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?

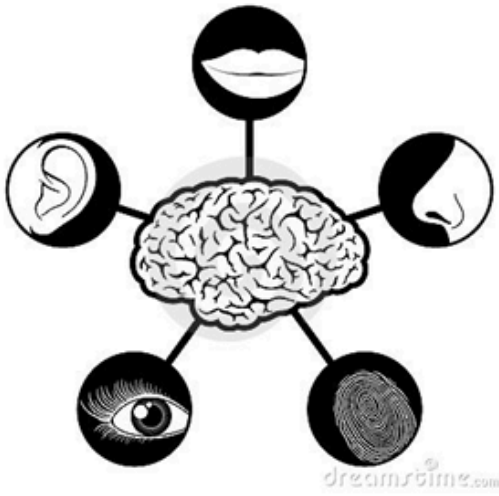
Data (quantitative vs qualitative)

Quantitative Data

are made with instruments such as rulers, balances, graduated cylinders, beakers, and thermometers.
These results are **measurable.**
(numbers)

Qualitative Data

use your senses to observe the results.



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

Example 1:

Oil Painting

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

Qualitative data:

- robust aroma
- frothy appearance
- strong taste
- burgundy cup

Example 2:

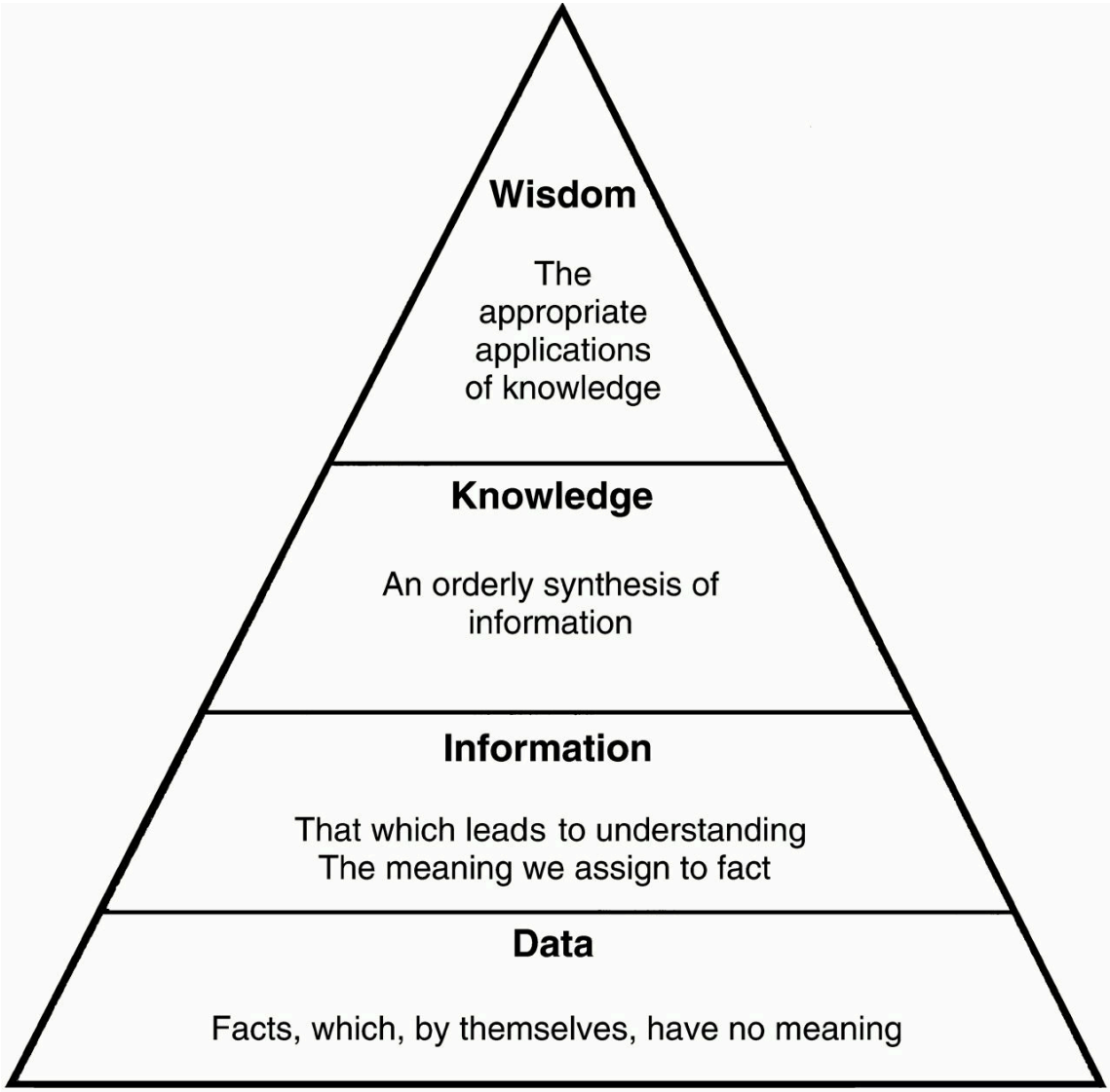
Latte

Quantitative data:

- 12 ounces of latte
- serving temperature 150° F.
- serving cup 7 inches in height
- cost $4.95

Data, information, knowledge, and wisdom

- **Data**
  - raw value
  - (e.g. here's his income)
- **Information**
  - set of data with meaning
  - (e.g. here's his credit application)
- **Knowledge**
  - interpretation of information in a context
  - (e.g. he is a risky credit applicant)
- **Wisdom**
  - Application of knowledge
  - (e.g. don't approve this loan to him)

