Big Data
A Systems Perspective

Alan Wagner

April 25, 2013

Outline

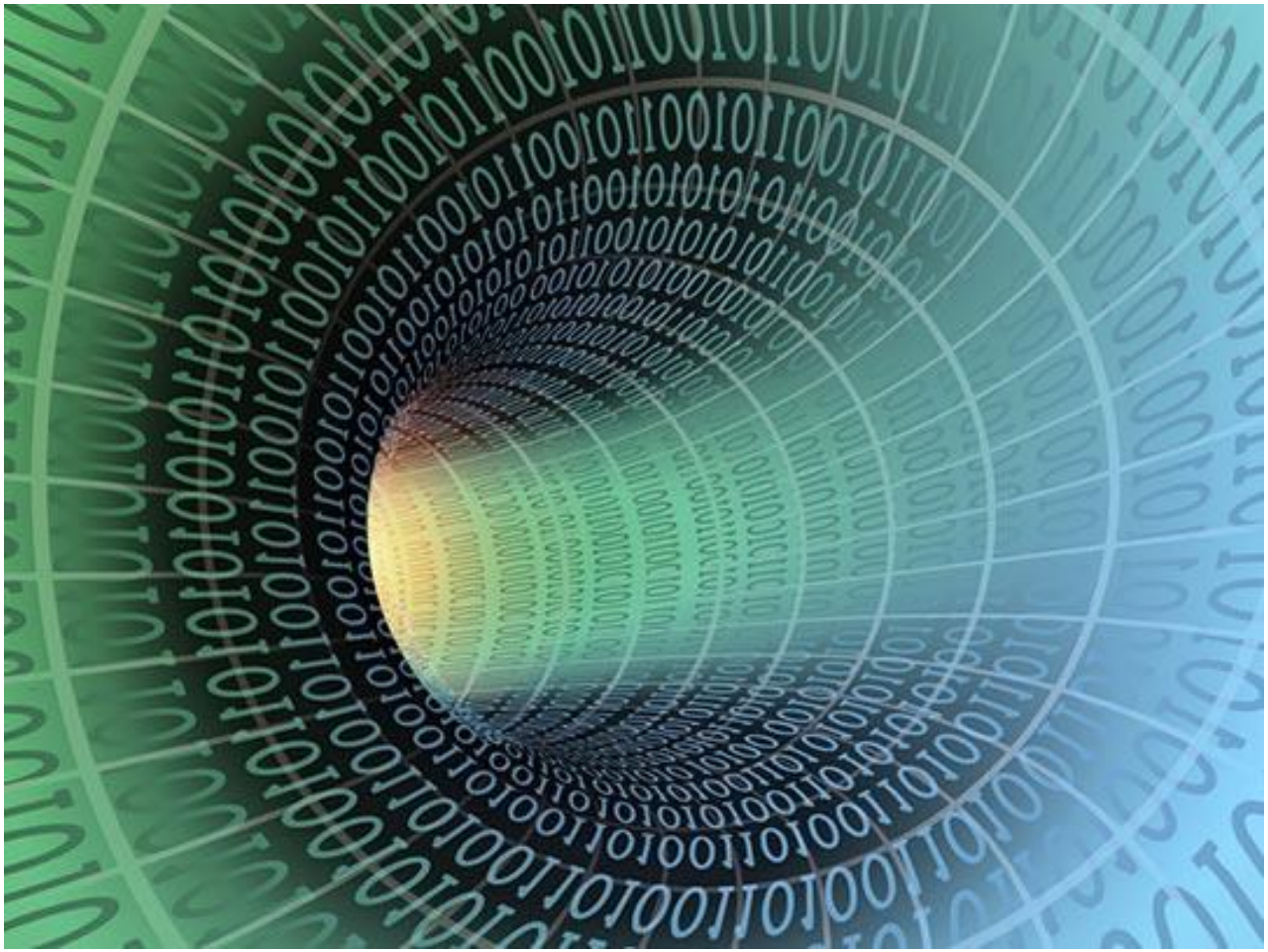
I. Why “Big Data” now

II. What is the problem?

III. Experiences

IV. Big Data and Universities

What is “big data”



