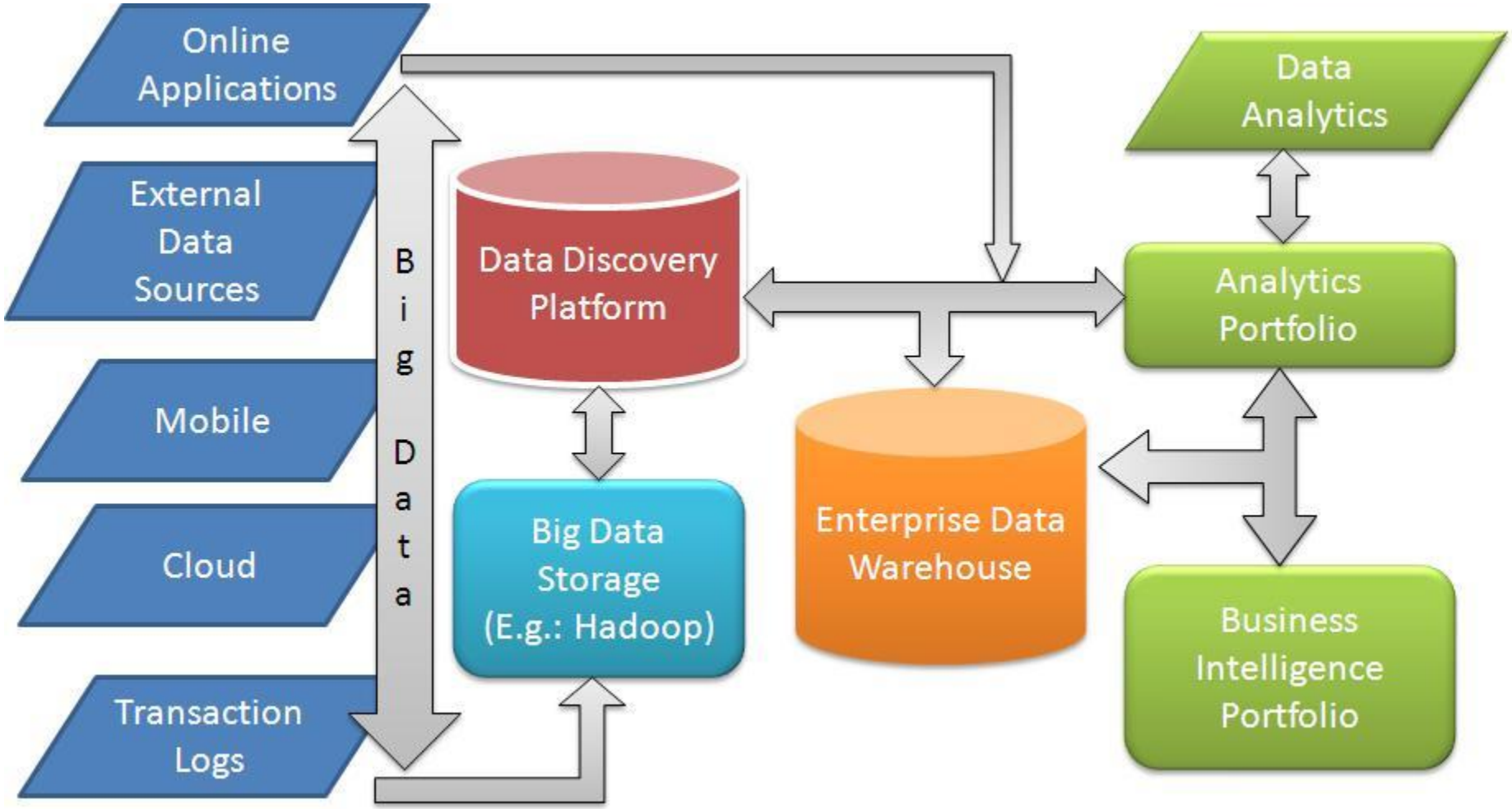
# BIG DATA

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

# Big Data
# Teknologi & Infrastructure

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

# Ecosystem Big Data



# Big Data :: Ecosystem

- Source :
  - Online Application , External Data Source, Mobile , Cloud, Transaction Logs
- Big Data Storage
  - Tools untuk menyimpan data berskala besar (big data)
  - Processing & Extracting Data
  - Contoh: Hadoop
  - Tools : Map/Reduce Architecture
- Data Discovery Platform ( Data Analytics )
  - Patterns
  - Answers Questions Business
  - Data as Gold Mining

