Introduction to Big Data

A Few Quotes to Start Us Off

Being locally relevant has always been the core of success in retailing, going back 100 years to the town general store whose owners knew what their customers wanted, liked, and would like to try.

(Stephen Quinn, 2012)

I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

(Alan Turing, 1950)

Learning Objectives

- **What is data?**
- **How is data stored and analyzed?**
- **What is big data?**
- **How is big data stored and analyzed?**

What is data?

Data (quantitative vs qualitative)

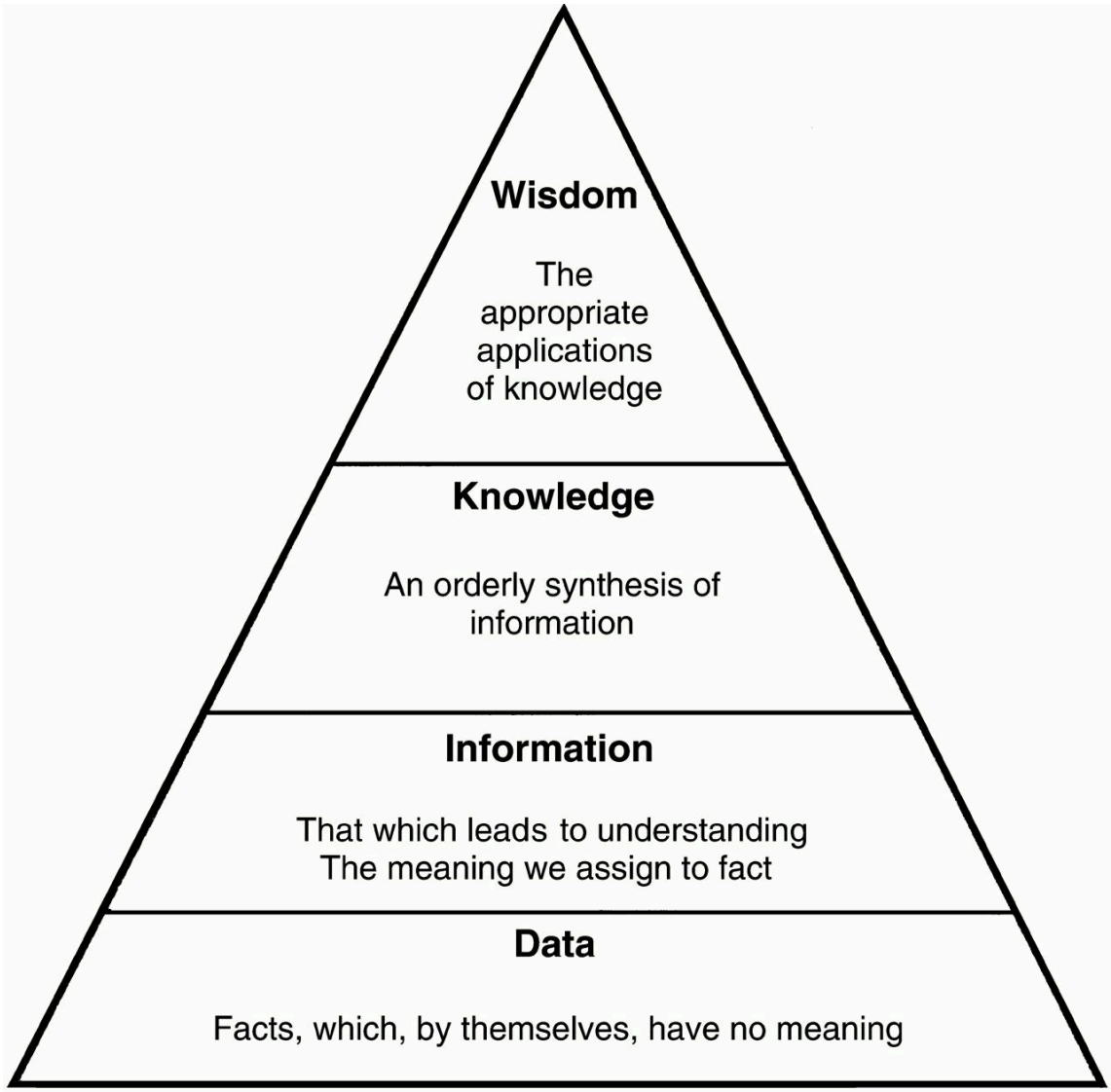
- **Quantitative** (numerical)
  - Observable and measurable
  - Structured and objective
  - Examples:
    - Income
    - Height

- **Qualitative** (descriptive)
  - Observable but not measurable
  - Unstructured and subjective
  - Examples:
    - Favorite color
    - Zip code

Example 1: *Oil Painting* 	**Example 1:** *Oil Painting* 
Qualitative data: • blue/green color, gold frame • smells old and musty • texture shows brush strokes of oil paint • peaceful scene of the country • masterful brush strokes	**Quantitative data:** • picture is 10" by 14" • with frame 14" by 18" • weighs 8.5 pounds • surface area of painting is 140 sq. in. • cost $300
Example 2: *Latte* 	**Example 2:** *Latte* 
Qualitative data: • robust aroma • frothy appearance • strong taste • burgundy cup	**Quantitative data:** • 12 ounces of latte • serving temperature 150° F. • serving cup 7 inches in height • cost $4.95

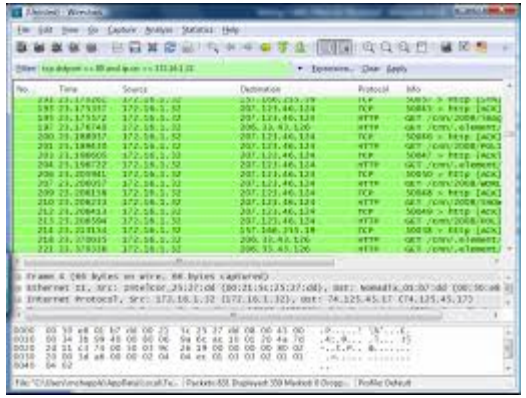
Data, information, and knowledge

- **Data**
  - raw value
  - (e.g. income)
- **Information**
  - set of data with meaning
  - (e.g. credit application)
- **Knowledge**
  - interpretation of information in a context
  - (e.g. risky credit applicant)



