Data Management for Data Science

Master of Science in Data Science
Facoltà di Ing. dell'Informazione, Informatica e Statistica
Sapienza Università di Roma

AA 2018/2019

An Introduction to Big Data

Domenico Lembo
Dipartimento di Ingegneria Informatica,
Automatica e Gestionale A. Ruberti

Availability of Massive Data

- Digital data are nowadays collected at an unprecedent scale and in very many formats in a variety of domains (e-commerce, social networks, sensor networks, astronomy, genomics, medical records, etc.)

- This is has been made possible by the incredible growth in recent years of the capacity of data storage tools and of the computing power of electronic devices, as well as by the advent of mobile and pervasive computing, cloud computing, and cloud storage.

Exploitability of Massive Data

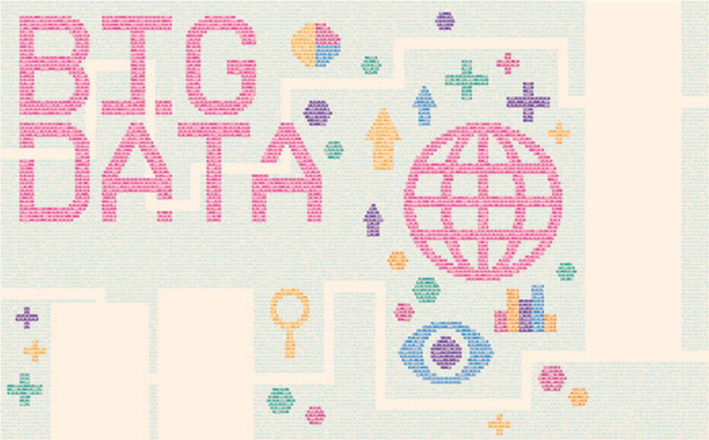
- How to transform available data into information, and how to make organizations' business to take advantages of such information are long-standing problems in IT, and in particular in information management and analysis.

- These issues have become more and more challenging and complex in the "Big Data" era

- At the same time, facing the challenge can be even more worthy than in the past, since the massive amount of data that is now available may allow for analytical results never achieved before

Be careful!

- "Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media" (Tim Harford, journalist and economist, March, 2014)*



* http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz3EvSLWwbu

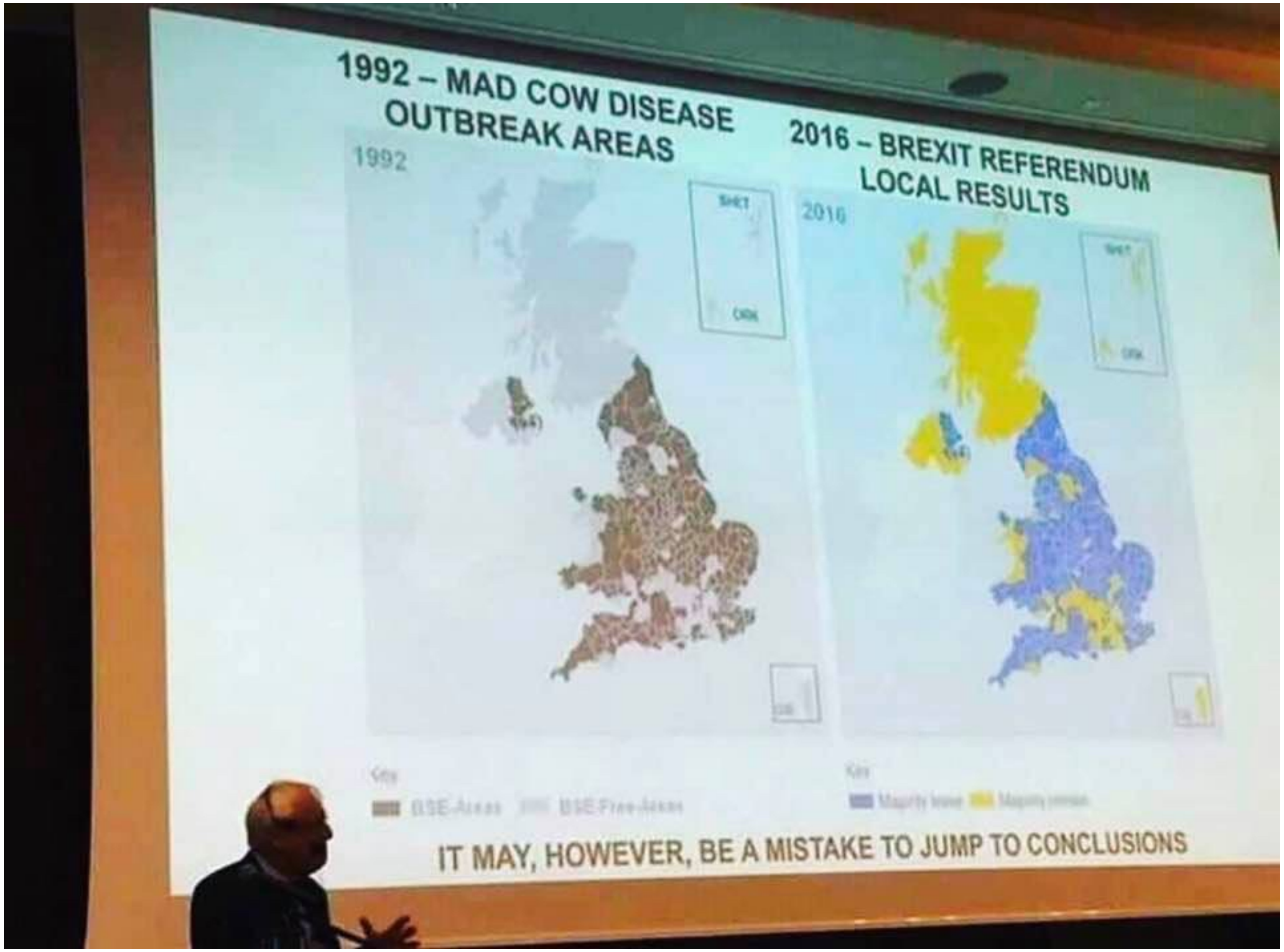
Moore's Law for #BigData: The amount of nonsense packed into the term "BigData" doubles approximately every two years (Mike Pluta, Data Architect, on Twitter August 2014).
https://twitter.com/mikepluta/status/502878691740090369