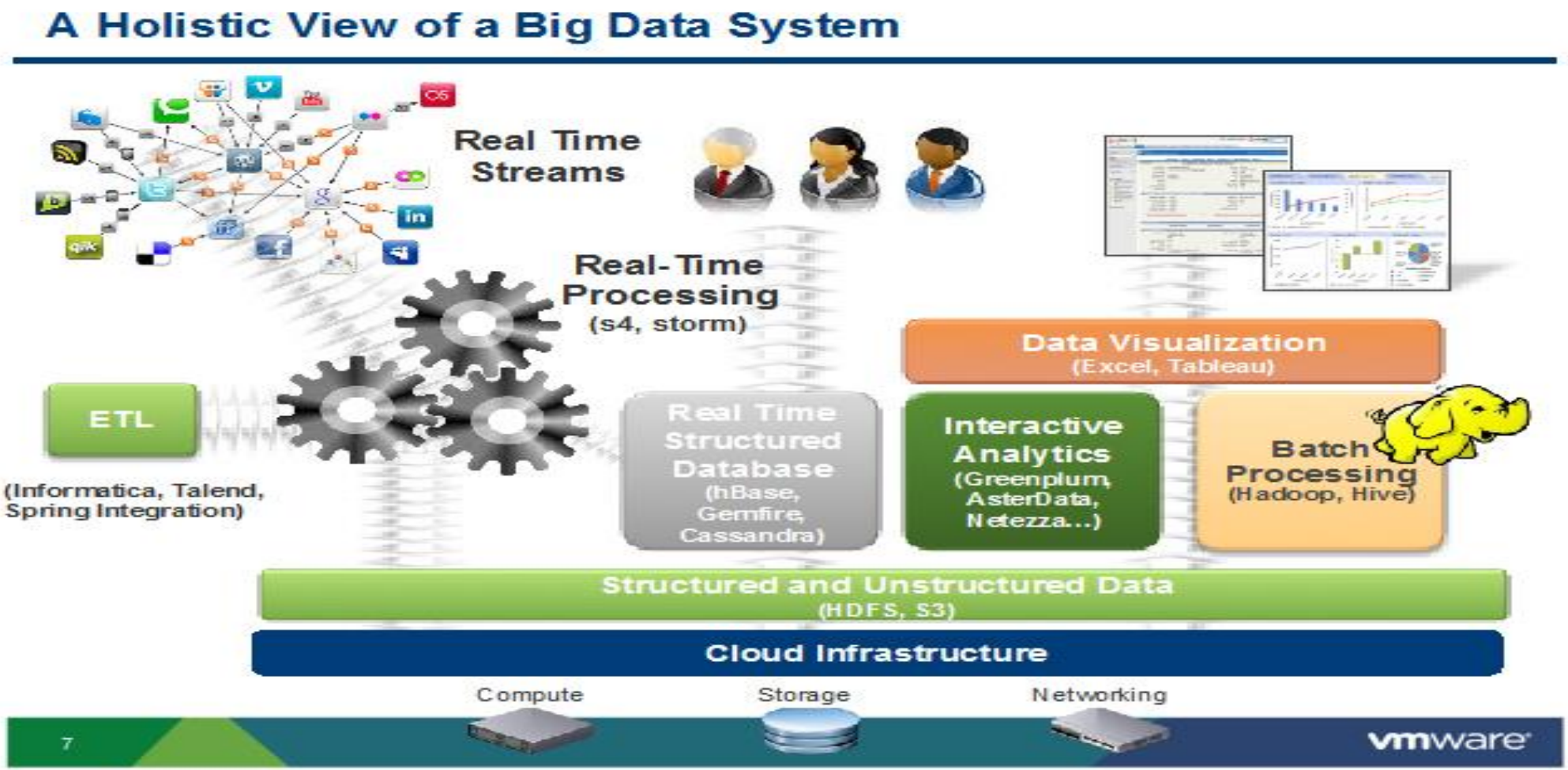
# Big Data :: Ecosystem

- Enterprise Data Warehouse
  - Integrasi Big Data dengan RDBMS ( Tradisional Database )
  - Mensupport informasi bagi organisasi
  - Platform data terintergrasi mudah di akses, digunakan, diubah untuk pengolahan informasi strategis
- Business Intellegence Portofolio
  - Analisa dari kinerja/laporan/data lampau dan kebutuhan terkini organisasi
- Data Analytics Portofolio
  - Data Science
  - Model-model prediksi data untuk meningkatkan mutu organisasi