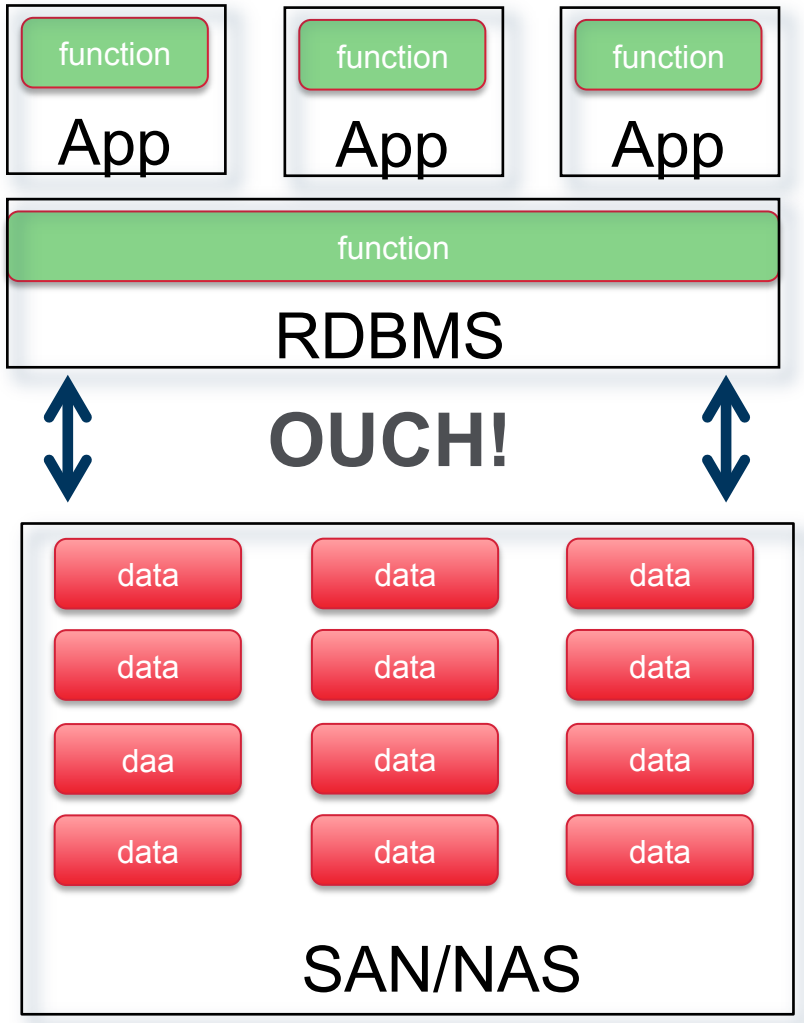
Data sources

- **Finance**:
  - CRM, ERP, billing, …
- **Social media**:
  - Twitter, LinkedIn, Facebook, …
- **Security**:
  - System logs, packet captures, audit logs, …
- **Pharma**:
  - Clinical trials, clinical practice, …
- . . .



How is data stored and analyzed?

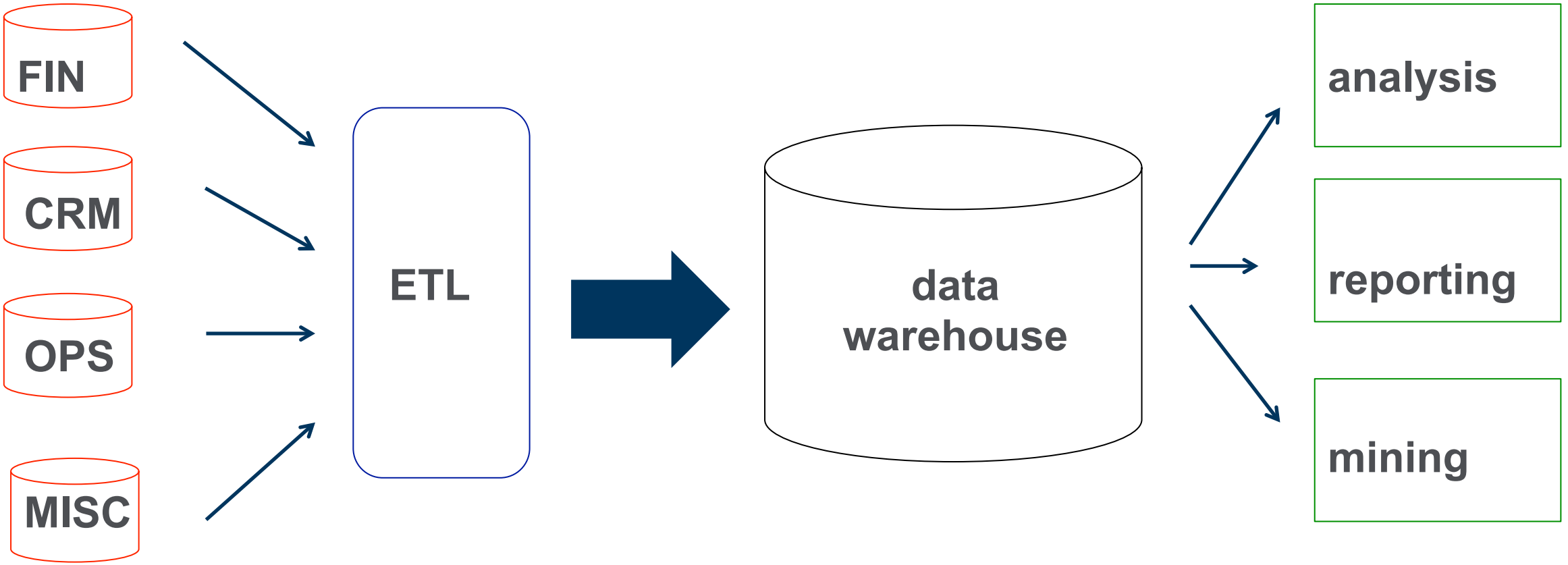
Data storage



DAS VS NAS VS SAN

	DAS	NAS	SAN
Storage Type	Sectors	Shared files	Blocks
Data Transmission	IDE/SCSI	TCP/IP, Ethernet	Fiber Channel
Access Mode	Clients or servers	Clients or servers	Servers
Capacity (Bytes)	10^9	10^9-10^{12}	>10^{12}
Complexity	Easy	Moderate	Difficult
Management Cost (per GB)	High	Moderate	Low