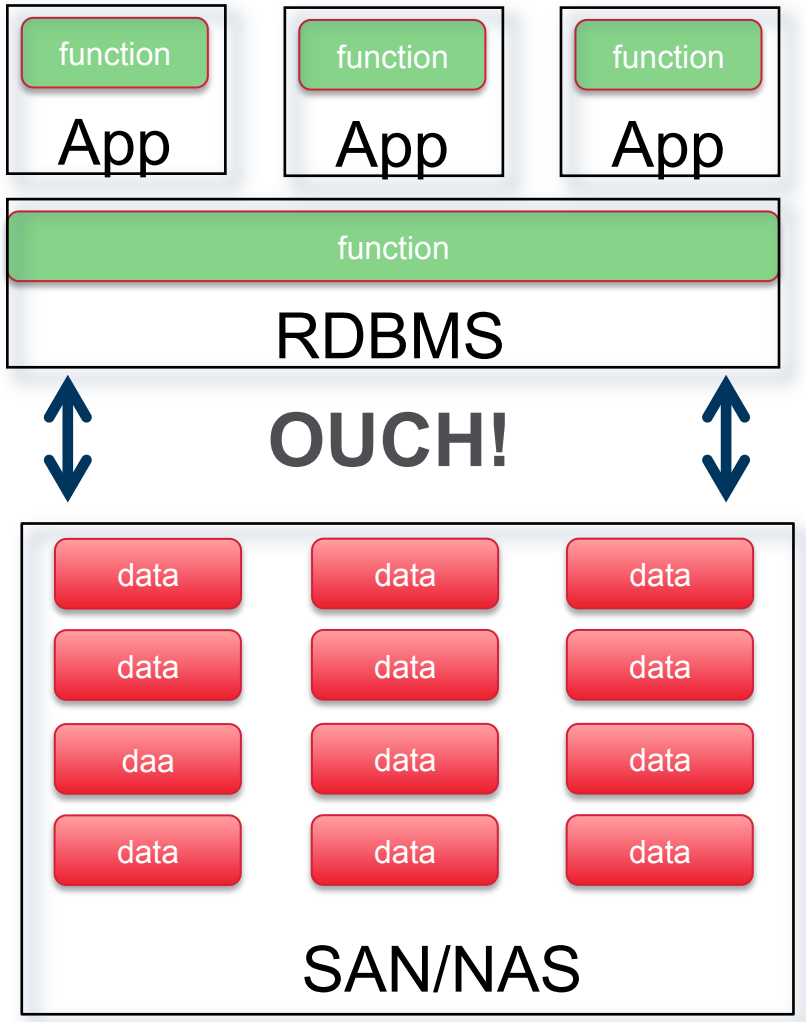


Data sources

- **Finance**:
  - CRM, ERP, billing, …
- **Social media**:
  - Twitter, LinkedIn, Facebook, …
- **Security**:
  - packet captures, audit logs, …
- **Biotech**:
  - sequence alignment, gene prediction, phylogenetics
- **IoT**
  - Sensors, wearables, logs, …

How is data stored and analyzed?

Data storage



ORACLE

Microsoft SQL Server

PostgreSQL

MySQL

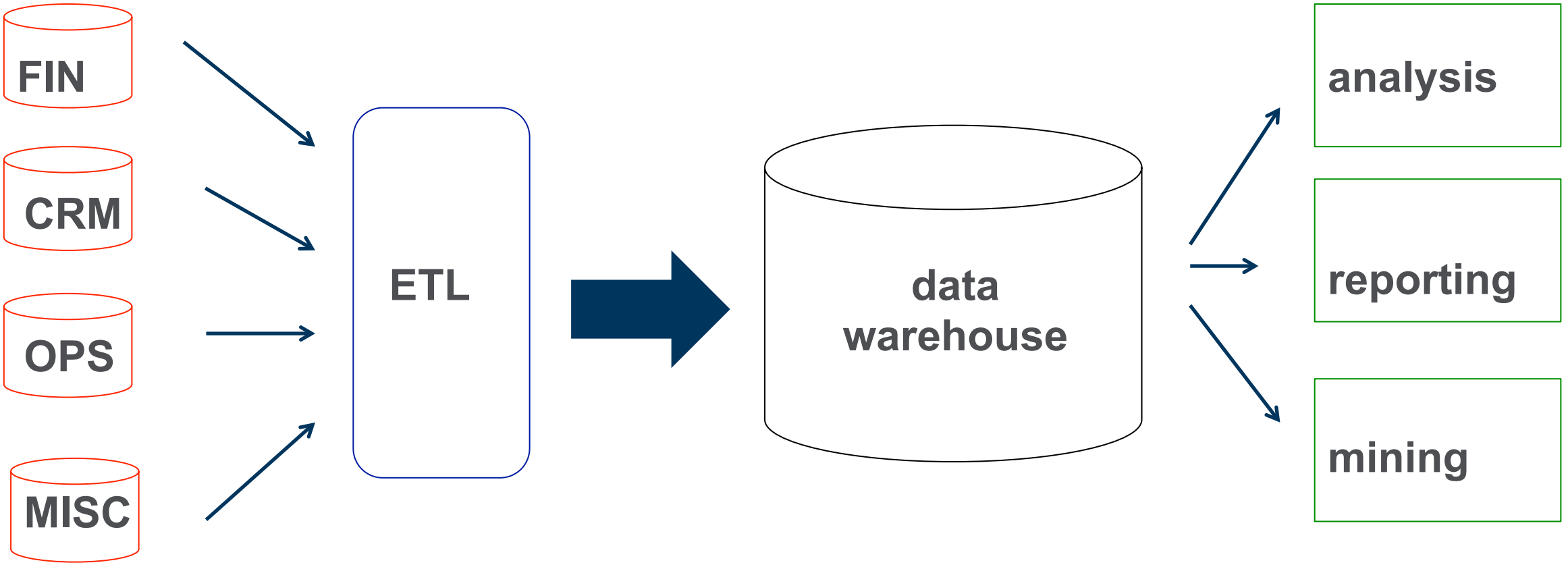
SYBASE

IBM DB2

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	10^3–10^{12}	$>10^{12}$
Complexity	Easy	Moderate	Difficult
Management Cost (per GB)	High	Moderate	Low

