WWW.IARIA.ORG

PANEL
ALLDATA/IMMM/KESA/MMEDIA

Computational Models for Big Data Processing

Moderator: Gary Weckman, Ohio University, USA

Getting lost

- Open Data
- Linked Data
- Big Data
- Small Data
- Public Data
- Private Data
- Data sets
- Data accuracy
- 5V+ (volume, velocity, veridicity, ……

- Computational Models
  - Data Management
  - Semantic Aspects
  - Legal Aspects

Computational Models for Big Data Processing

- **1 - In every sector, there is a growing need for companies to extract knowledge from Big Data in a fast and reliable way;**
- **2 – To do so, alternatively to stochastic analysis methods, companies could use rule-based NLP tools and environments, such as for example NooJ (www.nooj-association.org, http://en.wikipedia.org/wiki/NooJ). It would be interesting to practically see the advantages and disadvantages of such an adoption;**
- **3 - Specific professionals should be formed to perform Knowledge Extraction, starting from undergraduate courses.**

[Mario Monteleone]

Today's Panelists

Moderator

- **Gary Weckman, Ohio University, USA**

Panelists

- **Iryna Lishchuk, Institut für Rechtsinformatik/Leibniz Universität Hannover, Germany**
- **Venkat Gudivada, East Carolina University, USA**
- **Pedro Martins, Universidade de Coimbra, Portugal**
- **Gary Weckman, Ohio University, USA**
- **Jolon Faichney, Griffith University, Australia**
- **Jedrzej Rybicki, Forschungszentrum Juelich GmbH, Germany**
- **Mario Monteleone, Università degli Studi di Salerno, Italy** *[input only]*

BIG DATA are here… their challenges, too

- Iryna: The legal side of big data, e.g., what legal issues may arise by processing of big data,e.g. privacy or copyright concerns
- Venkat: Though the parallel computing models such as MPI, OpenMP, CUDA, OpenACC, and OpenCL existed for long, they failed in effecting widespread adaptation. In contrast, Hadoop parallel computing framework quickly became a mainstream and widely used parallel computing model. Of late, Apache Spark is emerging as a replacement for Hadoop? Is Spark the right Hadoop replacement? What happened to Apache Storm? What is the future of MPI, OpenMP, CUDA, OpenACC, and OpenCL? shorter: Is Spark the right Hadoop replacement? What happened to Apache Storm? What is the future of MPI, OpenMP, CUDA, OpenACC, and OpenCL?
- Pedro: Some relevant discussion topics regarding parallelization tools and their advantages and disadvantages.
- Gary: Big data analytics, issues with finding patterns in Big Data and issues in modeling
- Jolon: Open Data allows organisations to share data. However, opening Data can provide a number of challenges. We propose Open Schemas as a step before Open Data to allow organisations to share schemas. Opening Schemas will motivate organisations to provide cleaner data models and better schema documentation
- Jedrzej: I would like to pledge for the (sometimes overlooked) interplay between the data management and resulting/possible computational models. This is an open question if it is actually possible to define the computation model first and then find out a possible data management strategy to fit it. Or is the order of things in case of Big Data pre-defined and one has to use (perhaps less sophisticated) computation model since nothing else works with distributed data.

Q & A

Qs & As

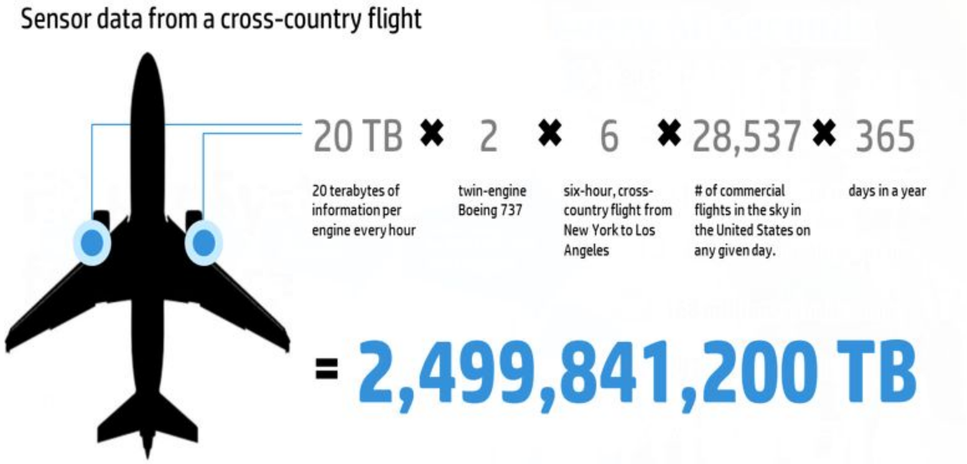
WWW.IARIA.ORG

Performance Challenges over Big and Small data

Pedro Martins

AllData 2016

What is BigData ?



What is BigData ?

Does not mean large volume of data, but how hard its analysis is, to reach a result in the desired time.

Big Data Landscape

Vertical Apps
Predictive Policing, bloomreach, MYRRIX

Log Data Apps
splunk, loggly, sumologic

Ad/Media Apps
rocketfuel, collective[i], bluefin, Recorded Future, Media Science, LuckySort, TURN, DataXu

Business Intelligence
ORACLE | Hyperion, SAP BusinessObjects, RJMetrics, Microsoft Business Intelligence, IBM, COGNOS, birst, MicroStrategy, Autonomy, QlikView, bime, DOMO, Chart.io, GoodData

Analytics and Visualization
tableau, Palantir, OPERA, metaLayer, METAMARKETS, dataspora, centrifuge, TERADATA ASTER, sas, TIBCO, KARMASPHERE, panopticon, pentaho, Datameer, platfora, ClearStory, CIRRO, alteryx, visual.ly, AYATA

Data As A Service
factual., kaggle, knoema, GNIP, DATASIFT, Windows Azure Marketplace, INRIX, LexisNexis, LOQATE

Analytics Infrastructure
Hortonworks, VERTICA, MAPR, cloudera, INFOBRIGHT, PARACCEL, EMC², GREENPLUM, NETEZZA, kognitio, DATASTAX, EXASOL, calpont

Operational Infrastructure
COUCHBASE, 10gen, TERADATA, HADAPT, TERRACOTTA, VoltDB, MarkLogic, INFORMATICA

Infrastructure As A Service
amazon web services, Windows Azure, infochimps, Google BigQuery

Structured Databases
ORACLE, MySQL, SQL Server, PostgreSQL, IBM DB2, SYBASE, memsql