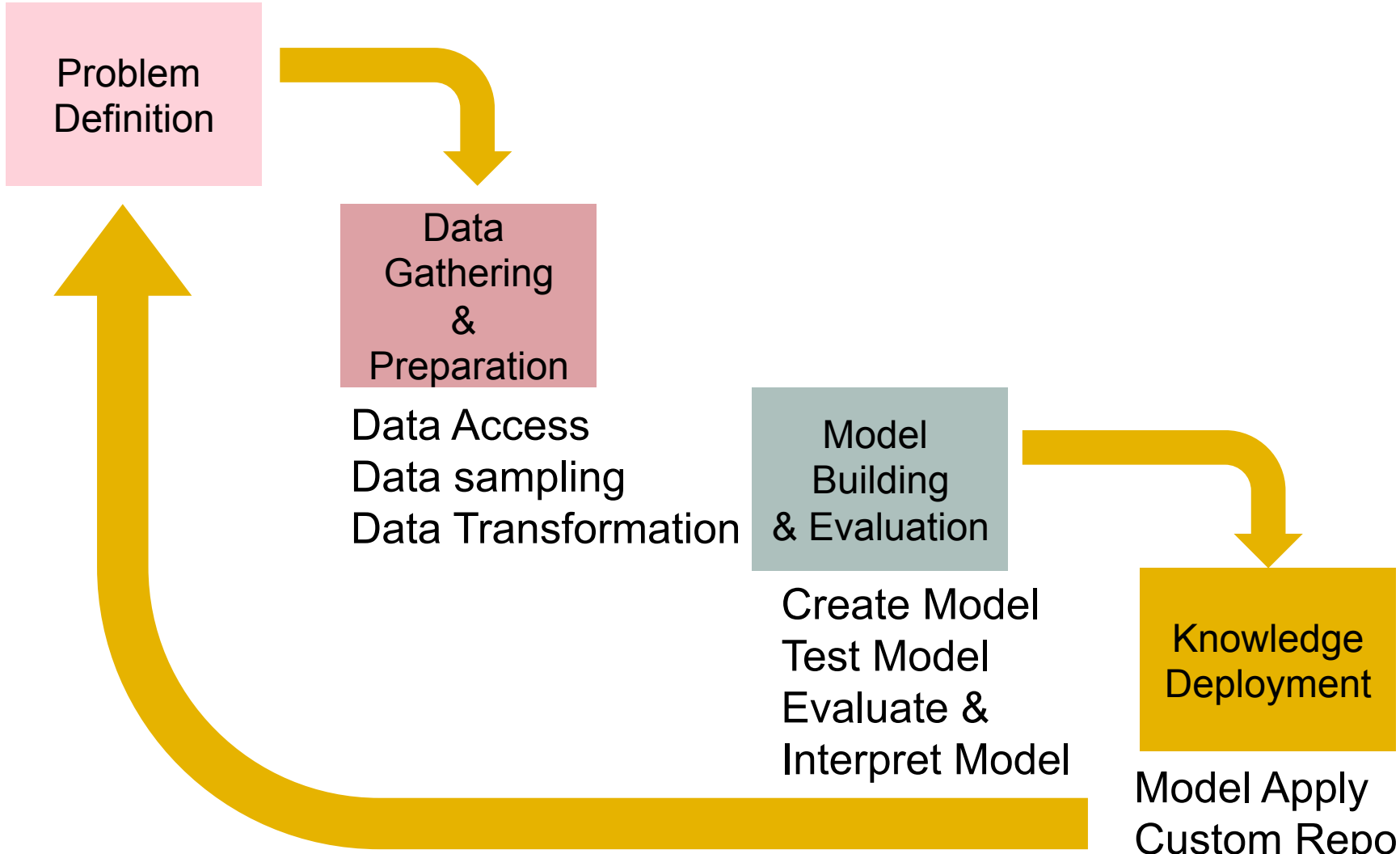
Data warehouse



Descriptive vs Predictive Analysis

Descriptive	Predictive
Demographic answers • How long's Mia been with us? • What's Mia's annual salary, average incentive in past 5 years? • Where does Mia live? What's her commute distance to the office?	**Predictive answers** • How long will Mia stay with us? • What other criteria will have an impact on Mia's retention? • What's the value potential on business impact Mia has over the next 2 years?
Performance answers • What hours did Mia work last week? • How many days was Mia absent last year? • What job promotions did Mia get the past 3 years? • Is Mia performing better versus last year? • What's the trend of Mia's merit rating in the past 3 years?	**Recommendations** • Which is the appropriate salary Mia most likely motivate to stay with us the next years? • What are the best incentives to get Mia stay with us? • Which training should we offer to Mia to make her perform even better?
Hindsight: rearview, reporting, dashboard, metrics, ratio's, slicing-and-dicing, tracking, monitoring,…	**Foresight: future-looking, likelihood, probability, hidden patterns, mathematical models, statistical forecasting,…**

Traditional Data Mining



What is big data?

Variety of Data Sources and Formats

Photos/Video

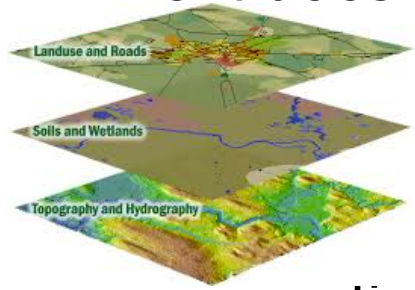
Customer data

Accounting & finance

Text docs

Social media

Log files



geospatial

Credit Card Transactions

Web user behaviour

Sensors

Big Data Definition

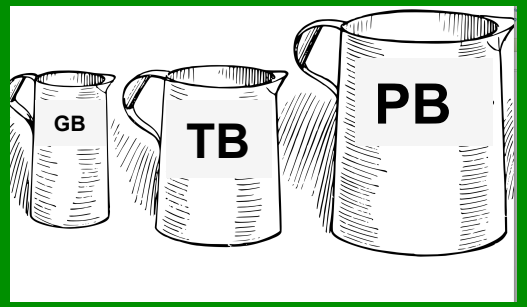
- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