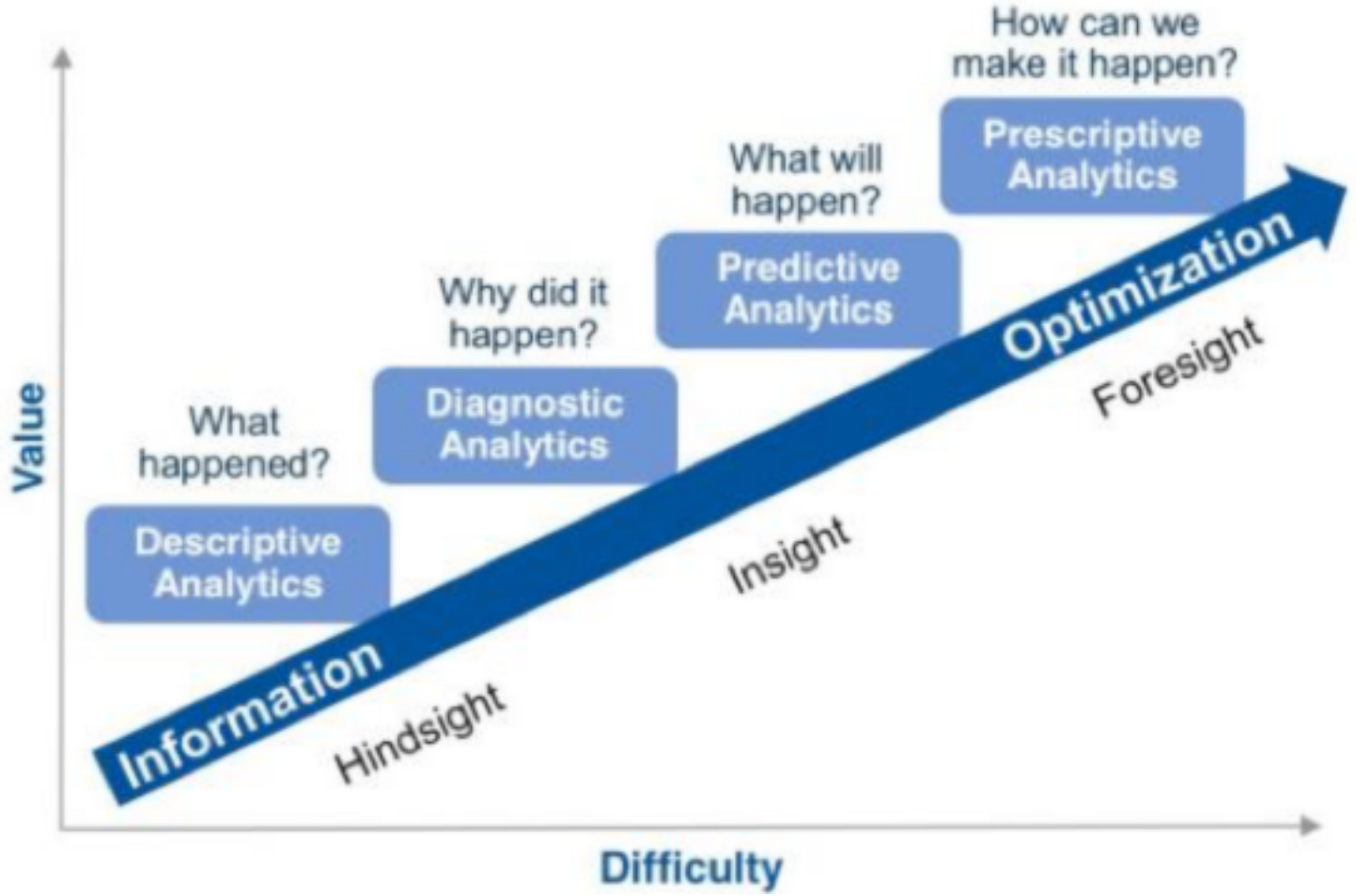
Traditional data warehouse



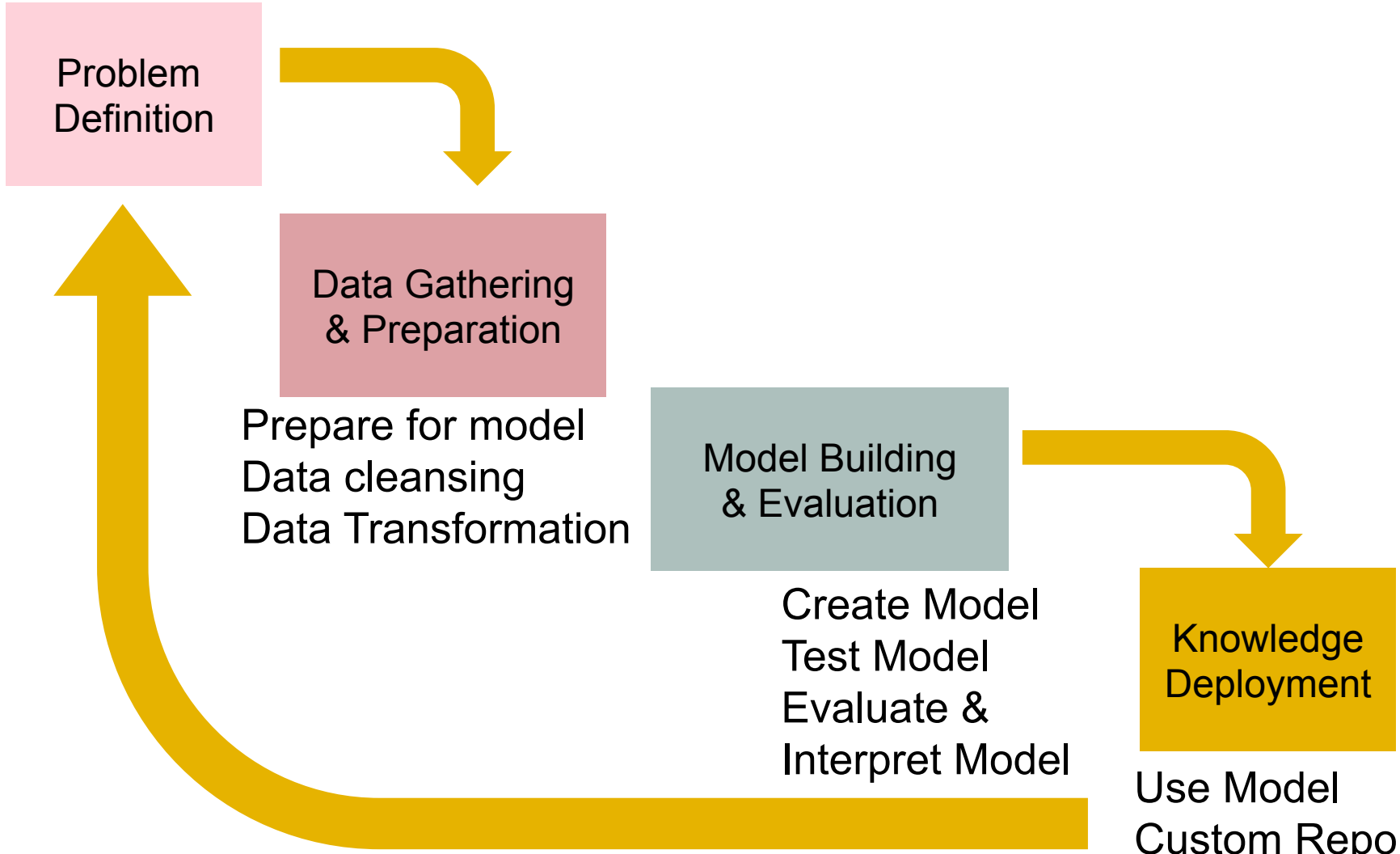
Problems with traditional warehouse

- Data must fit into a relational schema
- Data warehouse / data mart table schemas are fixed
- Updates to schemas are painful to users and developers
- Data warehouse platforms are expensive and don't scale well
- Data is traditionally silo'ed

Value in Analytics



Traditional Data Mining



What is big data?

Variety of Data Sources and Formats

Photos/Video

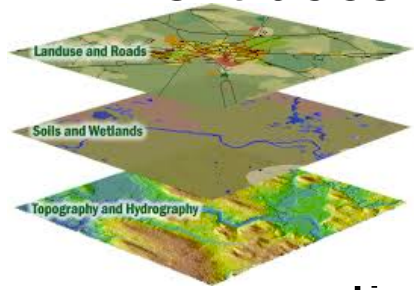
Customer data

Accounting & finance

Text docs

Social media

Log files



geospatial

Credit Card Transactions

Web user behaviour

Sensors

Big Data Definition

