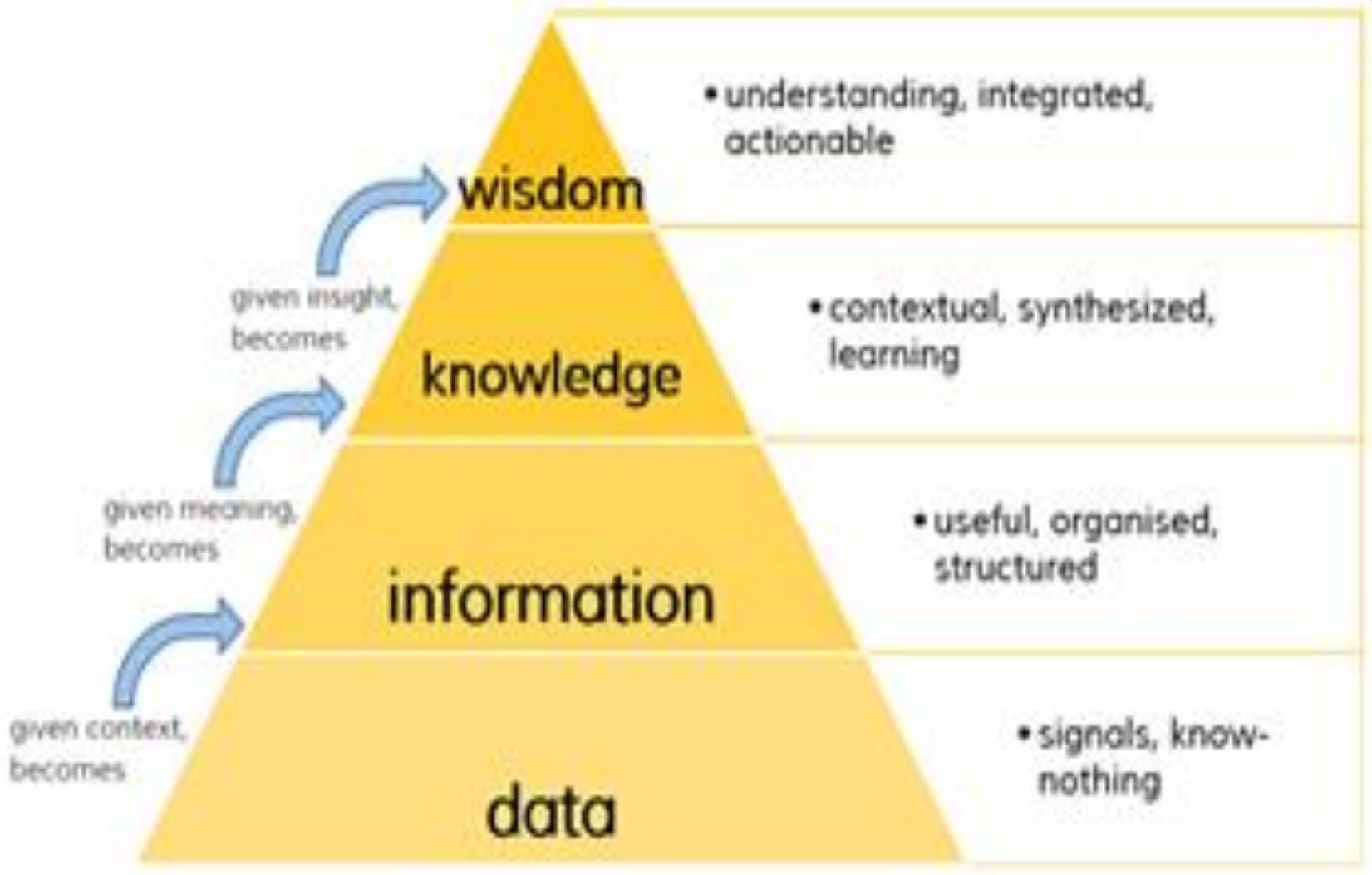
- Knowledge: adalah informasi yang dilengkapi dengan pemahaman pola hubungan dari informasi disertai pengalaman, baik individu maupun kelompok dalam organisasi.

Fungsi Informasi: I = i(D, S, T)

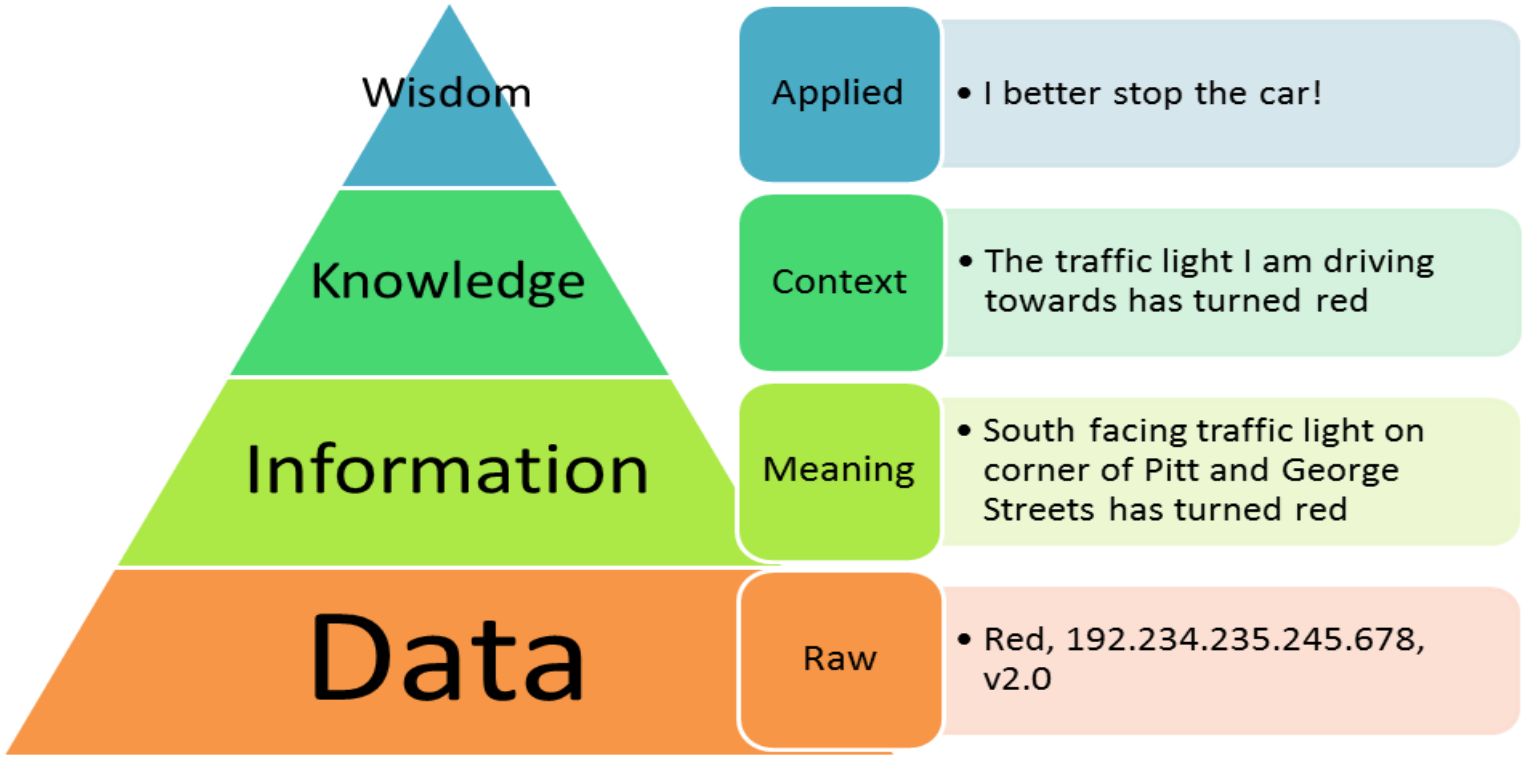
- **I :** Informasi
- **D**: Data
- **S** : Pengetahuan awal
- **T** : Waktu



Piramida : Data, Informasi, Knowledge, Wisdom



Piramida : Data, Informasi, Knowledge, Wisdom



Q?:

Berikan contoh keterkaitan data, informasi, knowledge dan wisdom. Dalam kehidupan sehari2 disekitar anda !

Contoh Kasus: Data, Informasi,Knowledge

- Data Kehadiran Pegawai:

NIP	Tanggal	Jam Masuk	Jam Pulang
1801	10/01/2019	07:30	16:30
1823	10/01/2019	07:22	16:30
1811	10/01/2019	07:46	16:52
1807	10/01/2019	08:01	19:20
1818	10/01/2019	07:15	16:30
1848	10/01/2019	07:39	16:50

Contoh Kasus: Data, Informasi, Knowledge

- Informasi akumulasi kehadiran pegawai dalam sebulan

NIP	Masuk	Alpa	Cuti	Sakit	Telat
1801	20			2	
1823	22				
1811	14	4			4
1807	15	2	3		2
1818	20		2		
1848	19		1	1	1

Contoh Kasus: Data, Informasi, Knowledge

- Pola Kehadiran Mingguan Pegawai

	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	8	0	1	0	4
Pulang cepat	0	1	1	1	7
Izin	3	1	0	0	5
Alpa	2	0	1	1	2

Pengetahuan tentang pola kebiasaan pegawai dalam jam datang/pulang kerja

Contoh Kasus: Data, Informasi, Knowledge

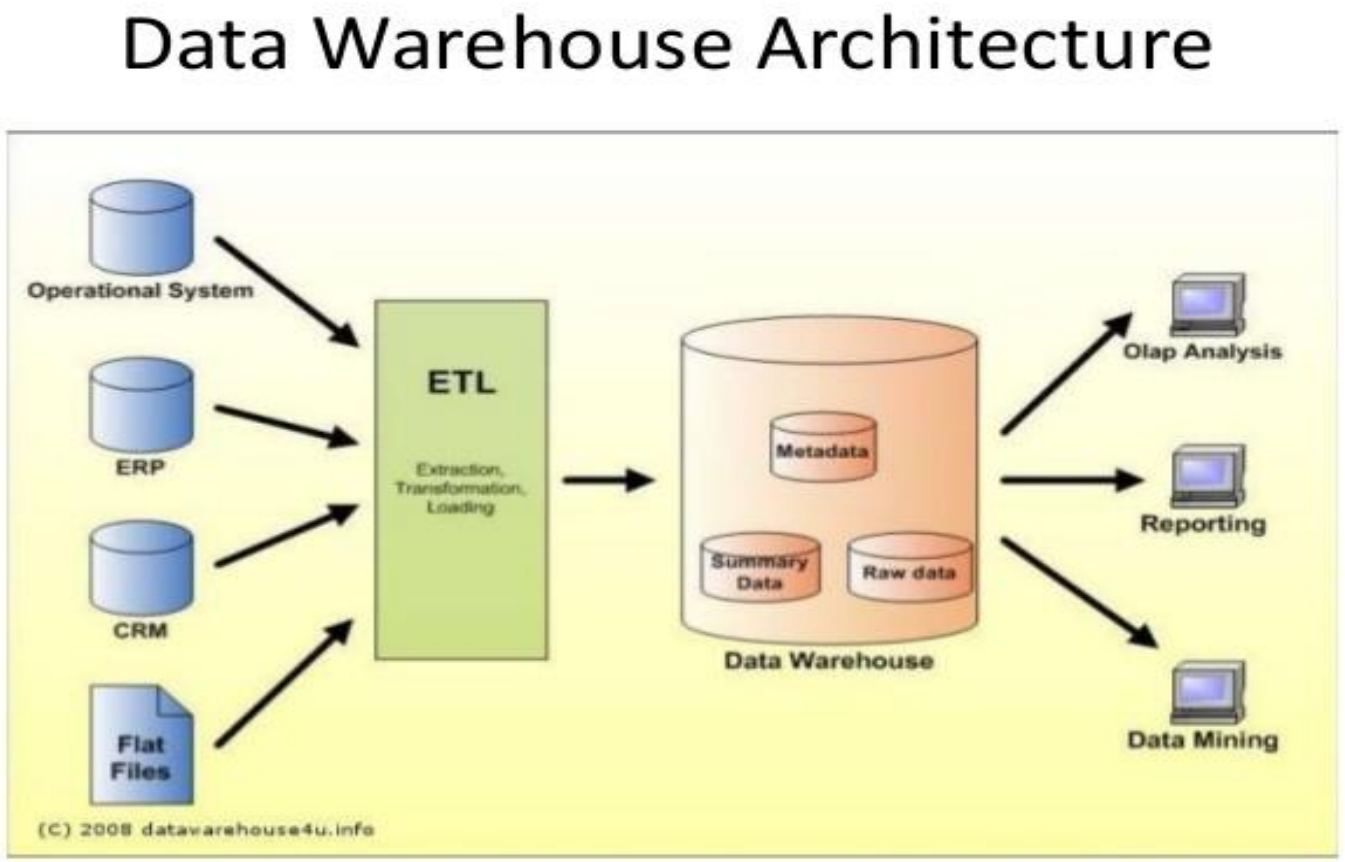
- **Wisdom**

  Kebijakan penataan jam kerja karyawan khusus untuk hari senin dan jumat

  Peraturan jam kerja:
  - Hari Senin dimulai jam 09:30
  - Hari Jumat diakhiri jam 15:30
  - Sisa jam kerja dikompensasi ke hari lain

