Big Data Management

Knowledge objectives

1. Give a definition of Big Data
2. Name eight features of cloud databases
3. Give a definition of Distributed Database
4. Recognize the problem of impedance mismatch
5. Name different kinds of NOSLQ databases
6. Recognize the main problems of NOSQL databases

Understanding Objectives

1. Estimate the cost of a distributed query
2. Transfrom the value in a schemaless database into a relational one

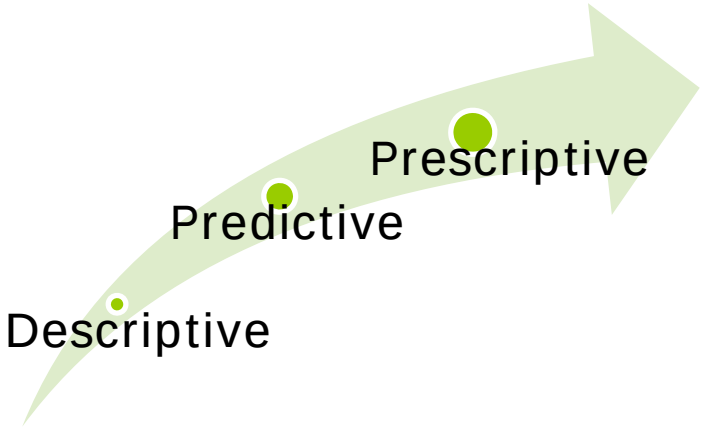
Motivation

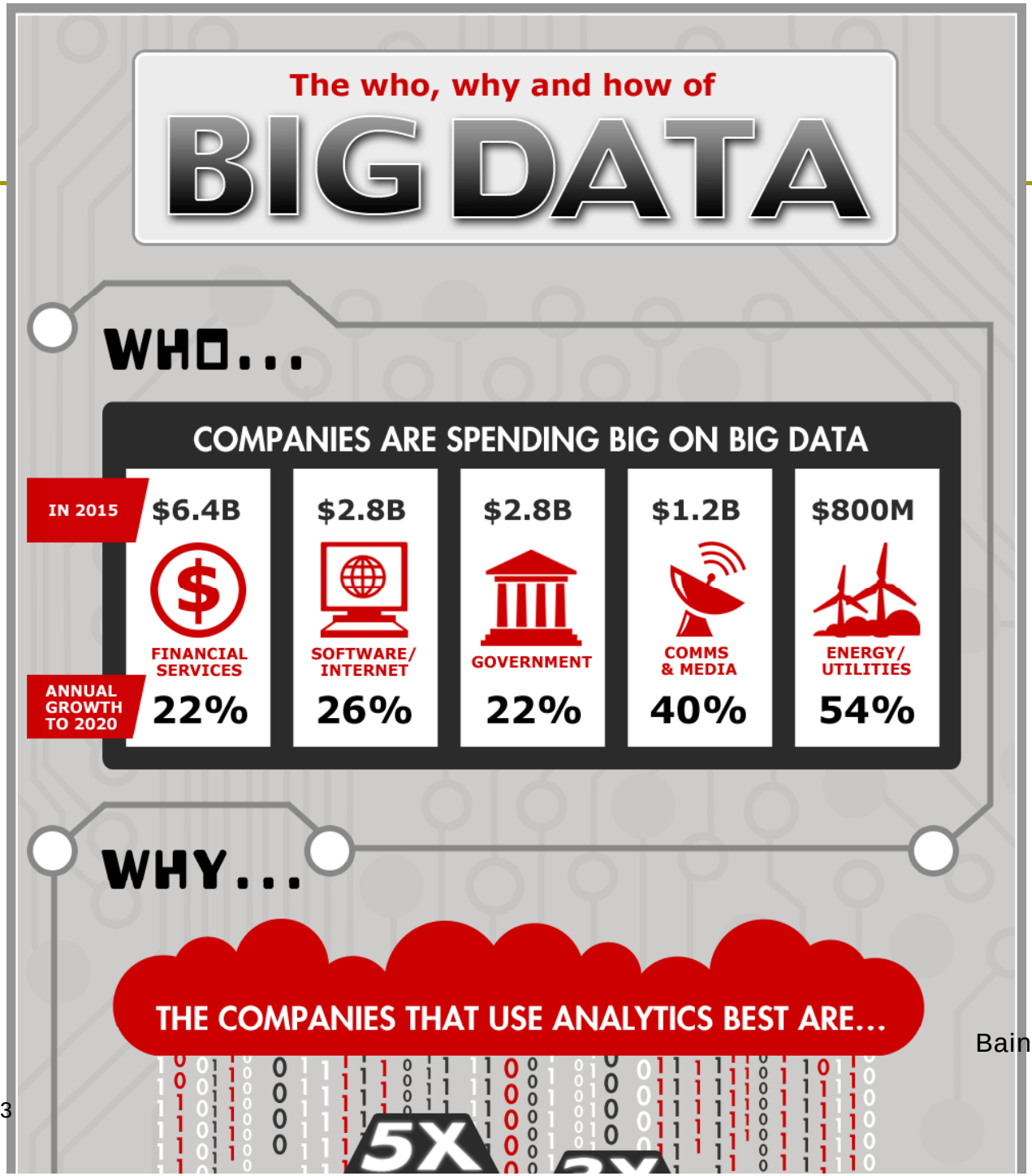
"*Without data you are just another person with an opinion*."

William Edwards Deming

"*It is a capital mistake to theorize before one has data*."