Five V's

- **Volume.** Data growing exponentially with the industry amid conflicting requirements of business intelligence and reporting vs. data mining and statistical analysis further compounding data management.
- **Variety.** Data that combine text data (social media, click stream, customer service records, etc.) with structured internal data (such as master and transactional data) and third party information (such as credit rating, econometric forecasts, maps or other proprietary databases).
- **Velocity.** Data consumption and movement from a myriad of sources that is generated or delivered at a rapid pace (such as click stream data, scanning for sensitive images in a crowd, etc.).
- **Variability.** This is not the statistical variability that represents signals in the data, but the many different meanings that data can have (e.g., text can be interpreted as number of words, or as semantics, or as concepts).
- **Value.** By addressing each characteristic individually and together holistically, an enterprise can deliver value (ROI) among its decision making processes.

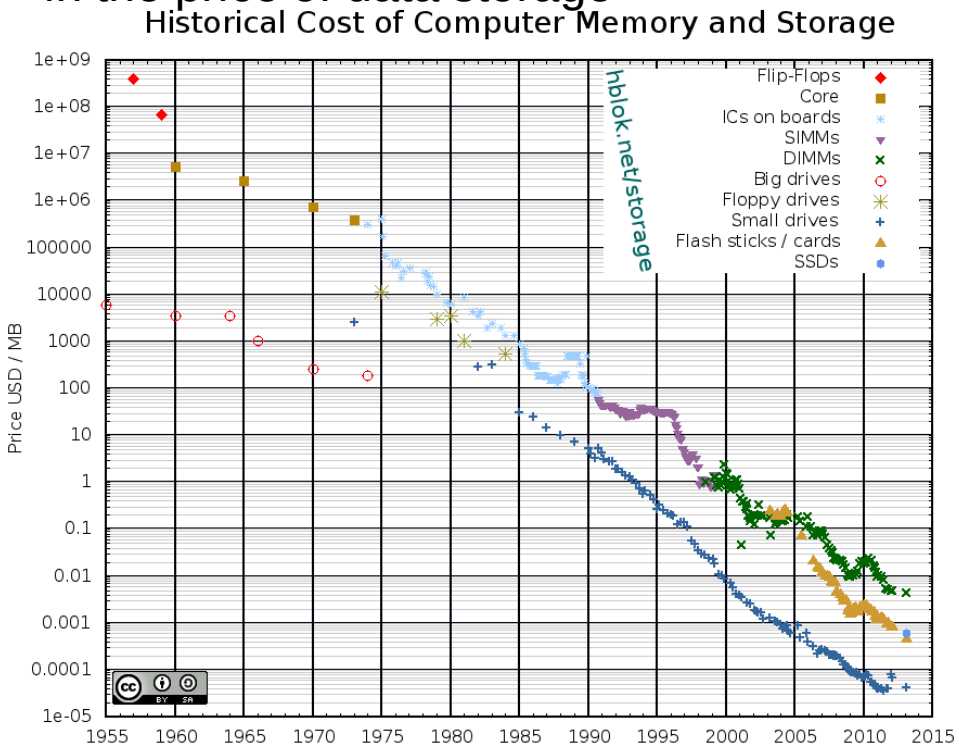
scalable analytics

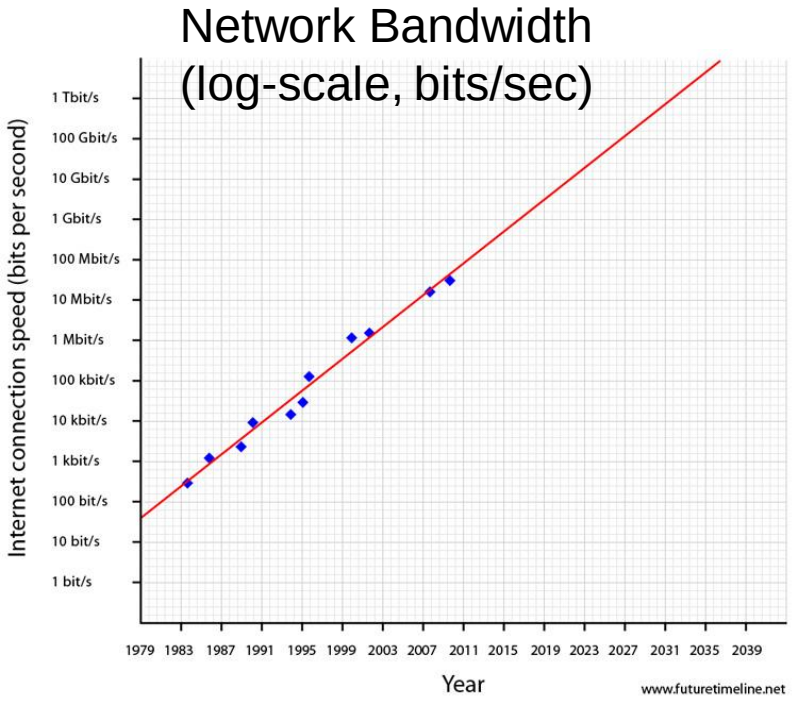
Why “Big Data”?