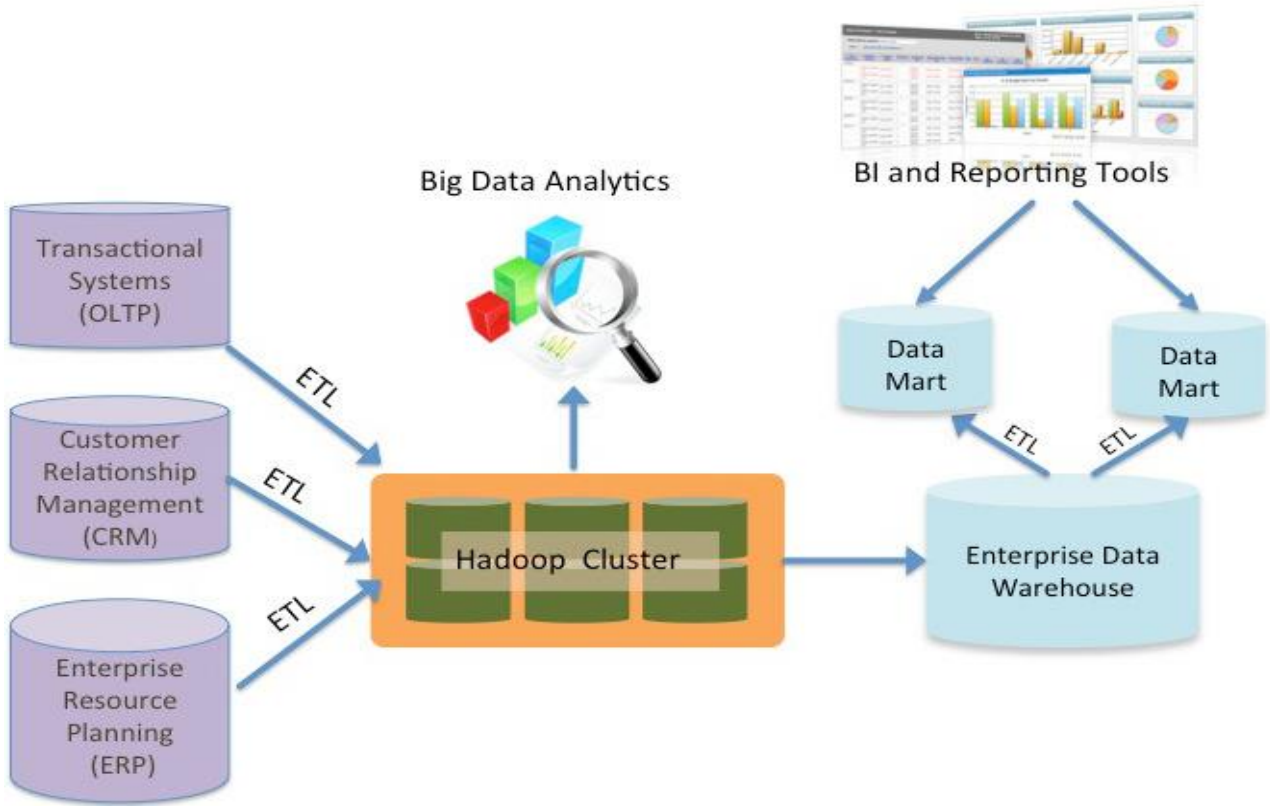
Data Processing : Data Warehouse

Data Processing: dengan sumber data yang terstruktur



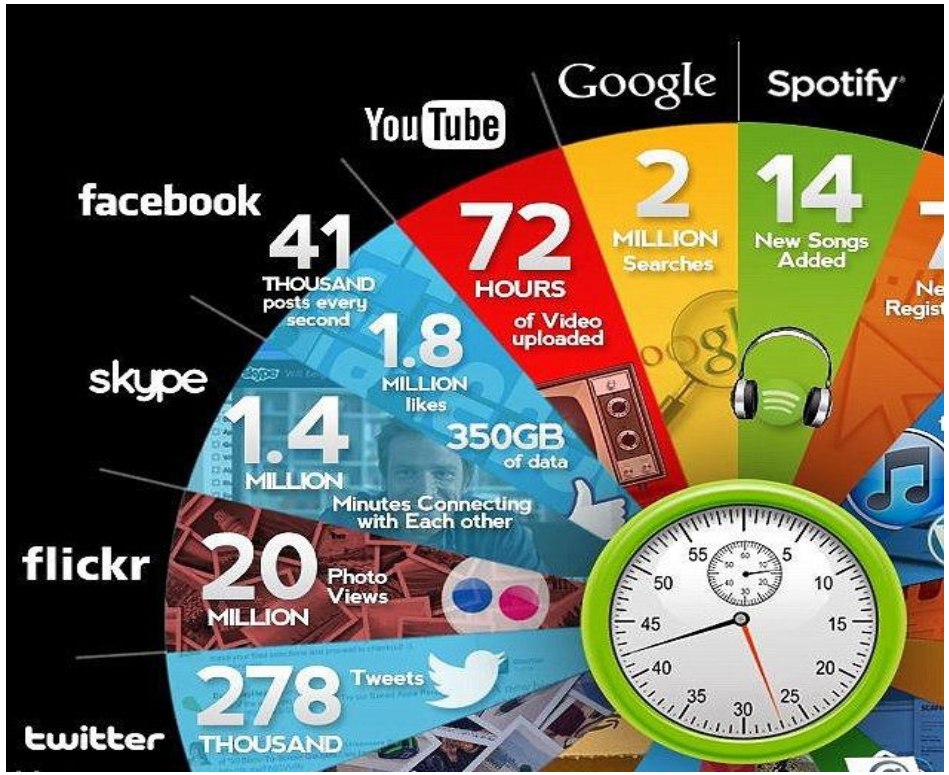
Data Processing : Data Warehouse & Big Data

Data Processing: dengan
sumber data yang terstruktur
& tidak terstruktur



Data Processing :: Problem Big Data

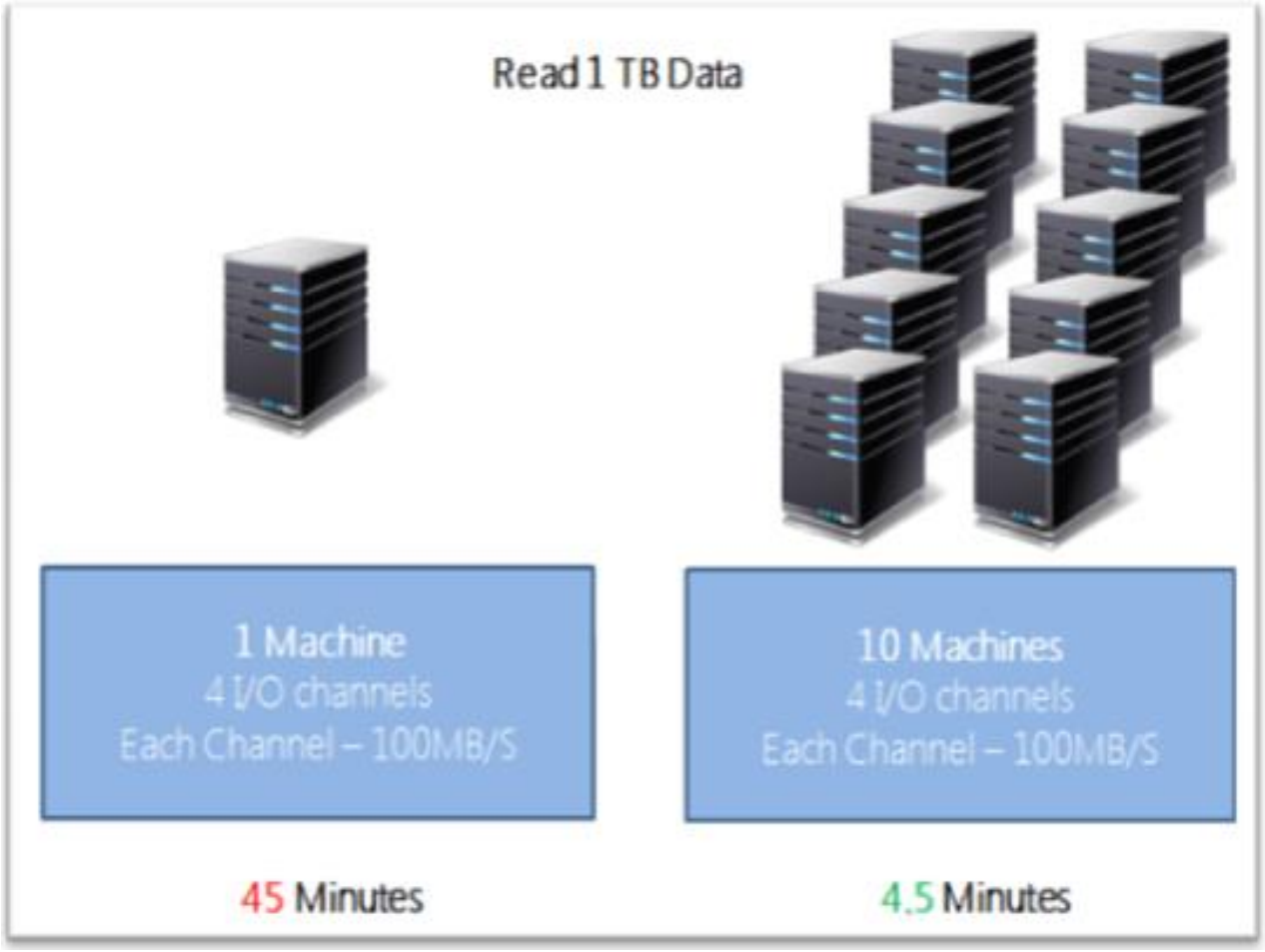
- Volume:
  - bertambah secara eksponensial. Saat ini **2015: 4,8** Zetta Bytes = **4.800** Peta Bytes = 4,8 juta Exa Bytes = **4,8** milyar Terra Bytes = **4,8** trilyun Giga Bytes. *)
- Velocity & Variety:



