Big Data Management

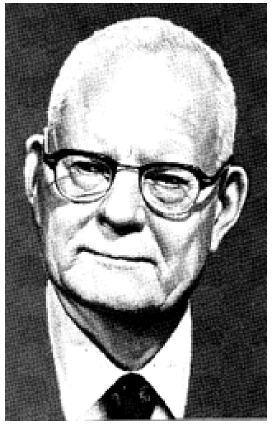
Big Data Management

Introduction to Big Data

Data driven decision making

The relevance of data

- "Without data you are just another person with an opinion."
  - William Edwards Deming (American engineer, statistician, professor and consultant



- "It is a capital mistake to theorize before one has data."
  - Sherlock Holmes (A Study in Scarlet)

We live in a data-driven society

Collect, store, combine and analyze any relevant data to gain competitive advantage

- Decision making
  - *To identify and choose alternatives based on values, preferences and beliefs of the decision-maker … every decision-making process produces a final choice.*

Wikipedia

- 90% of the world's data has been generated in the last two years
  - Data-driven decision making

Marr

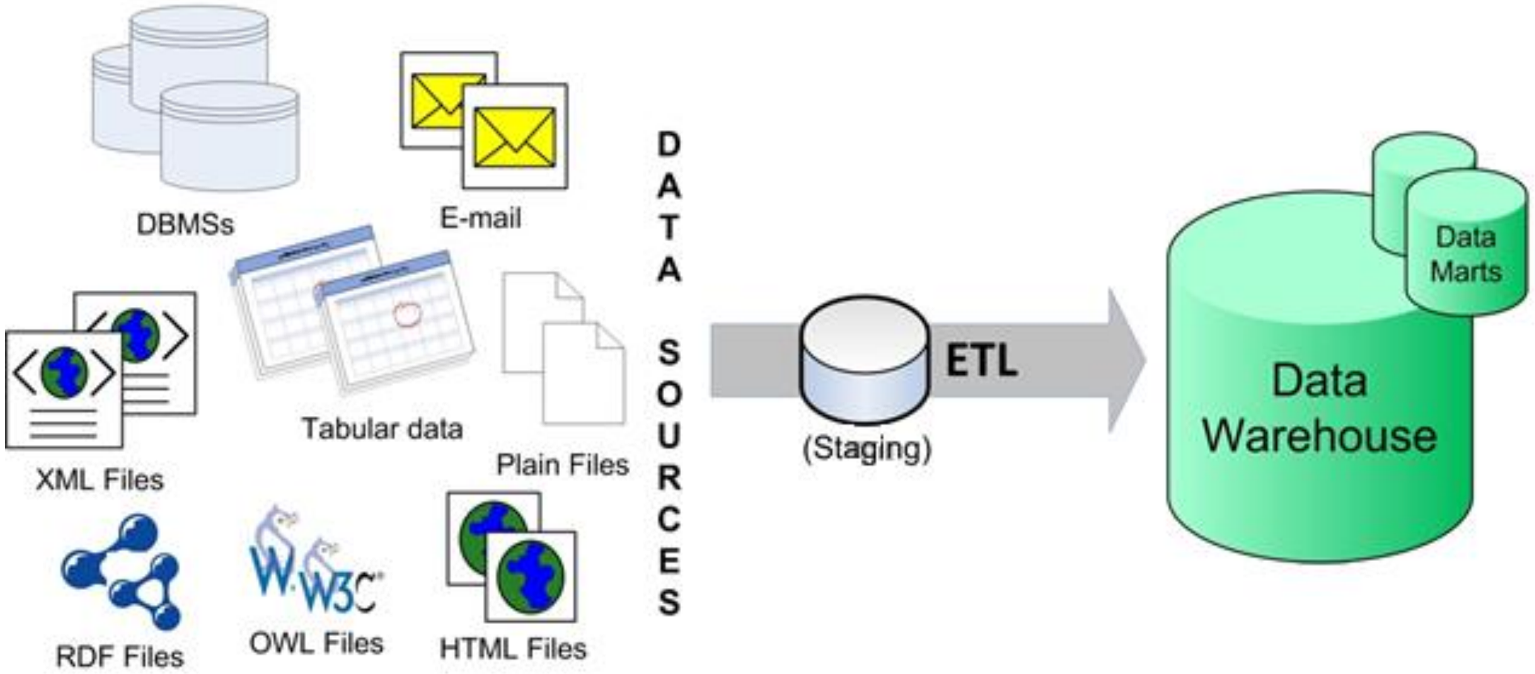
Data as the New Cornerstone

- We have witnessed the bloom of a new business model based on data analytics: *Data is not a passive but an active asset*
  - «*Data is the new oil!*» – Clive Humby, 2006
  - «*No! Data is the new soil*» – David McCandless, 2010
- Confluence of three major socio-economic and technological trends makes data-driven innovation a new phenomenon today:
  - The **exponential growth** in data generated and collected,
  - the **widespread** use of **data analytics** including start-ups and small and medium enterprises (SMEs), and
  - the emergence of a **paradigm shift in knowledge**
- Organizations must adapt infrastructures to leverage the data deluge (Digital data doubling every 18 months (IDC))

International Data Corp.'s (IDC)

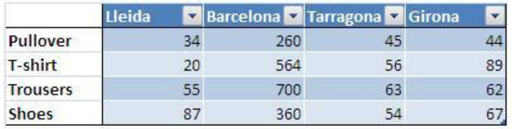
Business Intelligence: Data Management

- Well-established de facto standards:
  - Architecture: Corpotare Information Factory
  - Data Modeling: Multidimensional model

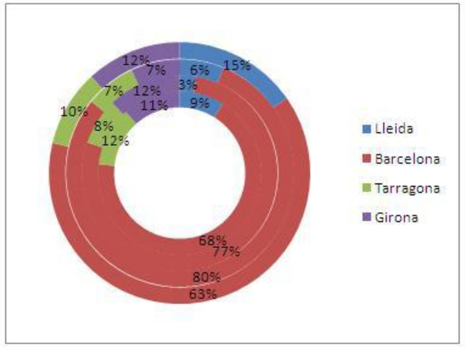


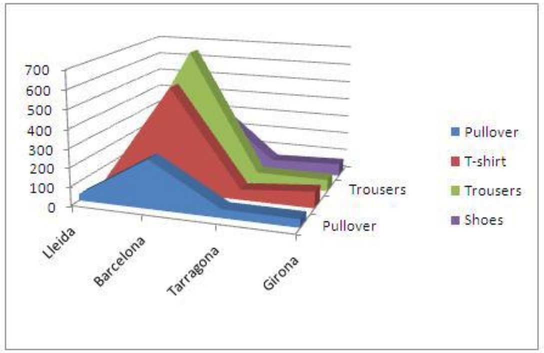
Business Intelligence: Analytics

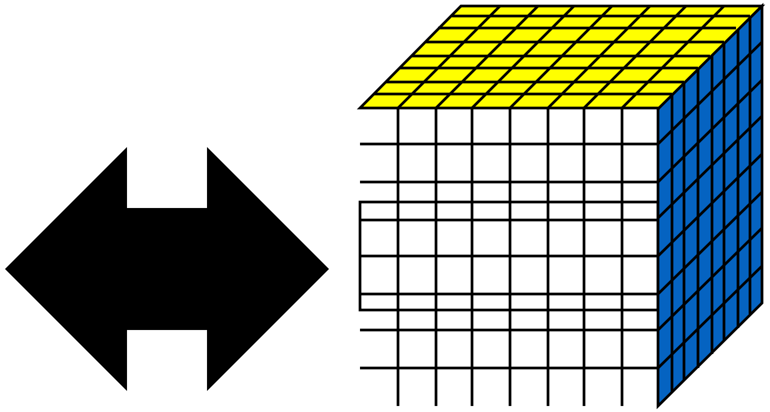
- Three different levels of detail
  - Querying & Reporting: Static report generation
  - OLAP: Dynamic navigation of data
  - Data Mining and Machine Learning: Inference of hidden patterns or trends