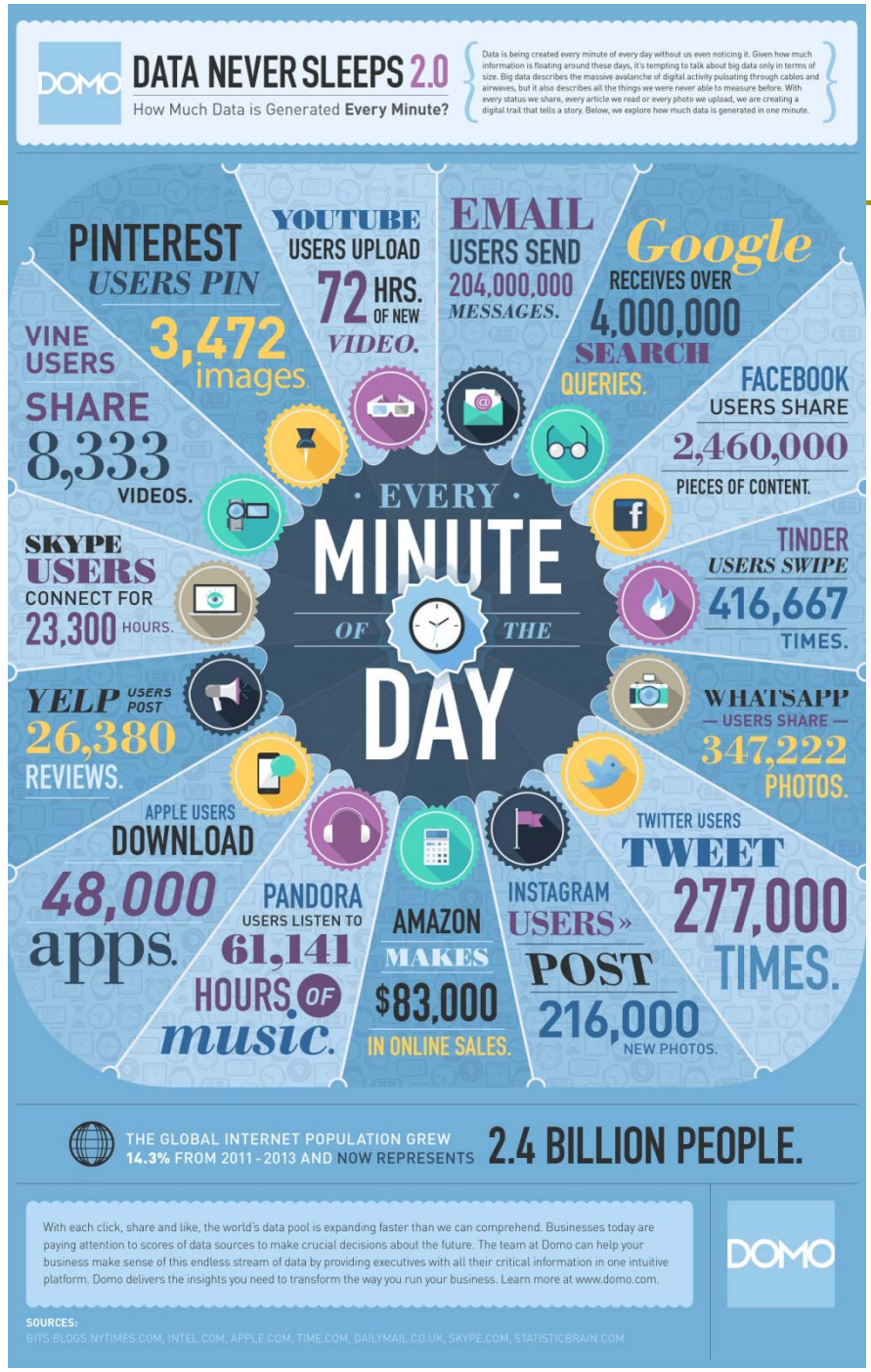
Sherlock Holmes (A Study in Scarlett)





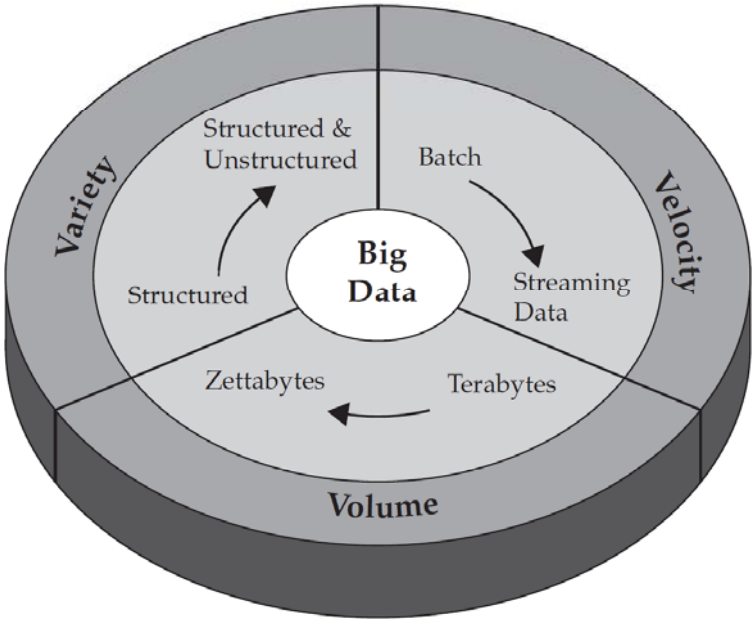
Bain & Company

It is estimated that in 2020 there will be more data than sand grains in the world (40 Zb)