Solusi Big Data :: DFS

1. Distributed File System ::

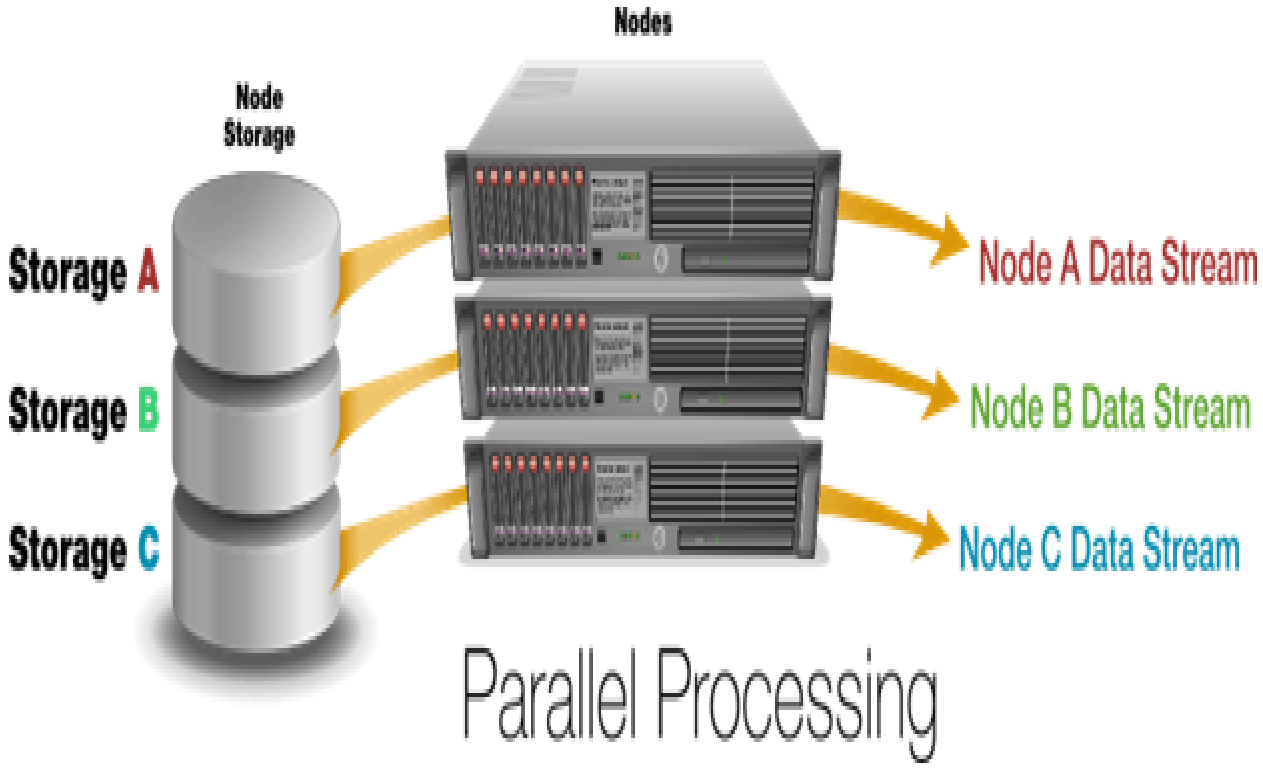
Membagi data menjadi blok-blok kecil dan load data ke beberapa mesin komputer dengan pararel processing



Solusi Big Data :: Pararel Processing

2. Pararel Processing ::

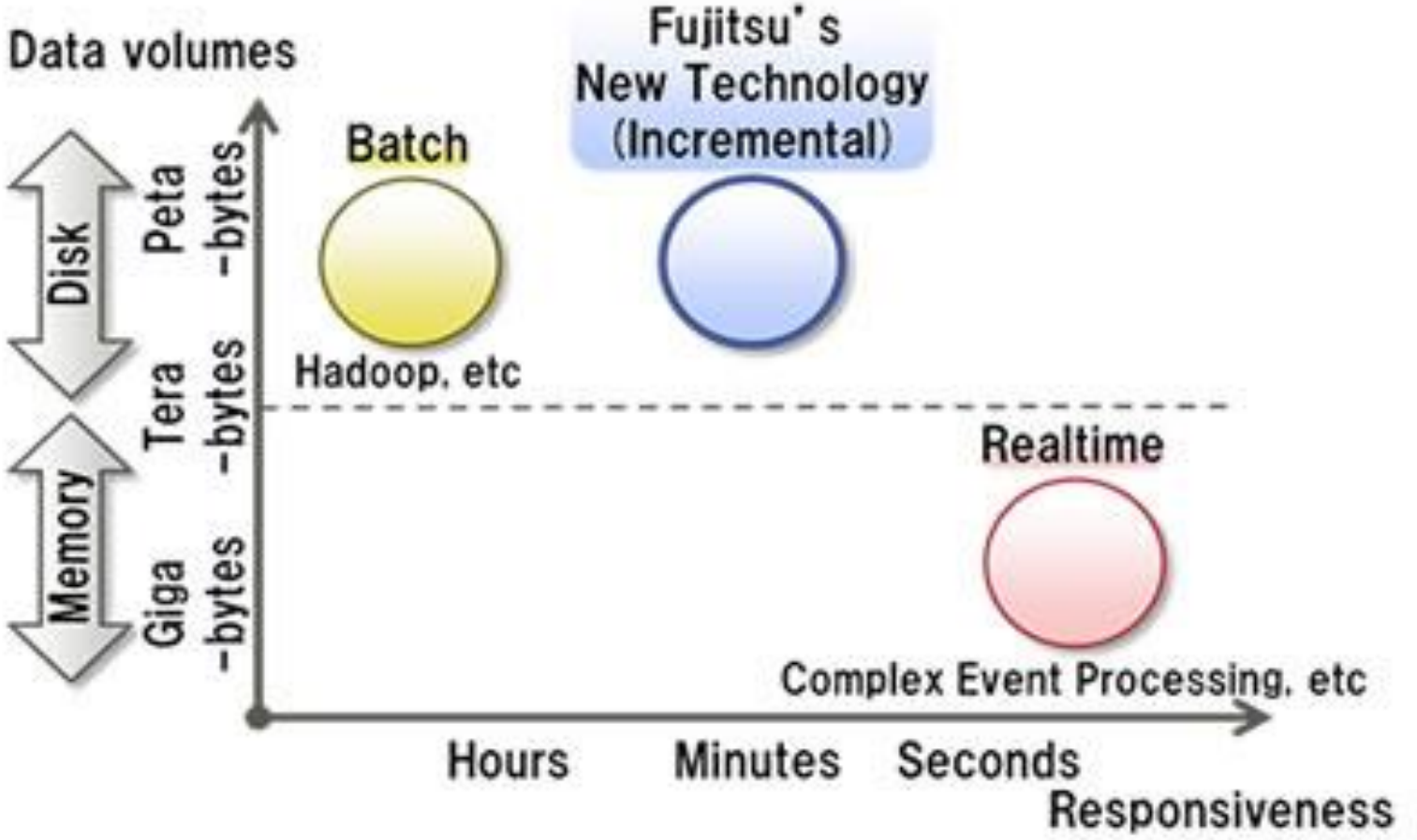
Data di bagi ke sejumlah N server, dan N Server akan melakukan proses analisas secara paralel, yang akan mereduksi waktu tunggu proses untuk menghasilkan laporan akhir dan hasil analisa data