	Lleida	Barcelona	Tarragona	Girona
Pullover	34	260	45	44
T-shirt	20	564	56	89
Trousers	55	700	63	62
Shoes	87	360	54	67

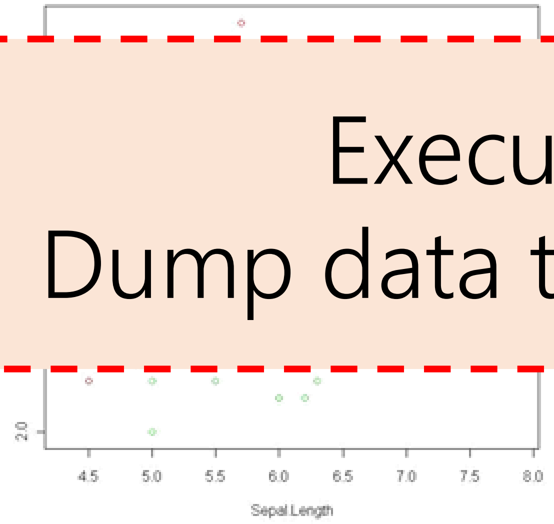


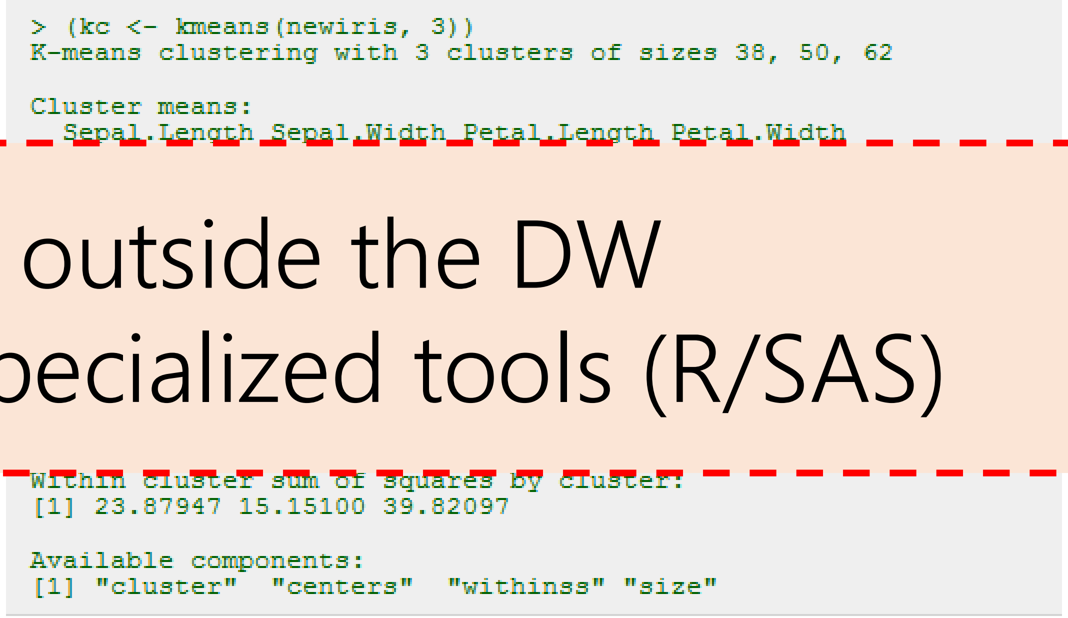




Business Intelligence: Analytics

- Three different levels of detail
  - Querying & Reporting: Static report generation
  - OLAP: Dynamic summarizations of data
  - **Data Mining and Machine Learning: Inference of hidden patterns or trends**