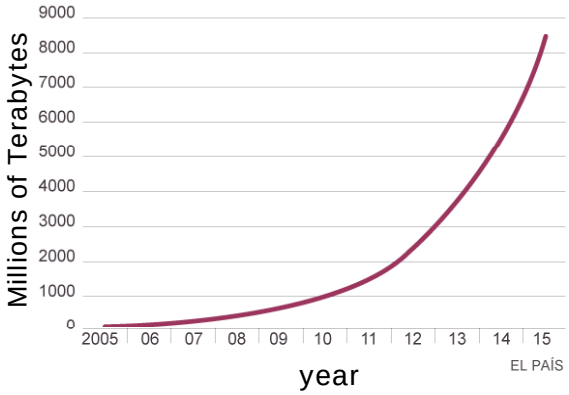


Big Data definition

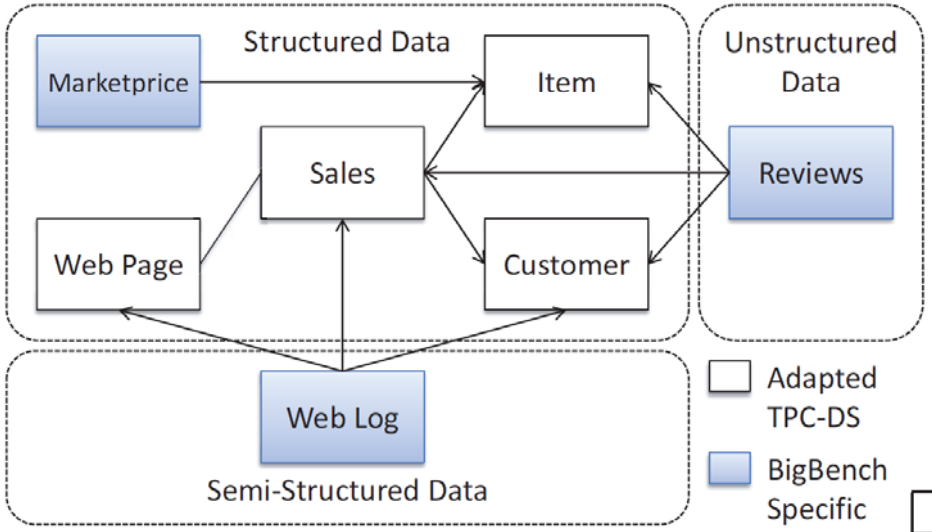
- Velocity
- Volume
- Variety
- ...
- Variability
- Validity/Veracity
- Value



From IBM "Understanding Big Data"



Bigbench



Query processing type	Total	Percentage(%)
Declarative	10	33.3
Procedural	7	23.3
Mix of Declarative and Procedural	13	43.3

Data sources	Total	Percentage(%)
Structured	18	60.0
Semi-structured	7	23.3
Un-structured	5	16.7

Analytic techniques	Total	Percentage(%)
Statistics analysis	6	20.0
Data mining	17	56.7
Reporting	8	26.7

Big Data sources

- Structured
  - Created (i.e., business data)
  - Provoked (e.g., customer feedback)
  - Transacted
  - Compiled (e.g., demographics)
  - Experimental (e.g., sampling customers)
- Unstructured
  - Captured (e.g., search words)
  - User-generated (e.g., social networks)

Types of Big Data Analyzed in Industry

	Manufacturing and Natural Resources	Media/ Communications	Services	Government	Education	Retail	Banking	Insurance	Healthcare	Transportation	Utilities
Transactions	73%	62%	67%	67%	54%	93%	83%	81%	75%	79%	80%
Log data	44%	57%	58%	59%	54%	40%	66%	61%	33%	71%	60%
Machine or sensor data	53%	38%	35%	33%	31%	27%	27%	48%	42%	50%	40%
Emails /documents	27%	43%	43%	41%	46%	27%	34%	39%	17%	29%	20%
Social media data	32%	52%	39%	26%	54%	73%	27%	13%	-	50%	-
Free-form text	17%	24%	28%	30%	31%	20%	34%	35%	67%	21%	40%
Geospatial data	27%	14%	19%	19%	38%	27%	27%	26%	8%	29%	40%
Images	19%	24%	17%	11%	38%	13%	5%	16%	25%	7%	-
Video	8%	29%	12%	7%	31%	13%	-	6%	8%	7%	-
Audio	10%	19%	8%	4%	8%	-	-	6%	-	-	-
Other	8%	14%	13%	15%	8%	7%	10%	16%	42%	14%	-
n =	*59*	*21**	*127*	*27**	*13**	*15**	*41*	*31*	*12**	*14**	*5**

Note: Highlighted cells indicate the top three data types by industry.
Multiple responses allowed

Source: Gartner (September 2013)

Big Data facets

- The Original
- as Technology
- as Data Distinctions
- as Signals
- as Opportunity
- as Metaphor
- as New Term for Old Stuff

Timo Elliott

Big Data related areas

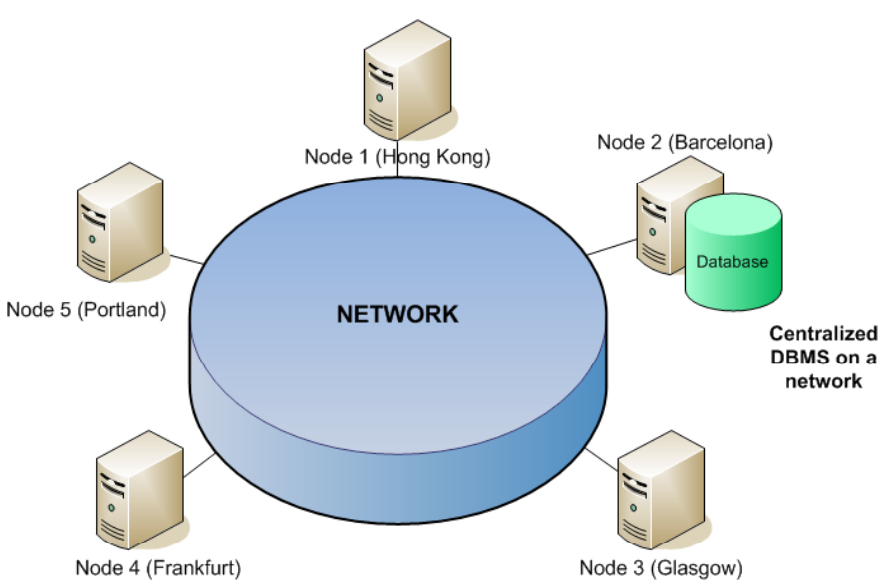
- Volume and Velocity
  - Declarative querying
  - Query optimization
- Variety and Variability
  - Data quality
  - Data integration
  - Web and text mining
  - Information retrieval
- Validity/Veracity
  - Data consistency
  - Uncertainty
  - Statistical reasoning
  - Data linkage (provenance)
- Value
  - Analytics
    - Data mining
    - Algorithmics
    - Automatic learning
    - Simulation
    - Privacy
  - Biologists
  - Linguistics
  - Chemists
  - Sociologists
  - Engineers