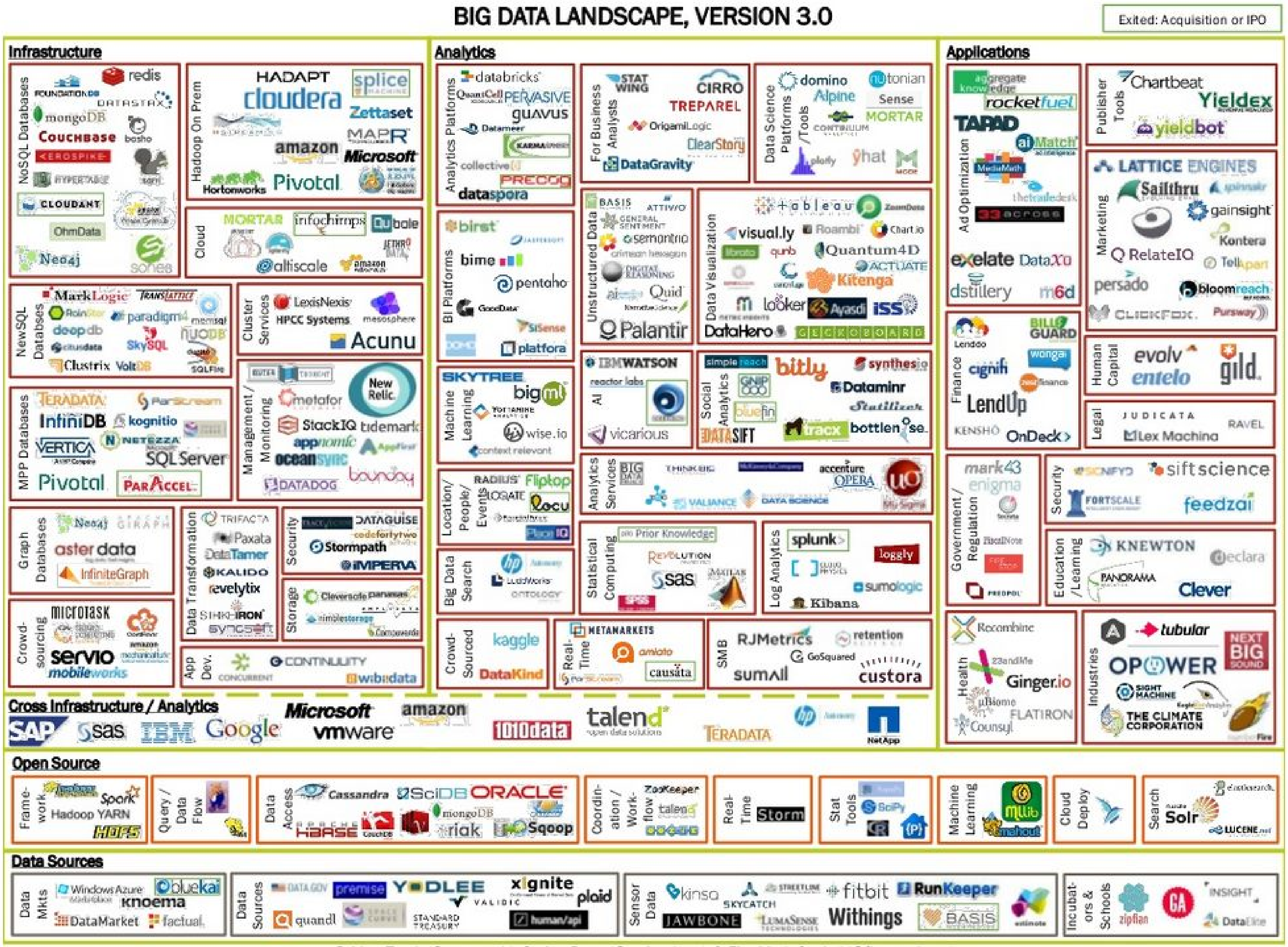
Technologies
hadoop, hadoop MapReduce, mahout, APACHE HBASE, Cassandra

 dave@vcdave.com blogs.forbes.com/davefeinleib

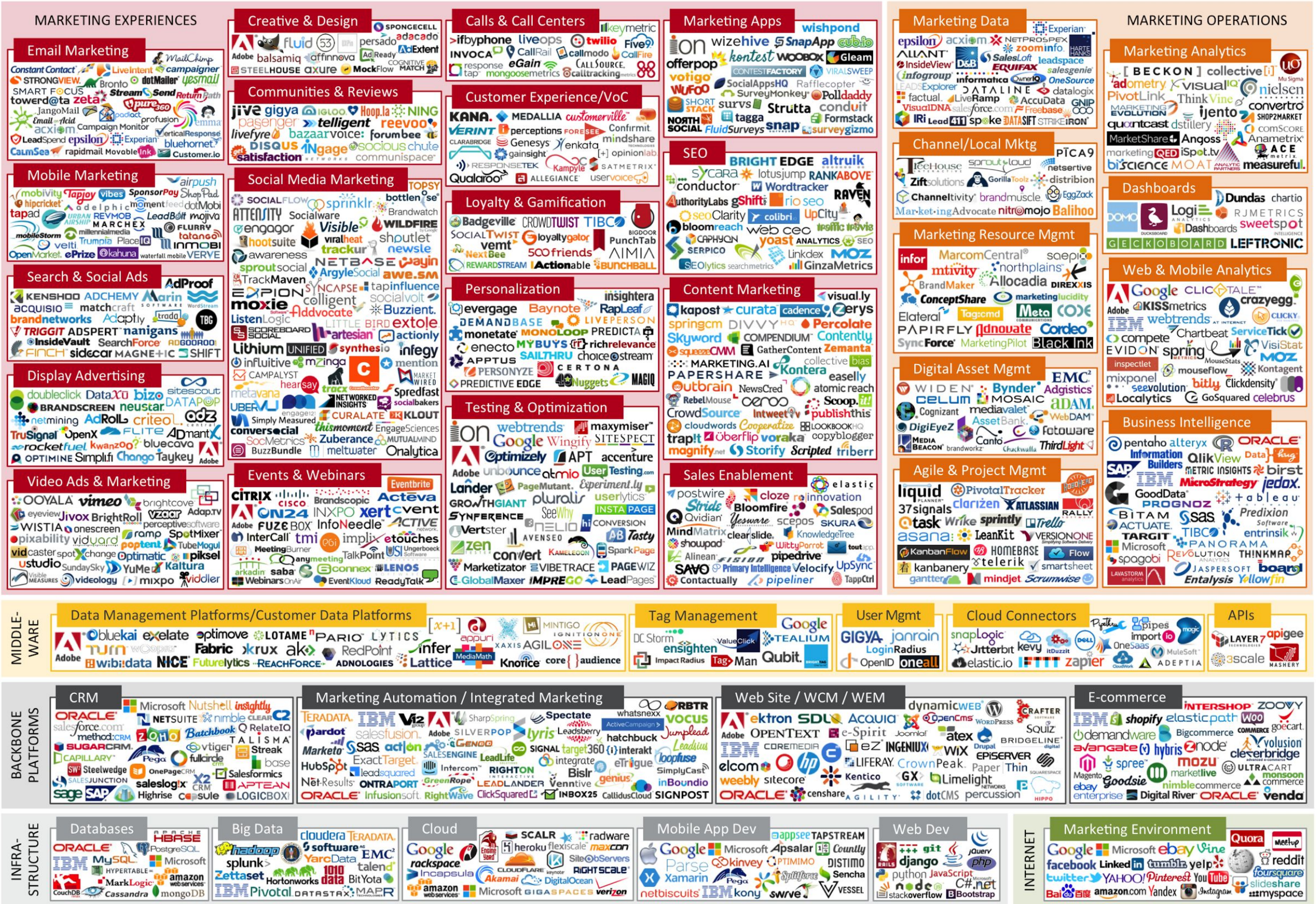
Big Data Landscape (Version 2.0)

© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

BIG DATA LANDSCAPE, VERSION 3.0



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)



by Scott Brinker @chiefmartec http://chiefmartec.com

How to achieve more performance ?

E (extraction)

- from data sources
- cell phone towers
- supermarket network

T (transformation)

- filter, data cleansing, validation, sort…

L (load)

- into any DW for later analysis

Q (query)

How to achieve more performance ?

Application solution?

Oracle, MySQL, Vertica, TeraData, MapReduce Architecture, etc

Is MapReduce the “holy grail”?

How to achieve more performance ?

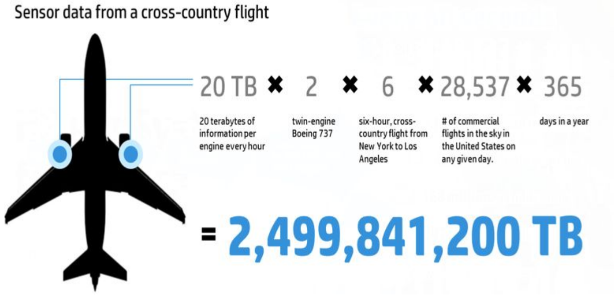
Application solution?

