Cloud & Big Data

Opportunities and Risks for Developing Countries

Patrick Valduriez

Zenith

informatics mathematics

UNIVERSITE MONTPELLIER

cnrs advancing the frontiers

LIRMM

Cloud & Big Data: the hype!

Cloud & Big Data: the hype!

Cloud and Big Data drive AfricaCom 2014

OPTIONS

SAVE | EMAIL | PRINT | PDF

AfricaCom 2014, being held on 11-13 November 2014 at the CTICC, has two key streams on its agenda - 'Cloud' and 'Big Data'.

Data is king - more so than ever. Knowing how to analyse the data retrieved through a myriad of transactions and transaction types, in order to serve that person/people group specific information that results in a desired response is both an art and a science - science fiction. This is now a business in its own right and involves sophisticated software systems than can take the bytes of data, process them quickly and dispatch with the objective of acquiring targets.

Whether we like it or not, we are all a part of the matrix - whether we have a Facebook account or not as a swiped credit card, any form of payment (other than cash), is recorded. Our personal details, likes and dislikes and even habits are all on file, somewhere. For marketers and brands alike, big data is an opportunity to increase sales. Big data is able to serve huge amounts of selected data to billions of people at exceptional speeds. It influences everyday purchase decisions, but the ability to keep up to date with the fast evolving developments in this field that will classify the winners and the losers of the information age.

Behind the Hype?

- Every one who wants to make big money
  - Intel, IBM, Microsoft, Oracle, Google, Facebook, Amazon, …
- In a big market
  - $18 billion in 2013, $24 billion in 2016
  - Source: International Data Corp. (IDC)

Behind the Hype?

Cloud & Big Data: how?

- *But do they have the same goals ?*

Outline of the Talk

- Cloud
- Big data
- Cloud & big data
- Opportunities and risks

Cloud

Cloud Computing

- The vision
  - On demand, reliable services provided over the Internet (the "cloud") with easy access to virtually infinite computing, storage and networking resources
- Effective!
  - Through simple Web interfaces, users can outsource complex tasks
    - Data storage, system administration, application deployment
  - The complexity of managing the infrastructure gets shifted from the users' organization to the cloud provider
- Capitalizes on previous computing models
  - Web services, utility computing, cluster computing, virtualization, grid computing
- Major players
  - Amazon, Microsoft, Google, IBM, Intel, etc.

Cloud Taxonomy

- Infrastructure-as-a-Service (IaaS)
  - Computing, networking and storage resources, as a service
  - Provides elasticity: ability to scale up (add more resources) or scale down (release resources) as needed
  - E.g. Amazon Web Services
- Software-as-a-Service (SaaS)
  - Application software as a service
  - Generalizes the earlier ASP model with tools to integrate other applications, e.g. developed by the customer (using the cloud platform)
  - Hosted applications: from simple (email, calendar) to complex (CRM, data analysis or social networks)
  - E.g. Safesforce Customer Relationship Management
- Platform-as-a-Service (PaaS)
  - Computing platform with development tools and APIs as a service
  - Enables developers to create and deploy custom applications directly on the cloud infrastructure and integrate them with applications provided as SaaS
  - E.g. Google Apps

Cloud Benefits

- Reduced cost
  - Customer side: the IT infrastructure needs not be owned and managed, and is billed only based on resource consumption
  - Cloud provider side: by sharing costs for multiple customers, reduces its cost of ownership and operation to the minimum
- Ease of access and use
  - Customers can have access to IT services anytime, from anywhere with an Internet connection
- Quality of Service (QoS)
  - The operation of the IT infrastructure by a specialized, experienced provider (including with its own infrastructure) increases QoS
- Elasticity
  - Easy for customers to deal with sudden increases in loads by simply creating more virtual machines (VMs) in data centers

What about Safety?

TECH

More: Cloud Computing Enterprise Amazon

Amazon's Cloud Crash Disaster Permanently Destroyed Many Customers' Data

HENRY BLODGET | APR. 28, 2011, 7:10 AM | 92,357 | 75

Big Data

Big Data: what is it?