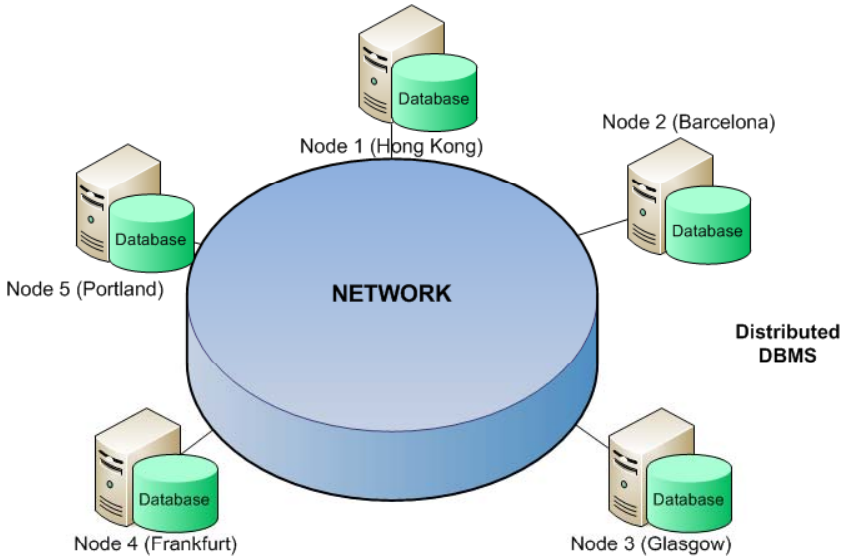
Key features of cloud databases

a) Quick/Cheap set up
b) Ability to horizontally scale
c) Ability to replicate & distribute (fragmentation)
d) Simple call level interface or protocol
   - No declarative query language
e) Weaker concurrency model than ACID
f) Efficient use of distributed indexes and RAM
g) Flexible schema
   - Ability to dynamically add new attributes
h) Multi-tenancy