- Larger Digital Footprint
- Reduced Disk and Network Costs
- Success of Major Internet Players
- Internet of Everything (digitalization)
- Internet Speed Expectation

Technology

Kryder's Law is showing the exponential decay in the price of data storage



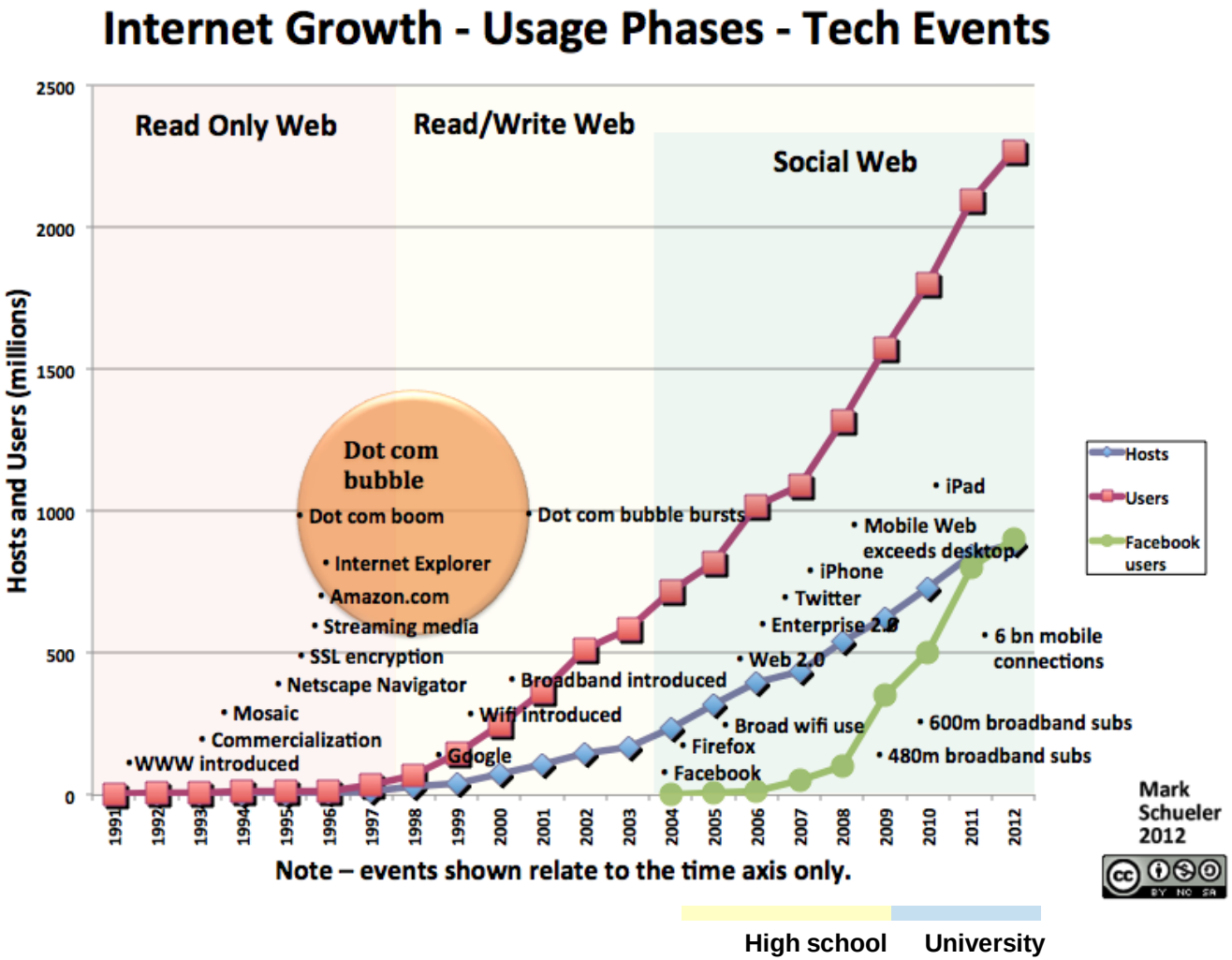


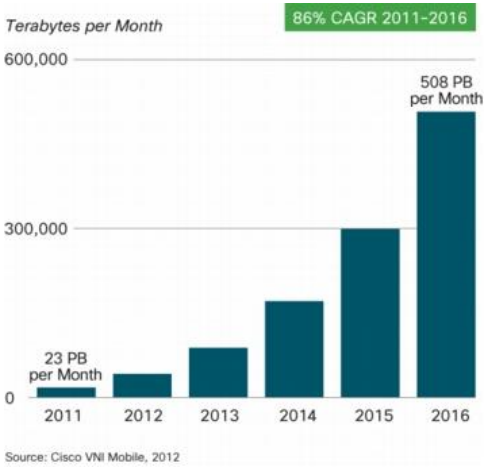
$50-100 for 1 Tbyte
Capacity increasing to 5-6Tbyte drives – 10-15 Tbyte drives 2020 (1-12 milliseconds)