- A buzz word!
  - With different meanings depending on your perspective
    - E.g. 10 terabytes is big for an OLTP system, but small for a web search engine
- A definition (Wikipedia)
  - Consists of data sets that grow so *large* that they become awkward to work with using on-hand data management tools
  - *But size is only one dimension of the problem*
- How *big* is big data?
  - Moving target: terabyte (10^{12} bytes), petabyte (10^{15} bytes), exabyte (10^{18}), zetabyte (10^{21})
  - Landmarks in DBMS products
    - 1980: Teradata database machine
    - 2010: Oracle Exadata database machine

Why Big Data Today?

- Overwhelming amounts of data
  - Exponential growth, generated by all kinds of networks, programs and devices
    - E.g. Web 2.0 (social networks, etc.), mobile devices, computer simulations, satellites, radiotelescopes, sensors, etc.
- Increasing storage capacity
  - Storage capacity has doubled every 3 years since 1980 with prices steadily going down
    - 1 Gigabyte (HDD): $400K in 1980, $10K in 1990, $1K in 1995, $10 in 2000, $0.02 in 2015
- Very useful in a digital world!
  - Massive data => high-value information and knowledge
  - Critical for data analysis, decision support, forecasting, intelligence, research, (data-intensive) science, etc.

Big Data Dimensions: the V's

- Volume
  - Refers to massive amounts of data
  - Makes it hard to store, manage, and analyze (big analytics)
- Velocity
  - Continuous data streams are being produced
  - Makes it hard to perform online processing
- Variety
  - Different data formats, different semantics, uncertain data, multiscale data, etc.
  - Makes it hard to integrate and analyze
- Other V's
  - Validity: is the data correct and accurate?
  - Veracity: are the results meaningful?
  - Volatility: how long do you need to store this data?

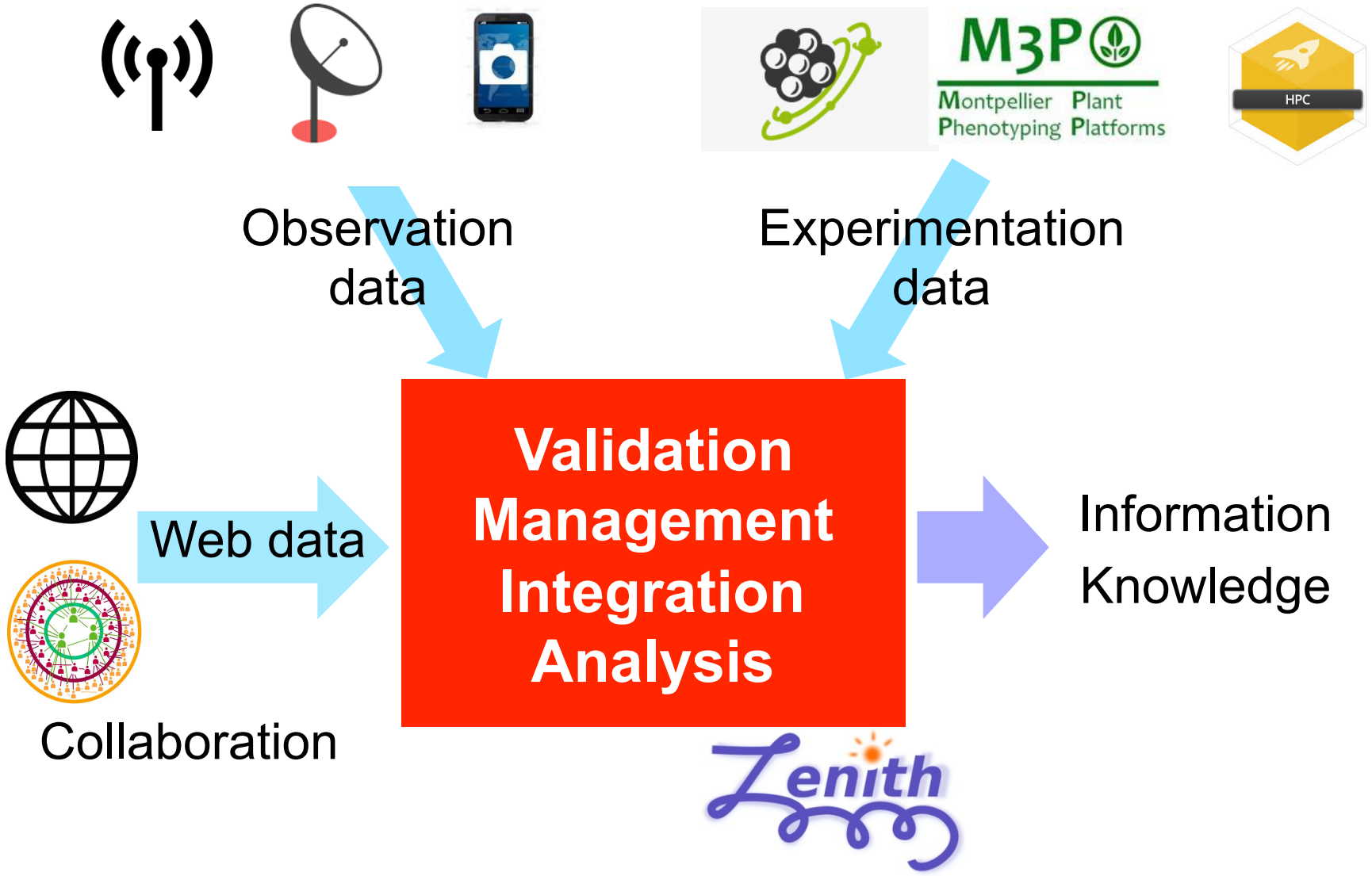
Big Data Analytics (BDA)