Big Data

"Big data" – the realization of greater business intelligence by storing, processing and analyzing data that was previously ignored due to the three Vs:

Volume



The volume of data is too large for traditional database software tools to cope with

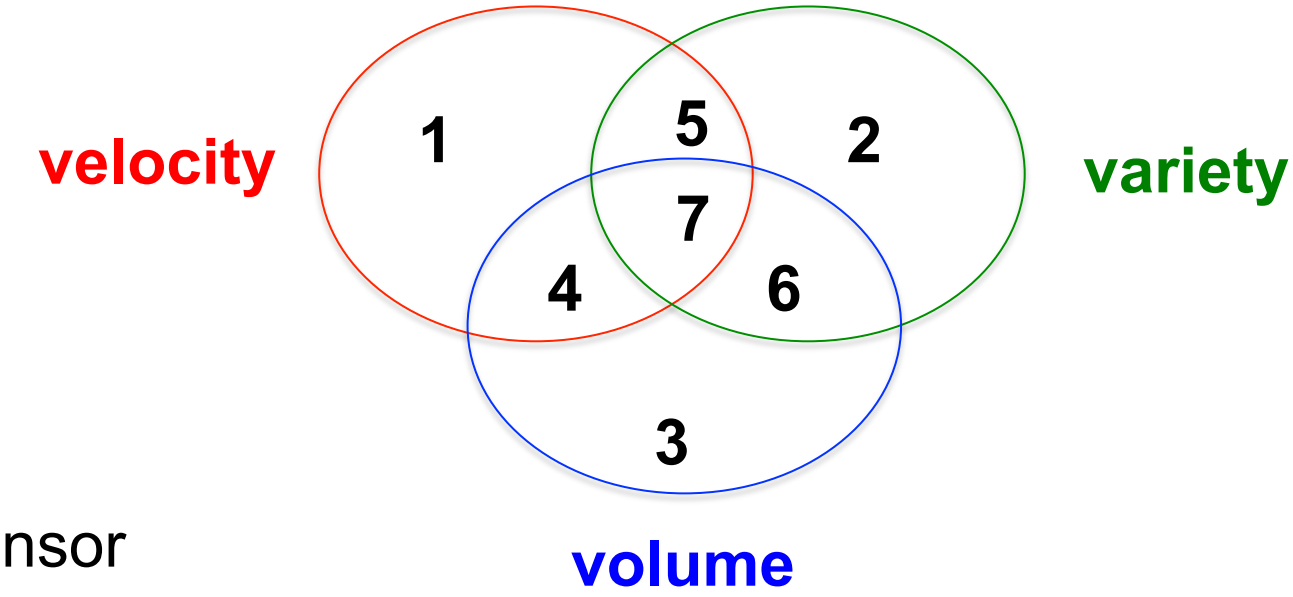
Velocity

The data is being produced at a rate that is beyond the performance limits of traditional systems

Variety

The data lacks the structure to make it suitable for storage and analysis in traditional databases and data warehouses

Summarize Big Data Using the 3 V's



Examples

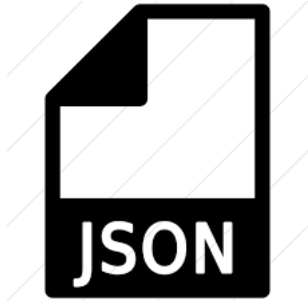
1 = small-scale sensor
2 = laptop
3 = image server
4 = large-scale sensor
5 = cell phone
6 = file server
7 = facebook, linkedin

Describe Variety of Big Data

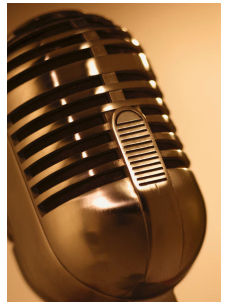
structured

Year	Total			On-Budget		
	Receipts	Outlays	Surplus or Deficit (–)	Receipts	Outlays	Surplus or Deficit (–)
1901	588	525	63	588	525	63
1902	562	485	77	562	485	77
1903	562	517	45	562	517	45