The Google Flu Trends*

- 2008: Google people publish "Detecting influenza epidemics using search engine query data" on nature (https://www.nature.com/articles/nature07634).
- They were able to track the spread of influenza across the US more quickly than the US Centers for Disease Control and Prevention (CDC).
- The tracking was essentially based on correlation between what people searched for online and whether they had flu symptoms.
- Four years later, with a similar experiments Google people overstimated the spread of influenza by almost a factor of two!
- "theory-free analysis of mere correlations is inevitably fragile if you have no idea what is behind a correlation".

* http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz3EvSLWwbu

Correlation vs Causality!



1992 – MAD COW DISEASE OUTBREAK AREAS

2016 – BREXIT REFERENDUM LOCAL RESULTS

IT MAY, HOWEVER, BE A MISTAKE TO JUMP TO CONCLUSIONS

Understanding Big Data is in fact difficult!

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISHBURNE
© marketoonist.com

"There are a lot of small data problems that occur in big data. They don't disappear because you've got lots of the stuff. They get worse!" (David Spiegelhalter, Cambridge University)

Thinking Big Data*

"Big Data" has leapt rapidly into one of the most hyped terms in our industry, yet the hype should not blind people to the fact that this is a genuinely important shift about the role of data in the world. The amount, speed, and value of data sources is rapidly increasing. Data management has to change in five broad areas: ***extraction*** *of data from a wider range of sources, changes to the logistics of data management with* ***new database and integration approaches****, the use of* ***agile principles in running analytics*** *projects, an emphasis on techniques for* ***data interpretation*** *to separate signal from noise, and the importance of well-designed* ***visualization*** *to make that signal more comprehensible. Summing up this means we don't need big analytics projects, instead we want the new data thinking to permeate our regular work."*

Martin Fowler

*http://martinfowler.com/articles/bigData/

Thinking Big Data

- Thus, roughly, *Big Data is data that exceeds the processing capacity of conventional database systems*

- But also *Big Data is understood as a capability that allows companies to extract value from large volumes of data*

- but, notice, this does not mean only extremely large, massive databases

- Besides data dimension, what characterizes Big Data are also the heterogeneity in the way in which information is structured, the dynamicity with which data changes, and the ability of quickly processing it

- This calls for new computing paradigms or frameworks, not only advanced data storage mechanisms

The Three Vs

To characterize Big Data, three Vs are used, which are the Vs of

- ***Volume***
- ***Velocity***
- ***Variety***

Volume

- Big data applications are characterized of course by big amounts of data, where big means extremely large, e.g., more than a terabyte (TB) or petabyte (PB), or more.
- There are various contexts in which these dimensions can be easily reached: chatters from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, GPS trails, financial market data, biological data, etc.
- Some more concrete examples:
  - Despite some Youtube statistics are available[1] the total storage capacity of Youtube it's not known, but realistically it should be no less than 1 EB (2016)
  - NSA data center: estimated storage capacity of at least 2,000 PBs (2013)[2]
  - Facebook: 300PB data warehouse (2014)[3]

[1] http://web.archive.org/web/20150217015601/http://www.youtube.com/yt/press/statistics.html
[2] http://www.forbes.com/sites/netapp/2013/07/26/nsa-utah-datacenter/#1b66cc7c3cd2
[3] https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/

Volume

- How many data in the world?