- Objective: find useful information and discover knowledge in data
  - Typical uses: forecasting, decision making, research, science, ...
  - Techniques: data analysis, data mining, machine learning, ...
- Why is this hard?
  - External data from various sources
    - Hard to verify and assess, hard to integrate
  - Different structures
    - Unstructured text, semi-structured document, key/value, table, array, graph, stream, time series, etc.
    - Hard to integrate
  - Low information density (unlike in corporate data)
    - Like searching for needles in a haystack
  - Simple machine learning models don't work
    - See next: "When big data goes bad" stories

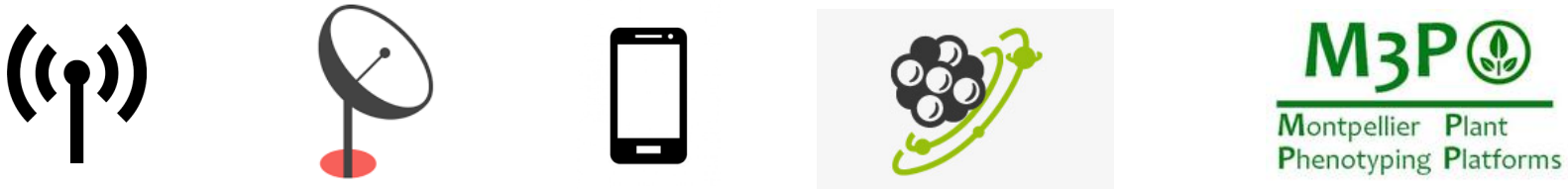
Some BDA Killer Apps

- Social network analysis
  - Modeling, simulation, visualization of large-scale networks
- Real-time processing and analysis of raw data from high-throughput scientific instruments
  - E.g. to detect changing external conditions
- Uncertainty quantification in data, models, and experiments
  - E.g. to measure the reliability of simulations involving complex numerical models
- Online fraud detection across massive databases
  - Applicable in many domains (e-commerce, banking, telephony, etc.)
- National security
  - Signal intelligence, anomaly detection, cyber analytics
  - Anti-terrorism, anti-crime
- Health care/medical science
  - Drug design, personalized medicine
  - Epidemiology
  - Systems biology