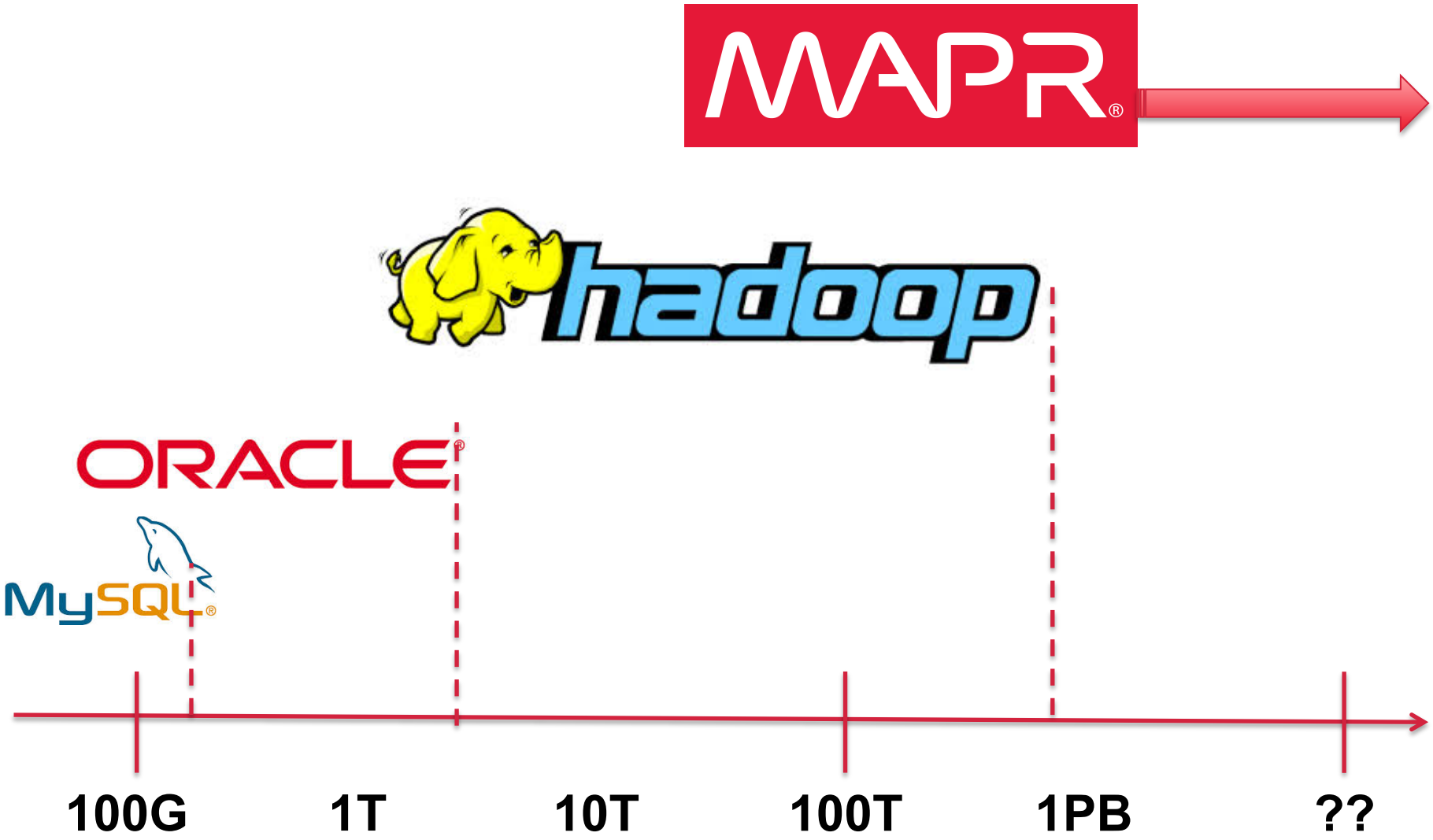
semi-structured



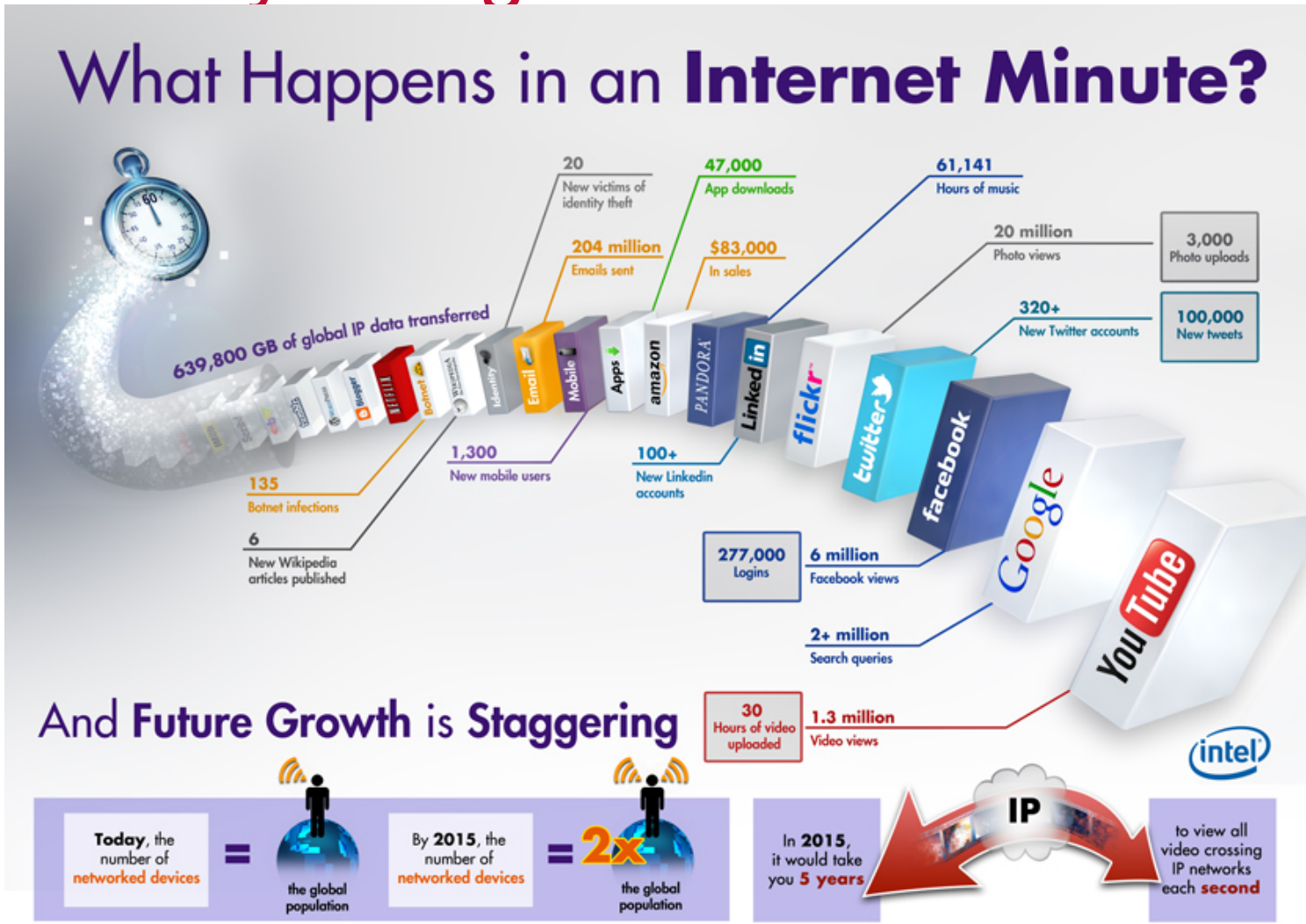
unstructured



Describe Volume of Big Data



Describe Velocity of Big Data



http://2renaissance.org/2012/08/27/will-the-internet-of-things-services-address-real-world-challenges/

How is big data stored and analyzed?