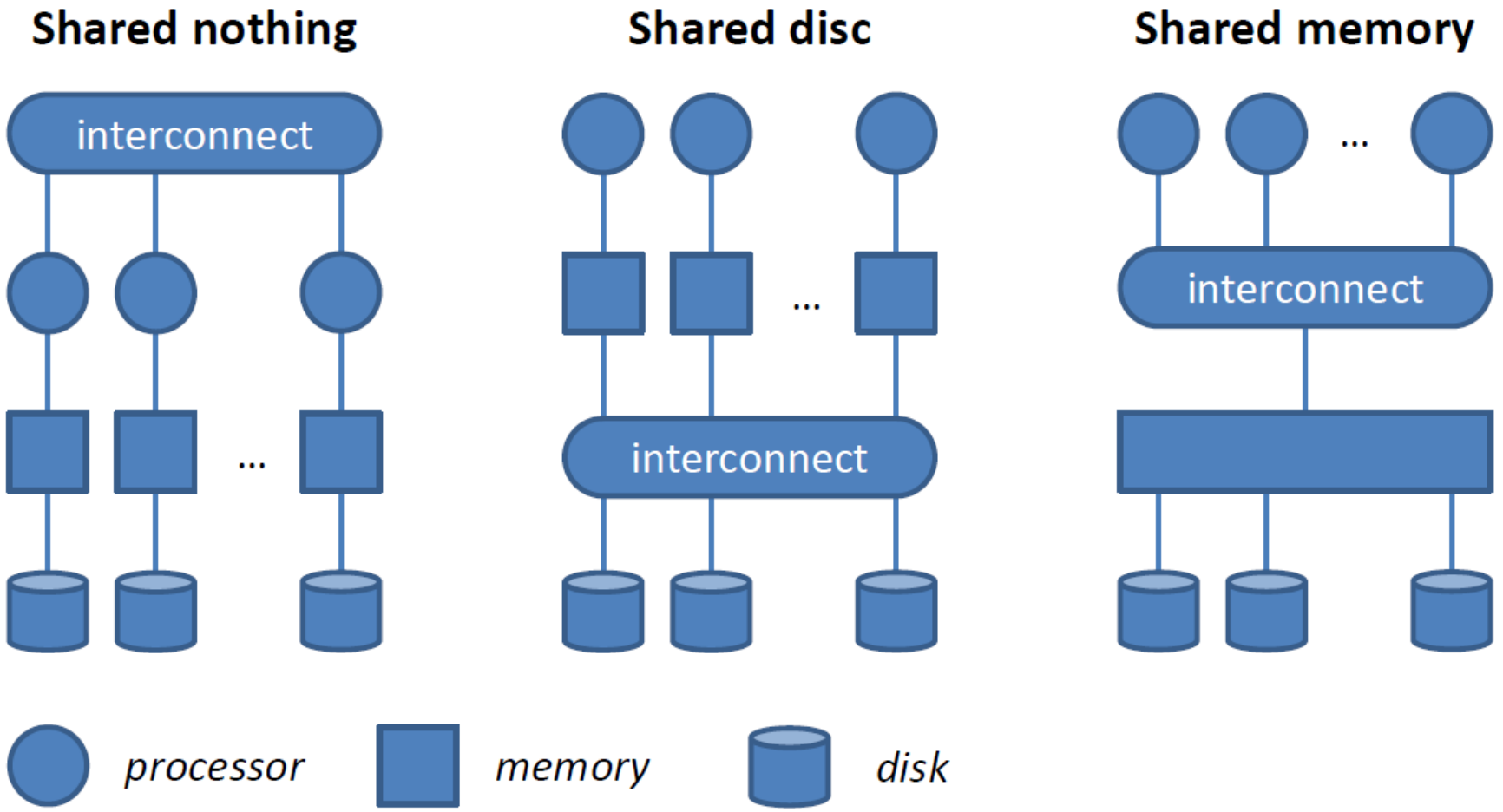
Distributed Database

- A distributed database (DDB) is a database where data management is distributed over several nodes in a network.
  - Each node is a database itself
    - Potential heterogeneity
  - Nodes communicate through the network





Parallel database architectures



D. DeWitt & J. Gray, “Parallel Database Systems: The future of High Performance Database Processing”, 1992
Figure from D. Abady

Activity

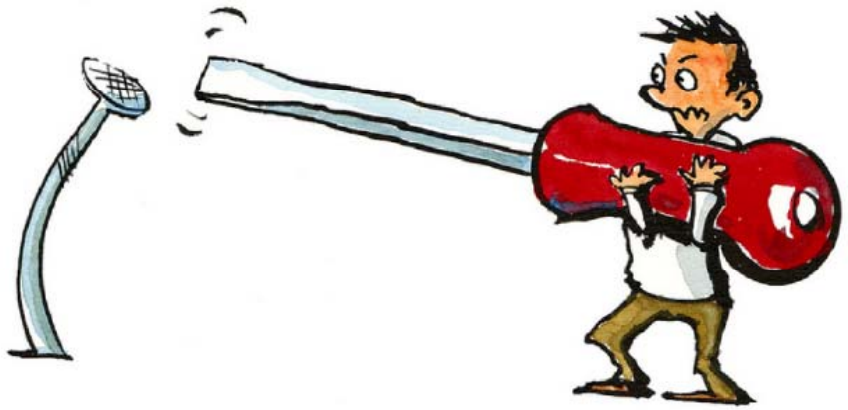
- *Objective: Recognize the benefits of distributing data*
- *Tasks:*
  1. *(5') Individually solve one exercise*
  2. *(10') Explain the solution to the others*
  3. *Hand in the three solutions*

- *Roles for the team-mates during task 2:*
  a) *Explains his/her material*
  b) *Asks for clarification of blur concepts*
  c) *Mediates and **controls time***

Impedance Mismatch

Of hammers and nails...

The Law of the Hammer

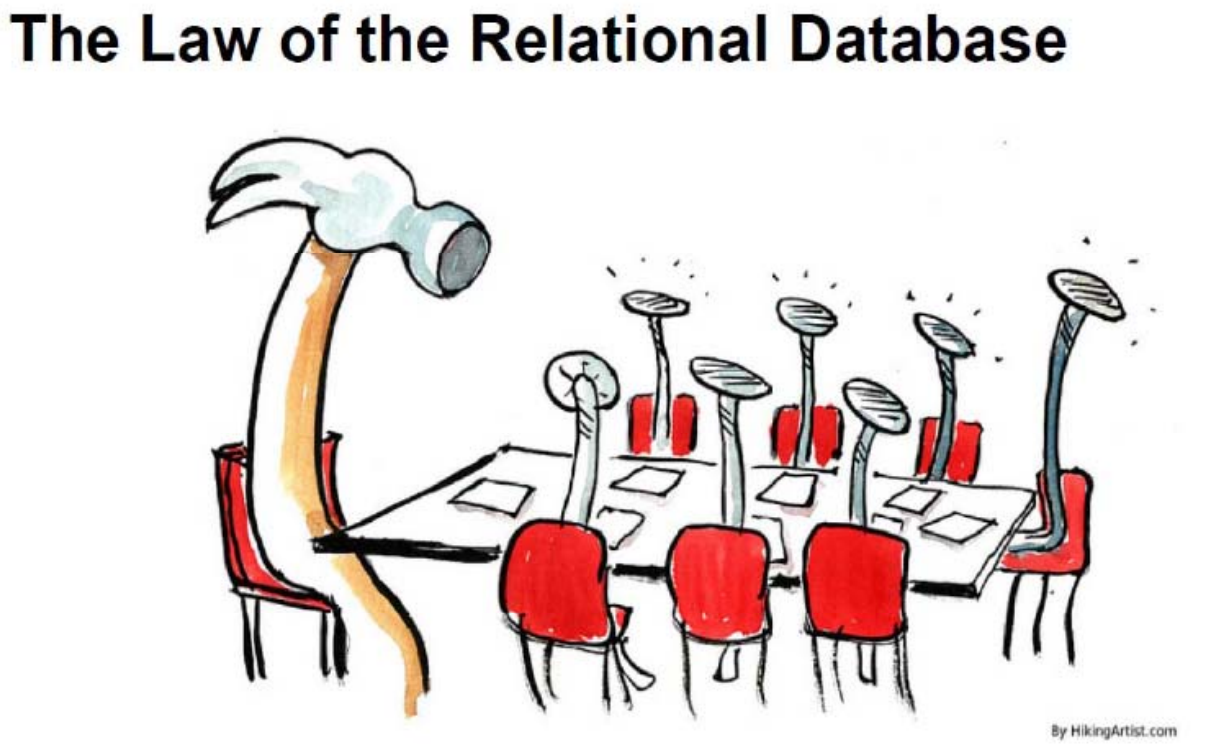


If the only tool you have is a hammer,
everything looks like a nail.