Example: data-intensive science



Example: data-intensive Science



The problem

"*Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse*"

The Office Science Data Management Challenge
USA DoE 2004

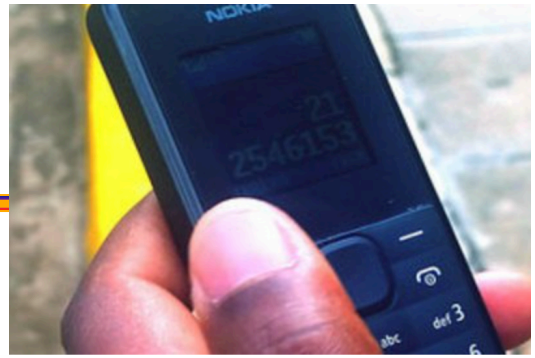
BDA Success Story 1

- CordMatch (Cytolon AG, Berlin)
- Challenge
  - Match cancer patients to cord-blood donors to rapidly retrieve stem cells for treatment
- Solution
  - Graph database (Neo4j) to store more than 2 million human leukocyte antigen (HLA) codes with 60 million direct relationships between the codes
  - A fast algorithm for single and double cord searches
- Results
  - Enables blood banks, transplant centers and registries to communicate
  - Customizable search options simplify the matching of cord blood units

BDA Success Story 2

- MySmartFarm (Cape Town, South Africa)
- Challenge
  - Help farmers make the best decisions about irrigation, pest control and fertilization
    - "The African smallholder farmer is producing a quarter to half of their potential" (W. von Loeper, MySmartFarm founder)
  - Opening up agriculture big data in Africa
- Solution
  - Database of agriculture data, including climate, soil information and disease (from Open Data sources)
  - A cloud-based analytics solution that promises to empower the agriculture industries
- Results
  - Winner of IBM's South Africa SmartCamp competition
  - Solution that will help farmers to triple or quadruple their production, while optimizing irrigation and saving water

BDA Success Story 3



- Echo Mobile (Nairobi, Kenya)
- Challenge
  - Alert the Sierra Leone government on new infections of the Ebola epidemic in different areas of the country
- Solution
  - A reporting SMS-based system that allows citizens and health workers to alert the Central Government Co-ordination Unit
  - The data is analyzed by a system developed by IBM Africa research lab.
- Results
  - Helped the government map the spread of Ebola and quickly respond to new infections while at the same time managing the epidemic in the affected communities

When Big Data goes bad

FORTUNE

November 5, 2013: 1:00 PM ET

How the models underlying today's supercomputing prowess are costing us its success.

By Joshua Klein

When Big Data goes bad - 1

The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance

List Price: ~~$70.00~~
Used: from **$35.54**
New: from **$1,730,045.91**

Have one to sell? Sell yours here

All | New (2 from $1,730,045.91) | Used (15 from $35.54)

Show: New | Prime offers only (0) — Sorted by: Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$1,730,045.91 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: ★★★★★ **93% positive** over the past 12 months. (8,193 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$2,198,177.95 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: ★★★★★ **93% positive** over the past 12 months. (125,891 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

When Big Data goes bad – 1

- Excerpts:

  What had happened was that two automated programs, one run by seller "bordeebook" and one by seller "profnath," were engaged in an iterative and incremental bidding war.

  Once a day profnath would raise their price to 0.9983 times bordeebook's listed price. Several hours later, bordeebook would increase their price to 1.270589 times profnath's latest amount.

When Big Data goes bad – 1

- Excerpts:

  What had happened was that two automated programs, one run by seller "bordeebook" and one by seller "profnath," were engaged in an iterative and

Problem: over simplified models, but reality is complex!

  bordeebook would increase their price to 1.270589 times profnath's latest amount.

When Big Data goes bad – 2

When Big Data goes bad – 2

A tiny company in Worcester, Mass., has paid the ultimate price for posting offensive T-shirts for sale online.