(e.g. Oracle, MySQL, Vertica, TeraData, MapReduce Architecture)

OR

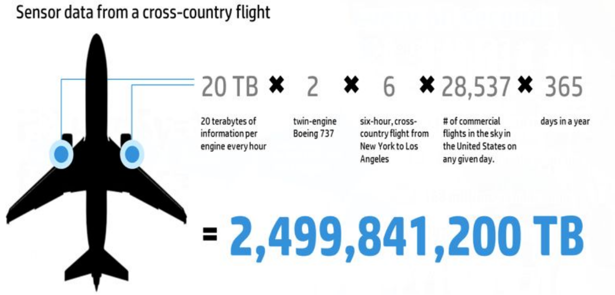
Divide to conquer?

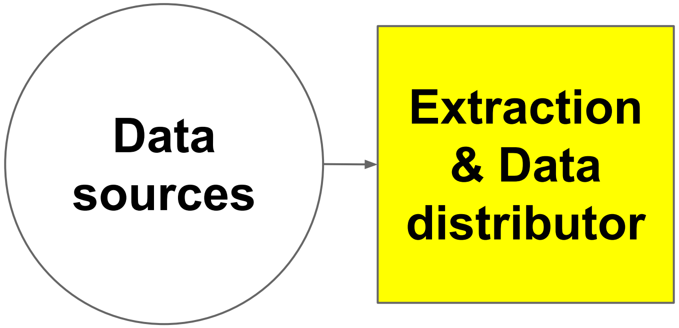
And this?



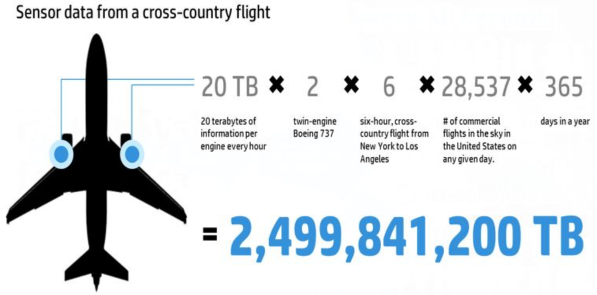
Data sources

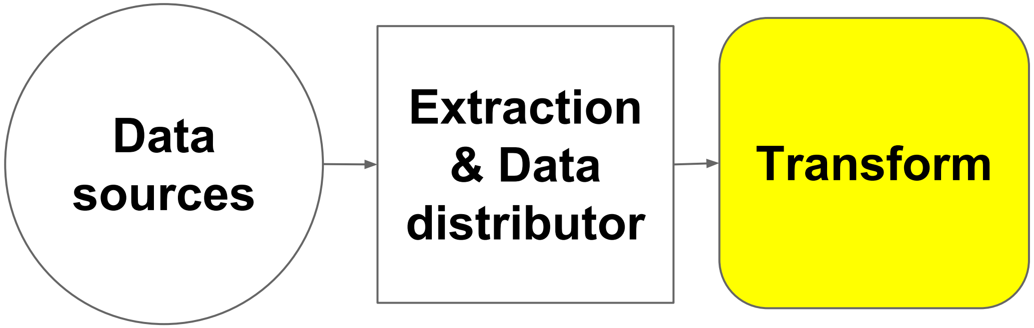
And this?


Sensor data from a cross-country flight
20 TB ✖ 2 ✖ 6 ✖ 28,537 ✖ 365
20 terabytes of information per engine every hour
twin-engine Boeing 737
six-hour, cross-country flight from New York to Los Angeles
# of commercial flights in the sky in the United States on any given day.
days in a year
= 2,499,841,200 TB

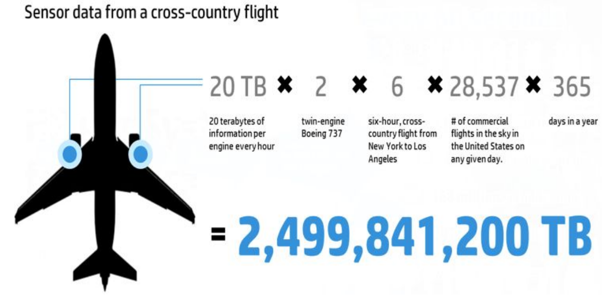

Data sources
Extraction & Data distributor

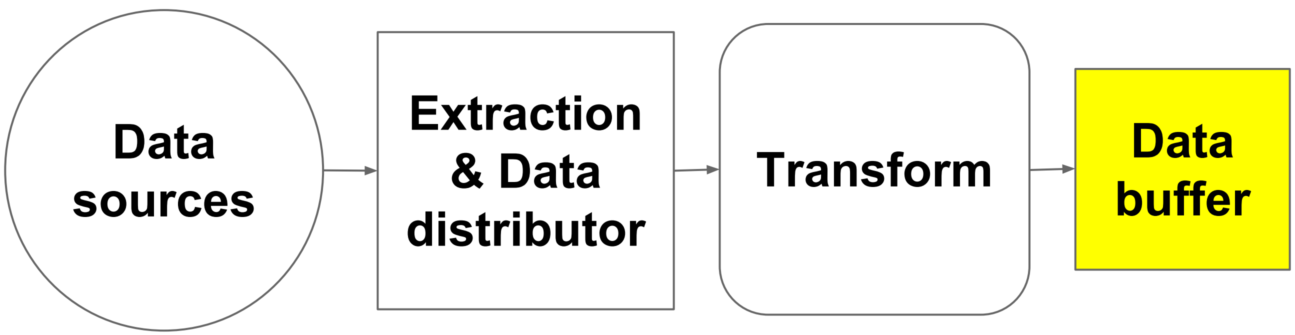
And this?


Sensor data from a cross-country flight
20 TB ✖ 2 ✖ 6 ✖ 28,537 ✖ 365
20 terabytes of information per engine every hour
twin-engine Boeing 737
six-hour, cross-country flight from New York to Los Angeles
# of commercial flights in the sky in the United States on any given day.
days in a year
= 2,499,841,200 TB

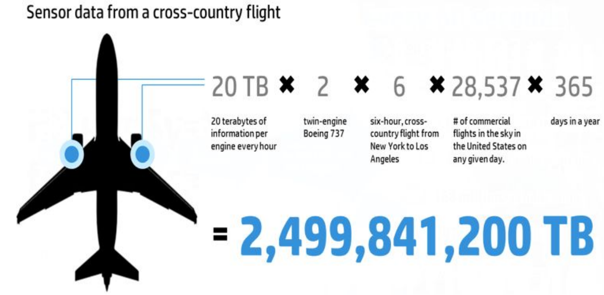

Data sources
Extraction & Data distributor
Transform

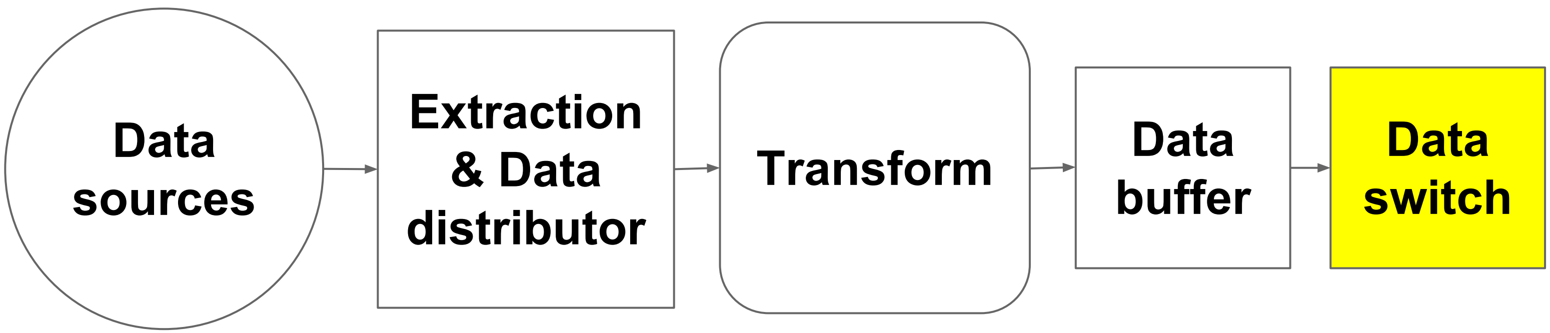
And this?