  According to IDC (International Data Corporation):
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006 (1EB=10^{18}B)
  - 500 Exabytes, 2009
  - 1.8 Zettabytes, 2011(1ZB=10^{21}B) [1]
  - 2.8 Zettabytes, 2012[1]
  - 4.4 Zettabytes, 2013
  - **175 Zettabytes by 2025**[2] (estimate)

Around 90% of world's data generated in the last 4 years.

The digital universe is doubling in size every two years[3]

Multiple of bits or bytes

Symbol	Name	Decimal value	Binary Value
k	kilo	1000	1024
M	mega	1000^2	1024^2
G	giga	1000^3	1024^3
T	tera	1000^4	1024^4
P	peta	1000^5	1024^5
E	exa	1000^6	1024^6
Z	zetta	1000^7	1024^7
Y	yotta	1000^8	1024^8

[1] http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html

[2] https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[3] https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Volume

- The sheer volume of data is enough to defeat many long-followed approaches to data management
- Traditional centralized database systems cannot handle many of the data volumes, forcing the use of clusters
- Data have to be necessarily distributed, and the number of sources providing information can be huge, much higher than the number considered in traditional data integration and virtualization systems

Velocity

- Data’s velocity (i.e., the rate at which data is collected and made available into an organization) has followed a similar pattern to that of volume
- Many data sources accessed by organizations for their business are extremely dynamic
- Mobile devices increase the rate of data inflow: data “everywhere”, collected and consumed continuously

Velocity

- Some examples:
  - Walmart: 1 million transaction per hour (2010)[1]
  - eBay: data throughput reaches 100 PBs per day (2013)[2]
  - Google processes 100 PBs per day (2013-14) [3]
  - Facebook: 600TB added to the warehouse every day (2014)[4]
  - 6000-8000 tweets per second every day (in 2019)[5]
- In 2013, it has been estimated that every minute of every day we created[6]:
  - More than 204 million email messages
  - 571 new Websites and 347 blog posts created
  - 72 hours of new YouTube videos
  - 1.8 millions of like on Facebook
  - 216.000 new photos on instagram
  - $83.000 spent on Amazon

[1] http://martinfowler.com/articles/bigData/

[2] http://www.v3.co.uk/v3-uk/news/2302017/ebay-using-big-data-analytics-to-drive-up-price-listings

[3] http://www.slideshare.net/kmstechnology/big-data-overview-2013-2014

[4] https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/

[5] http://www.internetlivestats.com/twitter-statistics/

[6] http://www.dailymail.co.uk/sciencetech/article-2381188/Revealed-happens-just-ONE-minute-internet-216-000-photos-posted-278-000-Tweets-1-8m-Facebook-likes.html

Velocity

- Processing information as soon as it is available, thus speeding the “feedback loop”, can provide competitive advantages
- Some examples of *Fast Data Processing*:
  - Customer Experience/Retail: online retails that are able to suggest additional products to a customer at every new information inserted during an on-line purchase (Click-stream analysis)
  - Financial Services Industry: Algorithmic trading using event process technology but also real-time data integration and analytics
  - Telecommunication: understand allocation of network resources based on traffic and application requirements, network usage patterns
  - Energy: real-time process of high volume of events to make important decisions in order to effectively and efficiently manage possible faults on the distribution network
  - Manufacturing: analyze real-time metrics to take corrective action before a failure occurs

Velocity

- **Stream processing** is a new challenging computing paradigm, where information is not stored for later batch processing, but is consumed on the fly

- This is particularly useful when data are too fast to store them entirely (for example because they need some processing to be stored properly), as in scientific applications, or when the application requires an immediate answer

Variety

- Data is extremely heterogeneous: e.g., in the format in which are represented, but also and in the way they represent information, both at the intensional and extensional level
- E.g., text from social networks, sensor data, logs from web applications, databases, XML documents, RDF data, etc.
- Data format ranges therefore from structured (e.g, relational databases) to semistructured (e.g., XML documents), to unstructured (e.g., text documents)

Variety

- As for unstructured data, for example, the challenge is to extract meaning for consumption both by humans or machines

- Entity resolution, which is the process that resolves, i.e., identifies, entities and detects relationships, then plays an important role

- In fact, these are well-known issues studied since several years in the fields of data integration, data exchange, and data quality. In the Big Data scenario, however, they become even more challenging