Fierce public backlash brought down Solid Gold Bomb, which made headlines in March for offering shirts that said "Keep Calm and Rape a Lot." The company closed its doors last week and let go its remaining three employees.

When Big Data goes bad – 2

- Excerpts:

Solid Gold Bomb, the company that made the shirt, wasn't necessarily aware that it was even selling it. Solid Gold Bomb's business isn't in artfully designing T-shirts. Instead, it writes code that takes libraries of words that slot into popular phrases (such as "Keep Calm and Carry On," which enjoyed a brief mimetic popularity online) to make derivations that get dropped onto a template of a T-shirt and automatically get posted as an Amazon item for sale.

Their mistake was overlooking a single word in a list of 4,000 or so others.

When Big Data goes bad – 2

- Excerpts:

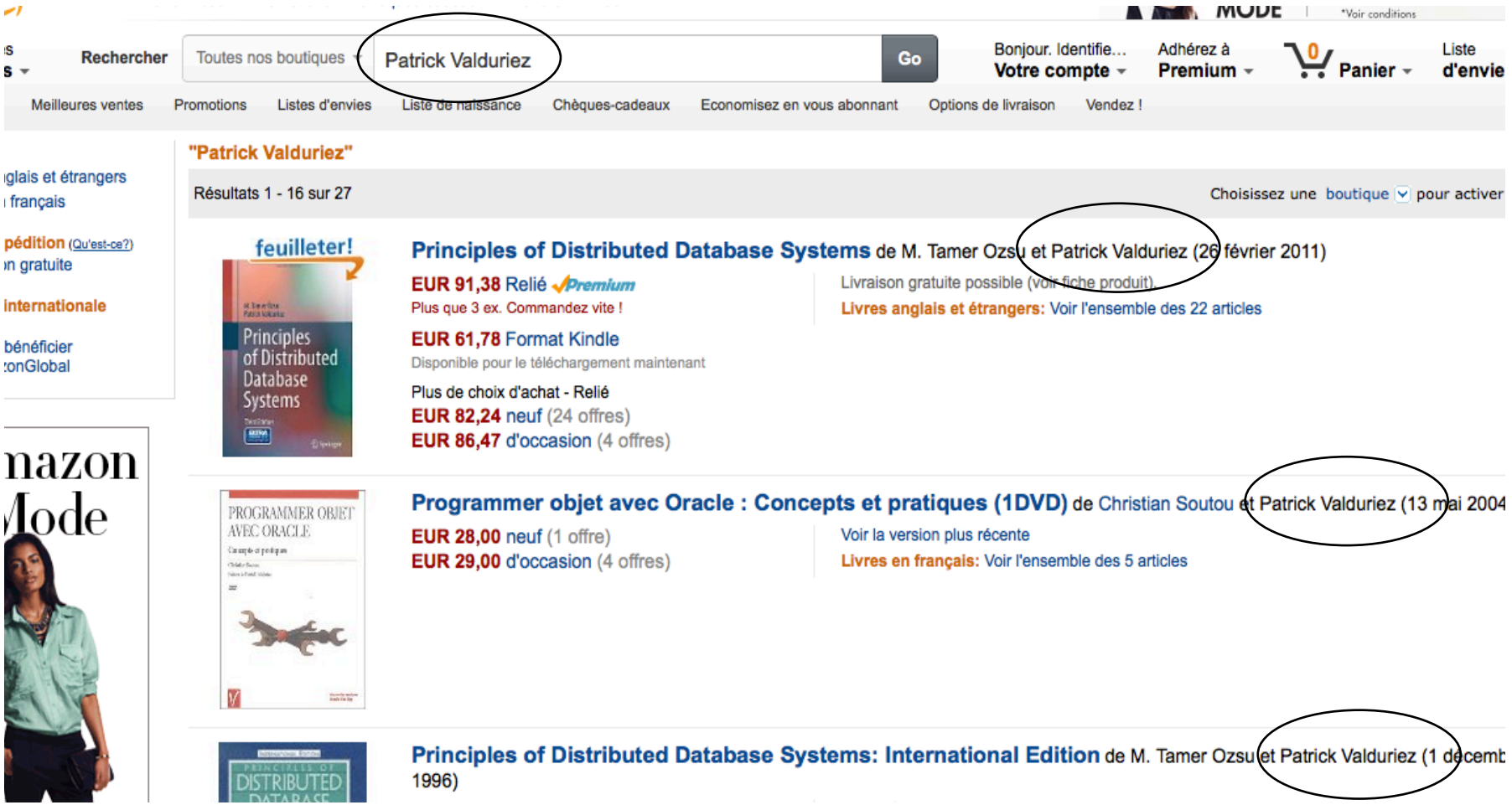
Solid Gold Bomb, the company that made the shirt, wasn't necessarily aware that it was even selling it. Solid Gold Bomb's business isn't in artfully designing T-shirts. Instead, it writes code that takes libraries of words that