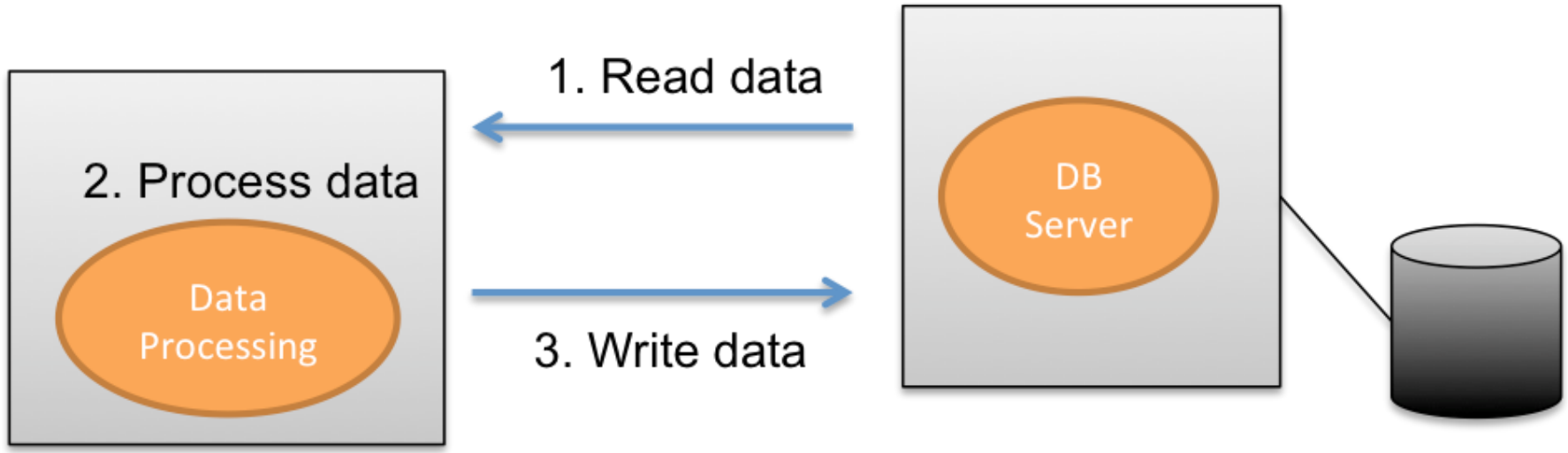
Solusi Big Data :: Pararel Processing

2. Pararel Processing ::

Mereduksi waktu tunggu proses

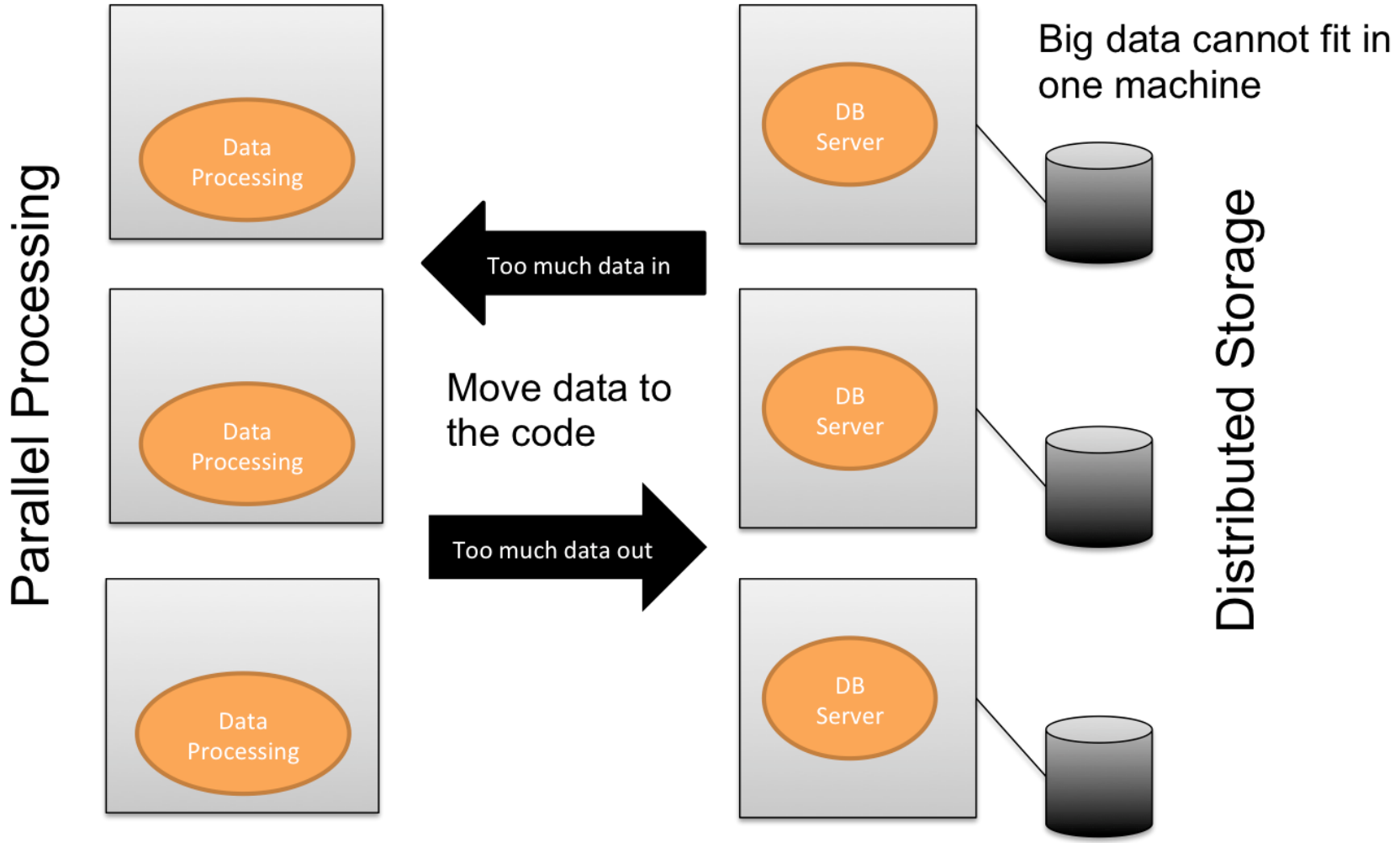


Data To Code :: Pendekatan tradisional

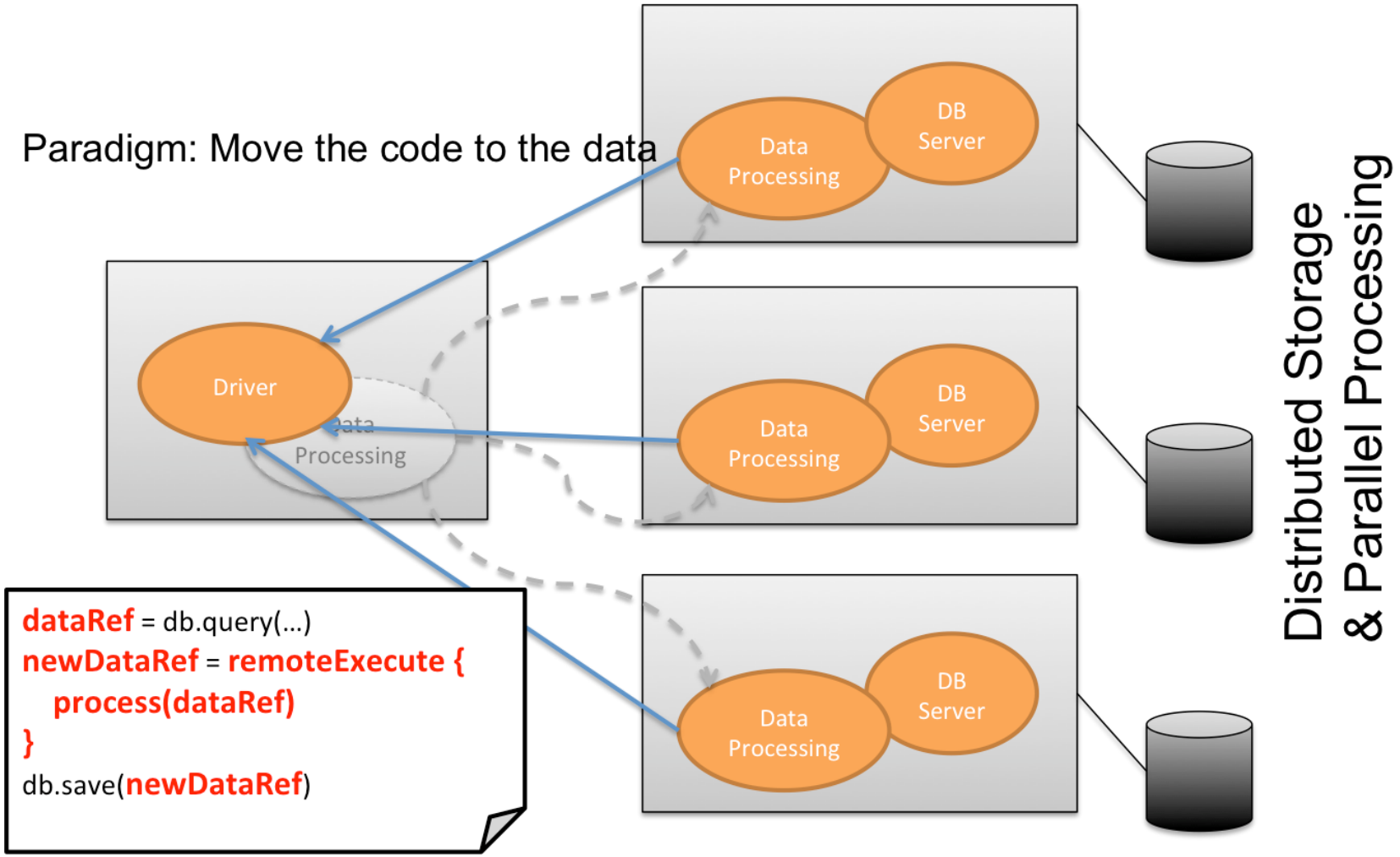



```
data = db.query(...)
newData = process(data)
db.save(newData)
```

Data To Code :: Pararel Processing



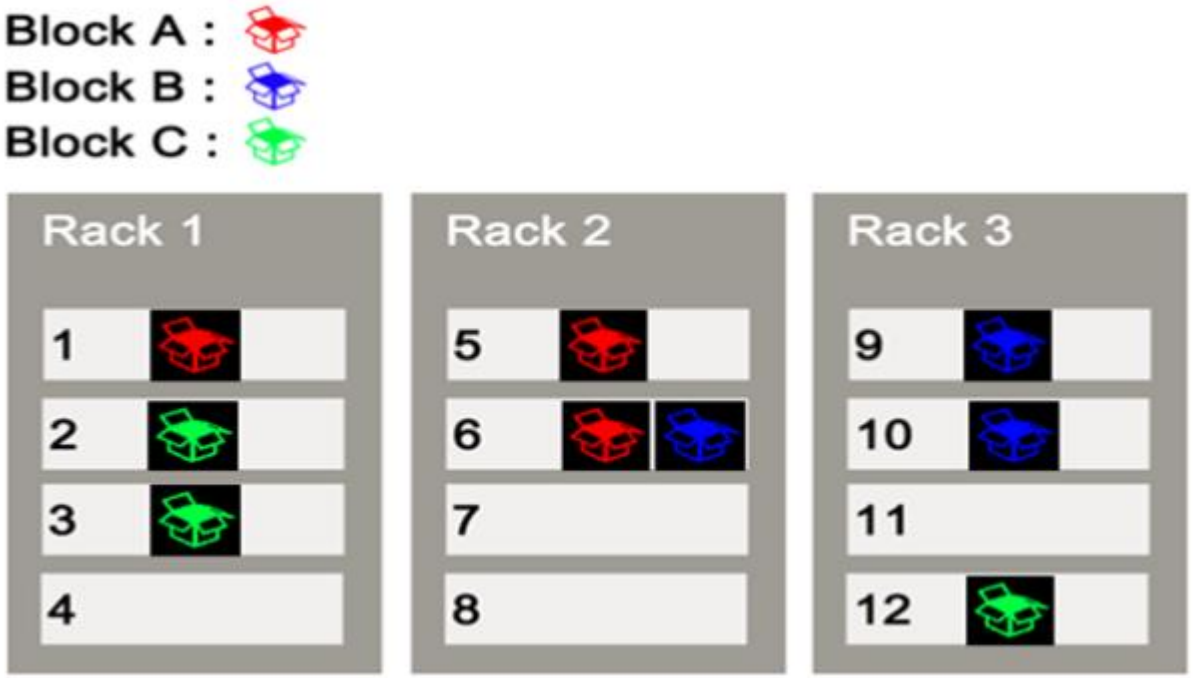
Data To Code :: Pararel Processing

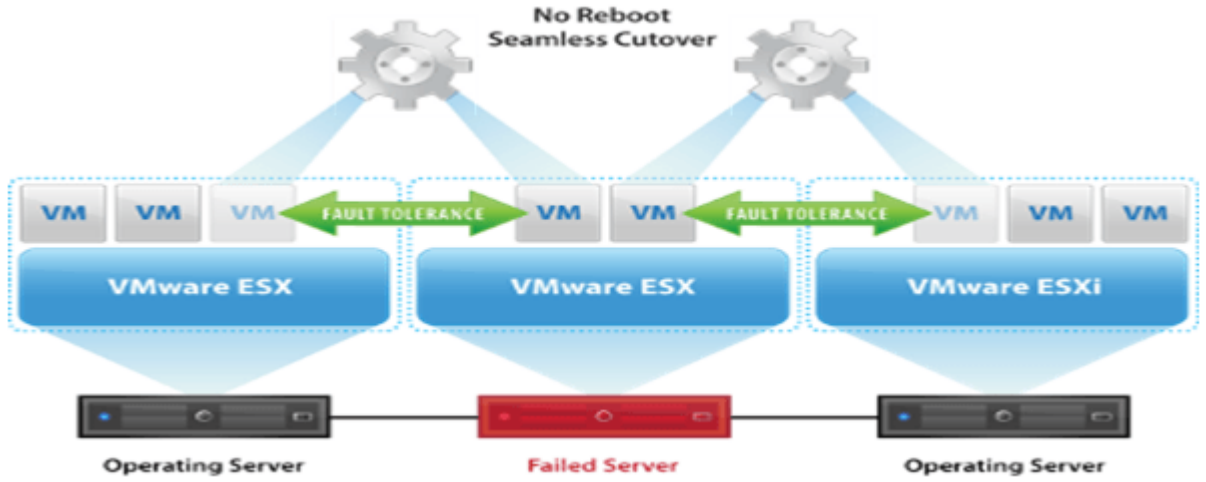


Solusi Big Data

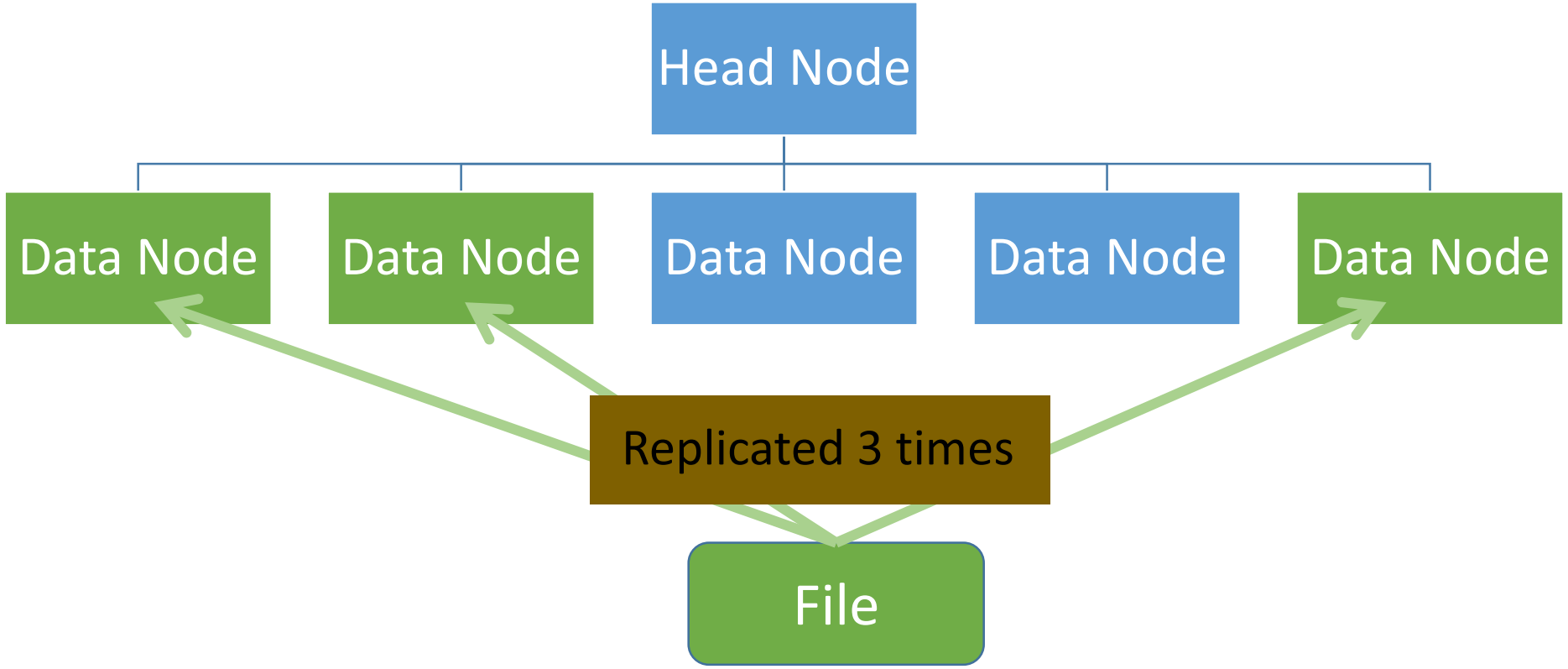
3. Fault Tolerance ::