SSD $.50-1.0 for 1Gbyte (100 microseconds)

10Mbits/second

90% servers today are 1Ge

Digital Footprint





Dramatic Increase in mobile data

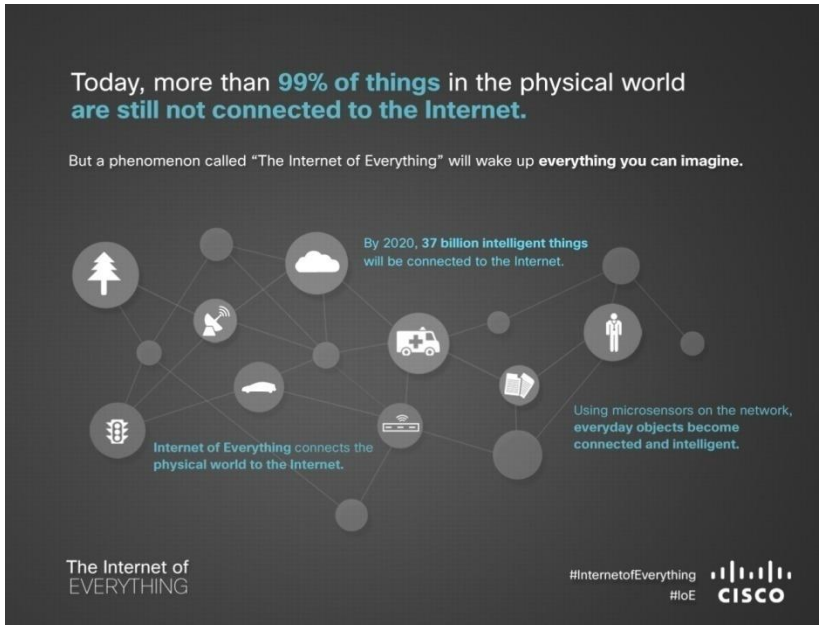
Success of Internet Companies

- YouTube (1B users, 4B views/day)
- Facebook (1B active users, 300M photos/day)
- Gmail (500M)
- Google+ (343M), 5M google-apps
- Hotmail(286M)
- LinkedIn (200M)
- Paypal (117M)
- Ebay (100M)
- Dropbox (100M)
- Steam (50M)
- Netflix (30M)
- Flickr (87M)