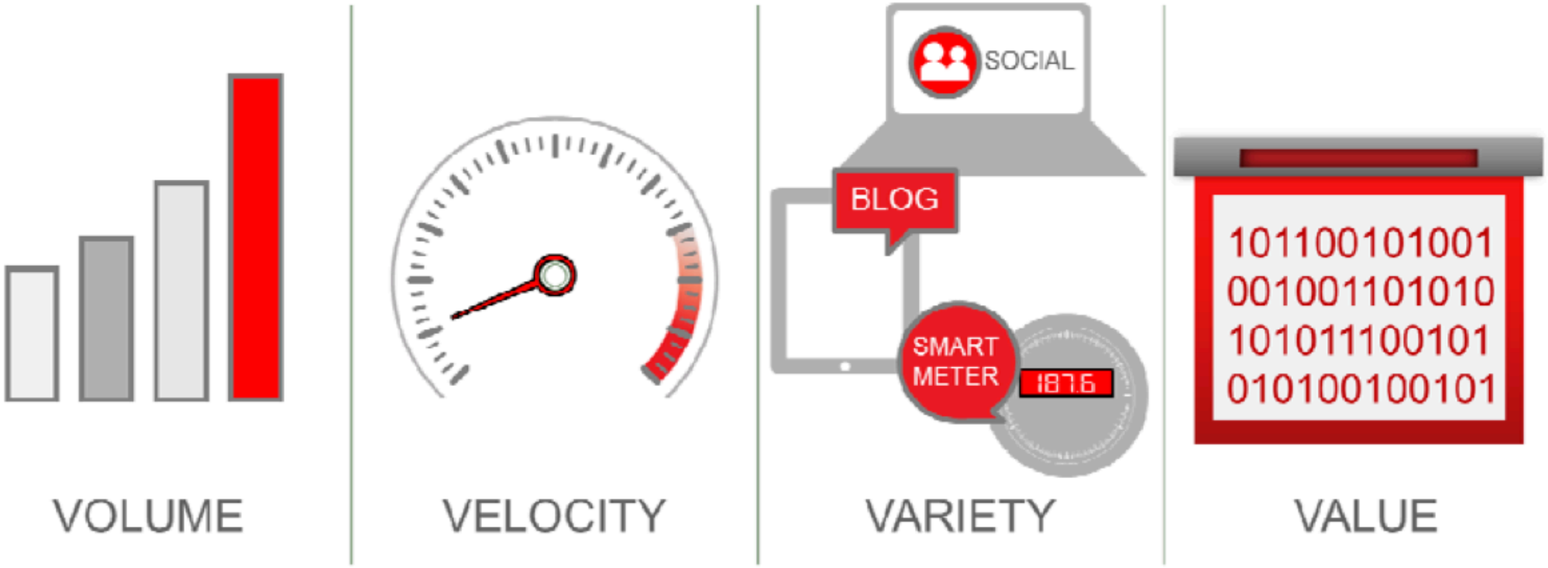
A fourth V: Veracity*

- Data are of widely different quality
- Traditionally data is thought of as coming from well organized databases with controlled schemas
- Instead, in “Big Data” there is often little or no schema to control their structure
- The result is that there are serious problems with the quality of the data

* The literature often mentions only *three* Vs and does not include veracity. However some authors tend to include veracity as a core characteristc of Big Data (alternatively, veracity is considered an aspect of variety)

Big Data: V^3+ Value

Big Data can generate huge competitive advantages!



The value of Data for organizations

- Although it's difficult to get hard figures on the value of making full use of your data, much of the success of companies such as Amazon and Google is credited to their effective use of data[1]

- Thus companies spend large amounts of money to reach this effective use: According to IDC, in 2017 big data and analytics software market reached $54.1 billion wordlwide, and it is expected to grow at a five-year CAGR (compound annual growth rate) of 11.2%. (analysis 2018-2022)[2]

- Thus various Big Data solutions are now promoted by all major vendors in data management systems

[1] http://martinfowler.com/articles/bigData/

[2] https://www.idc.com/getdoc.jsp?containerId=US44243318

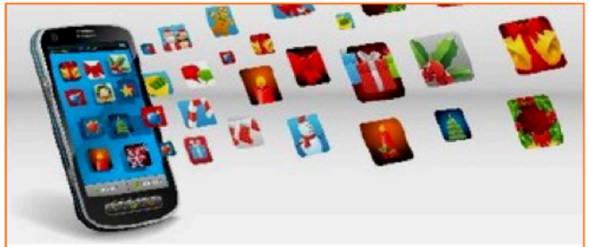
Potential value

US health care

- $300 billion value per year
- ~0.7 percent annual productivity growth

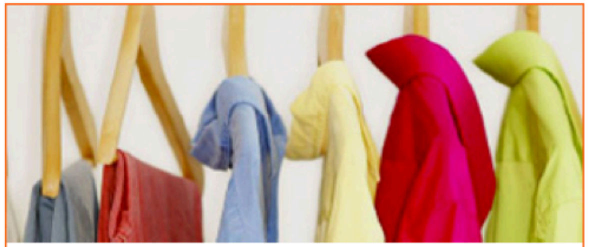
Europe public sector administration

- €250 billion value per year
- ~0.5 percent annual productivity growth



Global personal location data

- $100 billion+ revenue for service providers
- Up to $700 billion value to end users