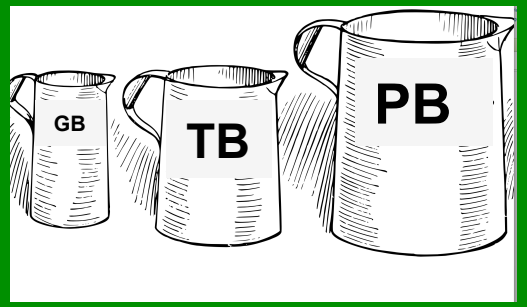
No *single* definition, but here's one to pick apart:

"***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.

Classic 3 V's of 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

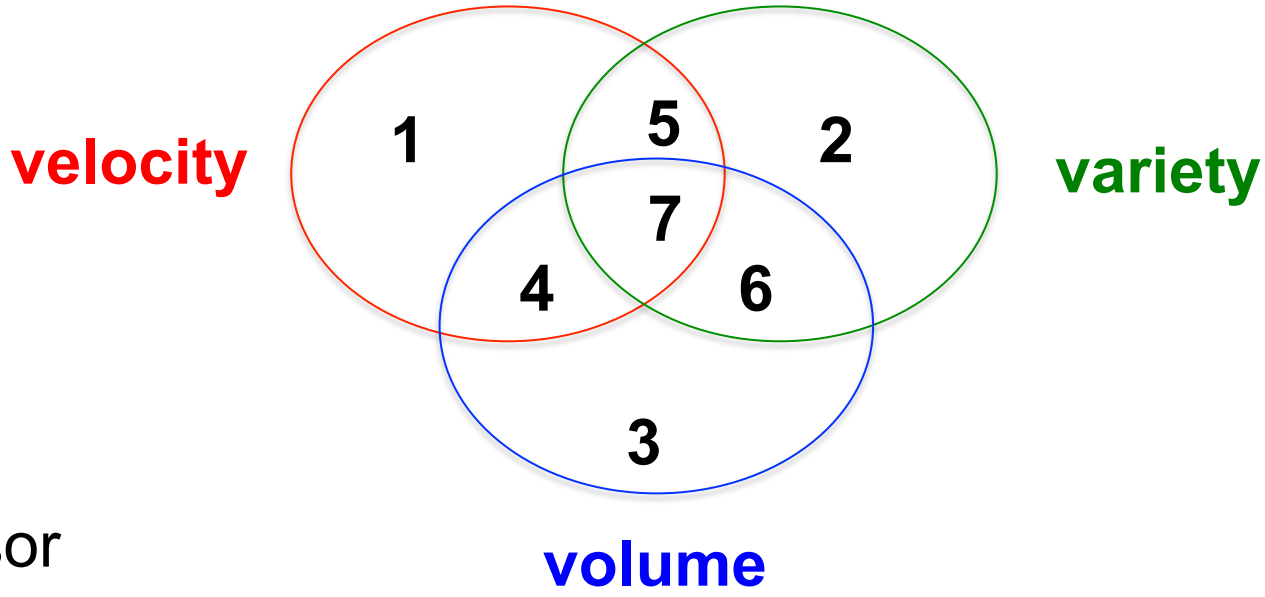
Velocity

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: ultimate big data

Now there are more V's in big data

- **Classic 3 V's:**
  - volume
  - variety
  - velocity

- **More V's**
  - value: how important is the data to the sponsors of a big data project
  - veracity: how believable is the data given its 3 V's?

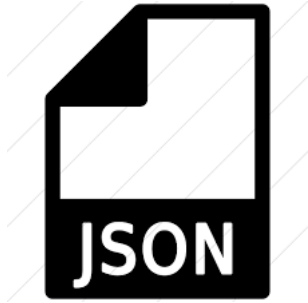
- **And even more V's**
  - variability: how often does the data change?