Salah satu alasan menggunakan Framework Big Data (Hadoop) adalah untuk mengerjakan suatu project karena memiliki tingkat toleransi kesalahan yang tinggi, walaupun proses data sudah dilakukan secara cluster, Big Data menawarkan solusi data dibuat replikasi di beberapa node



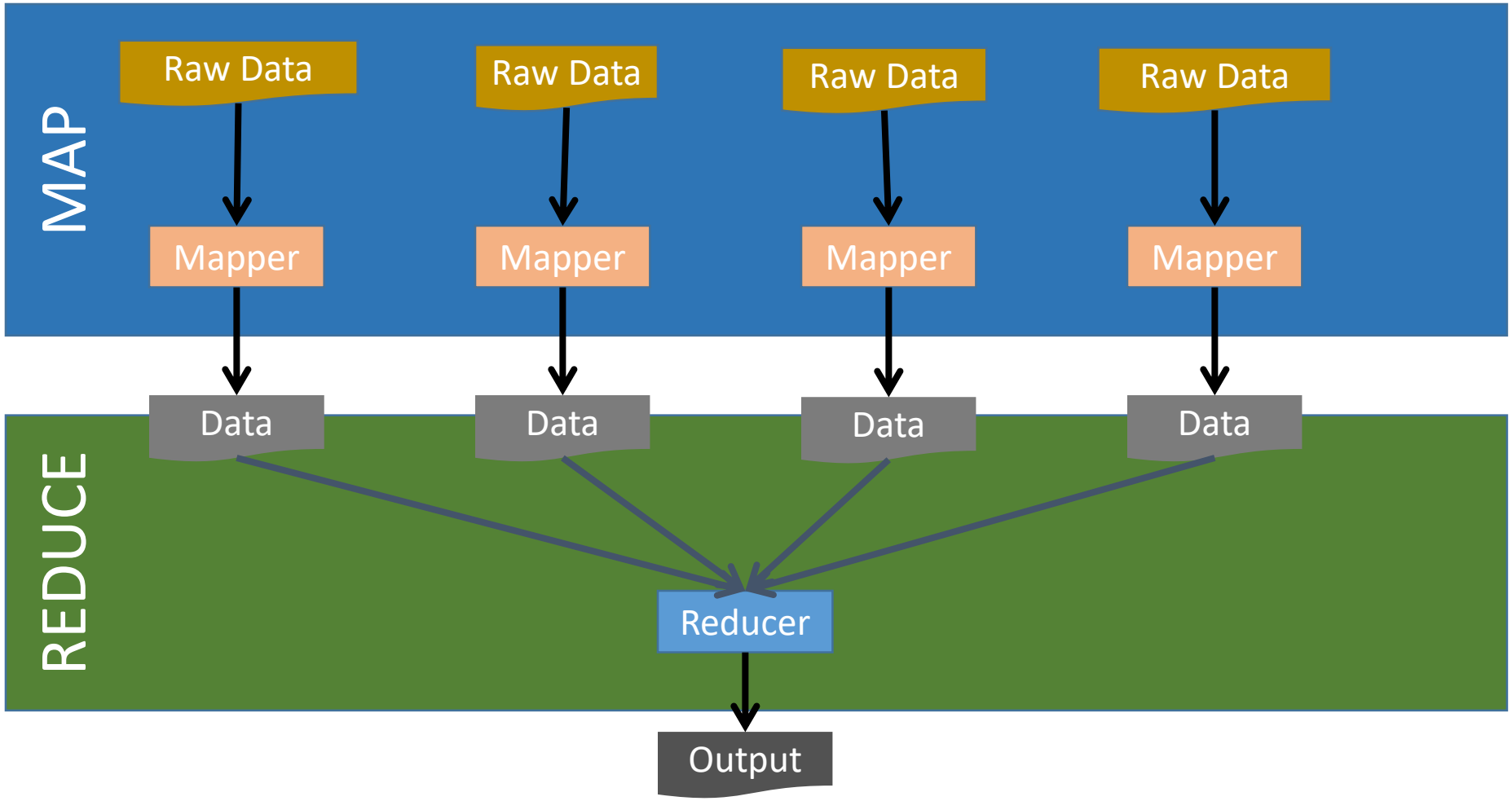


Hadoop File System :: HDFS



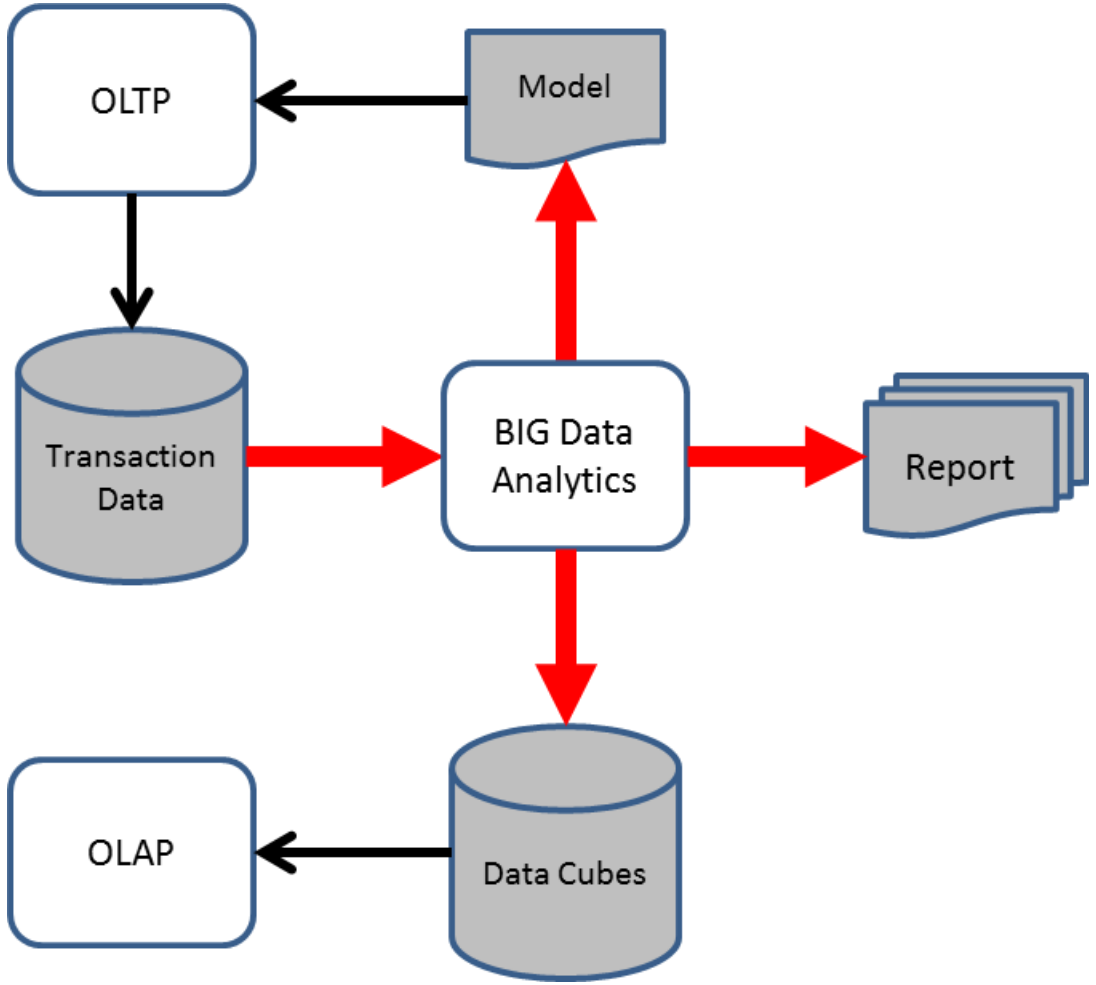
Read Optimised & Failure Tolerant

Map + Reduce = Extract, Load + Transform



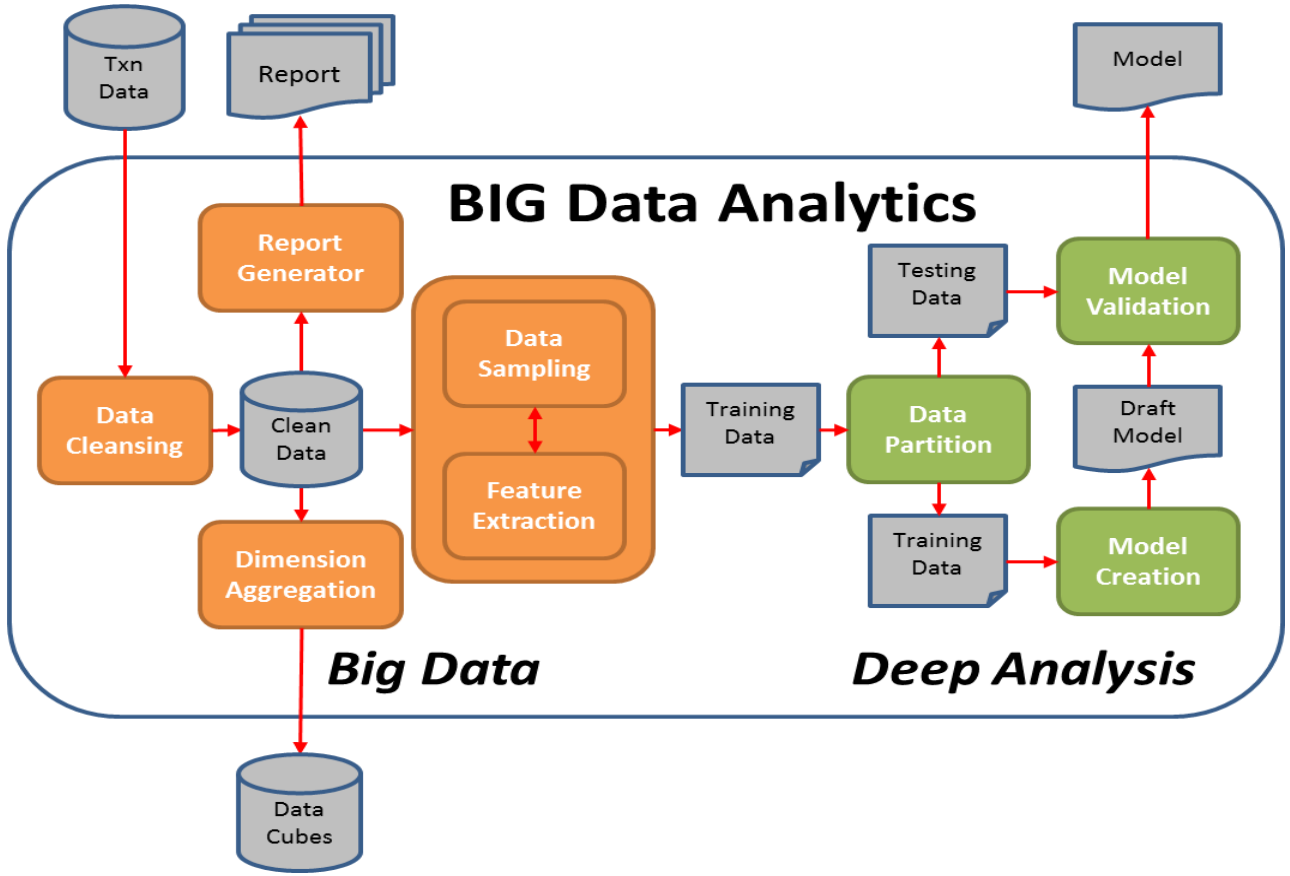
Data Processing :: Pipeline

- Big Data Analystics :
  - Online Time Processing
  - Online Analytic Processing