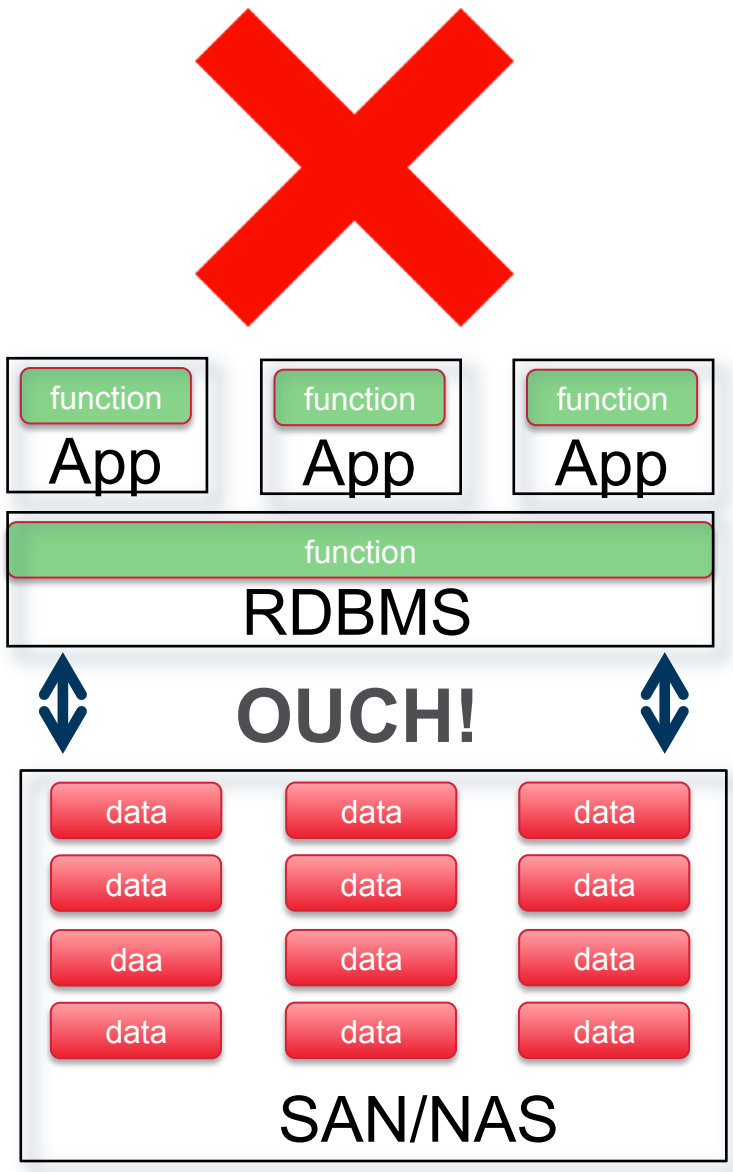
"Because RDBMSs can be beaten by more than an order of magnitude on the standard OLTP benchmark, then there is no market where they are competitive. As such, they should be considered as legacy technology more than a quarter of a century in age, for which a complete redesign and re-architecting is the appropriate next step."

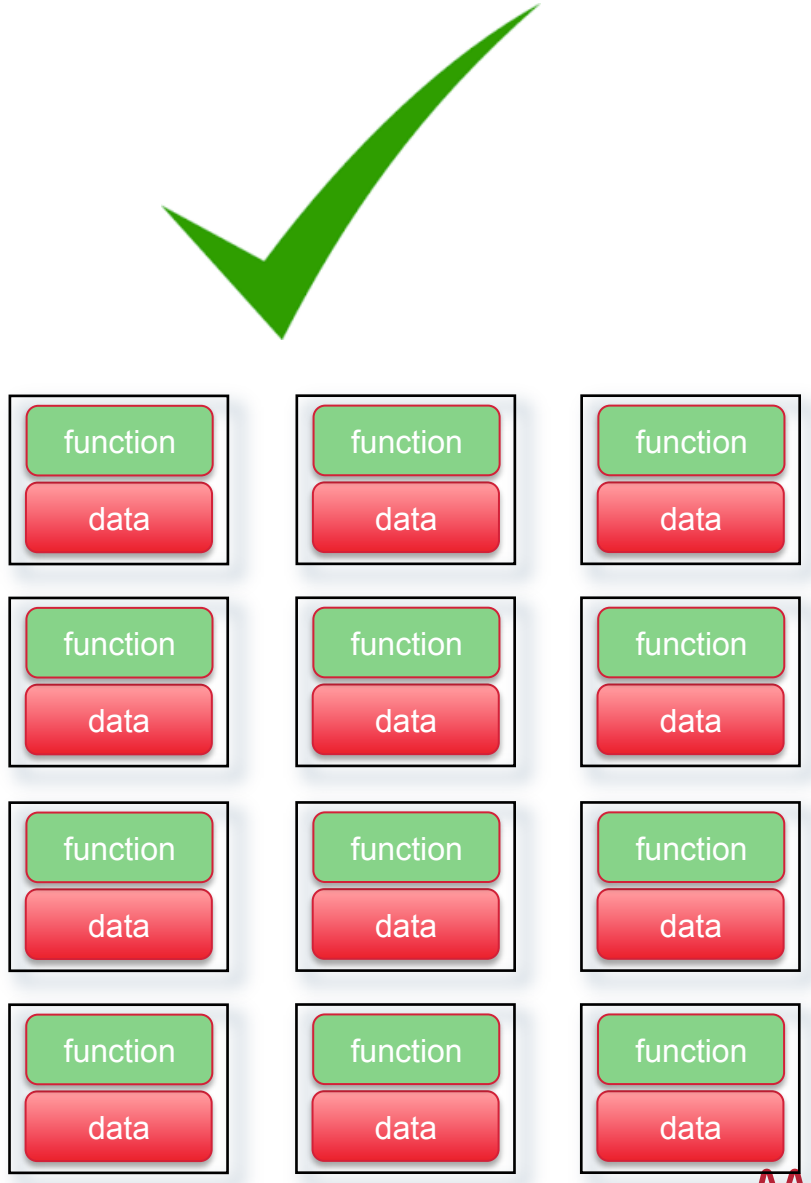
Michael Stonebraker (Creator of Ingres and Postgres)