US retail

- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing

- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

Demand for new data management solutions*

- In the scenarios we depicted it is not surprising that new data mangement solutions are demanded
- Indeed, despite the popularity and well understood nature of relational databases, it is not the case that they should always be the destination for data
- Depending on the characteristic of data, certain classes of databases are more suited than others for their management
- XML documents are more versatile when stored in dedicated XML storage systems (e.g., MarkLogic)
- Social network relations are graph by nature and graph databases such as Neo4J can make operations on them simpler and more efficient

* From: Edd Dumbill. What is Big data. In Planning for Big Data. O'Reilly Radar Team

Demand for new data management solutions*

- A *disadvantage of the relational database is the static nature of its schema*

- In an agile environment, the results of computation will evolve with the detection and extraction of new information

- Semi-structured **NoSQL databases** meet this need for flexibility: they provide some structure to organize data (enough for certain applications), but do not require the exact schema of the data before storing it

* From: Edd Dumbill. What is Big data. In Planning for Big Data. O'Reilly Radar Team

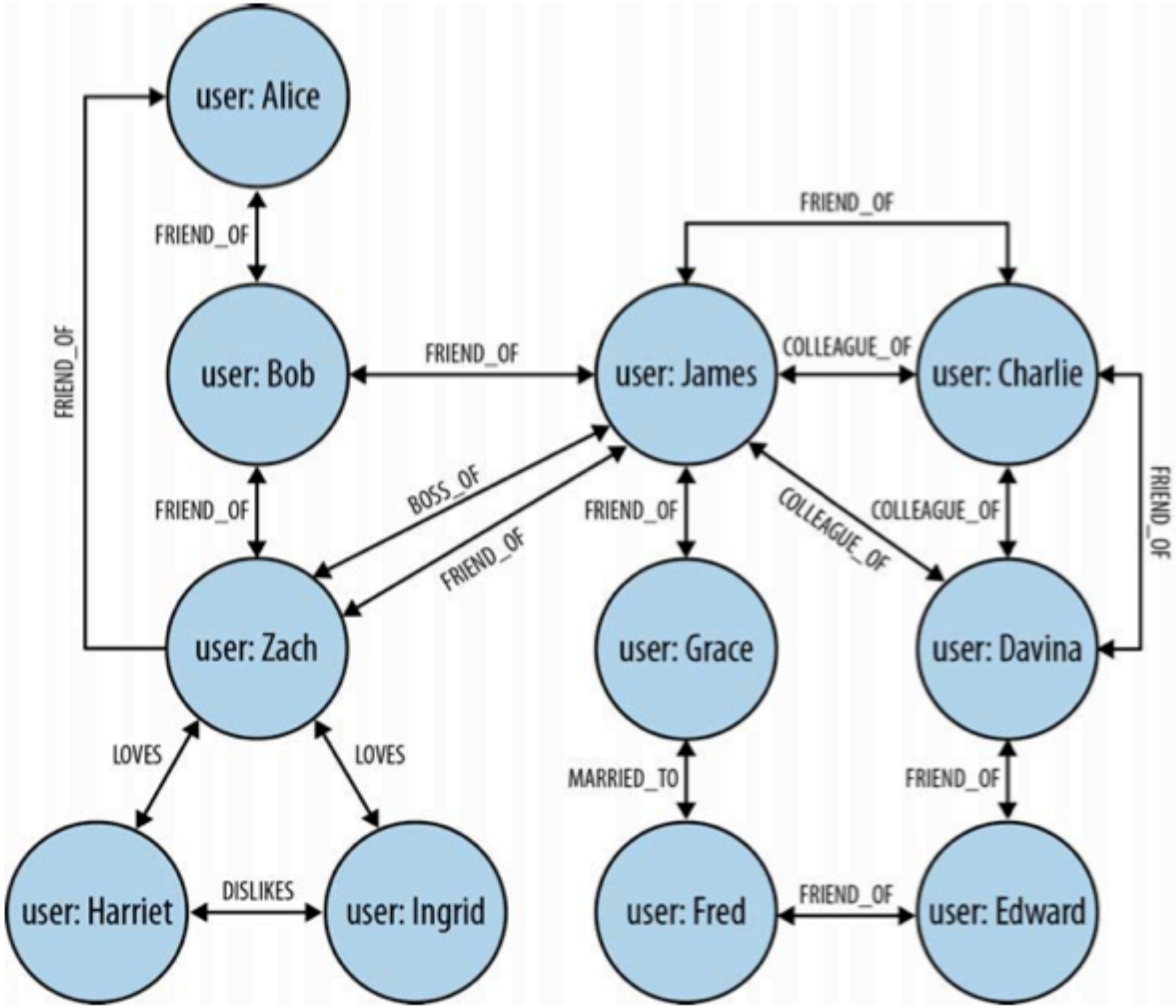
NoSQL databases*

Or better…not only SQL

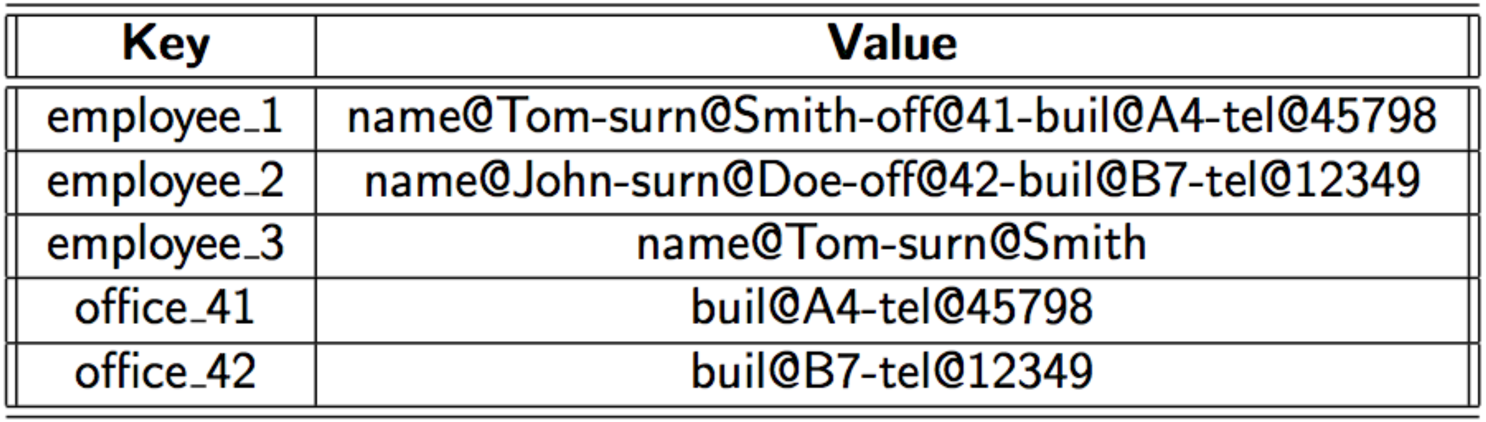
- The term "NoSQL" is very ill-defined. It's generally applied to a number of non-relational databases such as Cassandra, Mongo, Dynamo, Neo4J, Riak, and many others
- They embrace schemaless data, run on clusters, and have the ability to trade off traditional consistency for other useful properties
- Advocates of NoSQL databases claim that they can build systems that are more performant, scale much better, and are easier to program with

* From: Martin Fowler. NoSQL Distilled. Preface. (http://martinfowler.com/books/nosql.html)