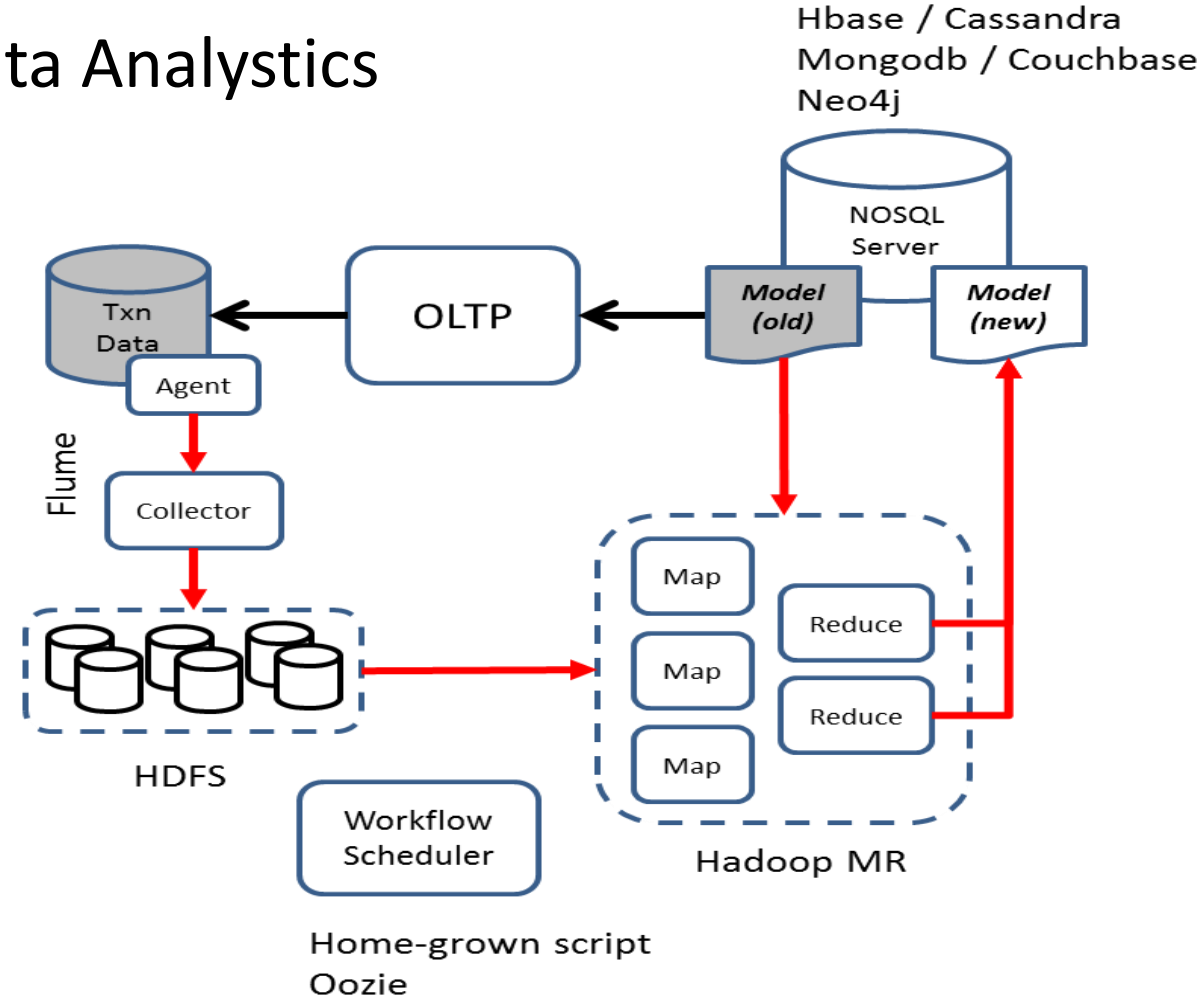
Data Processing :: Big Data + Deep Analysis