Sensor data from a cross-country flight
20 TB ✖ 2 ✖ 6 ✖ 28,537 ✖ 365
20 terabytes of information per engine every hour
twin-engine Boeing 737
six-hour, cross-country flight from New York to Los Angeles
# of commercial flights in the sky in the United States on any given day.
days in a year
= 2,499,841,200 TB

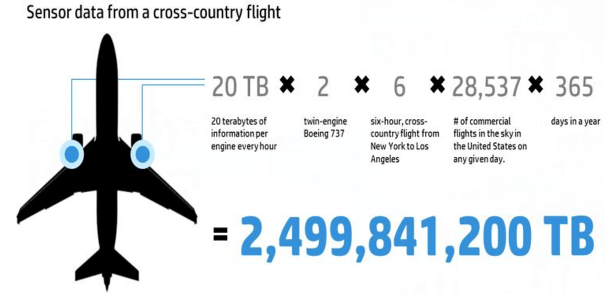

Data sources
Extraction & Data distributor
Transform
Data buffer

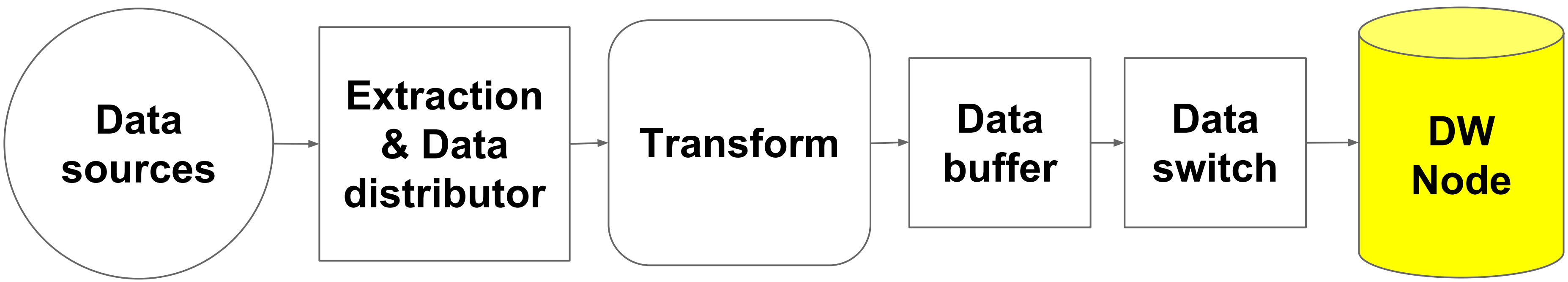
And this?


Sensor data from a cross-country flight
20 TB ✖ 2 ✖ 6 ✖ 28,537 ✖ 365
20 terabytes of information per engine every hour
twin-engine Boeing 737
six-hour, cross-country flight from New York to Los Angeles
# of commercial flights in the sky in the United States on any given day.
days in a year
= 2,499,841,200 TB


Data sources
Extraction & Data distributor
Transform
Data buffer
Data switch

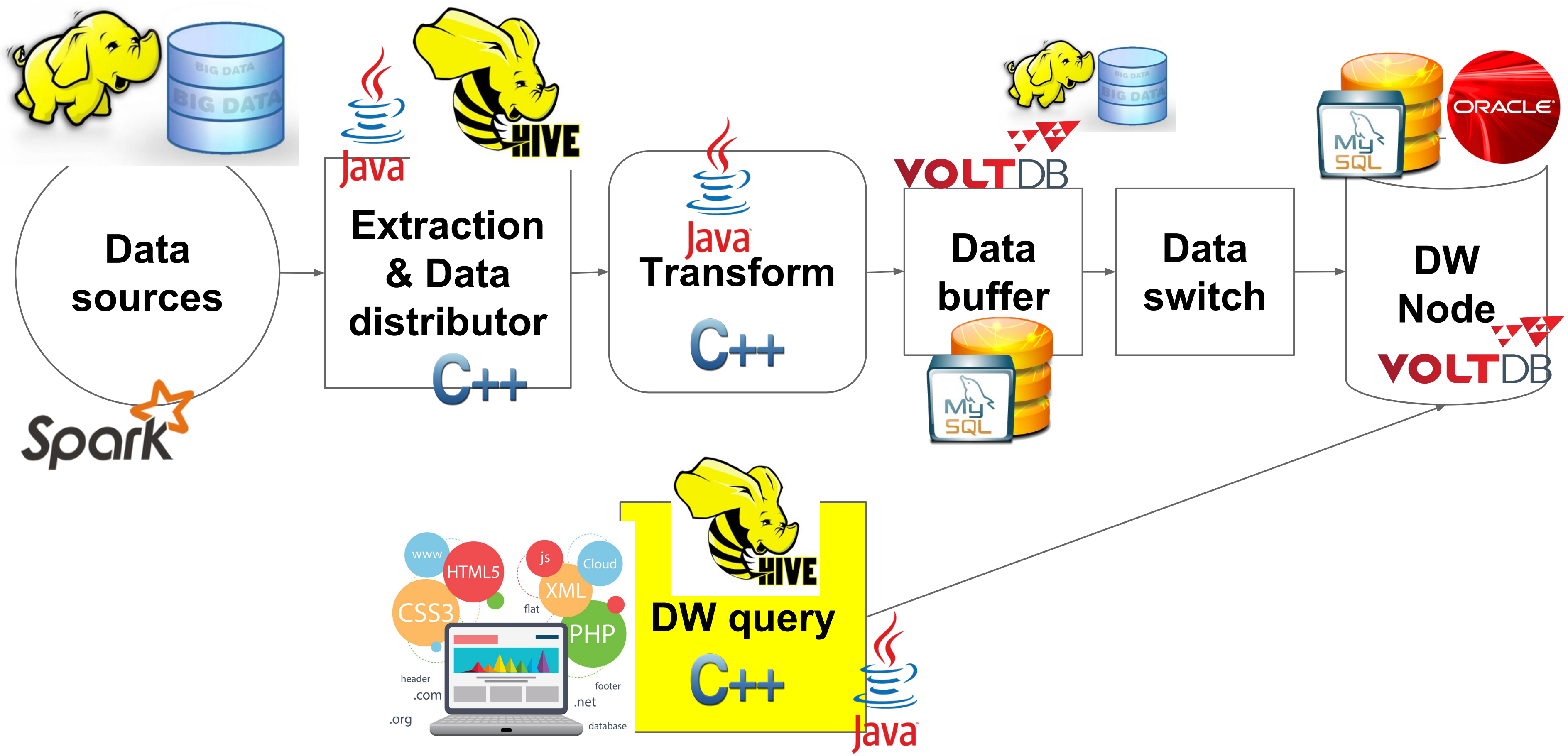
And this?