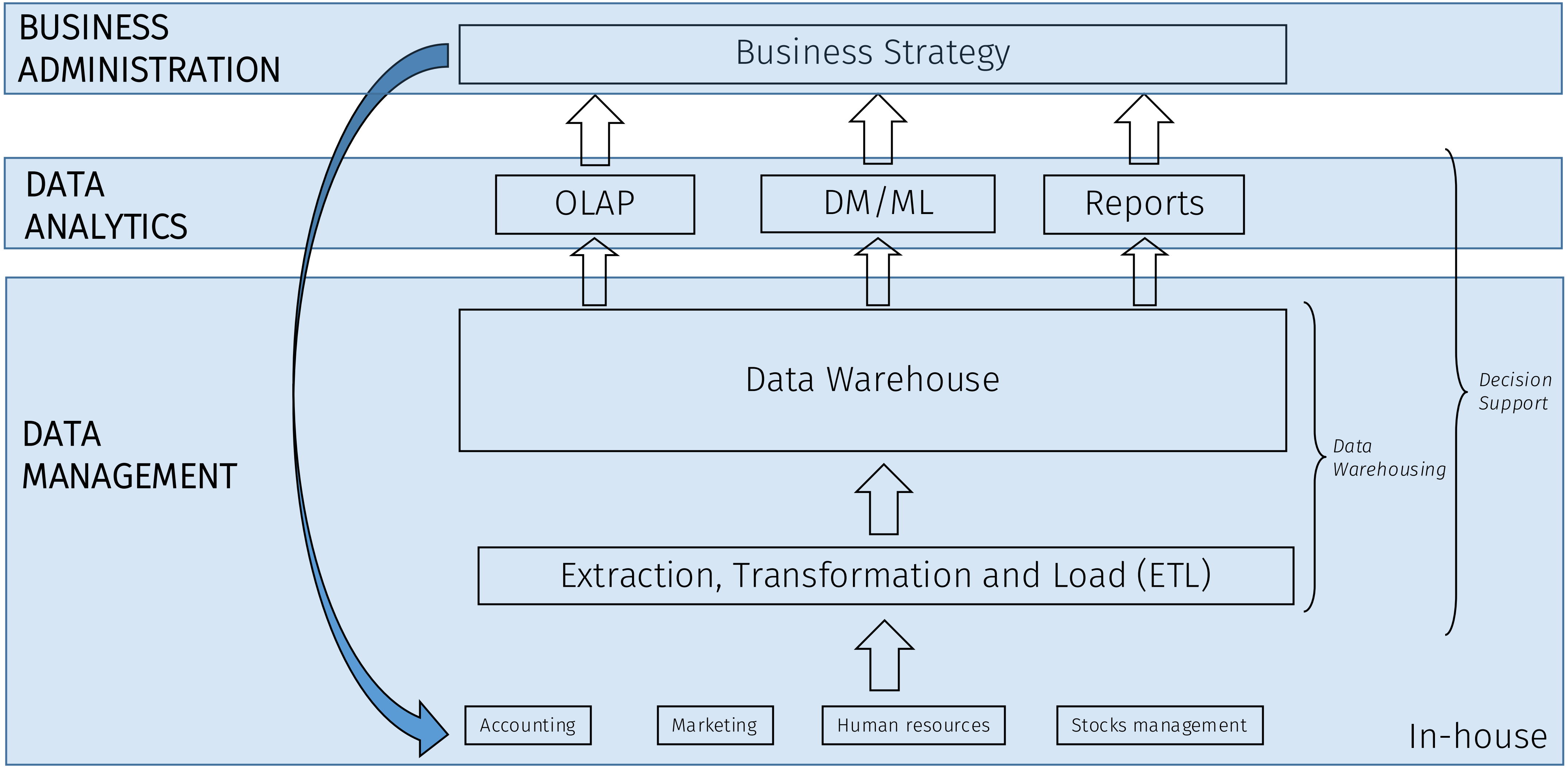


```
> (kc <- kmeans(newiris, 3))
K-means clustering with 3 clusters of sizes 38, 50, 62

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
```

```
Within cluster sum of squares by cluster:
[1] 23.87947 15.15100 39.82097

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

Executed outside the DW
Dump data to specialized tools (R/SAS)

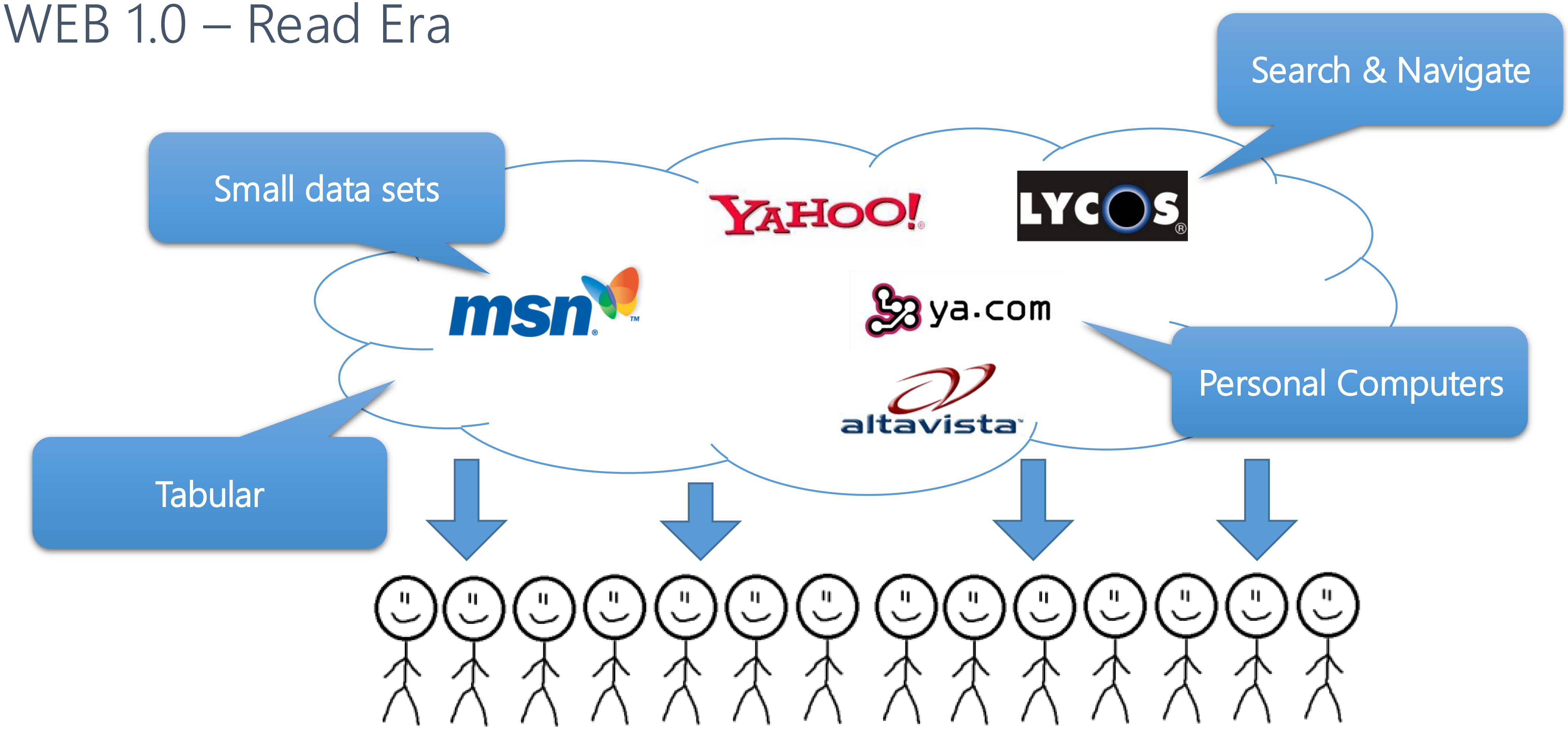
The Business Intelligence (BI) Cycle



Big Data

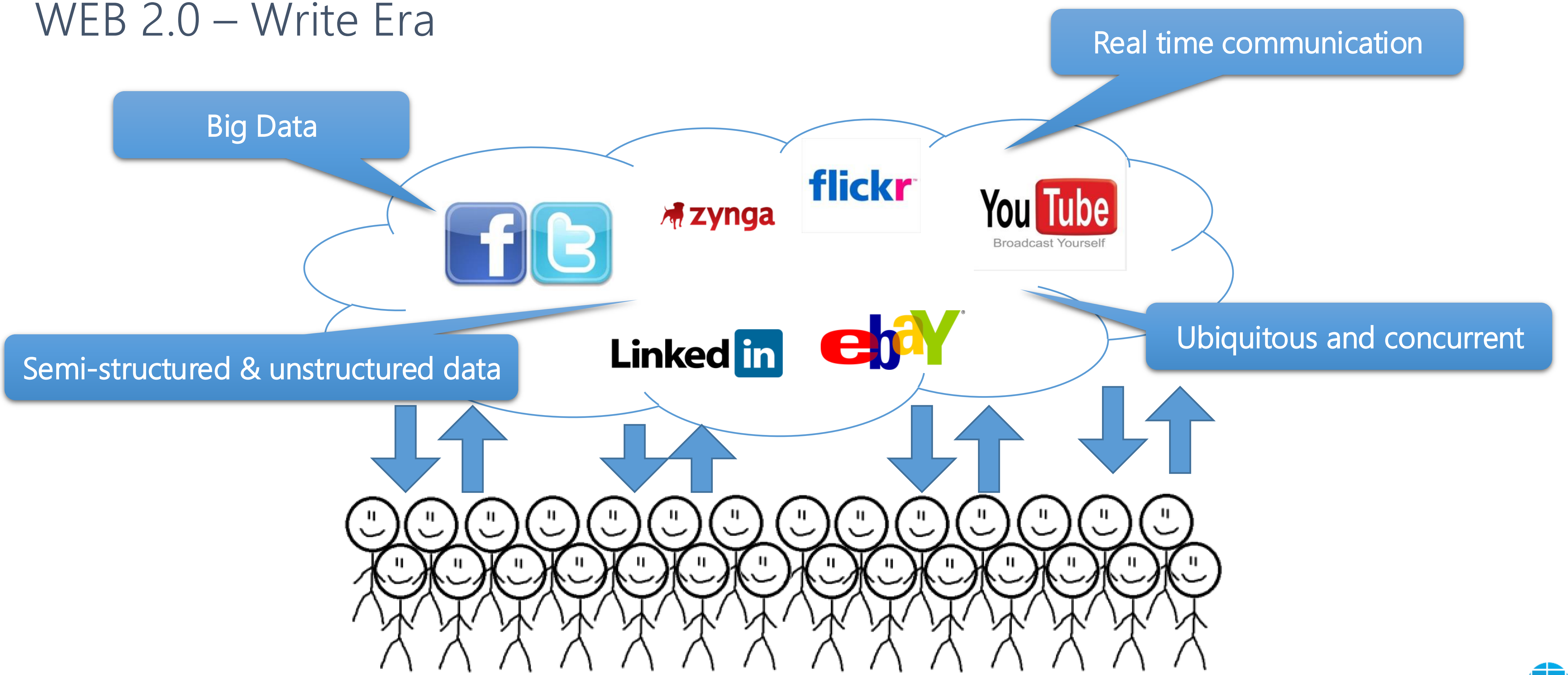
The end of an architectural era

WEB 1.0 – Read Era

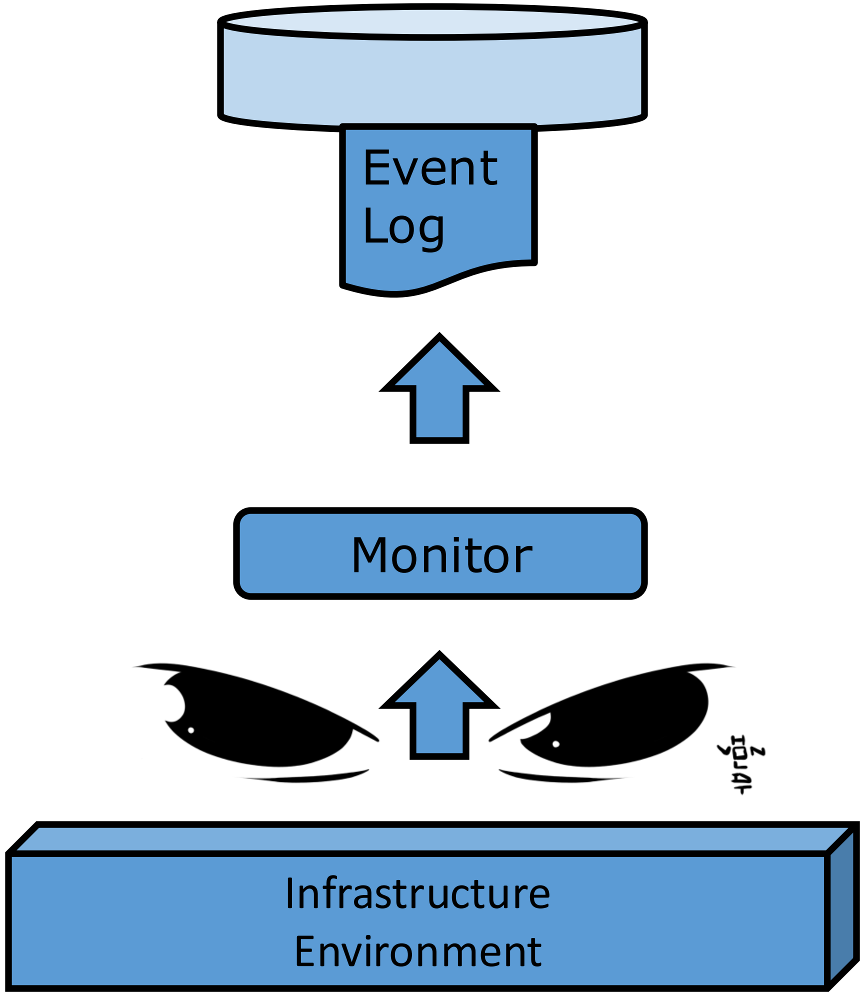


The end of an architectural era

WEB 2.0 – Write Era



Monitoring the infrastructure



Danish wind turbines

- One park:
  - 100+ turbines
- One turbine:
  - 500 sensors
  - More than 2500 derived data streams
- One sensor:
  - 8 bytes sampled at 100+Hz

100 turbines*2500 streams*100 samples/sec = $25 \cdot 10^6$ samples/second
8bytes*$25 \cdot 10^6$ samples/second*3600second/hour*24hours/day = 17.5TB/day
17.5TB/day*365 = 6+ PB/year/park

Having thousands of parks and storing 20+ years of history …