- Implementasi Big Data Analystics

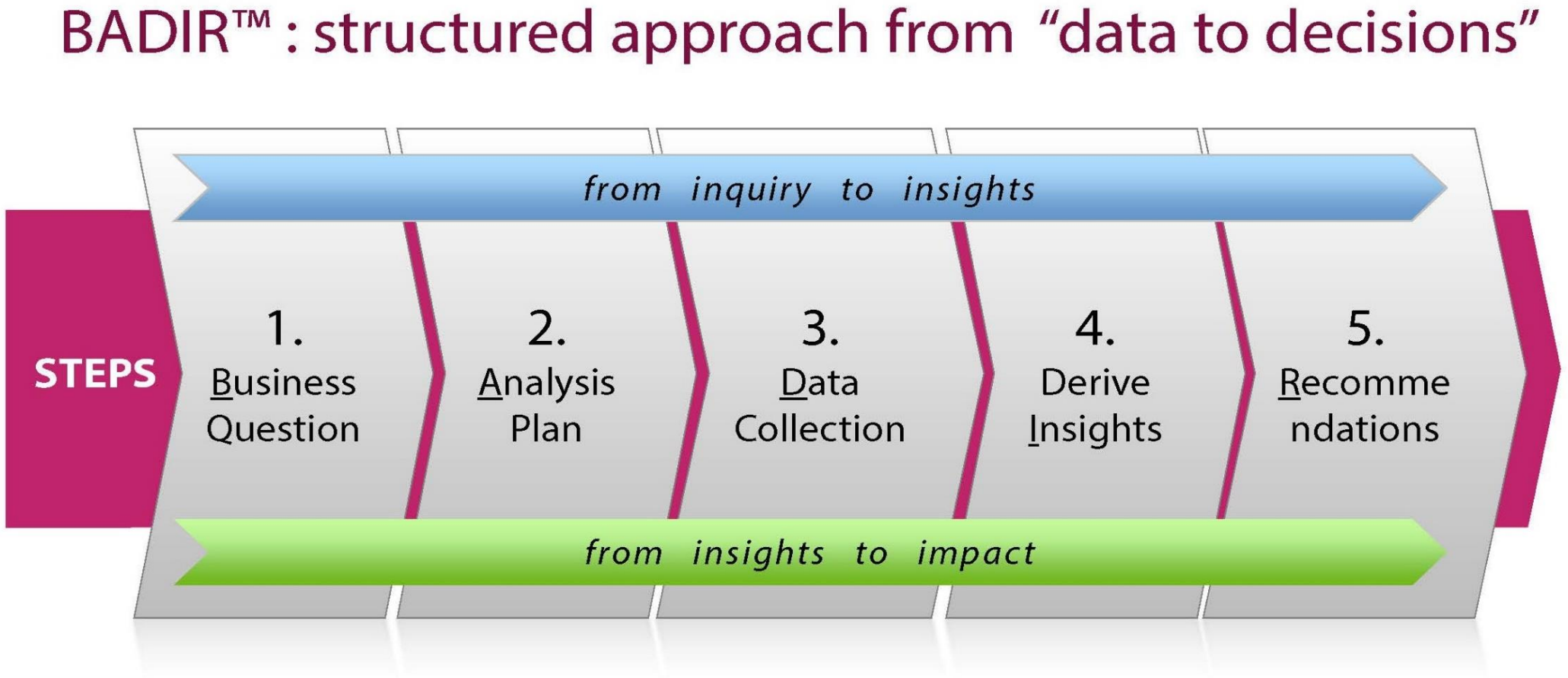


Data Processing :: Big Dat + Deep Analysis

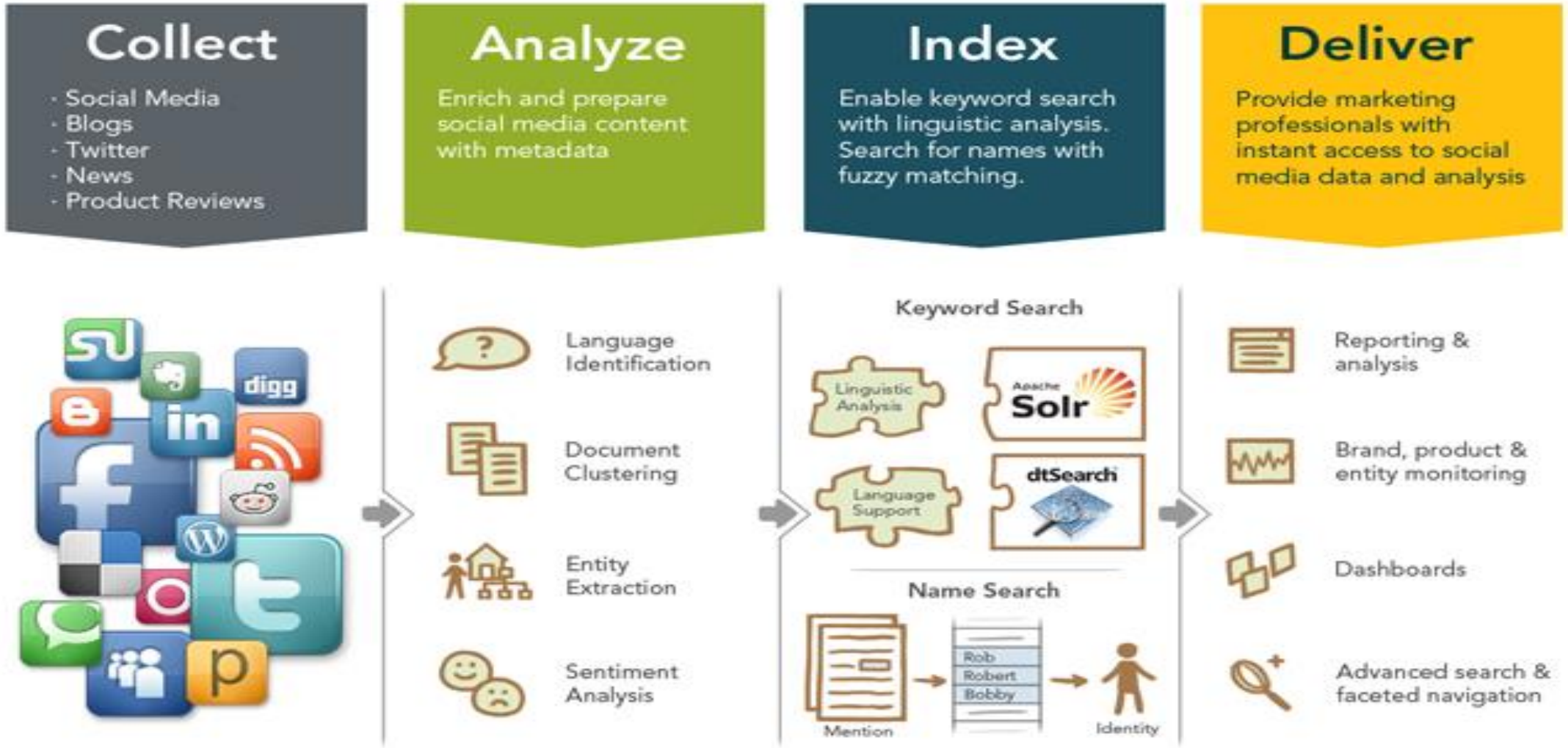
- Implementasi Big Data Analystics



Data Analytics :: Data to decisions

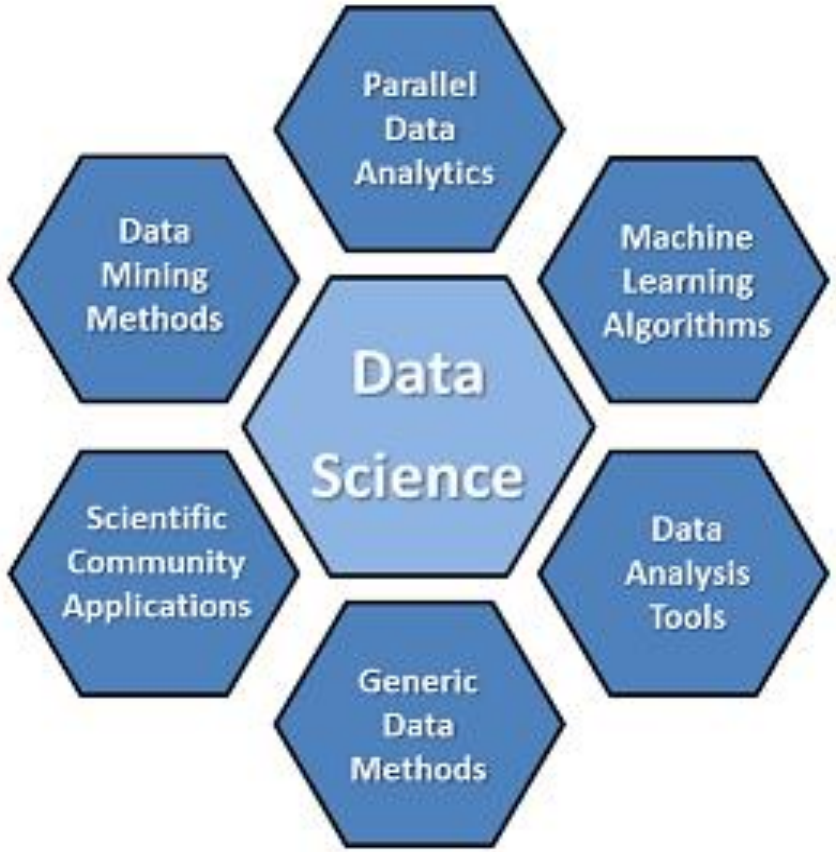


Collect – Analyze – Index - Deliver

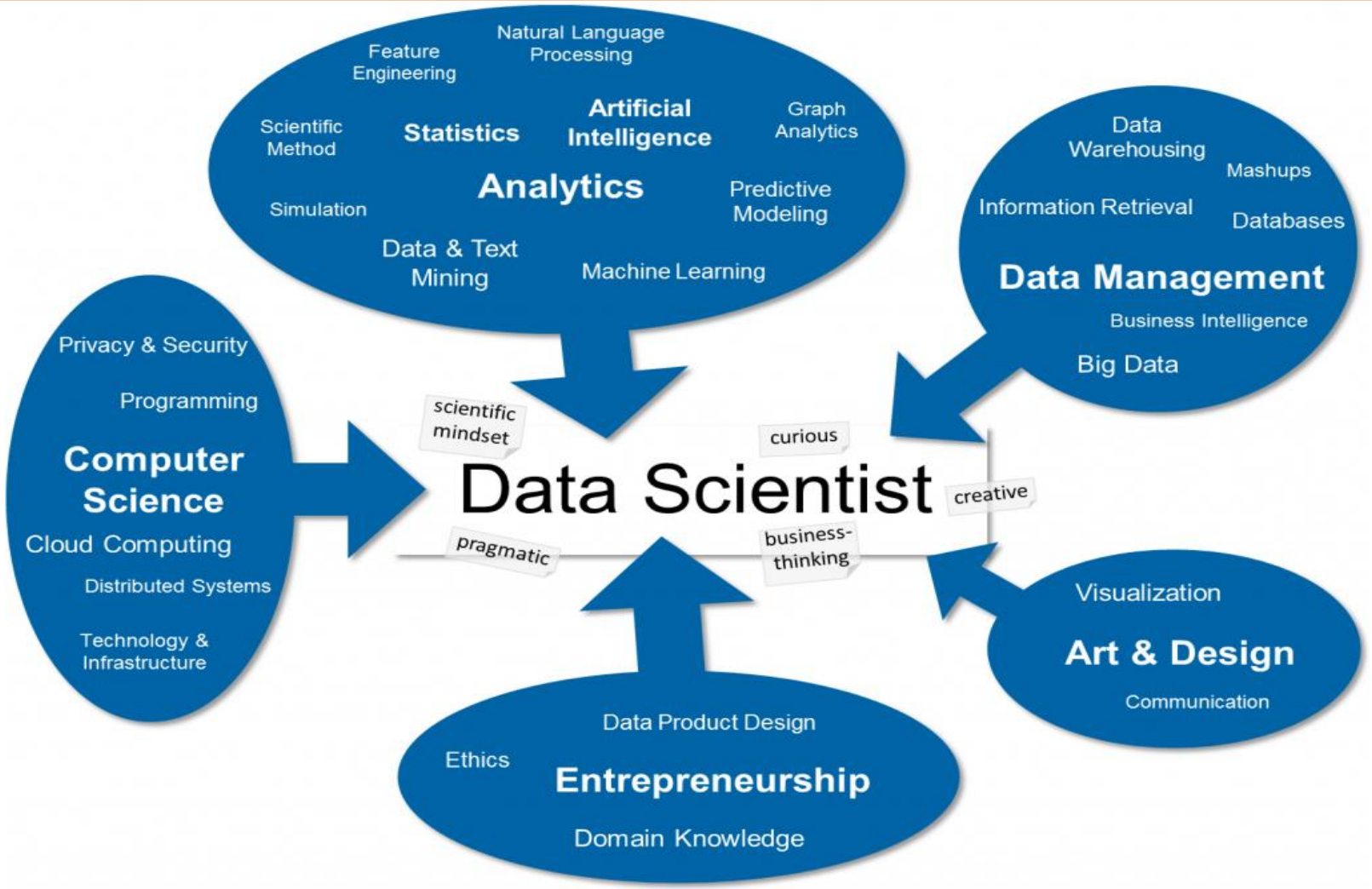


Data Analystics ?

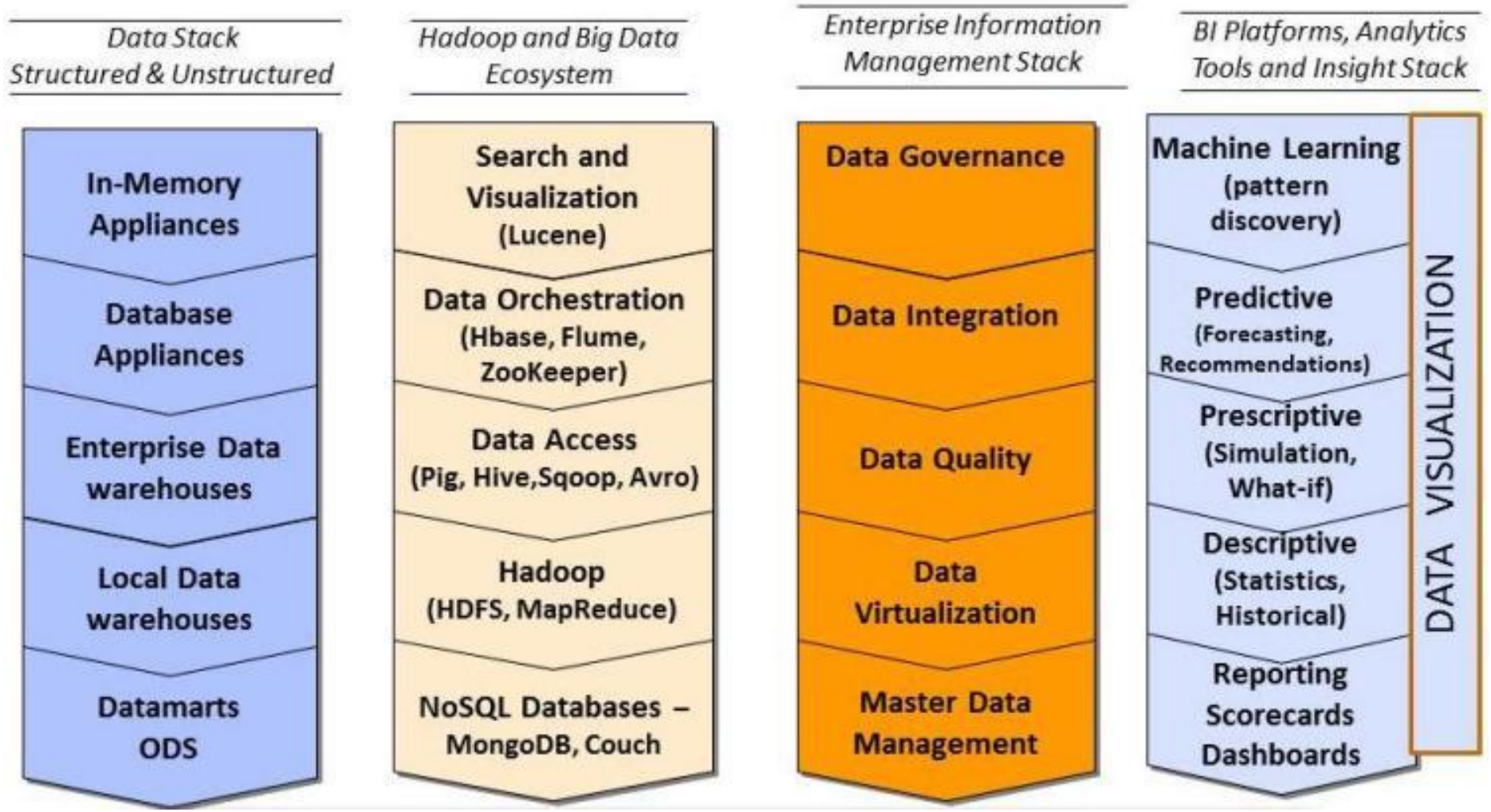
- Data Analytic (DA) adalah : bidang ilmu yang memeriksa data mentah (raw data) dengan tujuan menarik kesimpulan tentang informasi yang ada.
- Techniques & method analytics
  - *Statistics*
  - *Artificial Intelligence*
  - *Machine Learning*
  - *Data mining*
  - *Social Network Analysis*
  - *Text Mining and Web Analytics*
  - *Operational Research*
  - *Information Visualization*



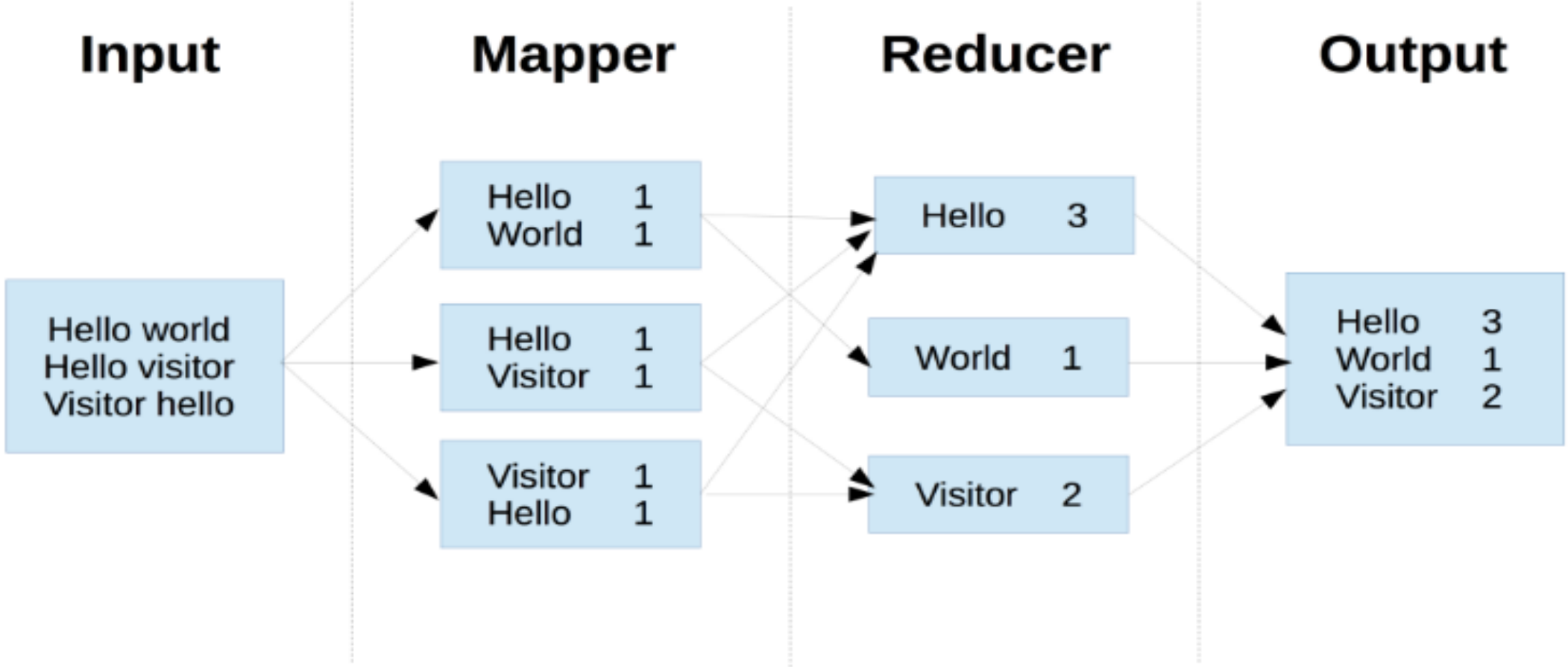
Data Scientist



Tools Data Analytics



Contoh Sederhana: Data Analytics



Sentiment Analysis

- Cabang ilmu pengetahuan di bidang Text Mining / Data Mining
- Dibagi menjadi 2 kategori:
  - **Coarse-grained sentiment analysis** - proses analysis pada level Dokumen. dengan melakukan klasifikasi orientasi **sebuah dokumen** secara keseluruhan. Orientasi ini ada 3 jenis : **Positif, Netral, Negatif**, orientasi dapat bersifat bersifat kontinu / tidak diskrit
  - ***Fined-grained sentiment analysis*** - *proses analysis* dengan melakukan klasifikasi pada sebuah **kalimat** dalam dokumen
- Contoh:
  - Saya tidak suka sayur-sayuran (negatif)
  - Layanan hotel yang baru saya kunjungi sangat memuaskan (positif)
- Resource data penelitian : SentiWordNet dan WordNet