# Big Data :: Fokus Pembelajaran

- ❑ Infrastructure
- ❑ Data Proccessing
- ❑ Data Analytics

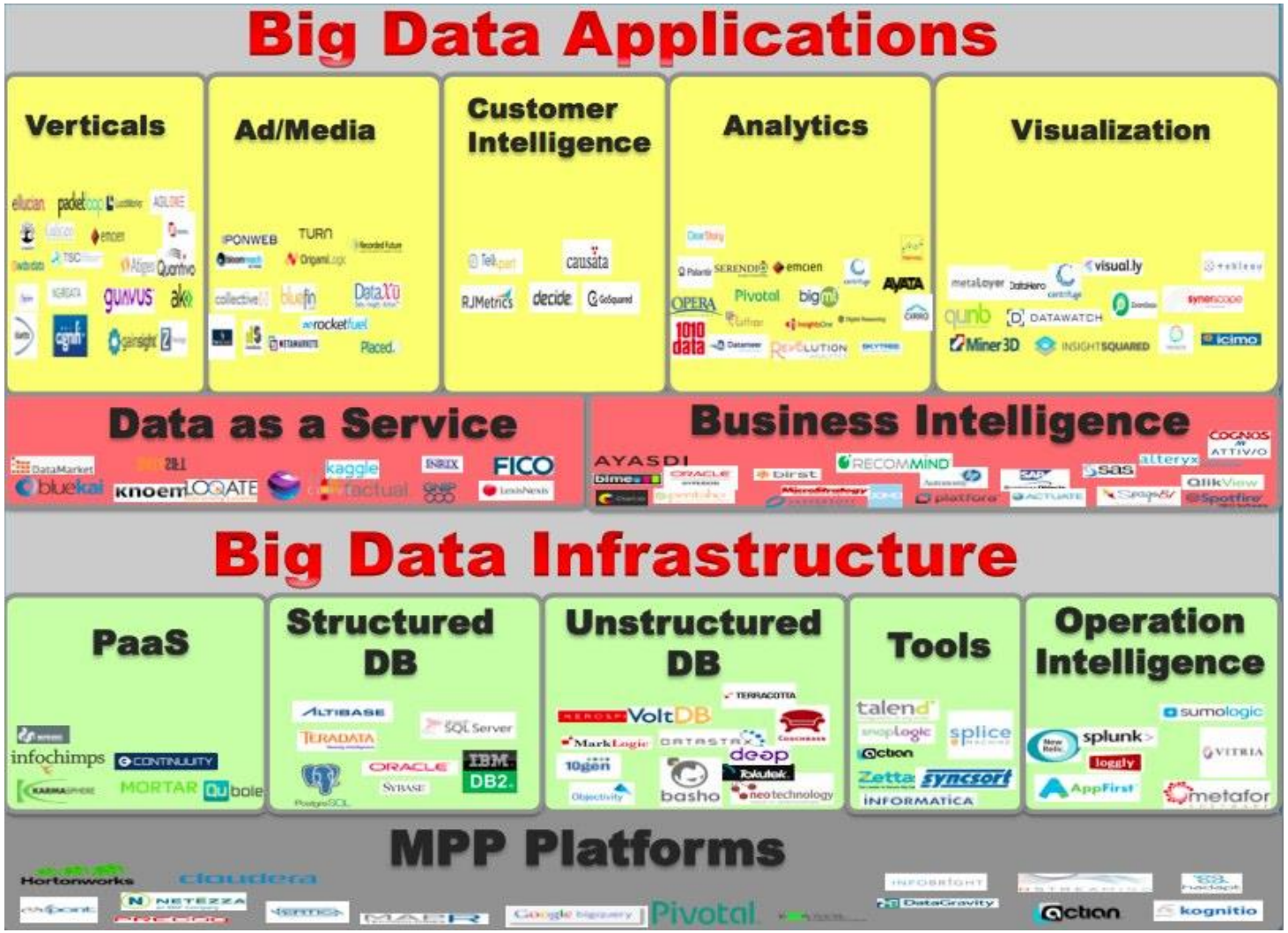
# Big Data :: Infrastucture




Infrastructure

# Big Data :: Infrastructure & Application

MPP Platform ::
Massive Parallel Processing Platofrm



# Big Data :: Infrastructure

- Server
  - Setup & Installation Server
  - Administration Server
  - Network Administration
  - Clouds Infrastructure
- Tools Data Processing
  - Realtime Structured Database (hbase, cassandra, gemfire)
  - Interactive Analytics (greenplum, asterdata,netezza)
  - Batch Processing ( hadoop, hive , hgrid )
- Security System

# Big Data :: Server Infrastucture

- Mengelola Big Data tidak harus dengan komputer super, tidak harus menggunakan komputer fisik, dan tidak harus mensetup sistem sendiri, karena ada penyedia Cloud.
  - ***Cloud Computing*** (komputasi awan): istilah bisnis penyewaan server dan/atau software di internet untuk menghemat biaya pengadaan dan perawatan ICT.
  - **Virtualisasi dan *Clustering*** merupakan teknologi lama yang dipakai untuk Cloud Computing dan Big Data.
  - Big Data dapat disediakan dengan Cloud dalam bentuk
    - SaaS
    - PaaS
    - IaaS
    - kombinasi dua atau ketiganya.

# SaaS :: Software as a Service

- **SaaS** menyediakan software aplikasi di internet (penyedia sistem cloud). Berbeda dengan perangkat lunak tradisional yang disediakan di komputer masing-masing, perangkat lunak SaaS terdapat di dalam jaringan dan hanya dipasang ketika digunakan.
- Contoh SaaS:
  - Email : gmail.com, yahoo.com,Zimbra
  - Document Sharing : docs.google.com, Quip
  - Project Mangement : Trello, Slack
  - Social Media : Facebook, LinkedIn, Path, Twitter
- SaaS khusus Big Data: Jkool www.jkoolcloud.com, dll.