Sensor data from a cross-country flight
20 TB ✖ 2 ✖ 6 ✖ 28,537 ✖ 365
20 terabytes of information per engine every hour
twin-engine Boeing 737
six-hour, cross-country flight from New York to Los Angeles
# of commercial flights in the sky in the United States on any given day.
days in a year
= 2,499,841,200 TB


Data sources
Extraction & Data distributor
Transform
Data buffer
Data switch
DW Node

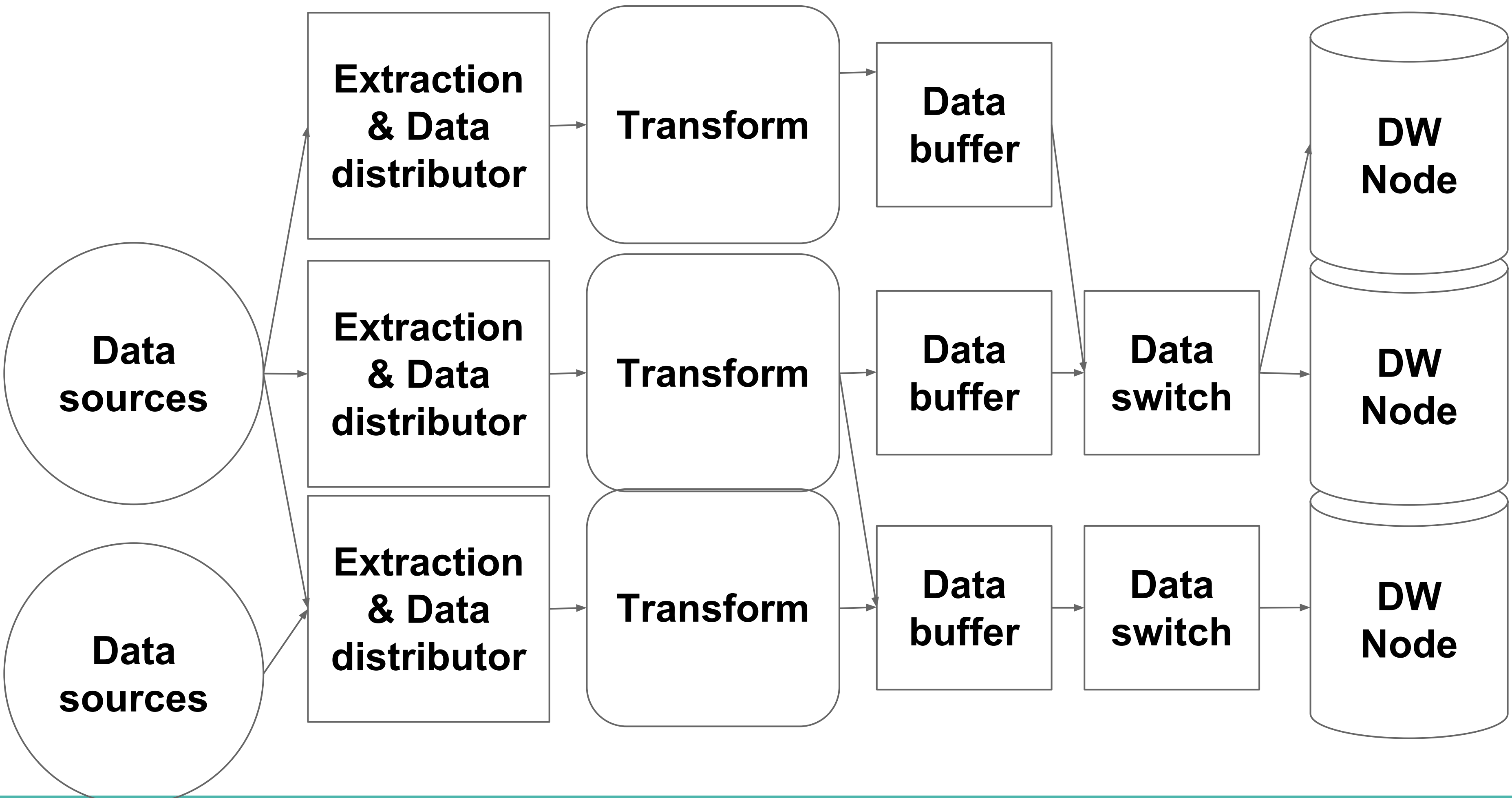
And this? - We call it AScale

(Automatic scaler, developed at University of Coimbra - Pedro Martins)



And this?

- A: scale-out (horizontally);



“The simplest explanation is usually the correct one”

William Ockham

Questions & Comments ?

Is MapReduce the “holy grail”?

Why not Map-Reduce architectures? (VLDB 2015)

Is MapReduce the “holy grail”?

Why not Map-Reduce architectures? (VLDB 2015)

- Little control of data flow
- Fault tolerance guarantees not always necessary
- Does not interface with existing analysis software

Is MapReduce the "holy grail"?

Why not Map-Reduce architectures? (VLDB 2015)

- Little control over storage (Append only file system)

- Little control over resource management

- Often used for "data dump" (irregular and unreliable)

Computational Models for Big Data Processing

Venkat N Gudivada
East Carolina University
Greenville, North Carolina
USA

High Performance Computing Models

- Availability of multiprocessor and multi-core chips and GPU accelerators at commodity prices
- Personal Supercomputers
- Cloud-hosted Cluster Computers for the masses
- Parallel computing for the elite (http://top500.org/)

High Performance Computing Models - Shared-memory Paradigm

- OpenMP (compiler directives)
- OpenACC (targets acceleration devices)
- CUDA (architecture and a programming model)
- OpenCL (heterogeneous platforms)
- Haskell Concurrent Programming Model

High Performance Computing Models - Distributed Memory Paradigm

- MPI (message passing)
- MapReduce (Hadoop Ecosystem – HDFS, MapReduce, Pig, Pig Latin, Hive, Cascading, Scalding, Cascalog, Storm, and Spark; Google Cloud DataFlow)
- Erlang Message Based Concurrent Programming Model

Future Trend

- Personal supercomputers (packaged and pre-installed applications)
- Parallel computing for the masses
- Exascale computing for the elite

Open Schemas
A step towards Open Data

Dr Jolon Faichney

School of Information and Communication Technology

Griffith University, Australia

j.faichney@griffith.edu.au

Dr Jolon Faichney

- Lecturer, Griffith University, Australia
- Founder, App Factory Student Enterprise
- Member of ODIQ Certificate Localisation Working Group

What is Open Data?

- Governments in democratic countries serve the public:
  - Therefore public should have access to data