Problem: context-independent model, but context does matter!

template of a T-shirt and automatically get posted as an Amazon item for sale.

Their mistake was overlooking a single word in a list of 4,000 or so others.

When Big Data goes bad (for me)



When Big Data goes bad (for me)

Object Technology de Mokrane Bouzeghoub, Georges Gardarin et Patrick Valduriez (5 juin 1997)

EUR 1,46 d'occasion (5 offres)

Livres anglais et étrangers: Voir l'ensemble des 22 articles

Ozzy Osbourne - Talking. de Patrick Valduriez (31 août 2003)

EUR 28,50 neuf (1 offre)

EUR 23,74 d'occasion (2 offres)

Livres anglais et étrangers: Voir l'ensemble des 22 articles

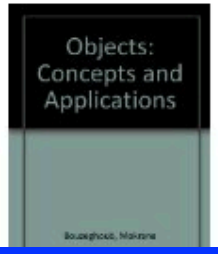
Ozzy Osbourne. Fucking Mad. Die Story zu seinen Songs. de Patrick Valduriez (30 juin 2003)

EUR 21,46 neuf (1 offre)

EUR 6,26 d'occasion (2 offres)

Livres anglais et étrangers: Voir l'ensemble des 22 articles

When Big Data goes bad (for me)



Object Technology de Mokrane Bouzeghoub, Georges Gardarin et Patrick Valduriez (5 juin 1997)

EUR 1,46 d'occasion (5 offres) | **Livres anglais et étrangers:** Voir l'ensemble des 22 articles

Problem: how do I get complete deletion of wrong information?

Ozzy Osbourne. Fucking Mad. Die Story zu seinen Songs. de Patrick Valduriez (30 juin 2003)

EUR 21,46 neuf (1 offre) | **Livres anglais et étrangers:** Voir l'ensemble des 22 articles

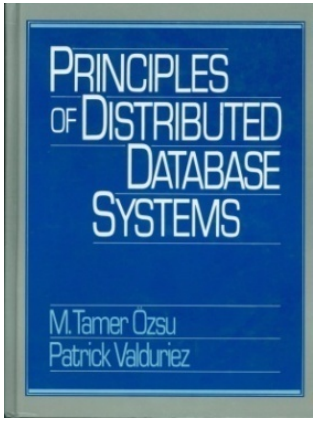
EUR 6,26 d'occasion (2 offres)

Cloud & Big Data

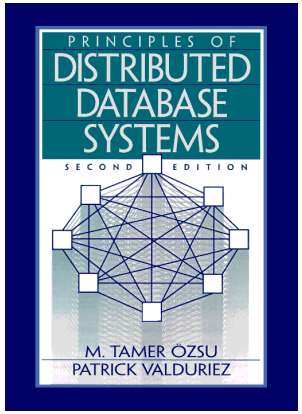
A Marriage of Convenience?