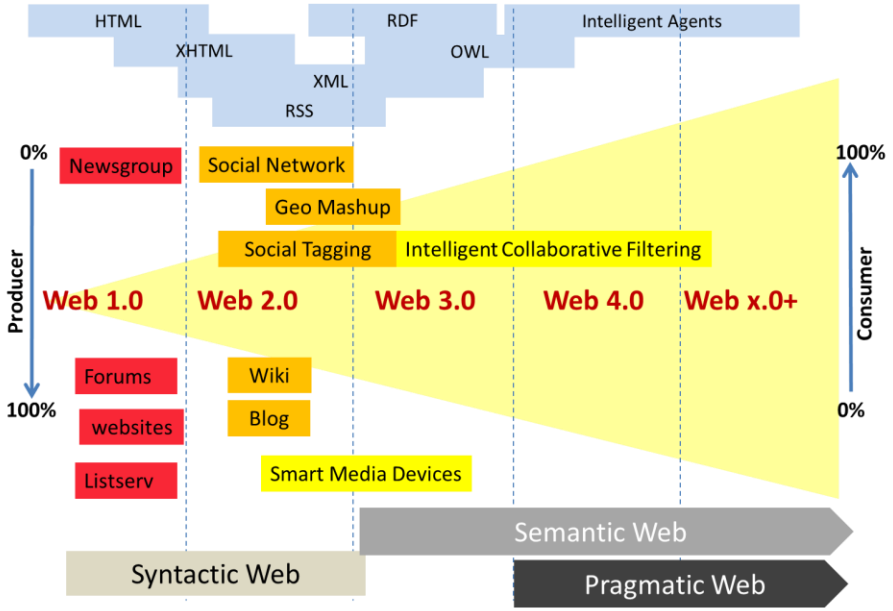
http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/

Internet of Everything – Web X.0

Physical world

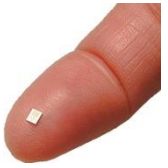


Virtual world



Digital breadcrumbs

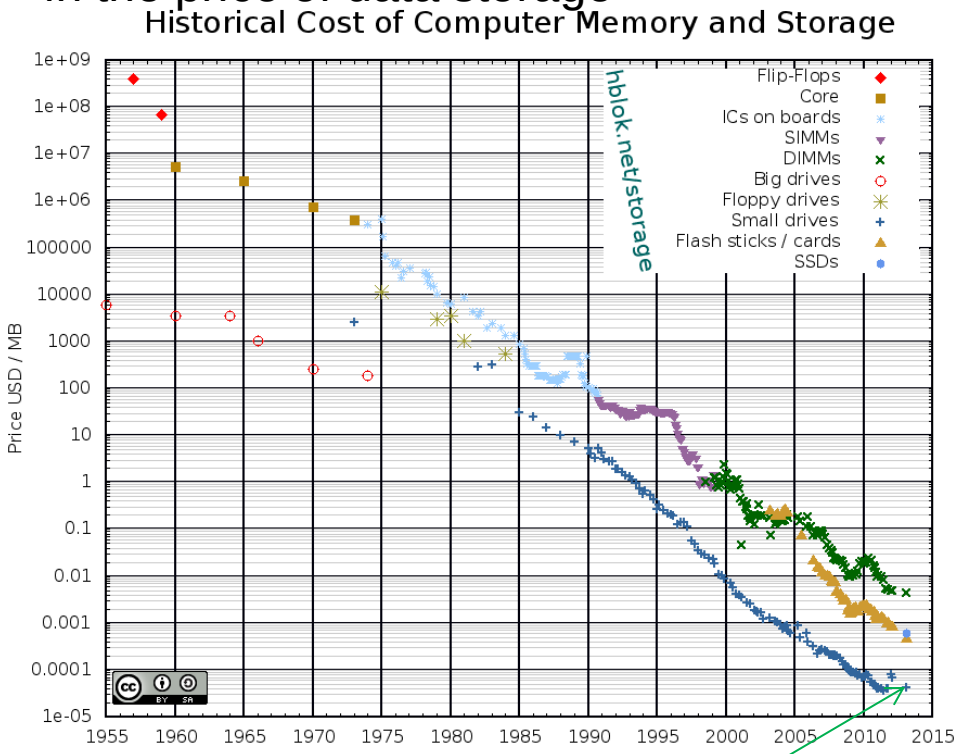
Macy's announced plans to begin its RFID rollout this year. Meanwhile, **J.C. Penney's** CEO, Ron Johnson, told a conference audience that he expects his firm to start affixing RFID tags to 100 percent of its merchandise, and to begin using the technology to enable self-checkout (**Walmart**, **American Apparel**)