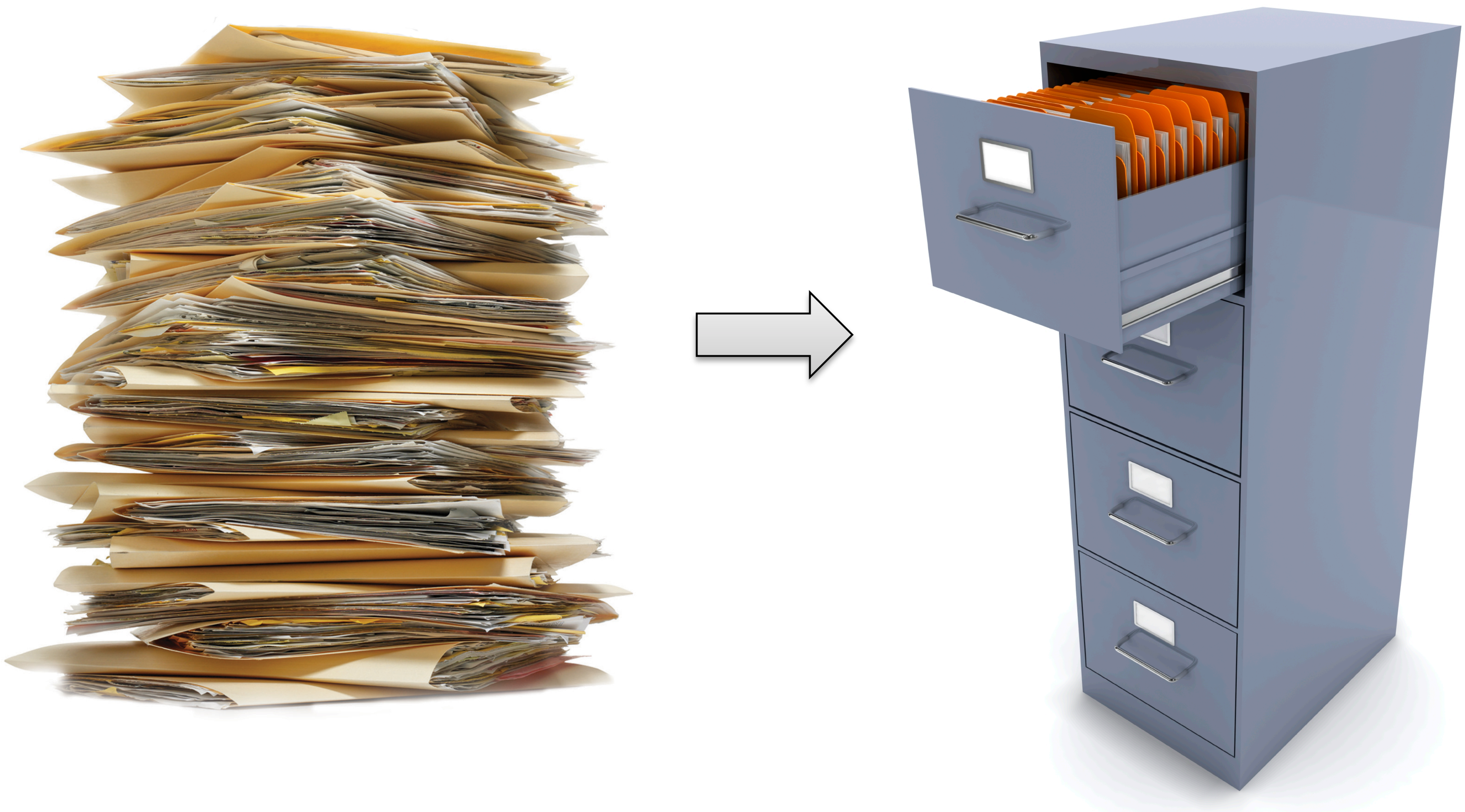
- Transparency
- Accountability
- Productivity

Open Data Institute

- Founded by Tim Berners-Lee and Nigel Shadbolt in 2012
- UK-based, worldwide
- Promote the concept of:

"Open by Default"

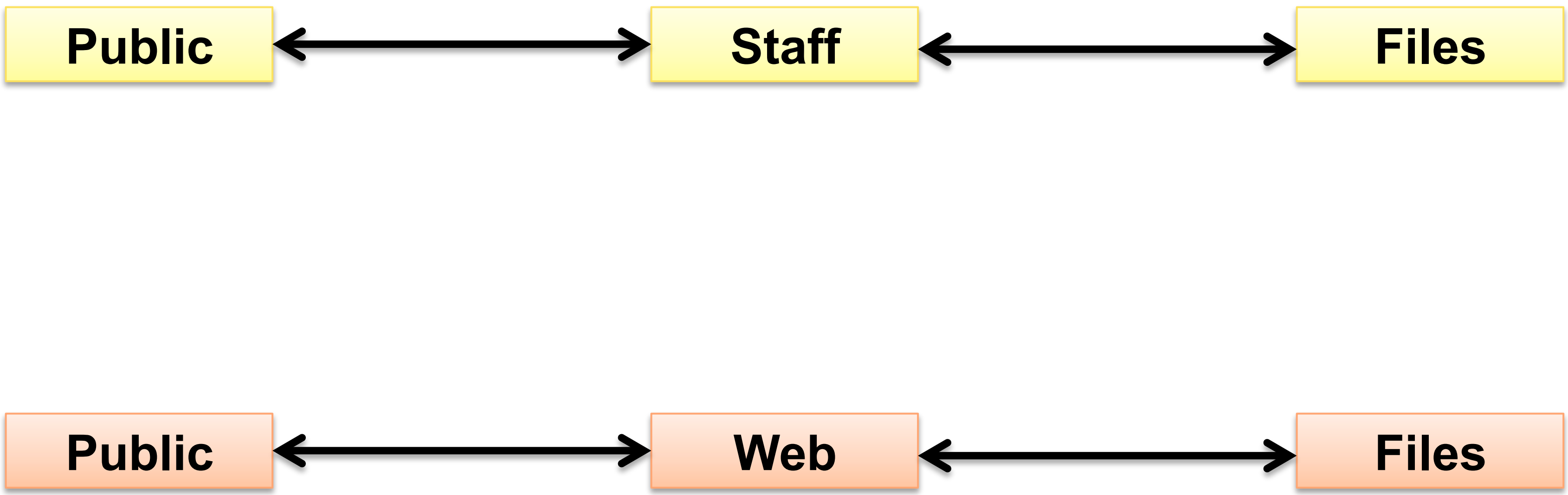
Traditional Government Data



Pre-Web Government Data

Previously, access to Government data required person-to-person interaction

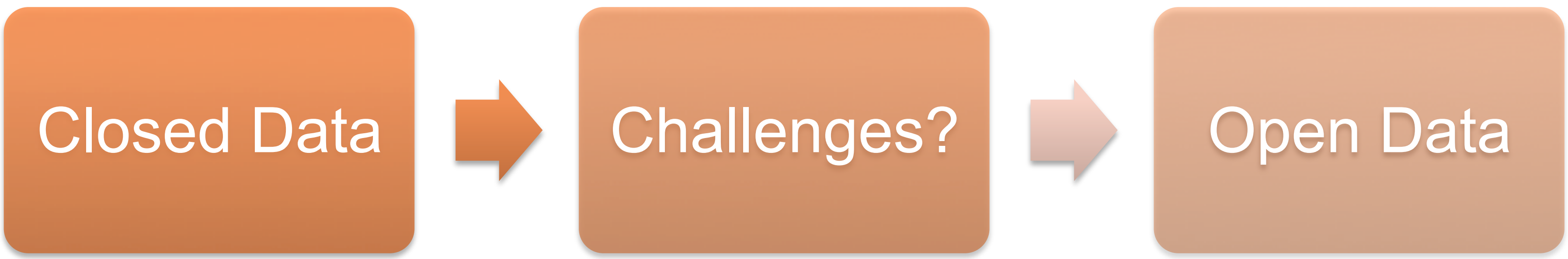
Post-Web Government Data



Open Data Challenges

- Cultural/policy challenges
  - Policy changes required
  - Resistance to data sharing within organisation
  - Exposure of poor/inaccurate quality data
- Technical challenges
  - Writing new software/APIs
  - Realtime data
- Privacy issues (de-identifying data)
- Resources/costs required

Open Data Challenges



Open Schema Approach

Closed Data → Open Schemas → Open Data

Open Schema Approach

- Encourage departments to Open Schemas as a first step towards Open Data
- Advantages:
  - Opening Data has policy challenges such as privacy
  - Can also be applied to data that won't be opened
  - Improves interoperation within organisation
  - Improves interoperation external to organisation
  - Work towards consistent schemas
  - Provides the public with an indication of the availability of data leading to demand-driven open data (DDOD)

Open Schema Approach

- Challenges:
  - Schemas can be messy
    - Not intended for public viewing
    - Evolved over many years and many people
  - Schemas may be poorly documented
  - Exposing schemas may be perceived as a security risk

- But
  - May motivate organisations to maintain well designed and documented schemas

Open Banking Example

- ODI's Open Banking Standard:
  - http://theodi.org/open-banking-standard

Data Documentation Initiative

- DDI:
  - The Data Documentation Initiative (DDI) is an international standard for describing statistical and social science data.
  - Documenting data with DDI facilitates interpretation and understanding -- both by humans and computers.
  - The freely available international DDI standard describes data that result from observational methods in the social, behavioral, economic, and health sciences.
  - Use DDI to **D**ocument, **D**iscover, and **I**nteroperate

Schema.org

- Google, Microsoft, Yahoo, Yandex
- Entities, Relationships, Actions

MedicalCondition

Thing > MedicalEntity > MedicalCondition

Any condition of the human body that affects the normal functioning of a person, whether physically or mentally. Includes diseases, injuries, disabilities, disorders, syndromes, etc.