New challenges for data management

VOLUME

¿ty

V

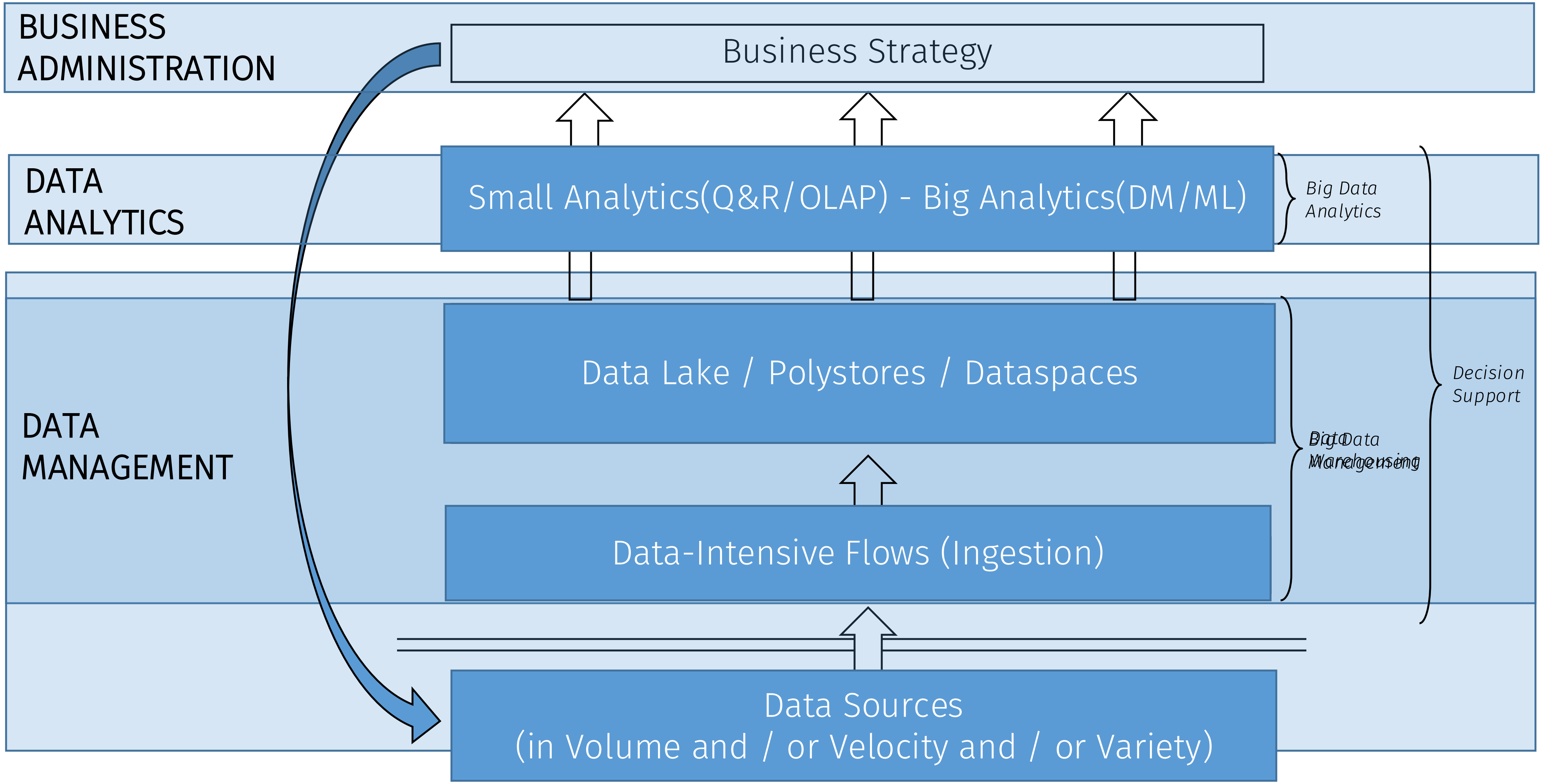
y

Value = f(V,V,V,V,V)

Variability

Variety

The Big Data Cycle



BUSINESS ADMINISTRATION

Business Strategy

DATA ANALYTICS

Small Analytics(Q&R/OLAP) - Big Analytics(DM/ML)

Big Data Analytics

DATA MANAGEMENT

Data Lake / Polystores / Dataspaces

Data-Intensive Flows (Ingestion)

Big Data Management

Data Warehousing

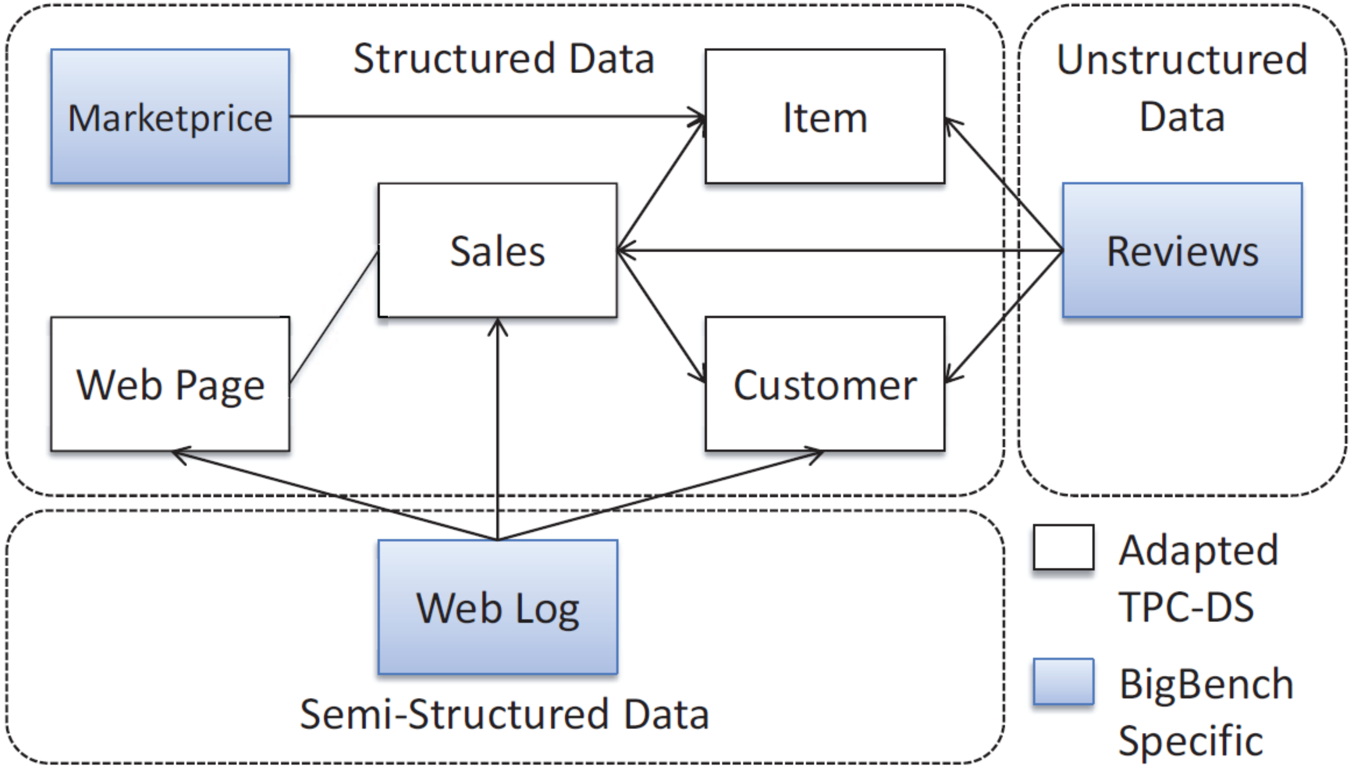
Decision Support

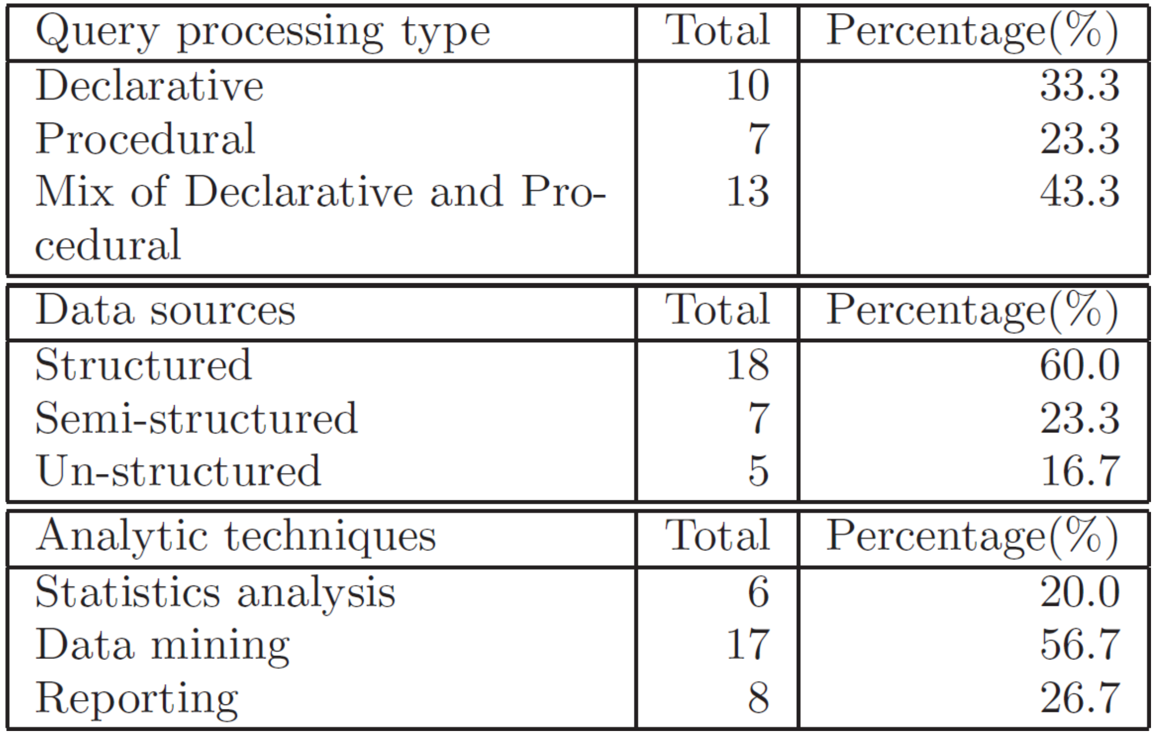
Data Sources
(in Volume and / or Velocity and / or Variety)

Big Data related areas

- Volume and Velocity
  - Distributed processing
  - Parallelism
  - Declarative querying
  - Query optimization
- Variety and Variability
  - Information retrieval
  - Web and text mining
  - Schema evolution
  - Data integration
- Veracity/Validity
  - Data quality
  - Uncertainty
  - Statistical reasoning
  - Data lineage and provenance
- Value
  - Analytics (ML)
  - Biology, Linguistics, Sports

Big Bench



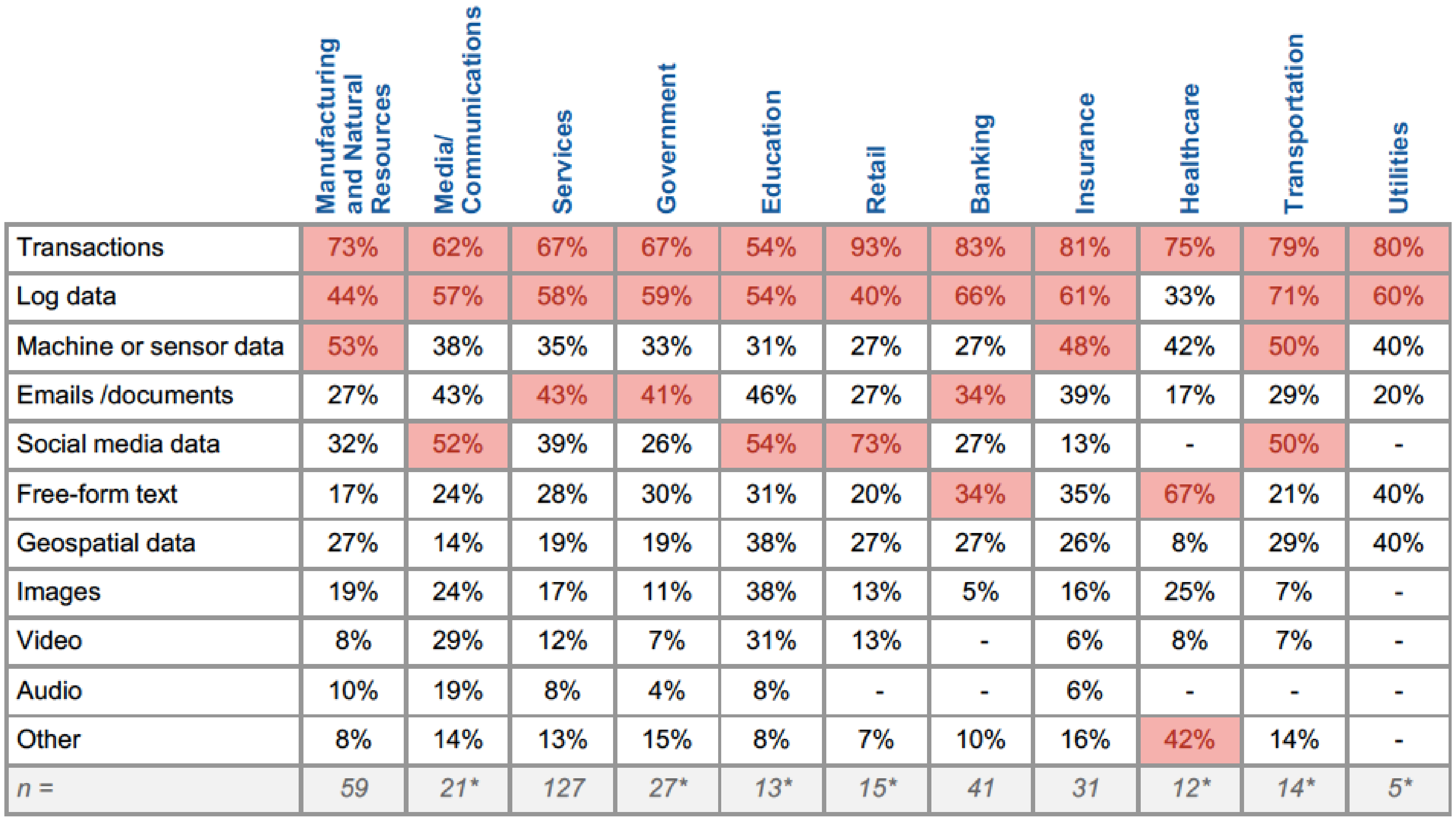