Abraham Maslow - The Psychology of Science - 1966

Petra Selmer, Advances in Data Management 2012

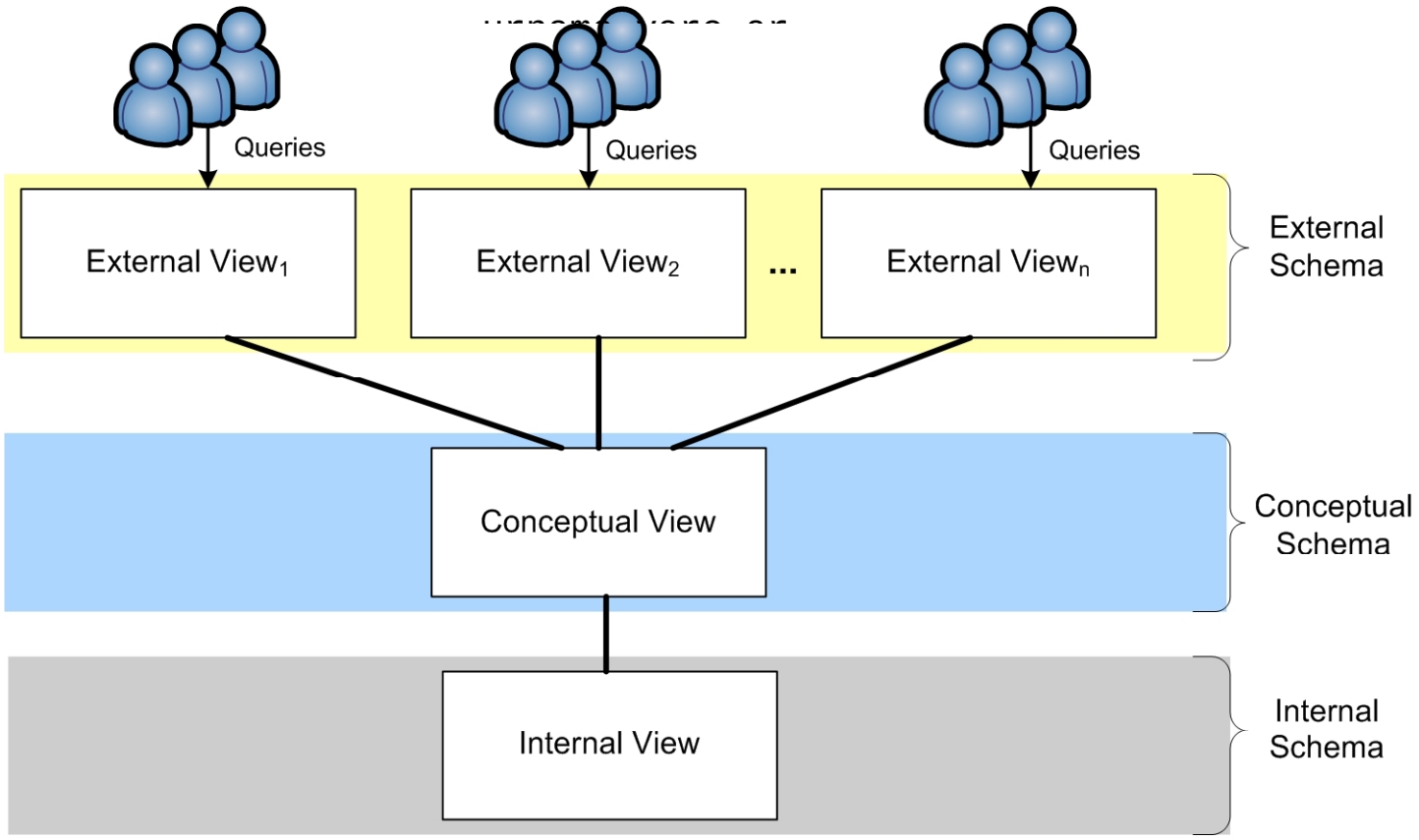
Impedance Mismatch



A Walk in Graph Databases - 2012

Petra Selmer, Advances in Data Management 2012

Schemaless Databases



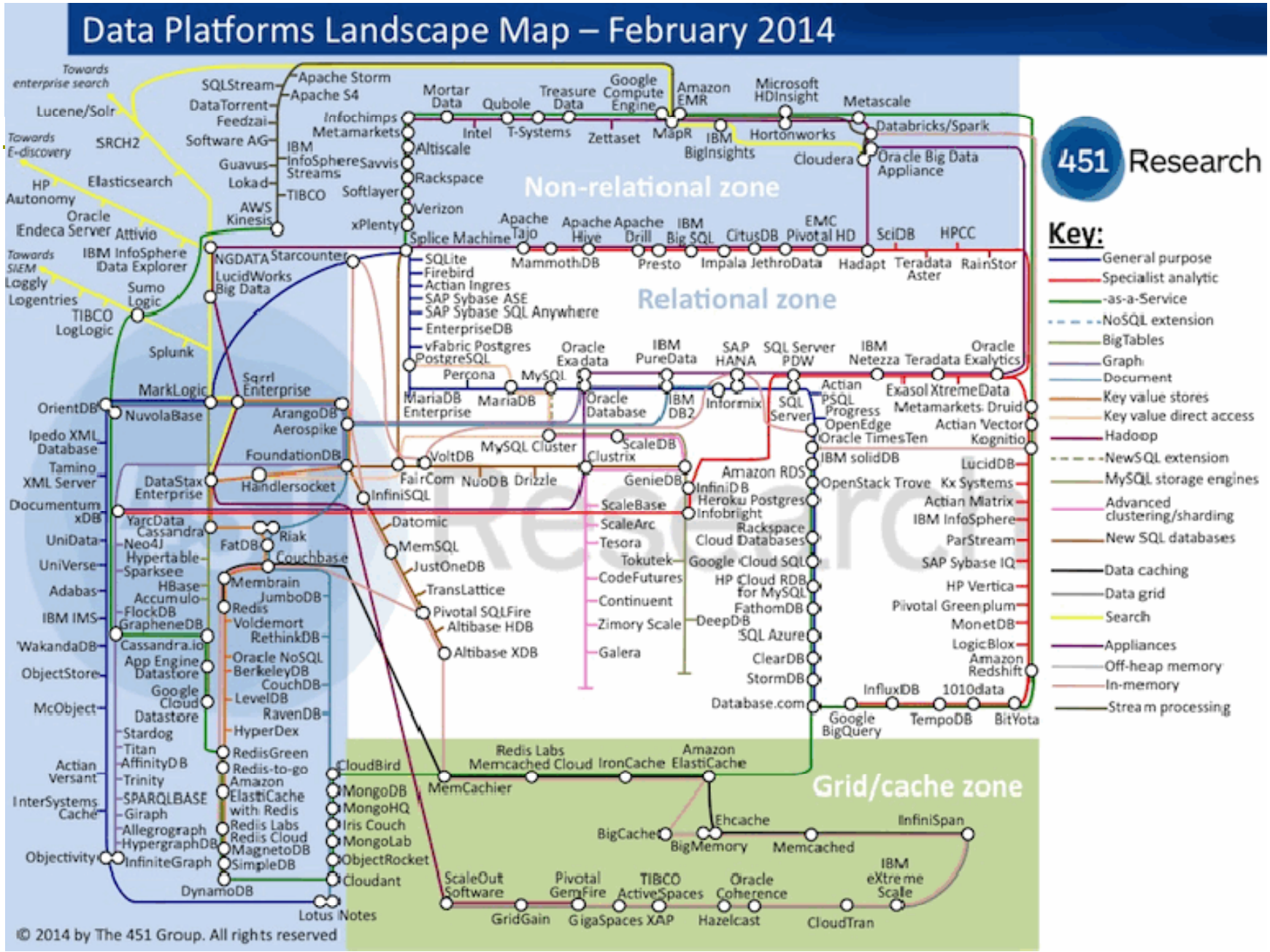
Different applications

Not Only SQL (different problems entail different solutions)

- OLTP
  - Object-Relational
    - Distributed databases
    - Parallel databases
- Scientific databases and other Big Data repositories
  - Key-value stores
- Data Warehousing & OLAP
  - MOLAP
  - Column stores
  - Multidimensional features
- Text / documents
  - Document databases
    - XML/JSON databases
- Stream processing
  - Stream processor
- Semantic Web and Open Data
  - Graph databases

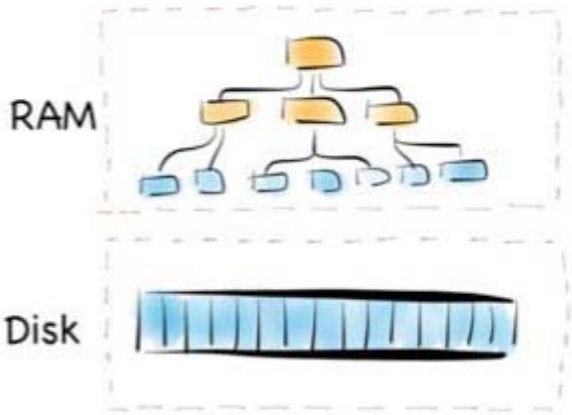
Schemaless Databases

- NOSQL solution for the impedance mismatch
- Several new data models were introduced