Usage: Between 100 and 1000 domains

[more...]

Property	Expected Type	Description
Properties from MedicalCondition		
`associatedAnatomy`	AnatomicalSystem or SuperficialAnatomy or AnatomicalStructure	The anatomy of the underlying organ system or structures associated with this entity.

Topics for Discussion

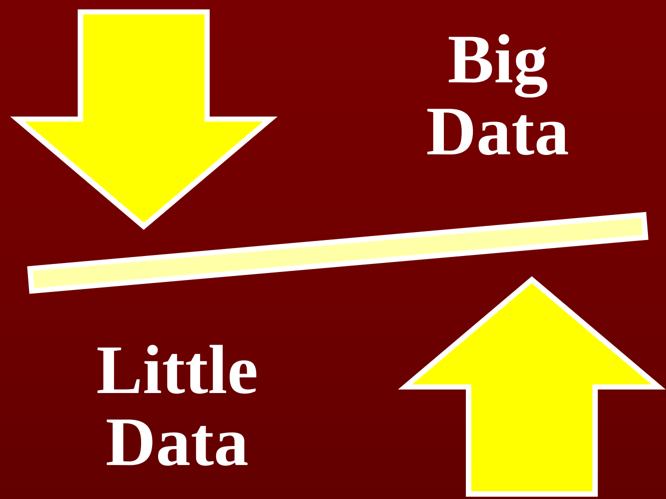
- Is "Open by Default" too ambitious/impractical?
- What challenges do organisations face in Opening Schemas?
- What benefits do Open Schemas provide?
- How can we work towards consistent schemas?

Topic: Computational Models for Big Data Processing

Gary R. Weckman
Ohio University

Big Data

- **Means in which data can be collected is more readily available than ever**
- **Big Data more relevant than ever because it can be used to improve decisions and insights within the domains it is used**
- **The term Big Data can be loosely defined as data that is too large for traditional analysis methods and techniques.**



Big: Random reduction
Little: Synthetic (SMOTE)
Imbalance Data
Missing Data

Big data analytics

issues with finding patterns in Big Data and issues in modeling

Big Data Issues

- **Due to the volume of Big Data**
  - **Traditional methods of analysis can function poorly**
  - **finding patterns that do not exist (Ratner, 2003)**

- **As datasets grow in size (leaning towards Big Data)**
- **spurious structures tend to also be discovered in**
- **datasets and interpreted as meaningful when they**
- **have no true meaning (Ratner, 2003).**

Table 5: Pattern Recognition Methods Applicable to Big Data

Topic	Reference and Examples
Anomaly Detection	A form of unsupervised classification where statistically rare events are of interest. Useful for removing background from data.
	Methods: PCA reconstruction error/residual analysis (Jackson & Mudholkar, 1979) (Jablonski, Bihl, & Bauer, 2015), Mahalanobis distance (De Maesschalck, Jouan-Rimbaud, & Massart, 2000) , RX detector (Reed & Yu, 1990)
Artificial Neural Networks	A variety of nonlinear classifiers that employ gradient descents and a variety of algorithms to classify and predict data
	Methods: Feedforward, recurrent, and self-organizing map methods (Young, Bihl, & Weckman, 2014) (Ward, Bihl, & Bauer, 2014)
Classification	Applying methods to create a model that accurately represents the data with respect to known classes.
	Methods: ANNs (Duda, Hart, & Stork, 2001), Classification Trees (Duda, Hart, & Stork, 2001), Discriminant Analysis (Dillon & Goldstein, 1984), Logistic Regression (Hosmer & Lemeshow, 2000),
Class Imbalance	Dealing with imbalances in classes, which can create bias in analytical methods
	Methods: Over-sampling, boosting, bootstrapping (Zhang J. , 2004), artificial sampling (Bui, 2004)
Clustering	Clustering, or unsupervised classification, refers to methods that search for underlying patterns in data
	Methods: Hierarchical (distance and linkage based) (Gordon, 1987) (Johnson, 1967) (Milligan & Cooper, 1987) (Milligan & Cooper, 1985), k-means (Jain, 2010), affinity propagation (Frey & Dueck, 2007), density based (Ester, Kriegel, Sander, & Xu, 1996), and other methods (Jain, 2010)
Crowd Sourcing	Employing multiple users to analyze data, contribute to a solution, or leverage their computer power
	Methods/Examples: Crowd sourced games to find patterns (Mavandadi, et al., 2012) (Martin, et al., 2013), multiple opinions for clinical data (Celi, Mark, Stone, & Montgomery, 2013), distributed projects (Schreiner, 2001)
Dimensionality Reduction Analysis	Dimensionality reduction through transforming data into a new space (feature extraction) or selecting subsets of original data features (feature selection).
	Methods: Principal Component Analysis (PCA) (Dillon & Goldstein, 1984), Factor Analysis (FA) (Dillon & Goldstein, 1984), Independent Component Analysis (Jain, Duin, & Mao, 2000), Kernel methods (Jain, Duin, & Mao, 2000), Stepwise, forward, and backward selection methods (Jain, Duin, & Mao, 2000), ANN signal to noise ratio feature screening (Bauer, Alsing, & Greene, 2000), input reduction (Young W. , Weckman, Thompson, & Brown, 2008), Wilk's Lambda (Dillon & Goldstein, 1984) (Eisenbeis, 1977), F-test (Bihl, Temple, Bauer, & Ramsey, 2015), and other methods (Jain, Duin, & Mao, 2000).methods (Jain, Duin, & Mao, 2000)
Imputation	Filling in missing observation through various methods, missing data can appear randomly through a dataset, in rows (e.g. survey responses), or in columns (possibly from a bad sensor or attribute)

MyHealthAvatar

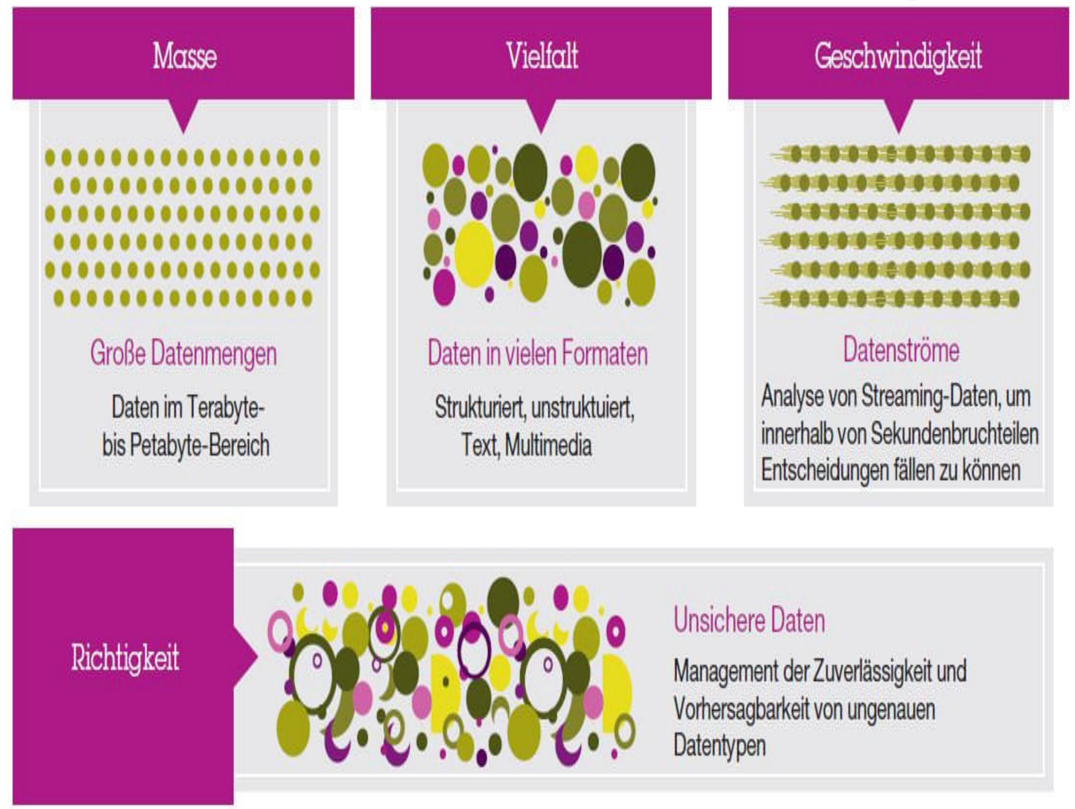
Computational Models for Big Data Processing
Legal Perspective