- Cloud and big data have different goals
  - Big data aims at added value and operational performance
  - Cloud targets flexibility and reduced cost
- But they can help each other by
  1. Encouraging organizations to outsource more and more strategic internal data in the cloud
  2. Get value out of it, e.g. by integrating their data with external data, through big data analytics at affordable cost

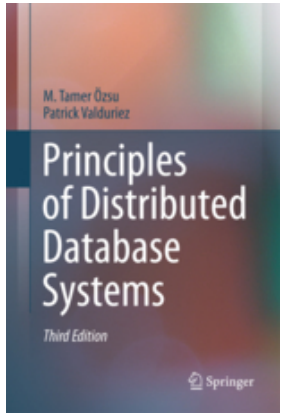
Principles of Distributed Data Management



Prentice Hall 1991
560p
distributed relational DBMS

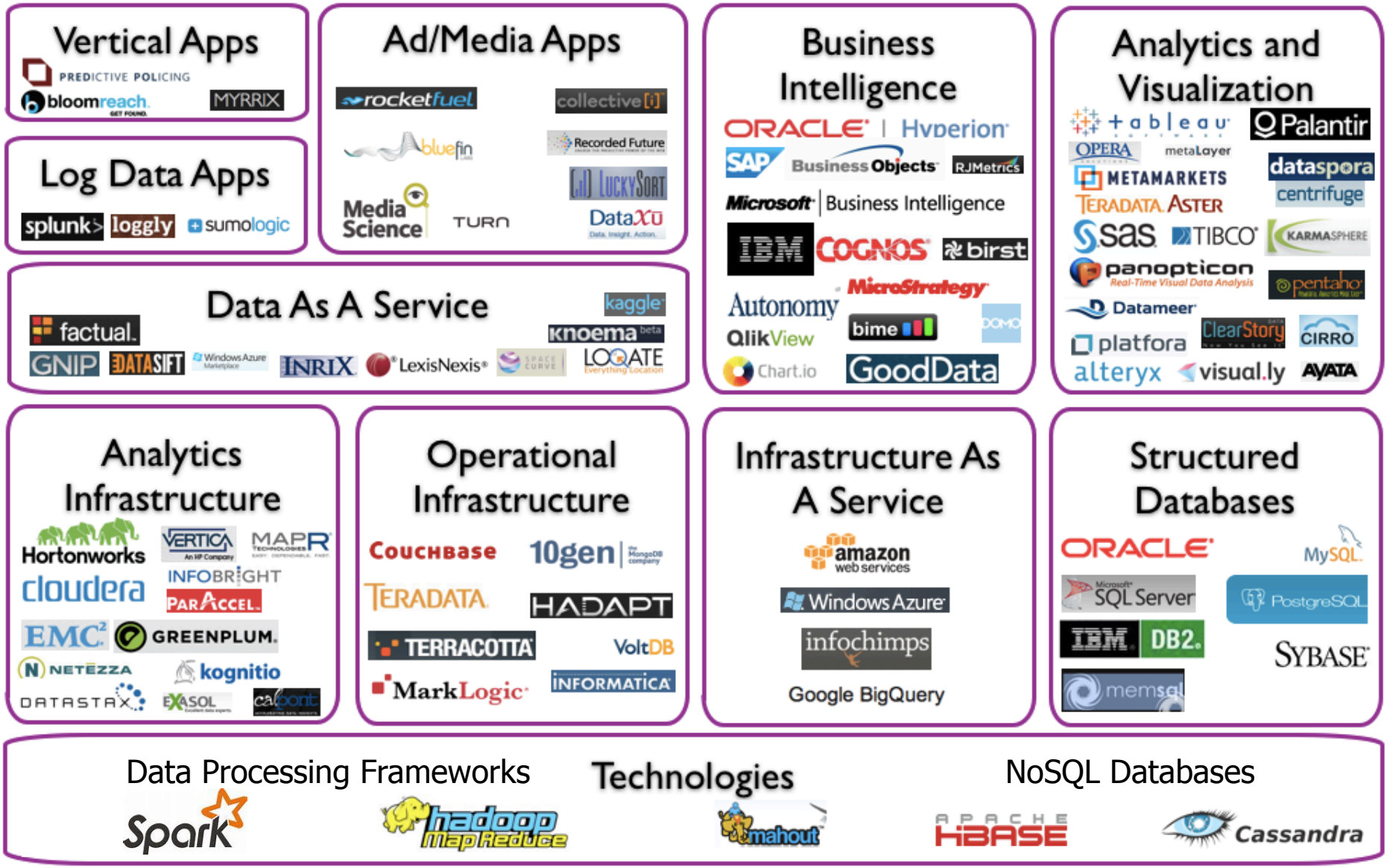


Prentice Hall 1999
660p
+ parallel DBMS



Springer 2011
850p
+ Cloud & big data

Cloud & Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

Cloud & Big Data Landscape

dave@vcdave.com

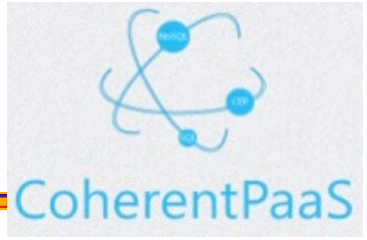
blogs.forbes.com/davefeinleib

A New Breed of Systems: multistores

- Also called *Polystores**
- Provide integrated access to multiple, heterogeneous cloud data stores such as NoSQL, HDFS and RDBMS
- Great for integrating structured (relational) data and big data
- Much more difficult than federated database systems
- A major area of research & development

*Michael Stonebraker (ACM Turing award 2015). The Case for Polystores. July 2015

CloudMdsQL Multistore



1. Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez. Integrating Big Data and Relational Data with a Functional SQL-like Query Language. *DEXA 2015 Conf.* (extended version invited to *TLDKS* journal).

2. Carlyna Bondiombouy. Query processing in cloud multistore systems. Ph.D. paper, *BDA 2015* Conf.

3. Carlyna Bondiombouy, Patrick Valduriez, Boyan Kolev. Query processing in cloud multistore systems: a survey. Inria Research Report, to appear, 2015.

4. Boyan Kolev, Patrick Valduriez, Carlyna Bondiombouy, Ricardo Jiménez-Peris, Raquel Pau, José Pereira. CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language. *Distributed and Parallel Databases*, online, 43 pages, 2015.

Opportunities and Risks for Developing Countries