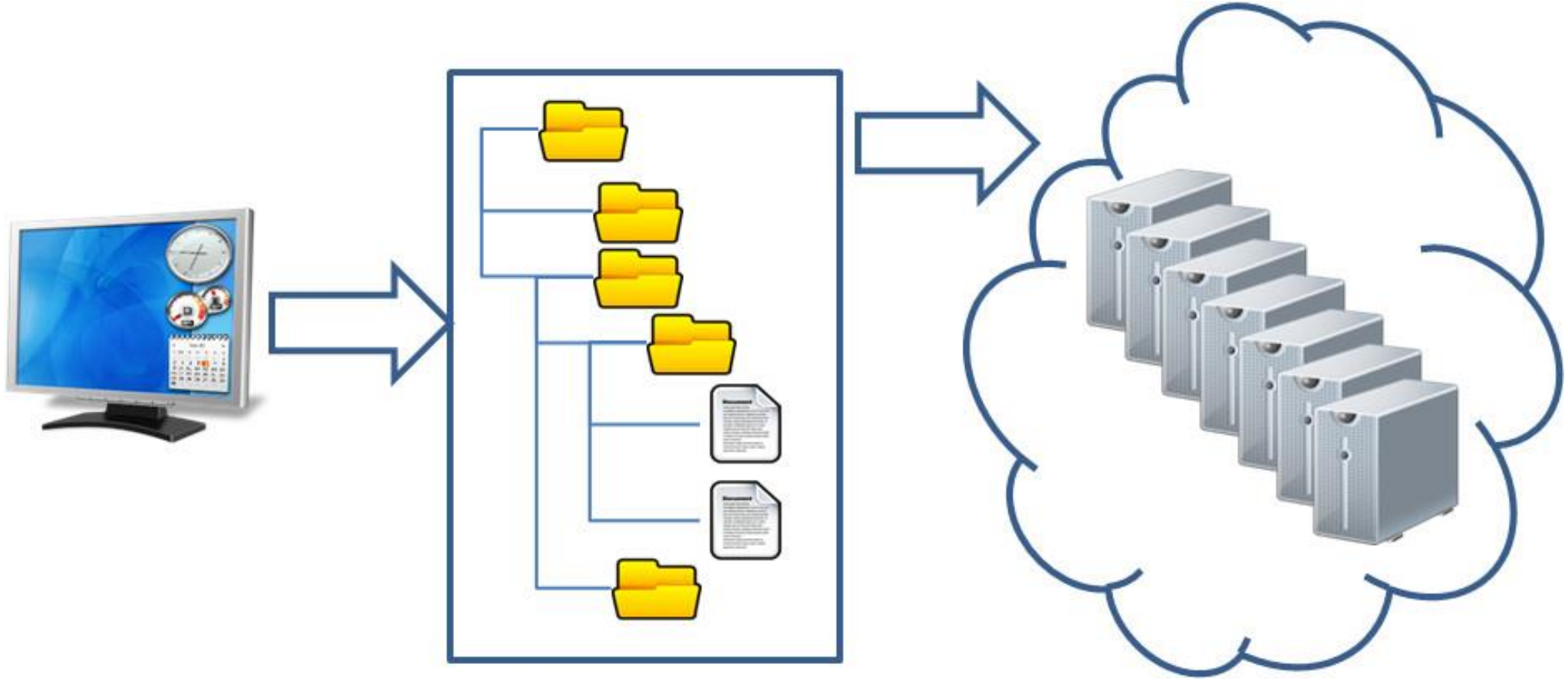
# PaaS :: Platform as a Service

- **PaaS** menyewakan sistem (SW dan HW) berupa platform untuk pengembangan aplikasi. PaaS menyediakan semua software yang dibutuhkan, seperti tool pemrograman dan database, termasuk untuk testing, deployment, dll.
- Contoh:
  - App Inventor untuk membuat aplikasi Android dengan web (ai2.appinventor.mit.edu),
  - Cloud 9 ( c9.io ) : platform Ubuntu + tools development : untuk pengembang software
  - OpenShift (RedHat) yang mendukung php, java, python, ruby, dll.,
  - Google Apps Engine, Amazon Web Services, Cloud Foundry, dll.

# IaaS :: Infrastructure as a Service

- **IaaS** menyewakan infrastruktur dalam bentuk komputer (virtual server, dns server, mail server, dll.), akses jaringan, penyimpanan (SAN: Storage Arena Network / NAS: Network-Attached Storage), cluster Big Data, dll.
- Contoh OSS: CloudStack, OpenStack, Eucalyptus, Proxmox (Linux distribution), OwnCloud, dll.
- Penyedia server: Amazon EC2, Ubuntu EC, Amazon Cloud Drive, Infinys (www.isi.co.id), dll.
- Free IaaS: Drive.google.com, Dropbox, dll.

# Cluster Big Data dan Cloud



# Google :: Big Data

- GFS (Google File System): Sistem (penyimpanan) file terdistribusi. Penyimpanan tidak dalam sebuah harddisk dalam sebuah komputer, tapi dalam banyak tempat penyimpanan secara menyebar *(distributed), clustering*.
- MapReduce: Arsitektur Program pengolahan data khusus untuk database terdistribusi, melalui sistem cluster, ribuan komputer.
- BigTable: sistem database Google, mendukung sistem file terdistribusi dan cocok untuk data tidak terstruktur yang diproses secara tumpukan *(Batch Processing),* bukan interaktif atau real-time.

# Hadoop :: Open Source Big Data

- HDFS (Hadoop Distributed File System): Seperti GFS, sistem (penyimpanan) file terdistribusi. Penyimpanan tidak dalam sebuah harddisk dalam sebuah komputer, tapi dalam banyak tempat penyimpanan secara menyebar *(distributed file system),* dan *clustering*.
- Hadoop MapReduce: Mengambil teknologi Google, Arsitektur Program pengolahan data khusus untuk database terdistribusi, melalui sistem cluster, ribuan komputer.
- Hadoop Base (HBase): seperti Google BigTable.

# Big Data :: Yahoo Server

# Hadoop at Yahoo (1)

- Yahoo! has more than 100,000 CPUs in over 40,000 servers running Hadoop, with its biggest Hadoop cluster running 4,500 nodes.
- Yahoo! stores 455 petabytes of data in Hadoop.
- That's big, and approximately four times larger than Facebook's beefiest Hadoop cluster.