Identify Ways to Scale to Process Big Data

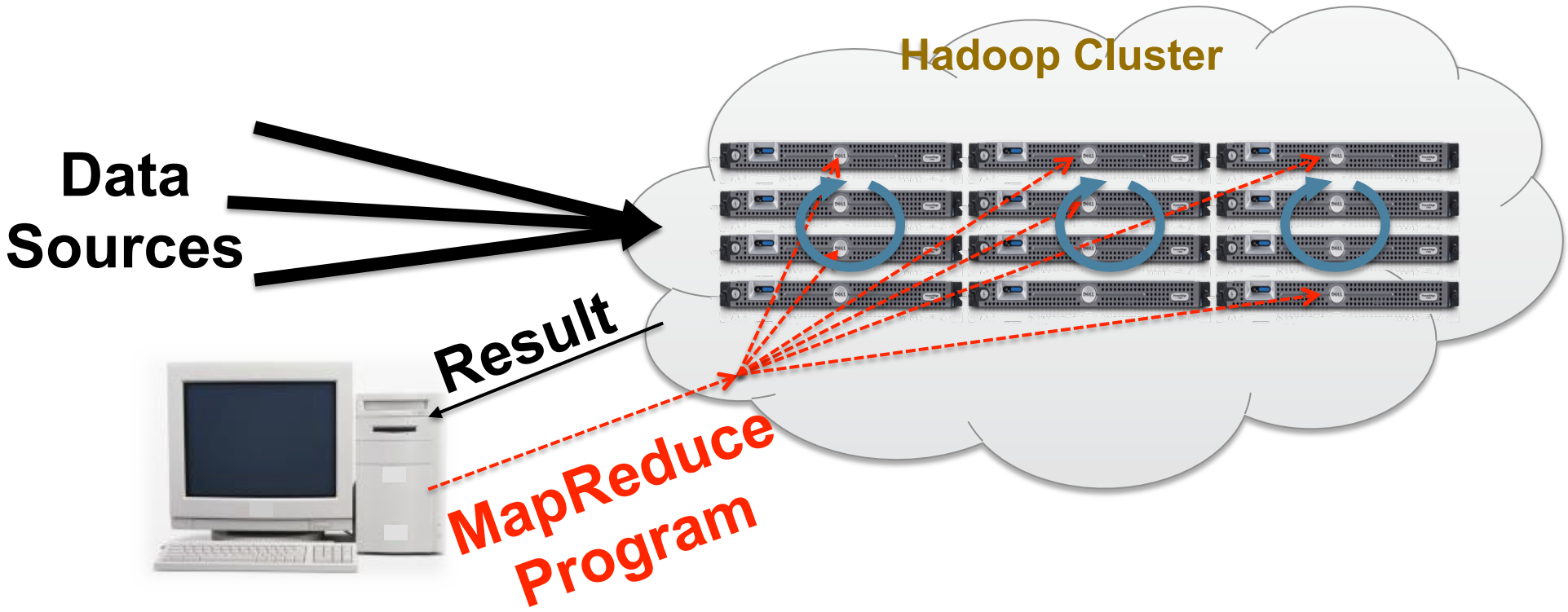
Paradigm	Example	limitations
Scale-up	Monolithic database	Infrastructure CAPEX/OPEX Availability/scalability Data gravity
Scale-out	Grid cluster	Synchronization overhead Programming complexity Specialized hardware
Sampling	Any approach	Lower accuracy Lower precision

Exploit Locality of Reference



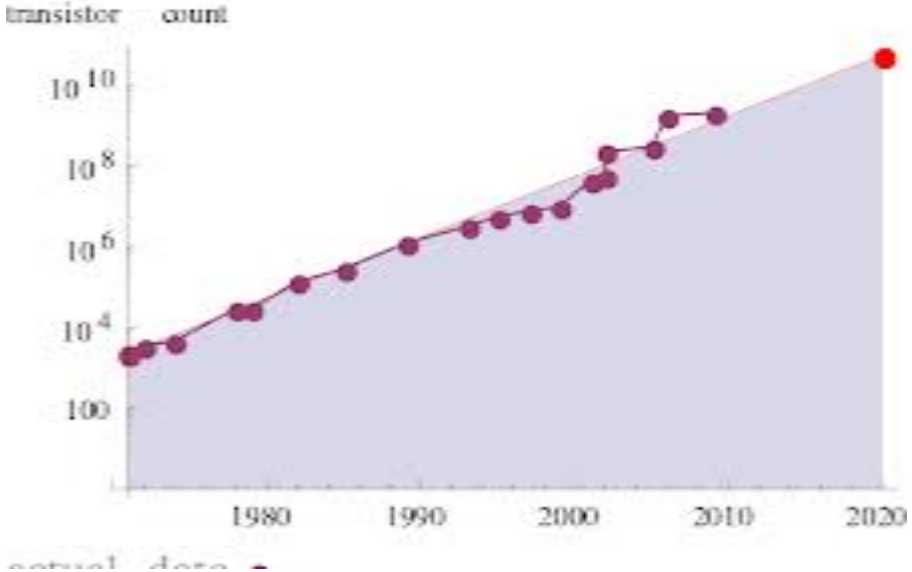


Distribute Data and Computation

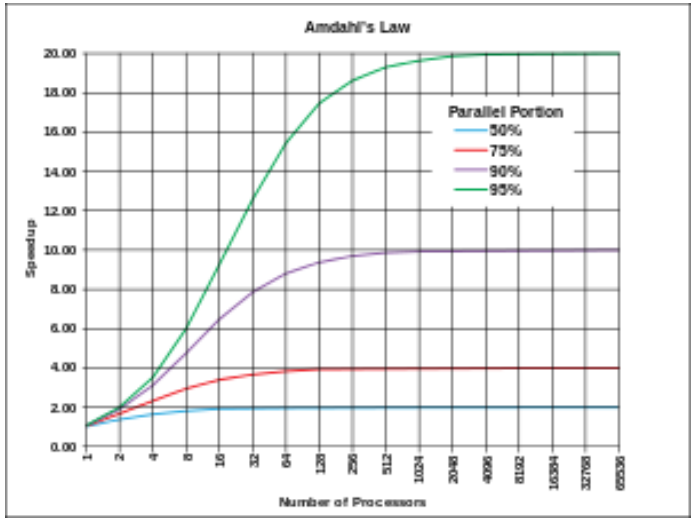


Discuss the Intersection of 3 Laws

Moore's Law and Kryder's Law



Amdahl's Law



Murphy's Law

Identify the Google White Papers

Distributed Storage Model

- **Google File System**
- Stores data on massive clusters of cheap machines
- Tolerates hardware failure
- Paper published in 2003.