Graph databases

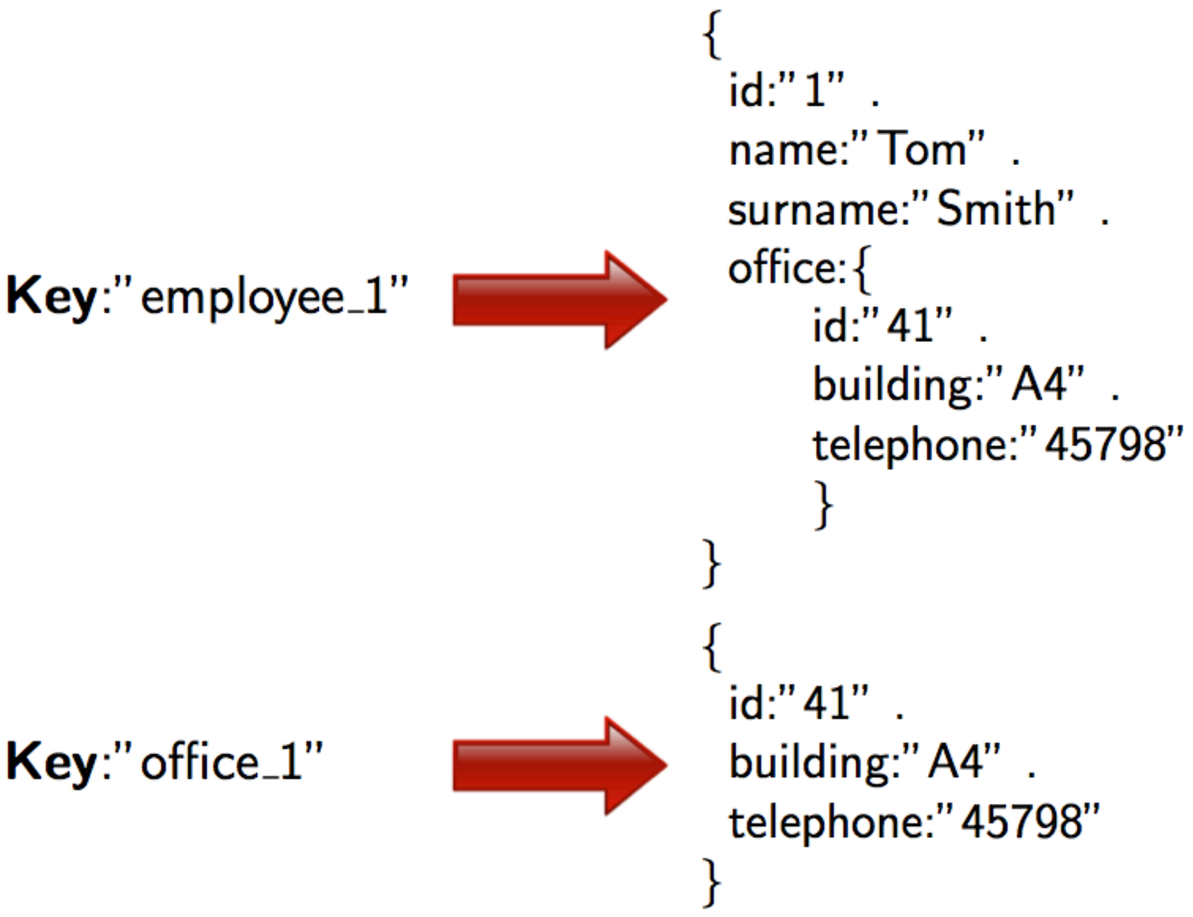


Key-values databases



Key	Value
employee_1	name@Tom-surn@Smith-off@41-buil@A4-tel@45798
employee_2	name@John-surn@Doe-off@42-buil@B7-tel@12349
employee_3	name@Tom-surn@Smith
office_41	buil@A4-tel@45798