Proses Sentiment Analysis

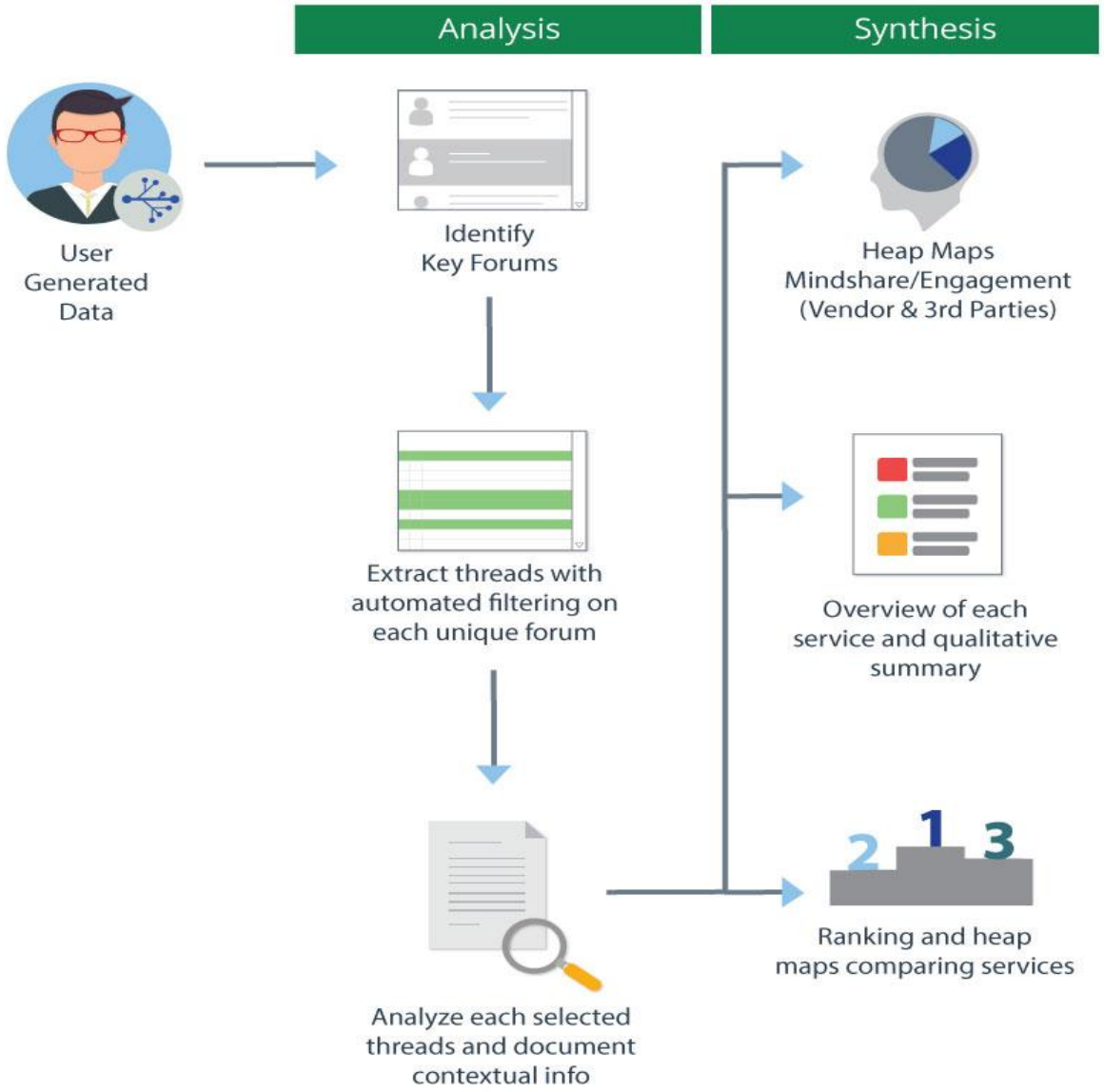
- Pada proses penelitian sentiment analysis dapat dibagi menjad 3 proses :
  - ***Subjectivity Classification***
    - menentukan kalimat yang merupakan opini :

      Kalimat: A bike has 2 wheels **VS** It is a good bike !
  - ***Orientation Detection***
    - *Menentukan orientasi/kecenderungan yang dimiliki : positif, negatif atau netral*

      Kalimat: It is a good bike ! **VS** ah, It is a bad bike
  - ***Opinion Holder and Target Detection :***
    - menentukan bagian yang merupakan **Opinion Holder** dan bagian yang merupakan **Target**

      Kalimat: ***Harry*** *said it is a good* ***bike***

Sentiment Analysis



Take a list of positive and negative words

Positive	Negative
Good	Bad
Great	Worse
Fantastic	Rubbish
Excellent	Sucked
Friendly	Awful
Awesome	Terrible
Enjoyed	Bogus

Match the two

Hotel Feedback

I had a fantastic time on holiday at your resort. The service was excellent and friendly. My family all really enjoyed themselves.

The pool was closed, which kind of sucked though.

Count them

Positive
Fantastic
Excellent
Friendly
Enjoyed

Negative
Sucked

4

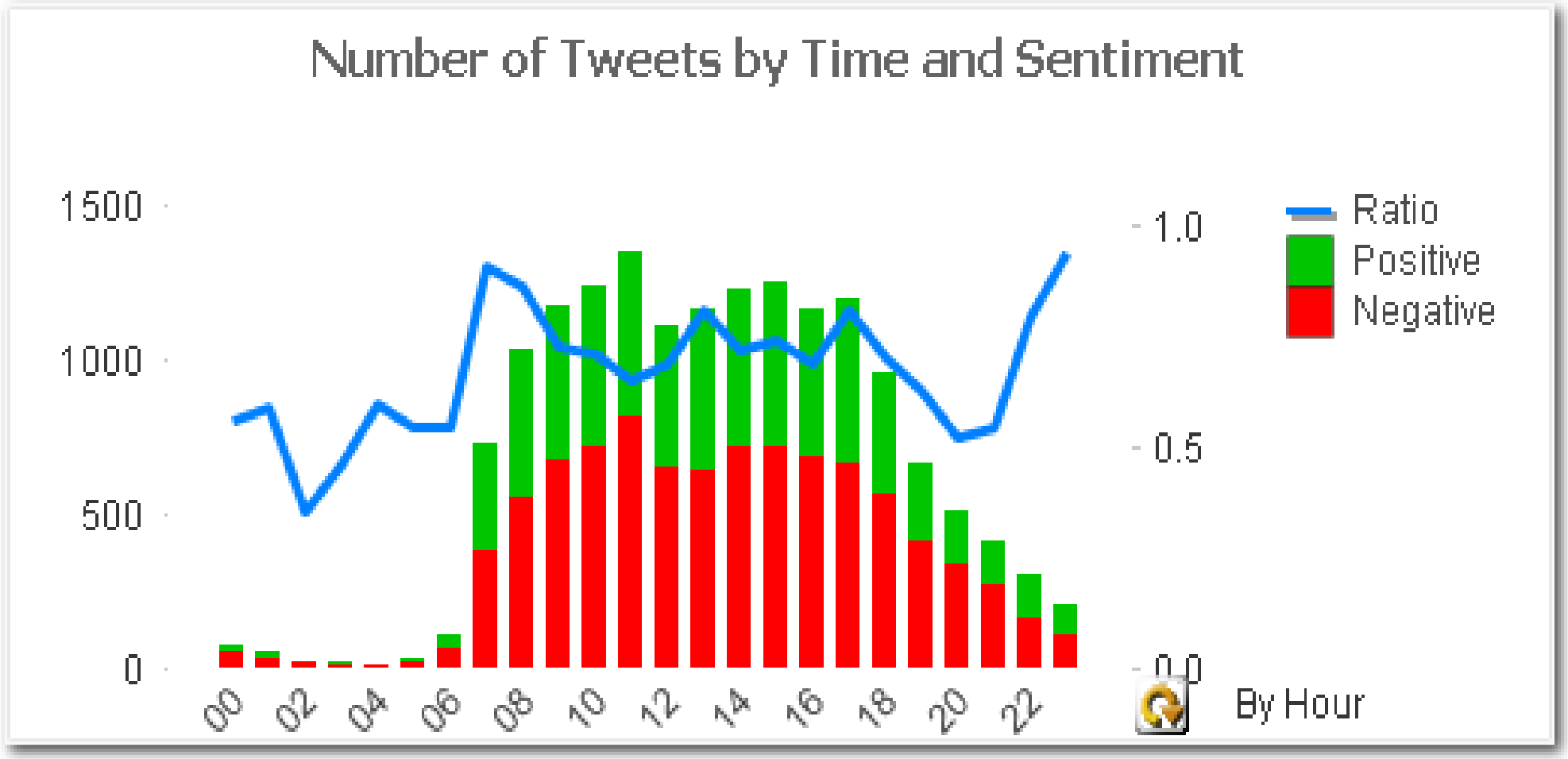
1

Subtract negative from positive

4 - 1 = 3

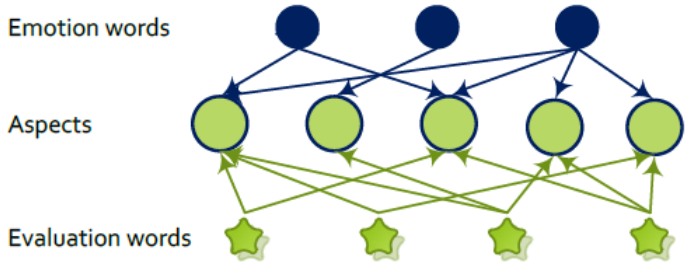
Overall sentiment: Positive

Sentiment Analysis - Twitter



Sentiment Analysis gets a lot more complicated…

Aspects, evaluation words and emotion words interaction



- An extracted aspect that is modified by many *evaluation words* is more likely to indicate an evaluative sentence.
- An extracted aspect that is modified by many *emotion words* is not a good indicator of an evaluative sentence.
- An evaluation word that does not modify *good* (high scored) aspects are likely to be a wrong evaluation word.
- The more evaluative the aspects are, the less emotional their associated emotion words should be.

$$asp(a_i) = \lambda \times \sum_{(i,j)\epsilon E_{va-a}} eva(va_j) - (1-\lambda) \times \sum_{(i,k)\epsilon E_{mo-a}} emo(mo_k) \quad (1)$$

$$eva(va_j) = \sum_{(i,j)\epsilon E_{va-a}} asp(a_i) \quad (2)$$

$$tmp(mo_k) = \sum_{(i,k)\epsilon E_{mo-a}} asp(a_i) \quad (3)$$

$$emo(mo_k) \propto -tmp(mo_k) \quad (4)$$

$$emo(mo_k) = -tmp(mo_k) + max = max - tmp(mo_k) \quad (5)$$

$$max = \max\{tmp(mo_1), tmp(mo_2), \ldots, tmp(mo_{|V_{mo}|})\} \quad (6)$$

Bing Liu @ AAAI-2011, Aug. 8, 2011, San Francisco, USA 160

Tugas

- Cari istilah di internet untuk :
  - Data Analytics
  - Data Scientist
- Penjelasan dilengkapi dengan gambar / diagram !!!