Distributed Compute Model

- **MapReduce**
- Sends compute to data on GFS, not vice versa.
- Vastly simplifies distributed programming.
- Paper published in 2004.

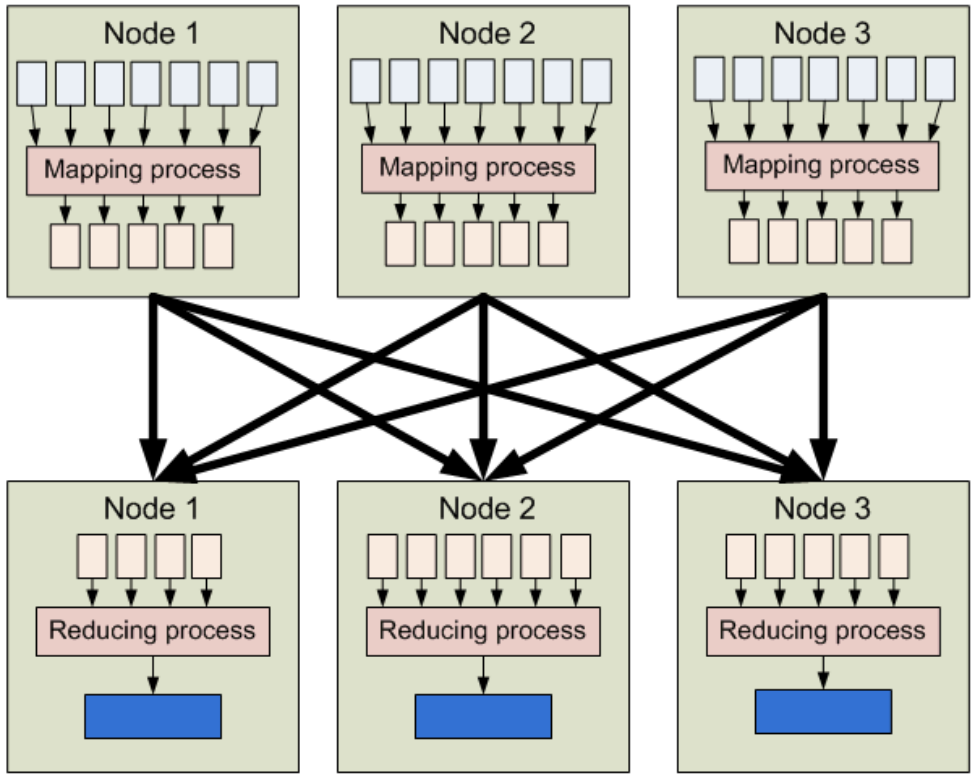
Runs on commodity hardware.
Costs scale linearly.

Describe the Hadoop Strategy

Distribute data
(share nothing)

Distribute computation
(parallelization without synchronization)

Tolerate failures
(no single point of failure)



http://developer.yahoo.com/hadoop/tutorial/module4.html

Hadoop Ecosystem

How is big data being used?

Big Data use cases

- **Verticals**
  - Advertising, media, and entertainment
  - Financial services
  - Government
  - Healthcare
  - Manufacturing
  - Oil and gas
  - Retail
  - telecommunications

- **Horizontals**
  - Data hub
  - Marketing optimization
  - Operational intelligence
  - Security and risk management

http://www.meltinfo.com/ppt/ibm-big-data

MapR Users & Applications

BOEING

- Intrusion detection & prevention
- Forensic analysis

McAfee SECURE

- Global threat analytics
- Virus analysis

verizon

- Customer revenue analytics
- ETL offload

Major Credit Card Issuer

- Recommendation engine
- Fraud detection and prevention

tsmc

- Manufacturing analysis
- Failure prediction

hp

- Clickstream analysis
- Quality profiling/field failure analysis

Leading Retailer

- Customer behavior analysis
- Brand monitoring

SILVERSPRING NETWORKS

- Smart meter analytics
- Power usage analysis

rubicon PROJECT

- Advertising exchange analysis and optimization

CISCO

- Customer sentiment
- Network analytics

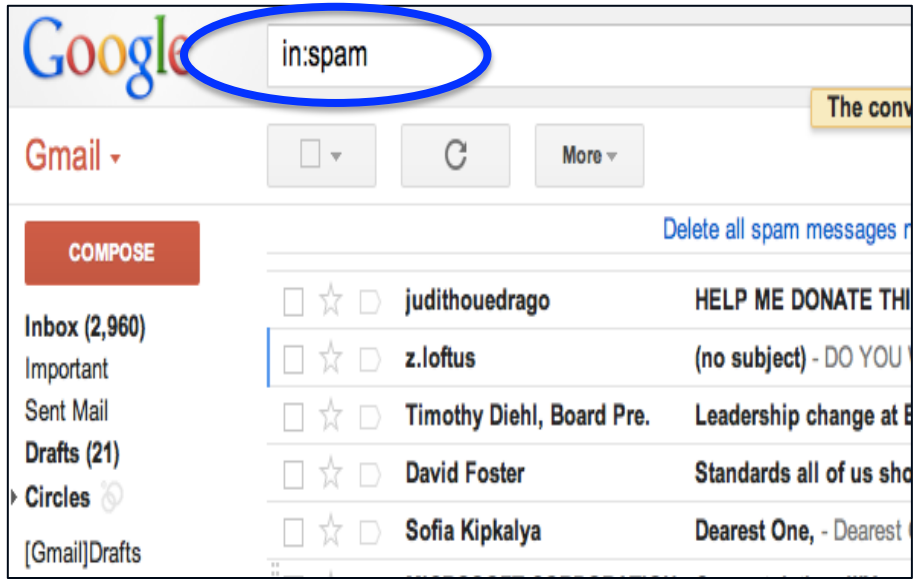
TRIGGIT

- Customer targeting
- Social media analysis

ancestry.com

- DNA-based relationship discovery
- Recommendation engine

classification



clustering

recommendation