- Source: http://www.techrepublic.com/article/why-the-worlds-largest-hadoop-installation-may-soon-become-the-norm/

# Hadoop at Yahoo (2)

- Yahoo! move email into Hadoop systems so that Yahoo! can analyze huge volumes of email for anti-spam purposes.
- Another example is Flickr photos. All photos are in Hadoop, so Yahoo! can run image recognition processes, but the main source of truth for photo serving is not in Hadoop.

- Source: http://www.techrepublic.com/article/why-the-worlds-largest-hadoop-installation-may-soon-become-the-norm/

# Big Data :: Other Platform

- Amazon: mengelola big data terkait pelayanan pelanggan, menjual jasa cloud untuk big data, seperti Amazon MapReduce dan Amazon EC2 Hadoop cluster.
- IBM: big data bertambah 2,5 exabyte per hari. Setahun Rp 13 trilyun pemasukan dari Big Data, a.l. mengolah big data untuk mengurangi kemacetan di Lyon Perancis.

# Open Source – Big Data

Contoh produk populer dan Open Source:

- Sistem operasi: Linux CentOS, Ubuntu, BlankOn, dll.
- Framework pengembangan: Apache Hadoop dengan Hadoop MapReduce, dll.
- Database NoSQL: Apache Cassandra, Apache HBase, MongoDB, dll.
- Tool dan Bahasa Pemrograman: Eclipse, Java, Hive (SQL), Pig, Python, R Project, dll.

# Open Source – Big Data

Produk Cloud untuk Big Data:

- Google Big Query, Google Compute Engine (Hadoop, Hive, Pig), Amazon Web Services MapReduce, Yahoo! Genome, dll.

Produk Hasil Modifikasi:

- Cloudera Enterprise (Hadoop for Enterprise), IBM Watson Foundations (Hadoop), IBM InfoSphere Streams, dll.

# Big Data :: Infrastructure – Security System

***Your unstructured or semi-structured data is at risk!***


Depending on the big data solution used (Hadoop - MapReduce, BigTable, MarkLogic, Cassandra, MongoDB, etc.) you might miss:

- ✓ Proper authentication
- ✓ Access control
- ✓ File system integrity
- ✓ Data validation
- ✓ OS hardening