  - Graph data model
  - Document-oriented databases
  - Key-value (~ hash tables)
  - Streams (~ vectors and matrixes)
- These *new* models lack of an explicit schema (defined by the user)
  - However, an implicit schema remains

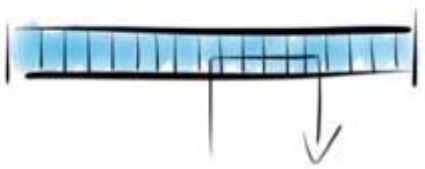

Data Platforms Landscape Map – February 2014
451 Research
Non-relational zone
Relational zone
Grid/cache zone
Key:
General purpose
Specialist analytic
-as-a-Service
NoSQL extension
BigTables
Graph
Document
Key value stores
Key value direct access
Hadoop
NewSQL extension
MySQL storage engines
Advanced clustering/sharding
New SQL databases
Data caching
Data grid
Search
Appliances
Off-heap memory
In-memory
Stream processing
© 2014 by The 451 Group. All rights reserved

Internal Structures

Riak, Mongo etc

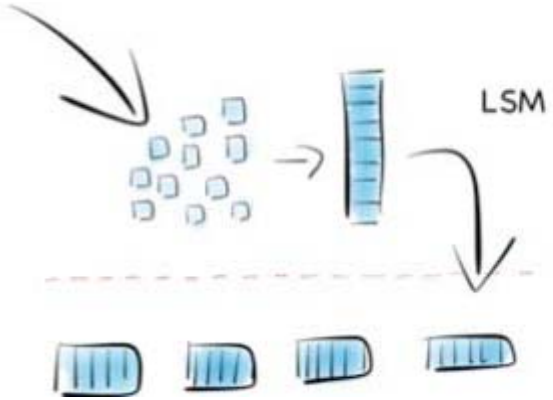


Kafka

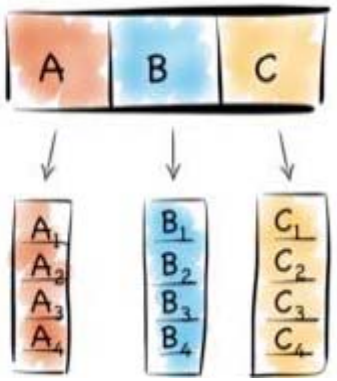


(Queues are Databases - 1995 Jim Gray)