Iryna Lishchuk, LL.M.
Institut für Rechtsinformatik
Leibniz Universität Hannover
Hannover, Germany

SEVENTH FRAMEWORK PROGRAMME

Big Data - Concept

- Volume
- Velocity
- Variety
- Veracity



http://www-935.ibm.com/services/de/gbs/thoughtleadership/GBE03519-DEDE-00.pdf

Personal Data - Concept

- Article 2, Data Protection Directive 95/46/EC

“*'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject');*

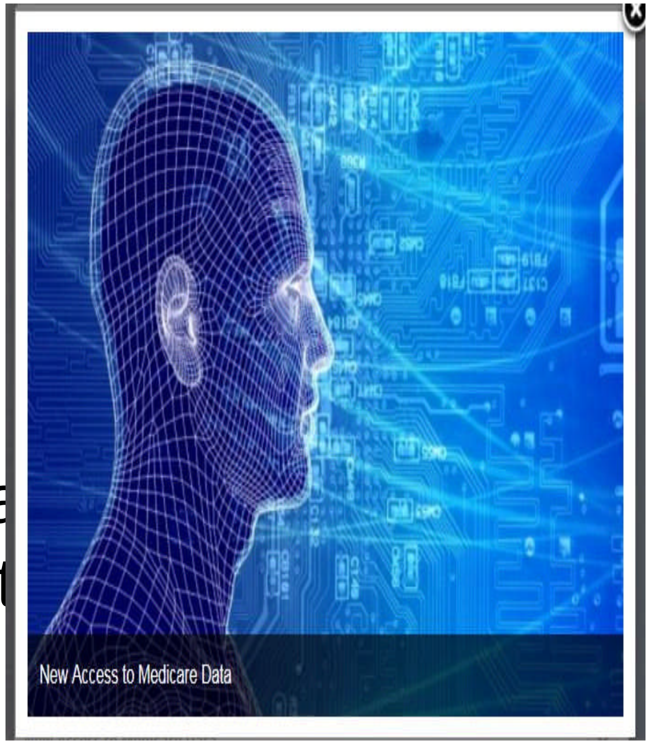
an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity“

Data Linking v. Data Protectior

→ Linking isolated data sets can make a person behind identifiable

→ Pure data procession can turn into procession of personal data

→ Procession of personal da
subject to legal requirement



SEVENTH FRAMEWORK PROGRAMME

Data Linking v. Data Protection

Article 2 Data Protection Directive 95/46/EC

Procession, collection and use of personal data subject to consent of the data subject, which is legitimate, when:

- Free
- Express
- Purpose related
- Informed

SEVENTH FRAMEWORK PROGRAMME

Data Linking v. Data Protection

MyHealthAvatar

→Linking of already available and for specific purpose collected data may constitute change of purpose

→Change of purpose requires for legitimation consent of the data subject

→ Data linking for a specific purpose may be justified if qualifies as compatible uses of data

SEVENTH FRAMEWORK PROGRAMME

Computational Models for Big Data Processing

Jędrzej Rybicki

23rd February 2016

Computer Science Archeology

Computer Science Archeology

```
shasum * | awk {'print $1'} |sort| uniq -c |grep -v " 1 "
```

Computer Science Archeology

```
shasum * | awk {'print $1'} |sort| uniq -c |grep -v " 1 "
```

- Unix pipelines

Computer Science Archeology

```
shasum * | awk {'print $1'} |sort| uniq -c |grep -v " 1 "
```

- Unix pipelines
- Ken Thomson et al. in PDP-11 (1972)

Computer Science Archeology

```
shasum * | awk {'print $1'} |sort| uniq -c |grep -v " 1 "
```

- Unix pipelines
- Ken Thomson et al. in PDP-11 (1972)
- limits on how big a single problem could be

Computer Science Archeology

```
shasum * | awk {'print $1'} |sort| uniq -c |grep -v " 1 "
```

- Unix pipelines
- Ken Thomson et al. in PDP-11 (1972)
- limits on how big a single problem could be
- no disks (or very small)

Data Management Models for Big Data Processing

Jędrzej Rybicki

23rd February 2016

Don’t store everything

Don't store everything

"[...] in theoretical work it is often possible to cherry pick assumptions to produce a given result."

— "Chameleons: The Misuse of Theoretical Models in Finance and Economics," by Paul Pfleiderer

Don't store everything

"[...] in theoretical work it is often possible to cherry pick assumptions to produce a given result."

— "Chameleons: The Misuse of Theoretical Models in Finance and Economics," by Paul Pfleiderer

- no problem was solved just by putting more data on it

Don’t store everything

“[...] in theoretical work it is often possible to cherry pick assumptions to produce a given result.”

— “Chameleons: The Misuse of Theoretical Models in Finance and Economics,” by Paul Pfleiderer

- no problem was solved just by putting more data on it
- monolith store vs. micro stores

Don’t store everything

“[...] in theoretical work it is often possible to cherry pick assumptions to produce a given result.”

— “Chameleons: The Misuse of Theoretical Models in Finance and Economics,” by Paul Pfleiderer

- no problem was solved just by putting more data on it
- monolith store vs. micro stores
- point of highest ignorance

Avoid web scale envy

- perhaps the big data companies/scientist are where they are because they started small?

Avoid web scale envy

- perhaps the big data companies/scientist are where they are because they started small?
- do you think that MapRed is good?

Avoid web scale envy

- perhaps the big data companies/scientist are where they are because they started small?
- do you think that MapRed is good?
- *no!* nothing else is possible with the selected data management solution

Avoid web scale envy

- perhaps the big data companies/scientist are where they are because they started small?
- do you think that MapRed is good?
- *no!* nothing else is possible with the selected data management solution
- ... but perhaps you can use something else

Avoid web scale envy

- perhaps the big data companies/scientist are where they are because they started small?
- do you think that MapRed is good?
- *no!* nothing else is possible with the selected data management solution
- ... but perhaps you can use something else
- graph algorithms: example of how data management can influence compuational models

$\Rightarrow$ vertex-oriented vs. pattern machting

Beware of difference between academia and private sector

- data sharing: crazy idea!

Beware of difference between academia and private sector

- data sharing: crazy idea!

$\Rightarrow$ sharing infrastructure?

Beware of difference between academia and private sector

- data sharing: crazy idea!

$\Rightarrow$ sharing infrastructure?

- data management for sharing might be different?

Beware of difference between academia and private sector

- data sharing: crazy idea!

$\Rightarrow$ sharing infrastructure?

- data management for sharing might be different?
- sharing with whom?

$\Rightarrow$ limit sophistication

Data Management Models influence Data Processing Models

Data Management Models for Big Data Processing:

- Don’t store everything
- Avoid web scale envy
- Academia vs. private sector
- Don’t be too sophisticated