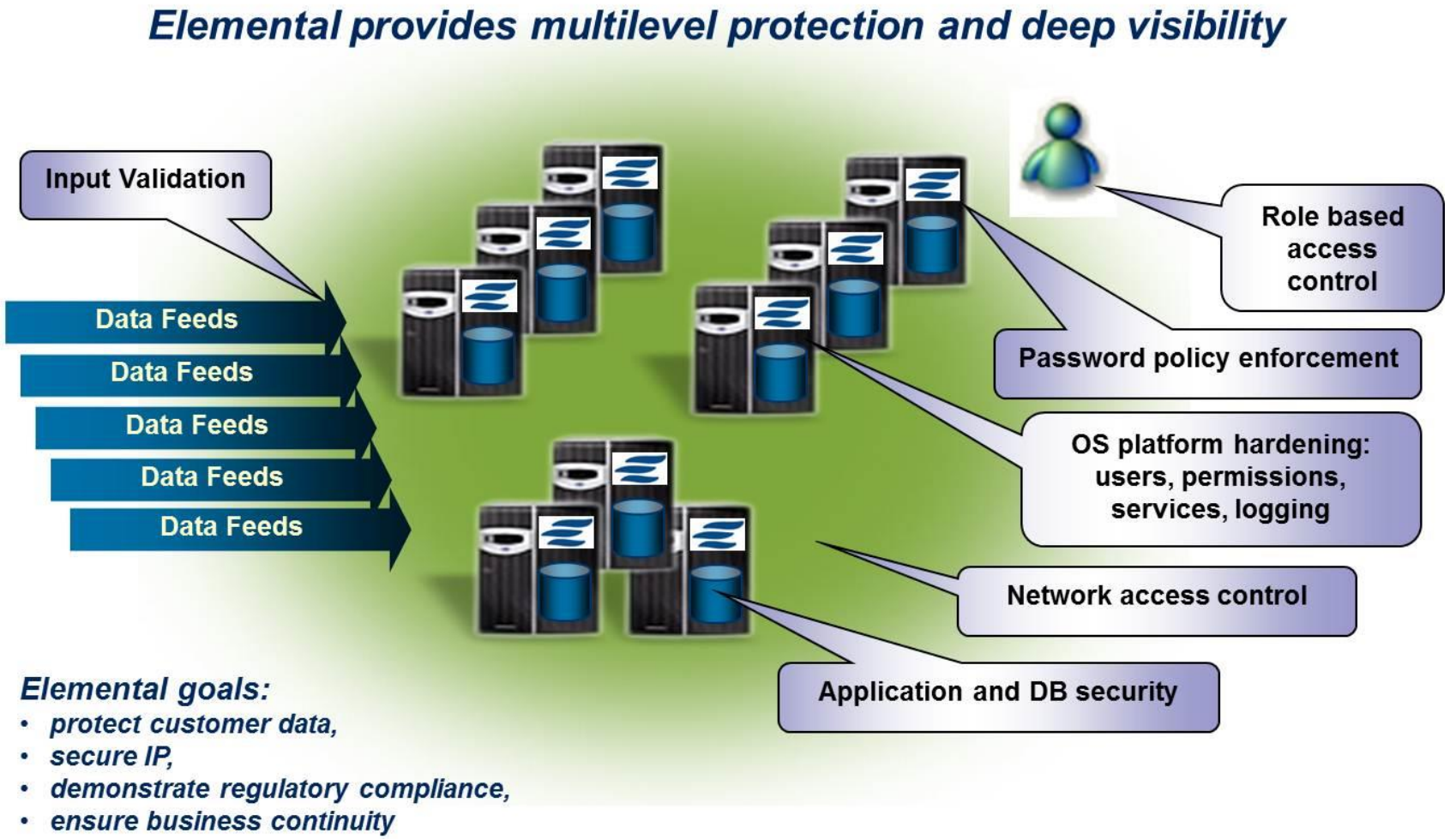
# Security Syetem :: Unsecure Data Feed

Sumber data2 yang belum aman

- Social Networks
- Data Repositories
- WebFeed :: RSS, RDF
- Log Files , Sensors Data
- Other Resources

- Alternaif Solusi
  - Proper authentication
  - Access Control
  - File System Integrity
  - Data Validation
  - OS Hardening

# Big Data :: Security System



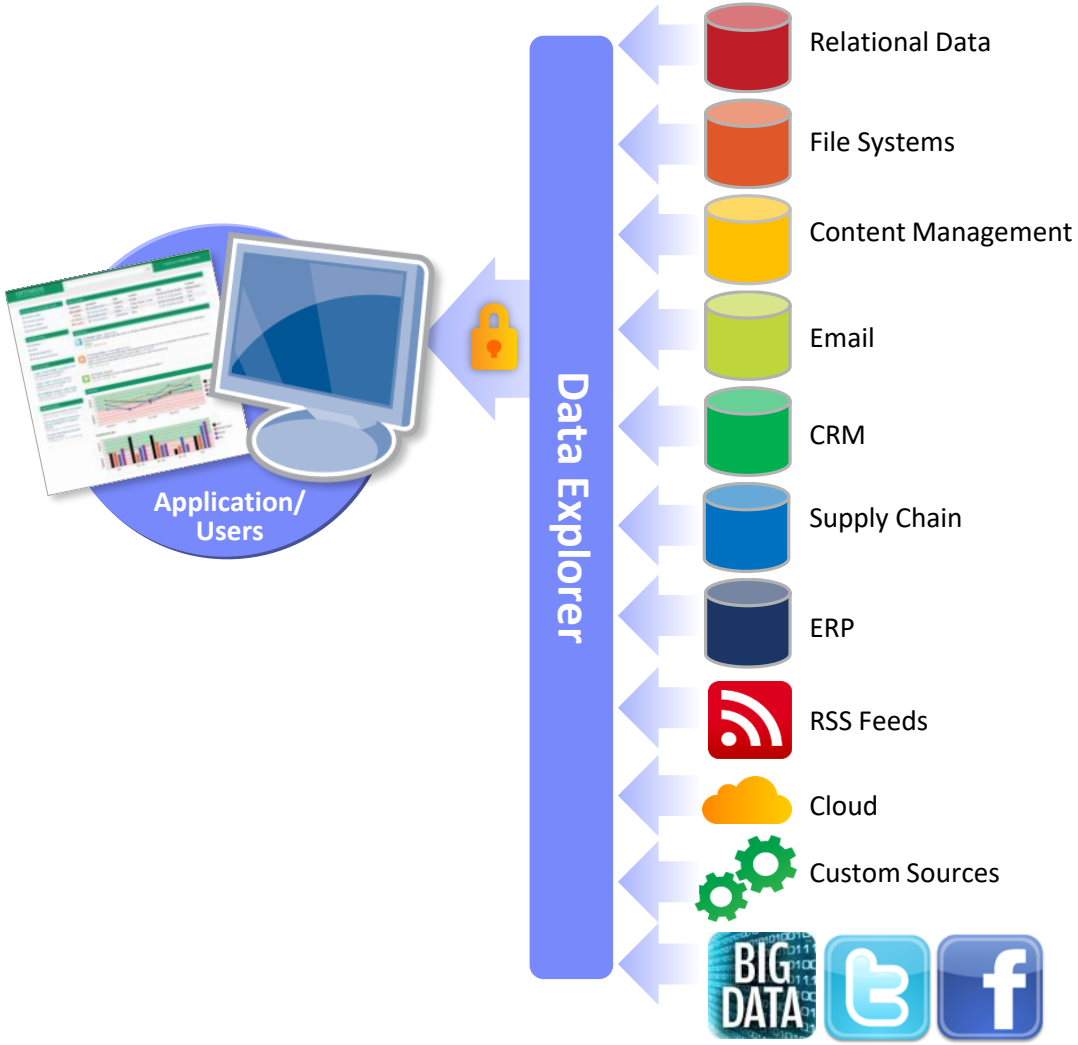
# Security System :: Build System

- **Passwords** - sebagian besar sistem NoSQL / Big Data systems tidak memiliki password atau menggunakan password default, karenanya semua orang mungkin dapat mengaksesnya. Buat sistem memaksa menggunakan password.
- **Input Validation** - Sistem NoSQL secara normal memungkinkan ter-ekspose SQL injection, bahkan SQL Injection dapat dilakukan melalu script JavaScript. Buat sistem aman dari sql injenction
- **Role-based Access Control** - Akses ke sistem didefinisikan dengan roles, buat sistem dengan tingkatan akses role
- **OS Hardening** - Sistem Operasi harus kuat dan terisolir, dengan fokus menerapkan proteksi berdasarkan area: *users, permissions, services, logging*., buat sistem yang kuat dan aman.
- **Persistent Control** - Secara konstan memonitoring dan kontinu dalam menerapkan tingkat keamanan pada server (host-level security policies).
- **Responsive to Change** - Sistem Akses kontrol yang secara otomatis mengikuti perubahan role dan kebijakan keamanan.
- **In-Line Remediation** - Sistem dapat melakukan perbaharuan konfigurasi, membatasi aplikasi dan perangkat, membatasi akses jaringan dari ancaman yang tidak berkepentingan