Query processing type	Total	Percentage(%)
Declarative	10	33.3
Procedural	7	23.3
Mix of Declarative and Procedural	13	43.3

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

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)

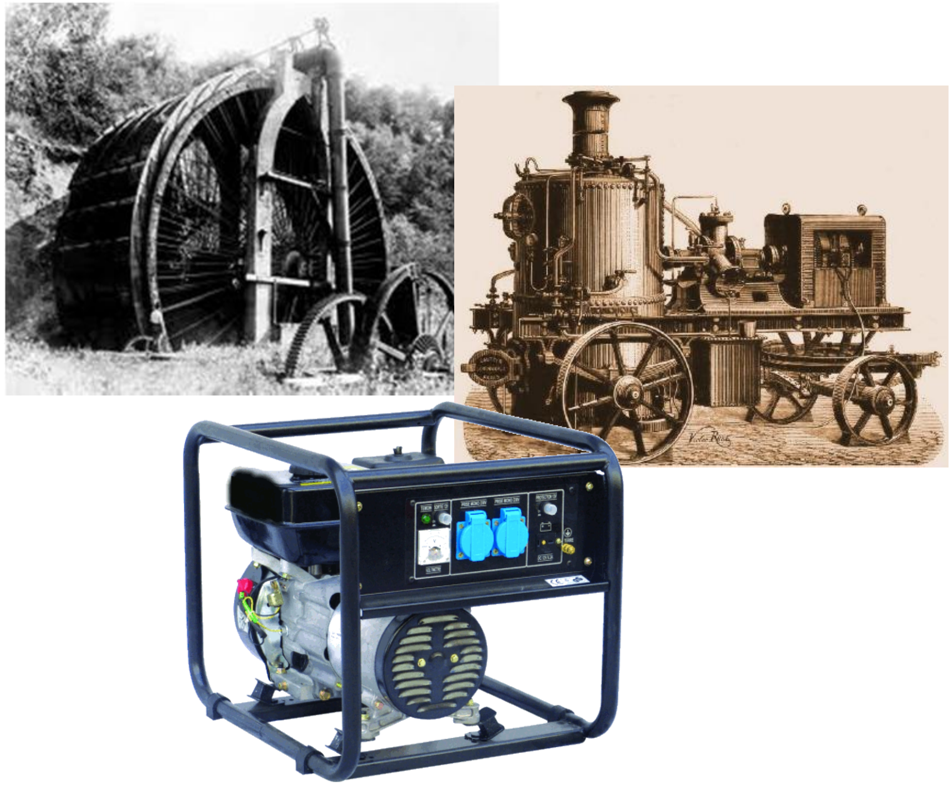
An orthogonal classification: kinds of data analytics

- Descriptive: Deterministically compute summarizations
  - Count, sum, average, min, max, etc.
  - Typical OLAP operations
- Predictive: Probabilistic by nature, try to forecast what may happen according to what have happened
  - Linear and non-linear regression,
  - Classification,
  - Clustering,
  - Association rules, etc.
- Prescriptive: Given the prediction(s) of a (several) model(s), understand why something is happening and undertake automatic action(s)
  - Examples:
    - Stock market (buy/sell shares)
    - Set Price (automatically increase/decrease)

Cloud Computing

Providing access to infrastructure

Analogy: Electricity as a Utility



Own production

Pay-per-use

Computation as a Utility

Private Data Centre

(*“Own production”*)