office_42	buil@B7-tel@12349

Document databases



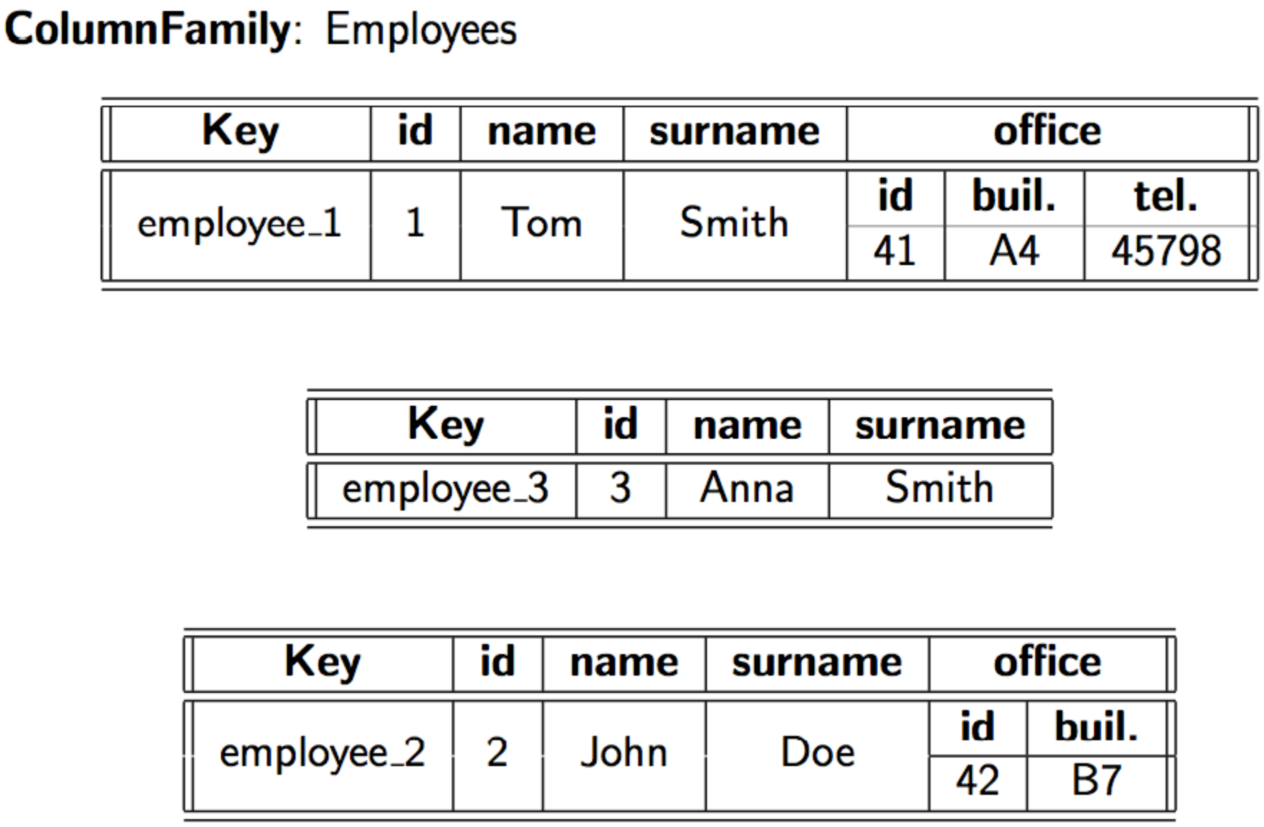
Key:"employee_1"

```
{
 id:"1" .
 name:"Tom" .
 surname:"Smith" .
 office:{
        id:"41" .
        building:"A4" .
        telephone:"45798"
        }
}
```