Describe Variety of Big Data

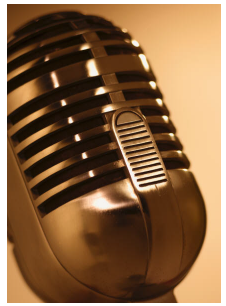
structured
well-defined data model

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
partially-defined data model

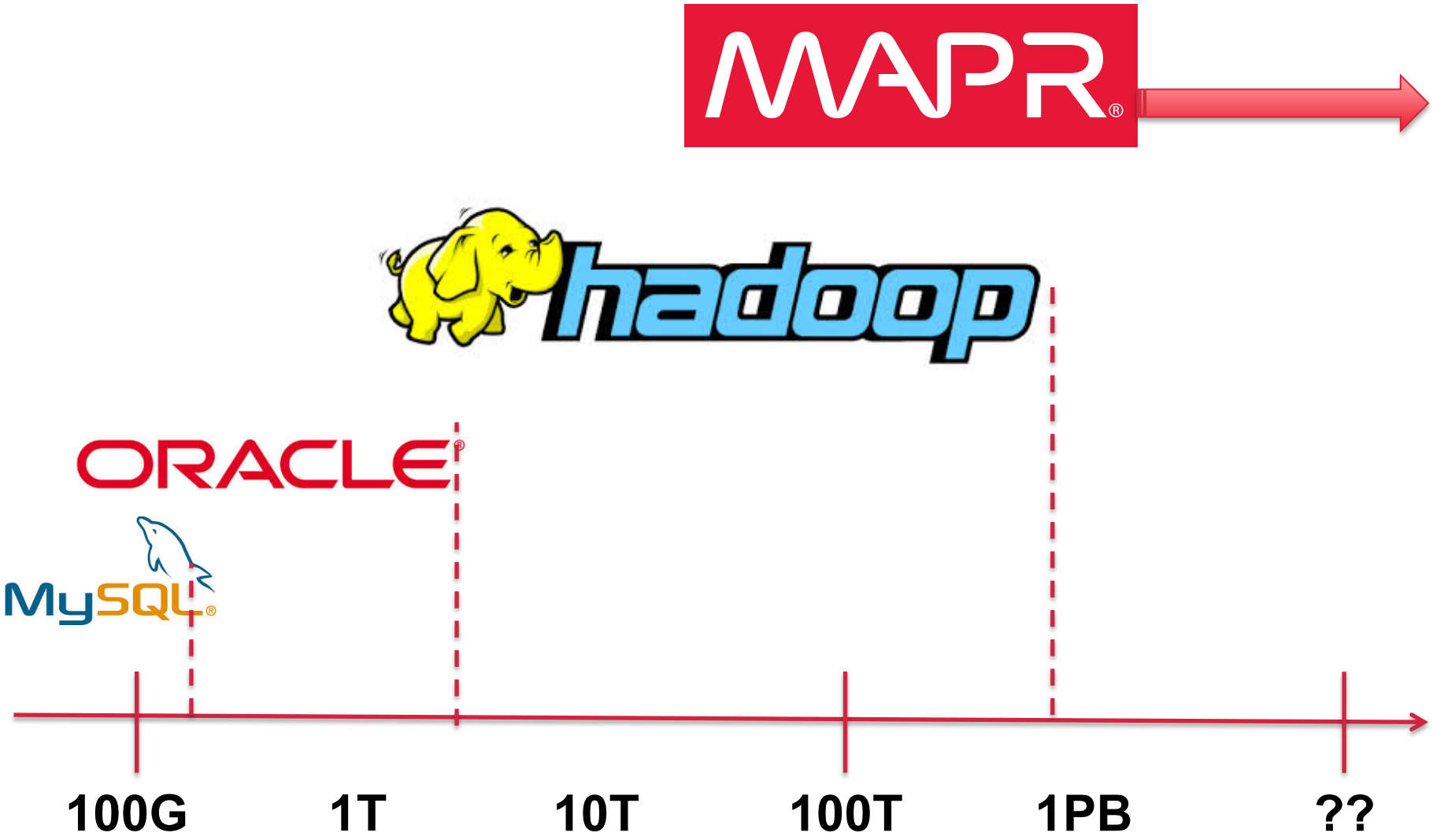


unstructured
undefined data model

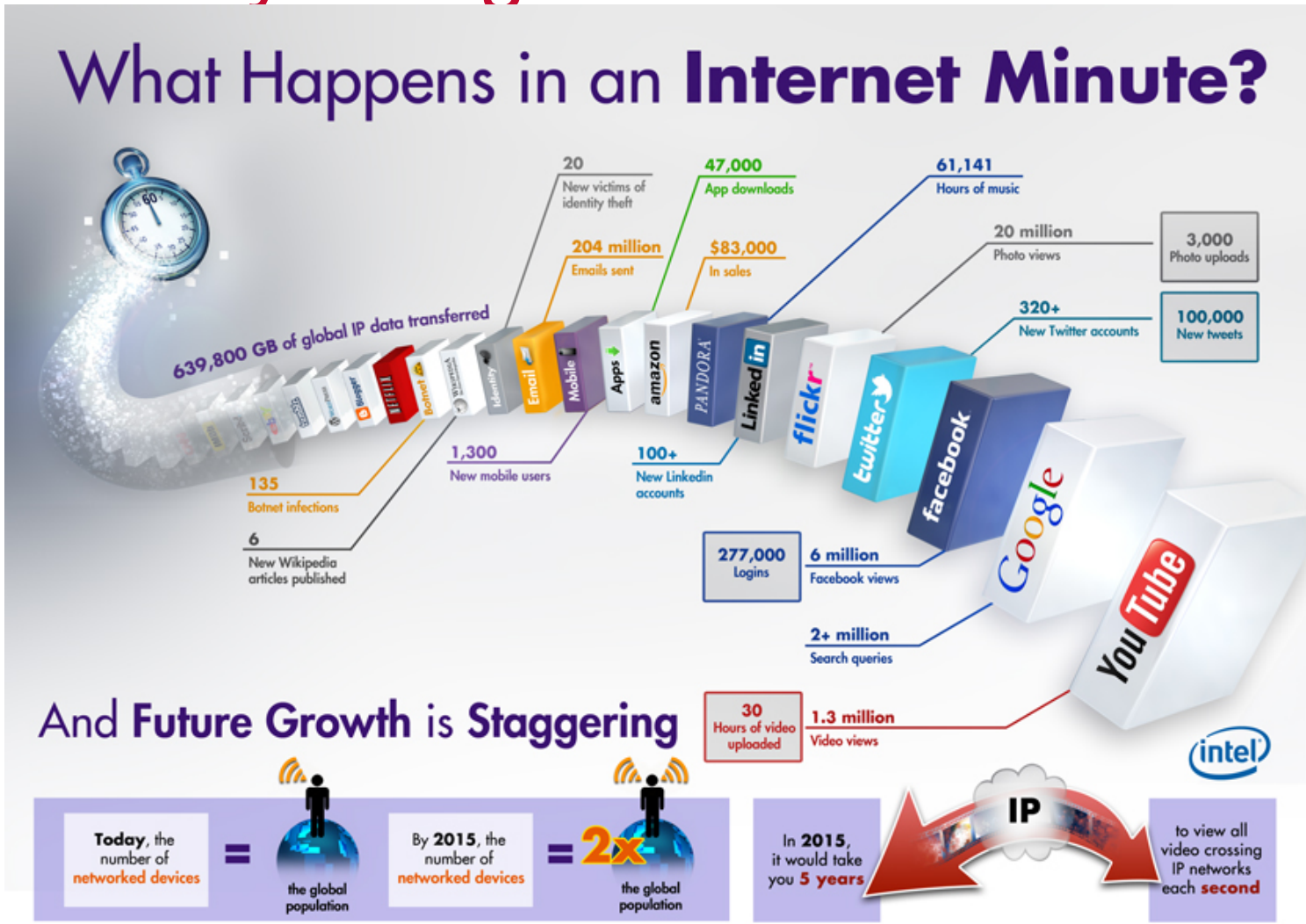


"data model": formalization of objects and relationships

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?