# Questions ?

- Sebutkan kategori pembagian pembelajaran Big Data (fokus)
- Sebutkan contoh  Cloud Big Data menggunakan konsep:
  - SaaS
  - PaaS
  - IaaS

# Ekplorasi Big Data : Value & Diagram




**Temukan, Visuasasikan & Pahami** semua **big data untuk meningkatkan** business knowledge (pengembangan bisnis) ::

- Efisiensi yang besar dari bisnis proses
- Hal baru (bisnis baru) dari kombinasi data dan analisa data dari berbagai pendekatan dan cara
- Kembangkan model bisnis baru yang dapat menghasilkan pertumbuhan / kenaikan pasar dan peningkatan pendapatan

Perhatikan teknologi yang digunakan !!!

# Keamanan Data !!

Pengembangan Keamanan / Intelegensi dengan memperbaiki solusi keamanan tradisional dengan melakukan analisa semua tipe dan sumber data yang ditangani

**Enhanced Intelligence & Surveillance Insight**

**Analyze data-in-motion & at rest to:**
- Temukan keterhubungan
- Deteksi pola (patterns) dan fakta-fakta
- Lakukan maitenance informasi keuangan

**Real-time Cyber Attack Prediction & Mitigation**

**Analyze network traffic to:**
- Temukan ancaman terbaru lebih dini
- Deteksi ancaman2 bersifat komplek
- Lakukan aksi secara real-time

**Crime prediction & protection**

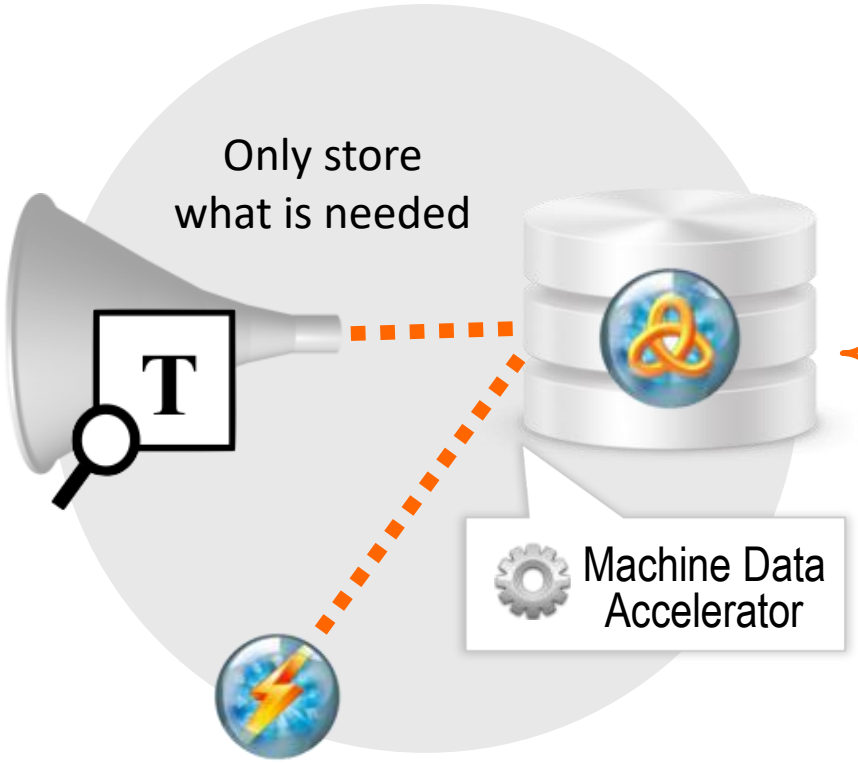
**Reduce Customer Churn**

**Analyze Telco & social data to:**
- Kumpulkan bukti-bukti criminal
- Pencegahan criminal activities
- Proaktif terhadap segala bentuk criminals
- Customer Retention

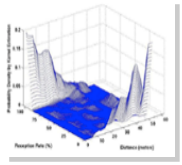
Perhatikan teknologi yang digunakan !!!