Key:"office_1"

```
{
 id:"41" .
 building:"A4" .
 telephone:"45798"
}
```

Column Family Databases



ColumnFamily: Employees

Key	id	name	surname	office		
employee_1	1	Tom	Smith	**id**	**buil.**	**tel.**
				41	A4	45798

Key	id	name	surname
employee_3	3	Anna	Smith

Key	id	name	surname	office	
employee_2	2	John	Doe	**id**	**buil.**
				42	B7

NoSQL databases*

- Is this the first rattle of the death knell for relational databases, or yet another pretender to the throne? Our answer to that is "neither"

- Relational databases are a powerful tool that we expect to be using for many more decades, but we do see a profound change in that relational databases won't be the only databases in use

- Our view is that we are entering a world of Polyglot Persistence where enterprises, and even individual applications, use multiple technologies for data management

* From: Martin Fowler. NoSQL Distilled. Preface. (http://martinfowler.com/books/nosql.html)

Multiple technologies for data management

As an exercise, let us ask google which is the database engine used by Facebook. We get the following tools[1]:

- MySQL as core database engine (in fact a customized version of MySQL, highly optimized and distributed) [2]
- Cassandra (an Apache open source fault tolerant distributed NoSQL DBMS, originally developed at Facebook itself) as database for the Inobx mail search
- Memcached, a memory caching system to speed up dynamic database driven websites
- HayStack, for storage and management of photos
- Hive, an open source, peta-byte scale data warehousing framework based on Hadoop, for analytics, and also Presto, an exabyte scale datawarehouse[3]

[1] https://www.techworm.net/2013/05/what-database-actually-facebook-uses.html

[2] http://www.datacenterknowledge.com/data-center-faqs/facebook-data-center-faq-page-2

[3] http://prestodb.io/

Data Warehouse

- A data warehouse is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more disparate sources
- According to Inmon[*], a data warehouse is:
  - **Subject-oriented**: The data in the data warehouse is organized so that all the data elements relating to the same real-world event or object are linked together
  - **Non-volatile**: Data in the data warehouse are never over-written or deleted once committed, the data are static, read-only, and retained for future reporting
  - **Integrated**: The data warehouse contains data from most or all of an organization's operational systems and these data are made consistent
  - **Time-variant**: For an operational system, the stored data contains the current value. The data warehouse, however, contains the history of data values

[*] Inmon, Bill (1992). Building the Data Warehouse. Wiley

Data Warehouse vs. Big Data

- *Are Data Warehouses (DWs) under the hat of Big Data*?
- The notion of **data warehousing dates back to the end of 80s**, and very many data warehouse and business intelligence solutions have been proposed since then
- BTW, Big Data and DWs have many points in common, at least w.r.t.
  - Volume: data warehouses store large amounts of data,
  - Variety: at least in principle, data warehouses integrate heterogeneous information
  - Veracity: data warehoses usually are equipped with data cleaning solutions, applied in the so-called extract-transformation-load (ETL) phase

Data Warehouse vs. Big Data

- Existing enterprise data warehouses and relational databases excel at processing structured data, and can store massive amounts of data, though at cost

- However, this requirement for structure imposes an inertia that makes data warehouses unsuited for agile exploration of massive heterogenous data

- The amount of effort required to warehouse data often means that valuable data sources in organizations are never mined

- Therefore, new computing models and frameworks are needed to make new DW solutions compliant with the Big Data ecosystem.

Map Reduce

- MapReduce is a programming framework for parallelizing computation

- Originally defined at Google

- Next, there have been various implementations

- A well-known open source distribution is Apache Hadoop

Map Reduce

A MapReduce program is constituted by two components

- **Map()** procedure (*the mapper*) that performs filtering and sorting (it decomposes the problem into parallelizable subproblems)
- **Reduce()** procedure (*the reducer*) devoted to solve subproblems

The MapReduce Framework manages distributed servers, which execute the various subtasks in parallel, and controls communication and data transfers between the various servers, as well as guarantees fault tolerance and disaster recovery.