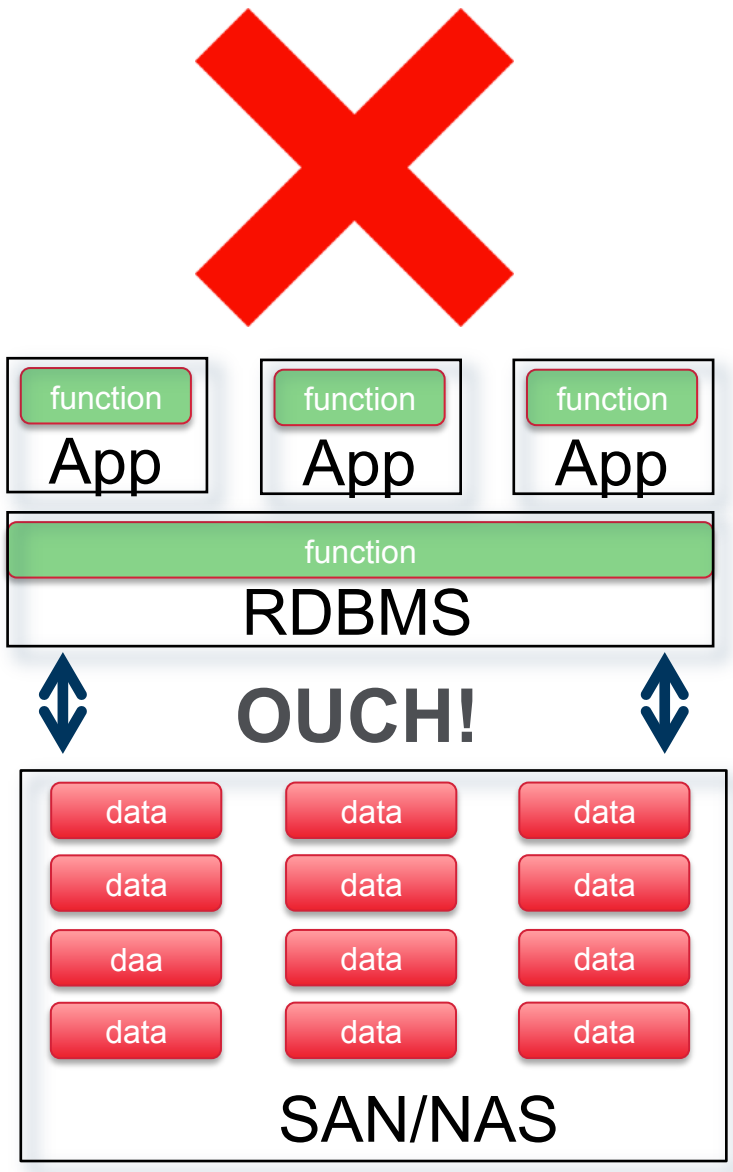
"Because RDBMSs can be beaten by more than an order of magnitude on the standard OLTP benchmark, then there is no market where they are competitive. As such, they should be considered as legacy technology more than a quarter of a century in age, for which a complete redesign and re-architecting is the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