Hbase, Cassandra, RocksDB etc



Redshift etc, Parquet (Hadoop)



Ben Stopford
Progscon & JAX Finance 2015

Polyglot Systems

- Federate different kinds of storage systems

Martin Fowler

http://martinfowler.com/bliki/PolyglotPersistence.html

NOSQL drawbacks

- No ACID
- No standard
- Low-level query

Michael Stonebraker

The Problem is Not SQL

- Relational systems are too generic…
  - OLTP: stored procedures and simple queries
  - OLAP: ad-hoc complex queries
  - Documents: large objects
  - Streams: time windows with volatile data
  - Scientific: uncertainty and heterogeneity
- … But the overhead of RDBMS has nothing to do with SQL
  - Low-level, record-at-a-time interface is not the solution

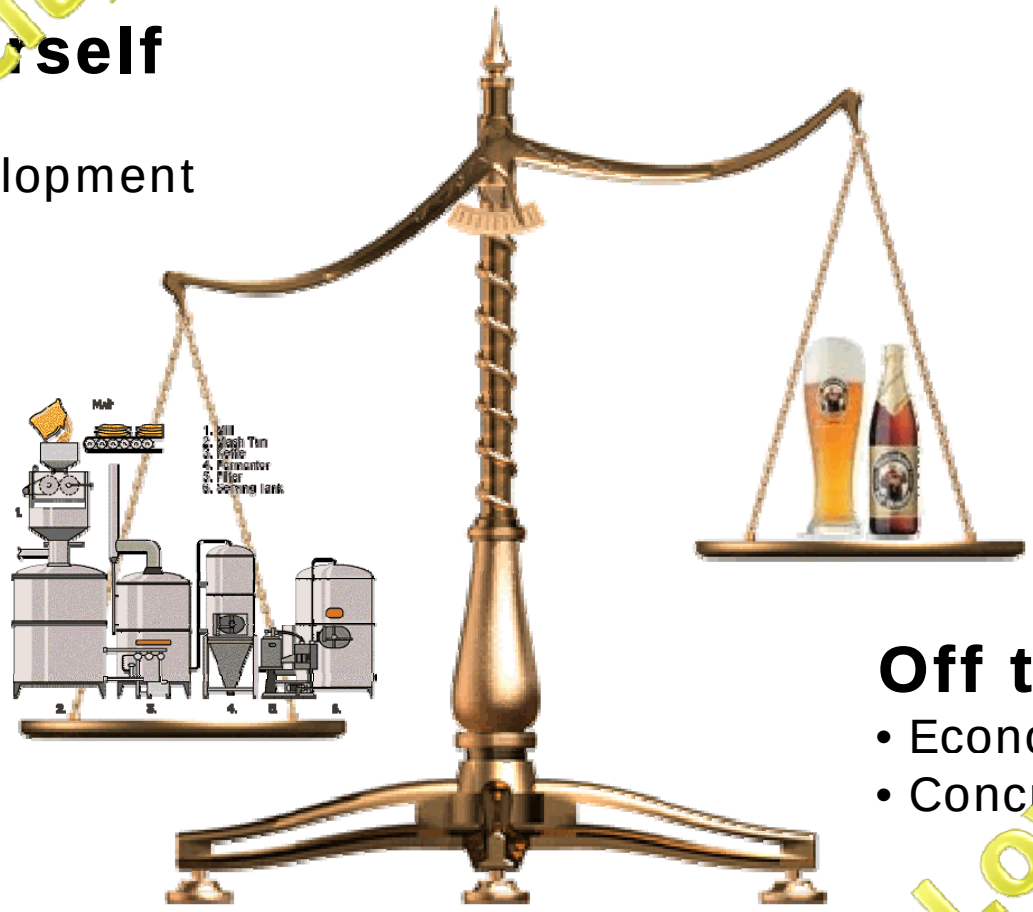
SQL Databases vS. NoSQL Databases
Michael Stonebraker
Communications of the ACM, 53(4), 2010

Brewery or bottled beer?

Specificity

Do It Yourself

- Expensive
- Ad hoc development



Low cost

Off the Shelf

- Economies of scale
- Concrete functionalities

Florian Waas analogy

Specific platforms

- Google BigTable
  - Published in 2006
  - Implemented by Hbase
    - Also Dynamo and Cassandra
- Google MapReduce
  - Published in 2004
  - Implemented by Hadoop
- MongoDB
  - Published in 2007
- Neo4J/Sparksee
  - Published in 2010/2008
- SAP HANA
  - Published in 2011
  - Prototyped in SanssouciDB

Summary

- Big Data definition
- Key features of cloud software (i.e., DBMS)
- Distributed Database definition
- Impedance Mismatch
- NOSQL main goals and features

Bibliography

- M. T. Özsu and P. Valduriez. *Principles of Distributed Database Systems*, 3rd Ed. Springer, 2011
- A. Ghazal et al. *BigBench: towards an industry standard benchmark for big data analytics*. SIGMOD Conference, 2013
- R. Cattell. *Scalable SQL and NoSQL Data Stores*. SIGMOD Record 39(4), 2010
- L. Liu, M.T. Özsu (Eds.). *Encyclopedia of Database Systems*. Springer, 2009