Public/Private Cloud

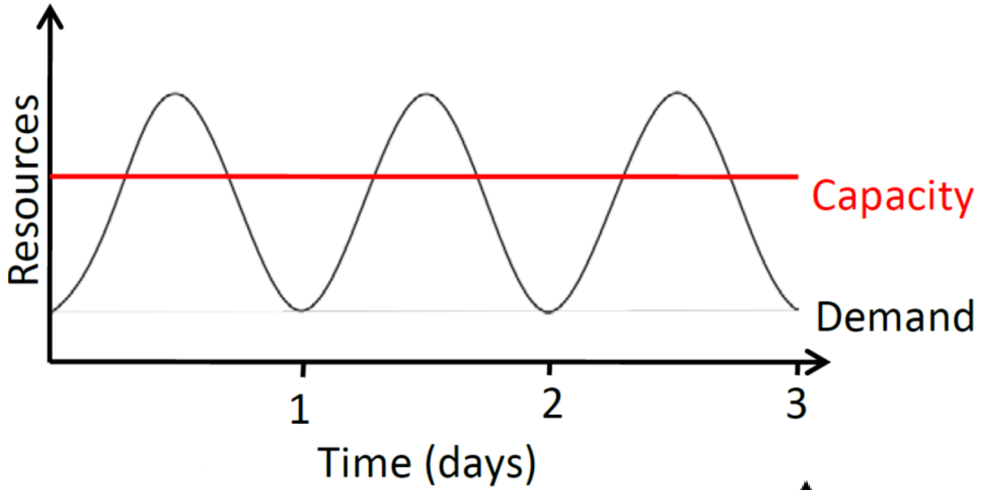
(Pay-per-use)

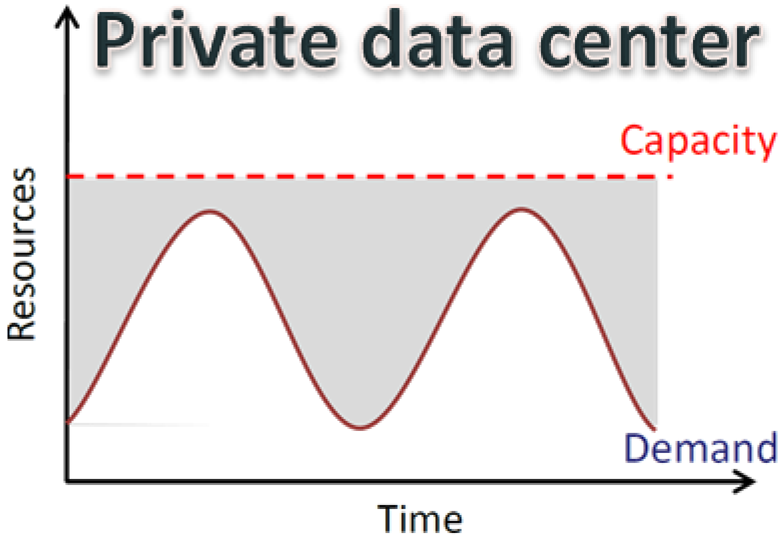
Cloud Computing definition

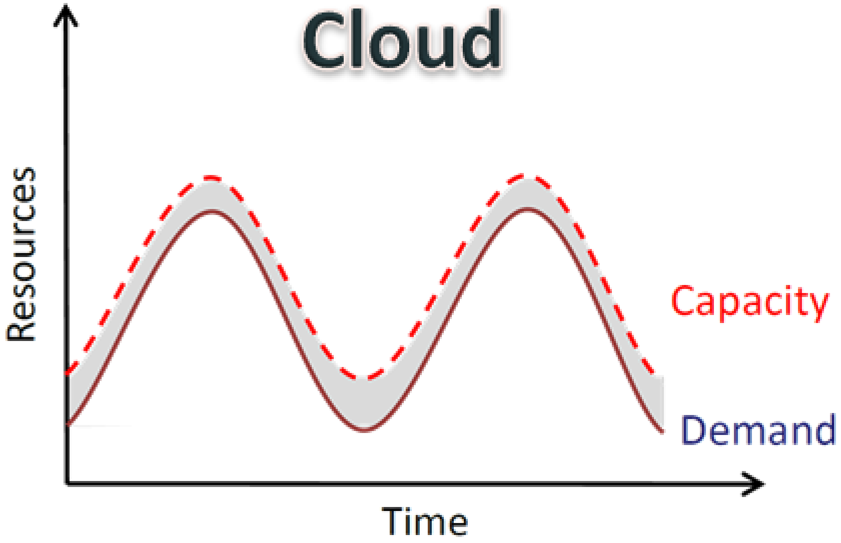
“Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”

NIST (National Institute of Standards and Technology)

Undercapacity Risk







D. Abadi

Novelty of cloud computing

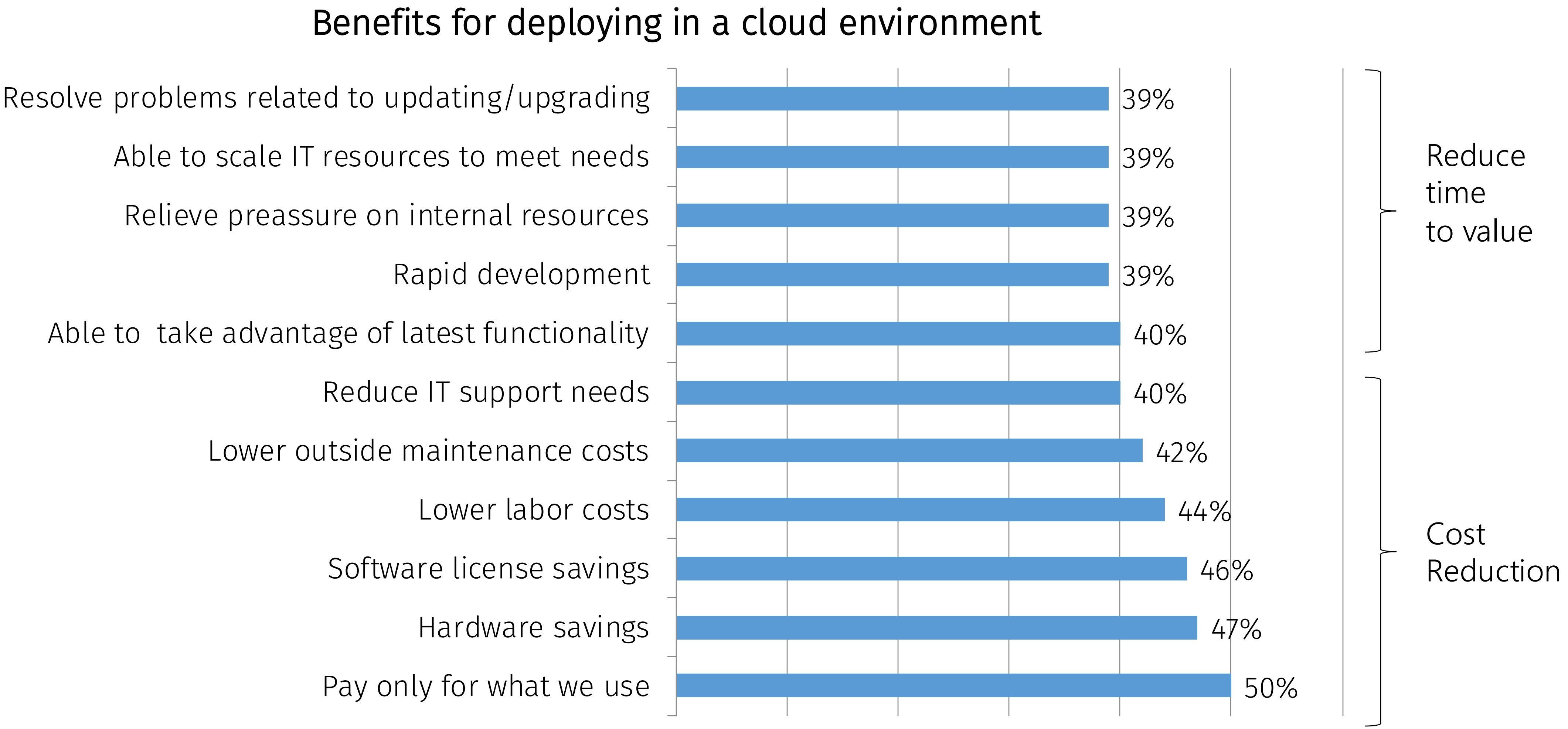
- Elimination of up-front commitment
- Illusion of infinite resources
- Pay-per-use (elasticity)
  - Cost is 5-7 times cheaper than in-house computing
- Service Level Agreements
  - E.g., Availability=uptime/(uptime+downtime)
    - Measured in terms of nines (99.99···9%)

Benefits of Cloud computing

- Reduce costs
  - Economy of scale in software development
  - Energetic efficiency
- Agility
- Flexibility
- Easier manageme
- Superior safety
- Better upgradeabilit
- More business

Big Data

Benefits of cloud computing

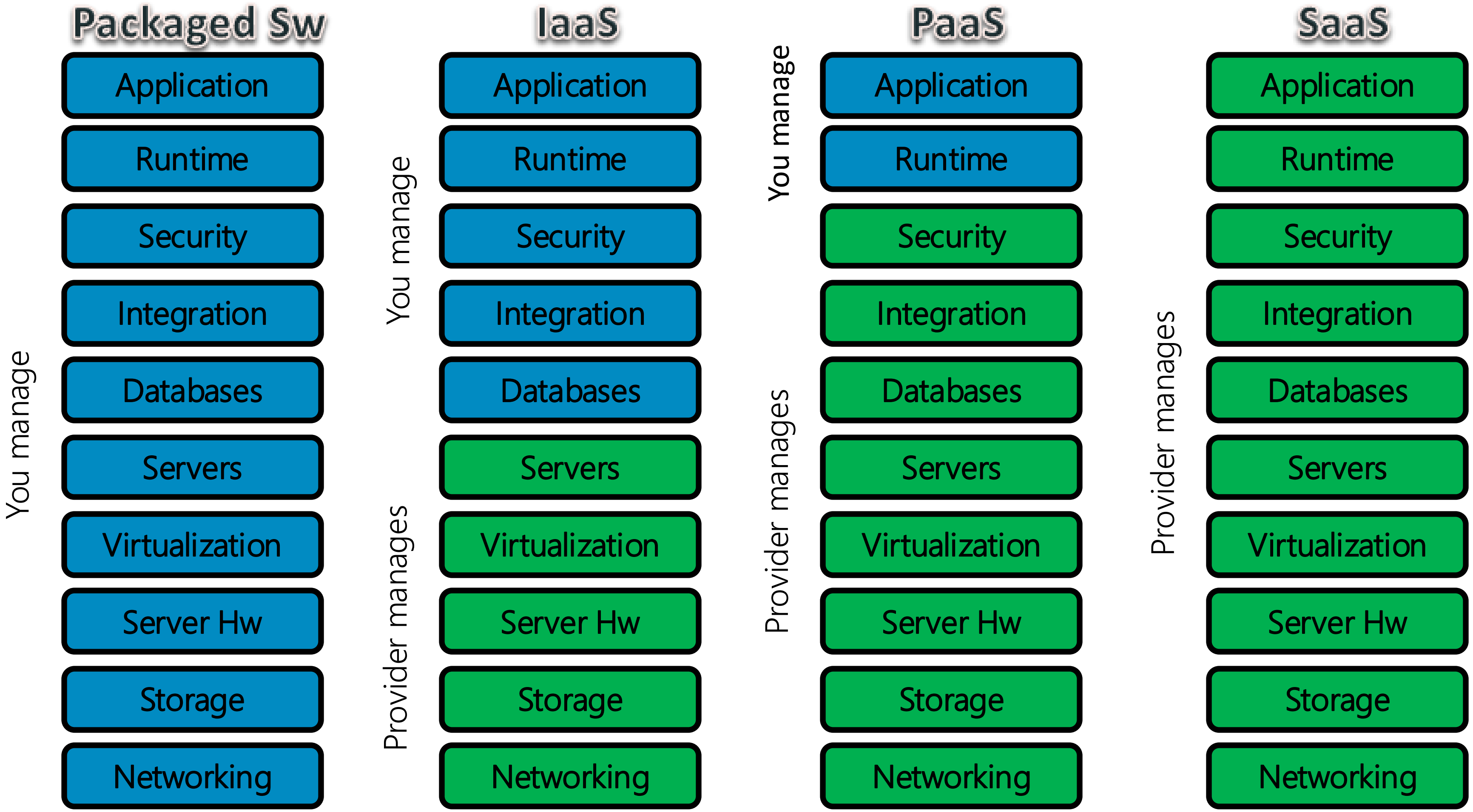


IBM global survey of IT and line-of-business decision makers 2012

Levels of Service

- The company outsources some responsibility to the service provider