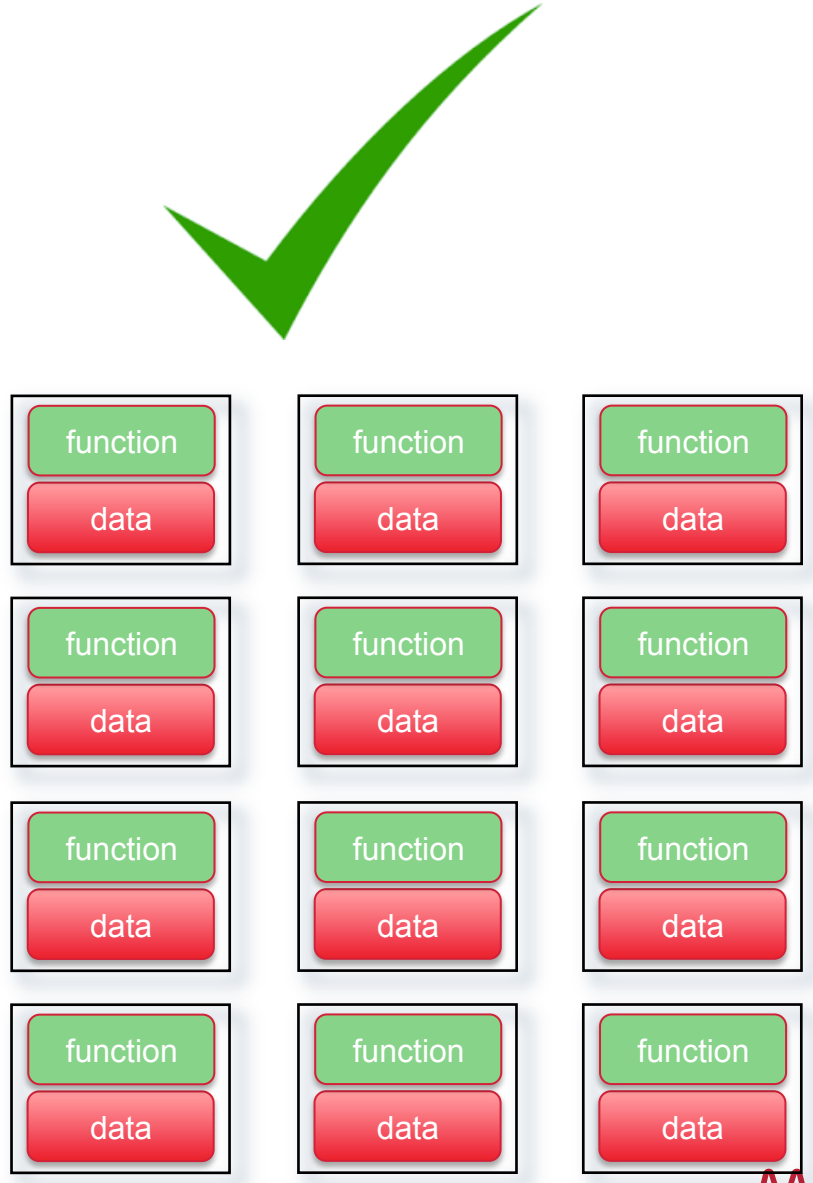
Michael Stonebraker (Creator of Ingres and Postgres)

Identify Ways to Scale to Process Big Data

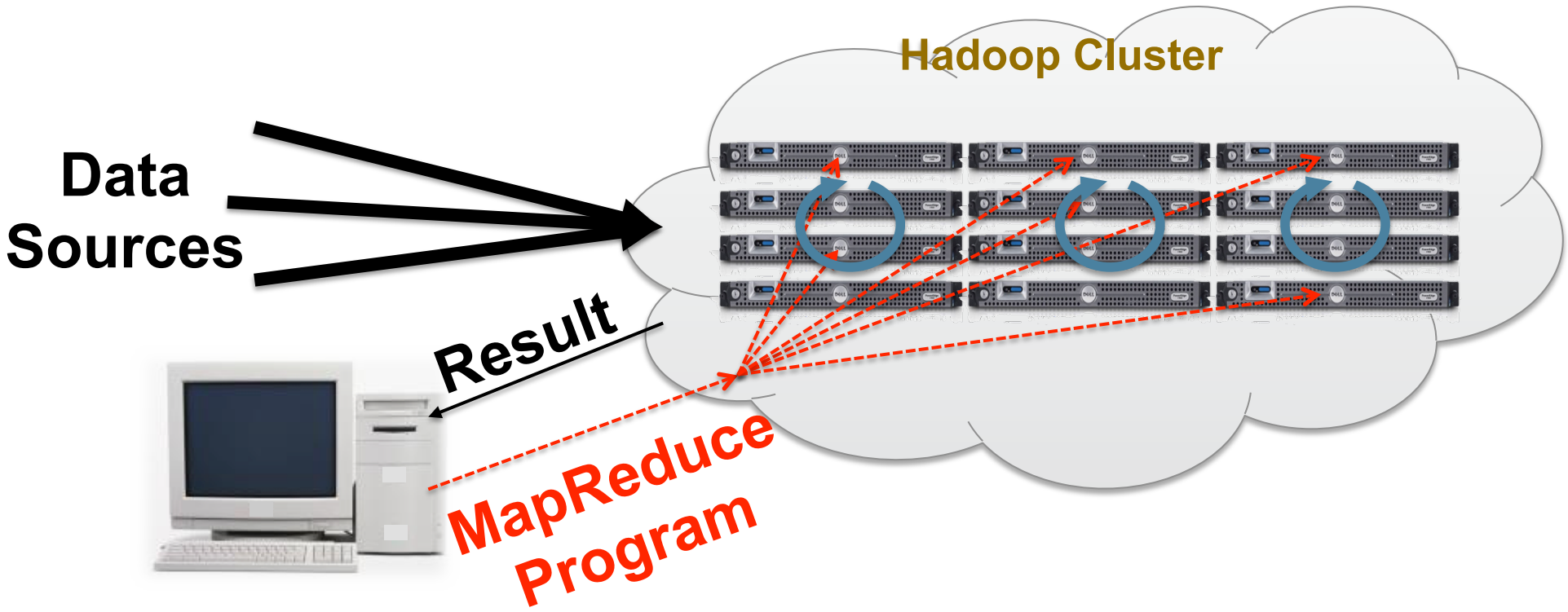
Paradigm	Example	limitations
Scale-up	Monolithic database	Infrastructure CAPEX/OPEX Availability/scalability Data gravity
Scale-out	Grid cluster	Synchronization overhead Programming complexity Specialized hardware
Sampling	Any approach	Lower accuracy Lower precision