# Big Data - Operations Analysis: Value & Diagram

Raw Logs and Machine Data




Indexing, Search



Statistical Modeling

Root Cause Analysis

Federated Navigation
& Discovery

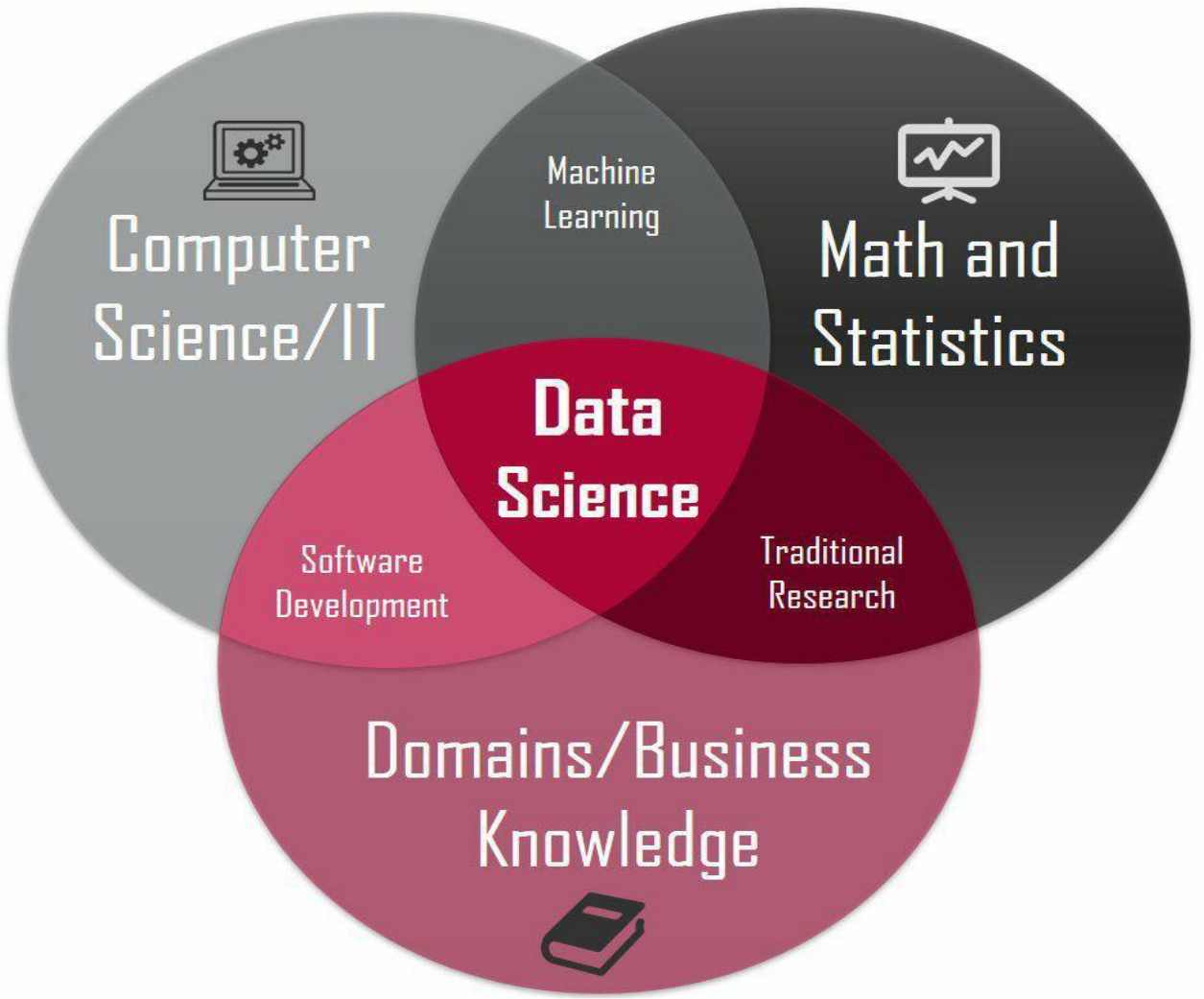
**Pengambil Keputusan**

Perhatikan teknologi apa yang akan digunakan !!!

# BigData :: Data Science

Tedsuka:
Akhir akhir ini sering harus menjelaskan tentang ilmuwan data atau "data scientist". Gambar ini cukup baik menjelaskan kaitan antara kompetensi teknologi informasi, matematika dan bidang usaha. Dengan terjadinya ledakan ketersediaan data dalam jumlah sangat besar karena meluasnya transaksi online dan IoT, bidang ini menjadi sangat strategis. Potensi data yang ada sangat besar untuk analisis dan pemahaman kondisi faktual untuk membuat strategi dan mengambil keputusan. Untuk itu diperlukan  #data science

Teddy Sukardi



# Teknologi Terkait Big Data

- *Virtualization* (Virtualisasi): Cara menjalankan komputer “maya” *(guest)* pada komputer “asli” *(host)*. Contoh: sebuah komputer menjalankan sistem operasi Linux, lalu menjalankan program mesin virtual untuk menjalankan sistem operasi lain (Linux, Windows, dll.)
- *Clustering*: Membuat kluster beberapa komputer digabung dalam jaringan menjadi “satu” komputer.
- *Distributed & Paralel Computing*: menjalankan sebuah proses (misal penyimpanan dan pengolahan data) pada banyak komputer dalam suatu cluster.

# Big Data dan Cloud

Mengelola Big Data tidak harus dengan komputer super, tidak harus menggunakan komputer fisik, dan tidak harus mensetup sistem sendiri, karena ada penyedia Cloud.

- *Cloud Computing* (komputasi awan): istilah bisnis penyewaan server dan/atau software di internet untuk menghemat biaya pengadaan dan perawatan ICT.
- Virtualisasi dan *Clustering* merupakan teknologi lama yang dipakai untuk Cloud Computing dan Big Data.
- Big Data dapat disediakan dengan Cloud dalam bentuk SaaS, PaaS, IaaS, atau kombinasi dua atau ketiganya.