  - Infrastructure as a Service (IaaS)
    - You get a server to connect through remote connection protocols (e.g., VPN, SSH, FTP)
    - Typically it covers the hardware (e.g., computers, network, virtualization)
  - Platform as a Service (PaaS)
    - You get software modules needed to run applications (e.g., databases, web servers, security)
  - Software as a Service (SaaS)
    - Software is there ready to be used (e.g., Google Docs, Dropbox)
  - Business as a Service (BaaS)
    - A whole business process is outsourced (e.g., Paypal, Amadeus)
- Levels are incremental: SaaS implies PaaS, and PaaS implies IaaS

Share of responsability



Packaged Sw

You manage

- Application
- Runtime
- Security
- Integration
- Databases
- Servers
- Virtualization
- Server Hw
- Storage
- Networking

IaaS

You manage

- Application
- Runtime
- Security
- Integration
- Databases

Provider manages

- Servers
- Virtualization
- Server Hw
- Storage
- Networking

PaaS

You manage

- Application
- Runtime

Provider manages

- Security
- Integration
- Databases
- Servers
- Virtualization
- Server Hw
- Storage
- Networking

SaaS

Provider manages

- Application
- Runtime
- Security
- Integration
- Databases
- Servers
- Virtualization
- Server Hw
- Storage
- Networking

Service providers

- Some of the strongest players in the market
  - Amazon Web Services (AWS)
  - Google Cloud
  - Microsoft Azure
  - IBM Cloud
  - Rackspace
  - Digital Ocean

Closing

References

- D. Abadi. Data management in the cloud: Limitations and opportunities. IEEE Data Engineering Bulletin 32(1), 2009
- C. Baun et al. Cloud Computing. Springer, 2011
- P. Mell and T. Grance. The NIST Definition of Cloud Computing. Special Publication 800-145, National Institute of Standards and Technology (September 2011)
- C. Baun et al. Cloud Computing. Springer, 2011
- M. Madsen. Cloud Computing Models for Data Warehousing. Third Nature Technology White Paper, 2012
- A. Ghazal et al. BigBench: towards an industry standard benchmark for big data analytics. SIGMOD'13
- NIST Cloud Computing Program, http://www.nist.gov/itl/cloud
- Gartner Reports. G00232650, G00175593, and G00219131