Exploit Locality of Reference



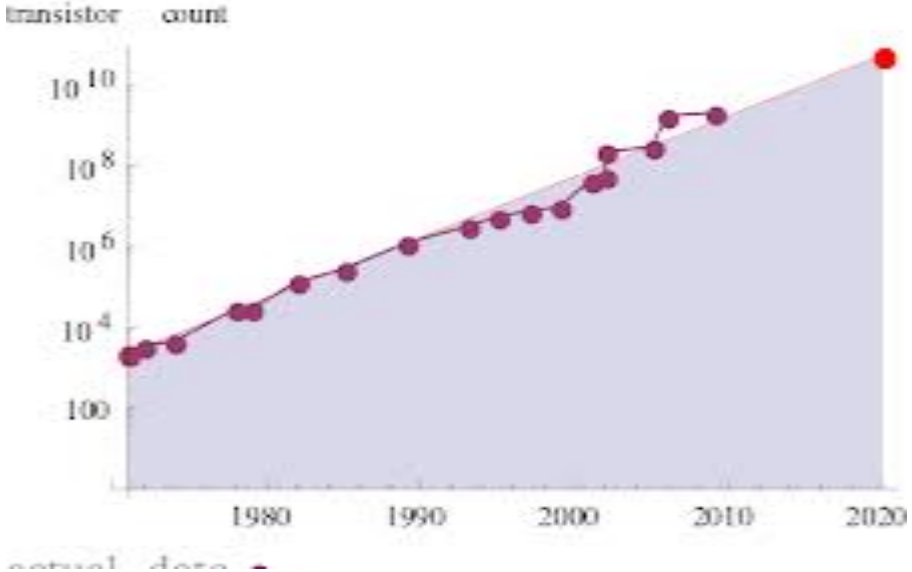


Distribute Data and Computation

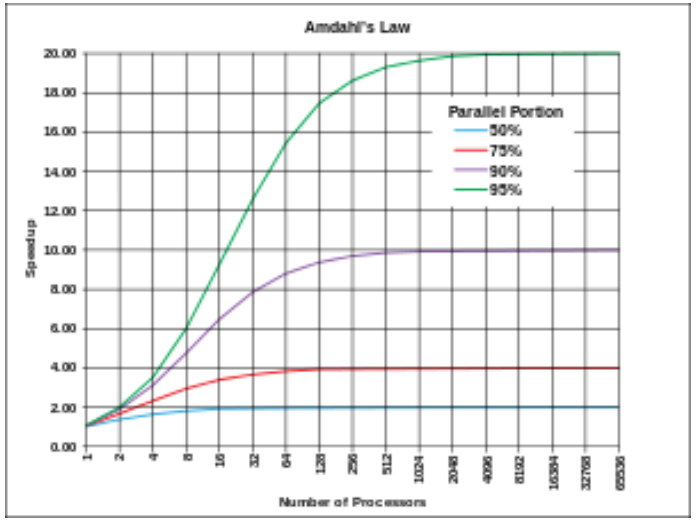


Discuss the Intersection of 3 Laws

Moore's Law and Kryder's Law



Amdahl's Law



Murphy's Law

Identify the Google White Papers

Distributed Storage Model

Google File System

- Distributes data across cluster of nodes, not central storage.
- Data distribution automaticlally managed
- Tolerates failures
- Paper published in 2003.

Distributed Compute Model

MapReduce

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

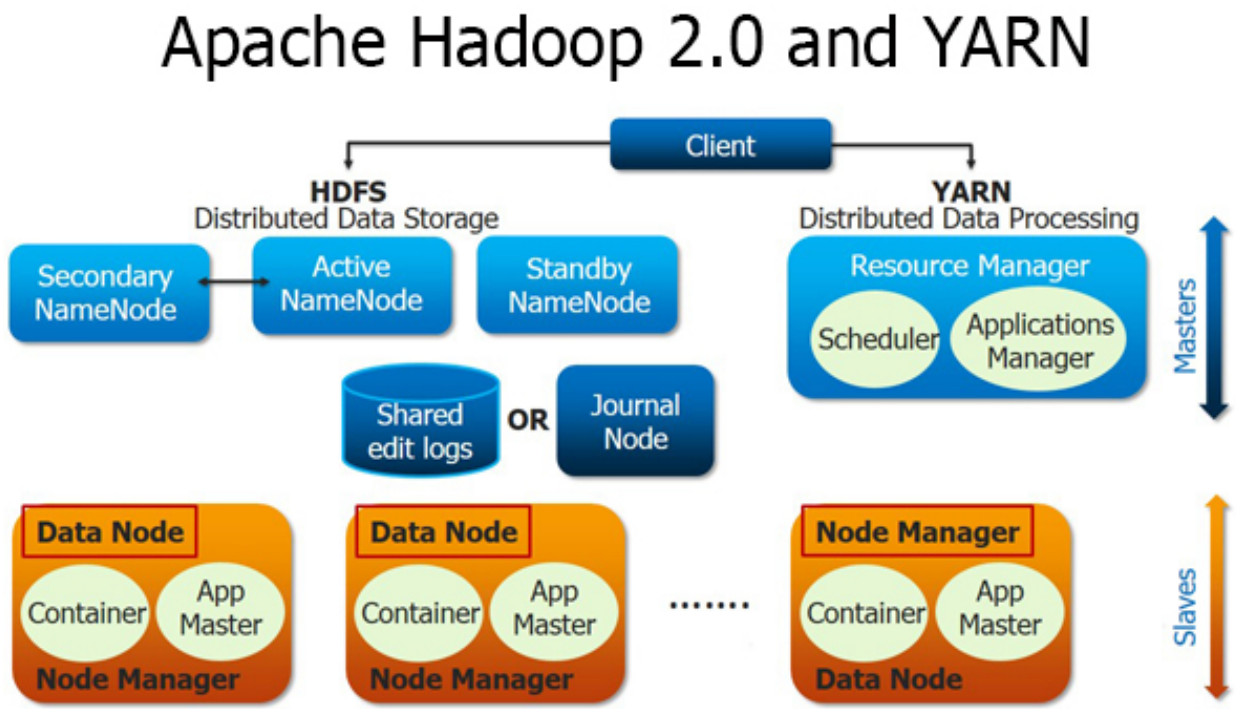
Runs on commodity hardware.
Costs scale linearly.

Describe the Hadoop Strategy

Distribute data
(share nothing)

Distribute computation
(parallelization without synchronization)

Tolerate failures
(no single point of failure)



Hadoop Ecosystem (partial view)