Cloud & Big Data for Development

- Note: Cloud & Big Data is not the magic bullet for age-old development challenges
  - Hunger, poverty, disease, education, political unstability, corruption, war, climate change, etc.

- But it can bring powerful tools to help fighting these challenges in new ways
  - By supplying better (precise, realtime) information to decision-makers, companies, citizen, etc.

Opportunities

- A very large base of mobile phone users
  - Who use advanced services, to tweet, buy/sell goods, search for jobs, transfer data, etc.
- Cloud services can leverage this user base to
  - Foster data crowdsourcing (see Ebola example) to produce timely, precise information
  - Unleash the creativity of African entrepreneurs in developing new big data applications for Africa and create an African big data ecosystem

Risks

- Lack of cloud infrastructure*
  - 39 data centers (SaaS): Angola (2), Kenya (3), Mauritius (7), Morocco (4), Nigeria (3), South Africa (19), Tunisia (1)
  - But only 6 cloud centers (Iaas, PaaS): Maursitius (1), South Africa (5)
- Dependency to remote cloud providers
  - Lower QoS/price ratio and loss of control over private & public data
- Lack of skills in big data: data capture, analytics, machine learning
  - Remember "when big data goes bad" stories
- Lack of vision and investment from governments, private sectors and international organizations

*Source: http://www.datacentermap.com

Where Regulation Can Help

- Impose guarantees by cloud providers on customer data (in addition to QoS)
  - Safety, privacy, complete deletion
  - Logging and auditing
  - Transparency wrt where the data is stored
- Balance the power of cloud providers
  - To limit dominant positions and improve QoA/cost
- Promote interoperability between cloud providers
  - Standard protocols
  - Open source software
- Promote open data, users education and the "right to know" about our data
  - Both legal and technical knowledge

AFRICOMM

e-Infrastructures and e-Services for Developing Countries