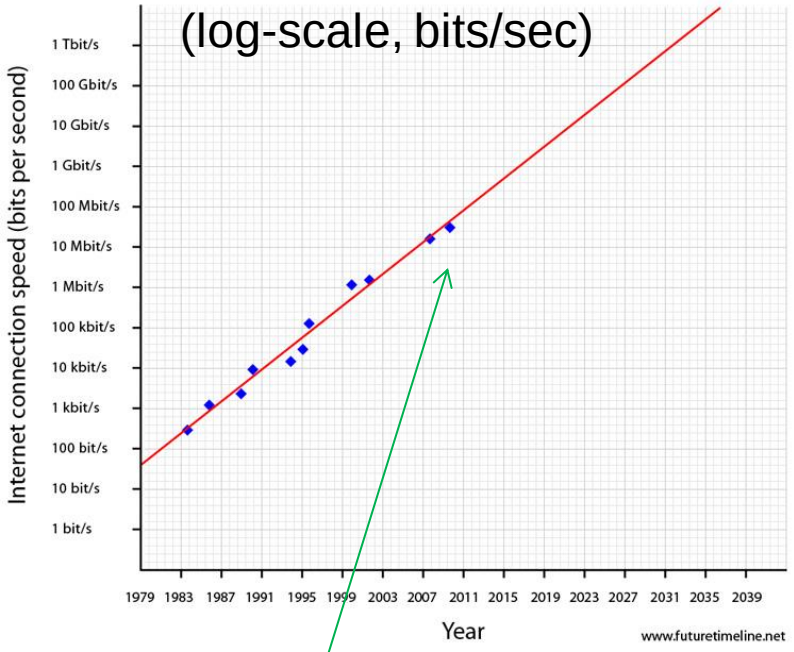
http://www.rfidjournal.com/articles/view?10263 Jan 2013

Technology

Kryder's Law is showing the exponential decay in the price of data storage



Network Bandwidth (log-scale, bits/sec)



$50-100 for 1 Tbyte
Capacity increasing to 5-6Tbyte drives
(10-15 Tbyte drives 2020 (1-12 milliseconds))
SSD $.50-1.0 for 1Gbyte (100 microseconds)

10Mbits/second

90% servers today are 1Ge

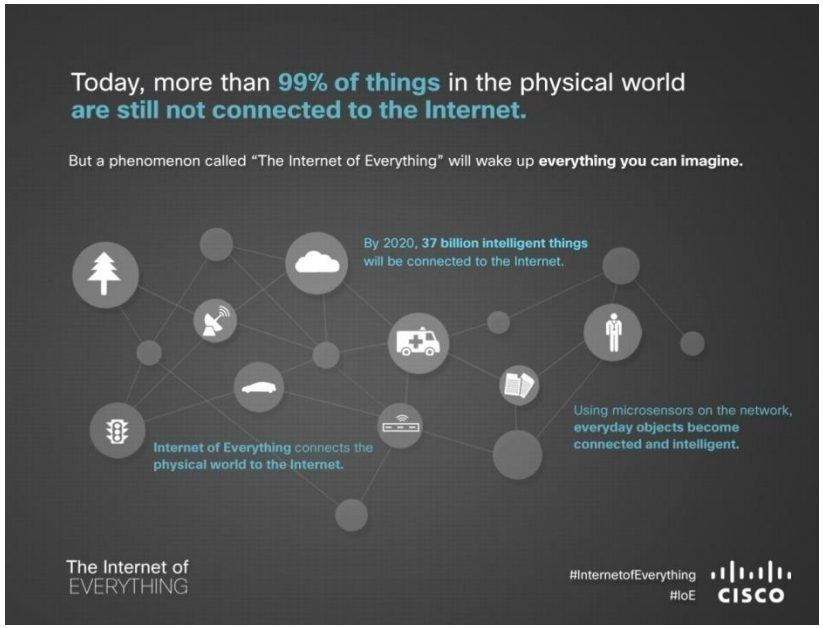
Success of Internet Companies

- YouTube (1B users, 4B views/day)
- Facebook (1B active users, 300M photos/day)
- Gmail (500M)
- Google+ (343M), 5M google-apps
- Hotmail(286M)
- LinkedIn (200M)
- Paypal (117M)
- Ebay (100M)
- Dropbox (100M)
- Steam (50M)
- Netflix (30M)
- Flickr (87M)

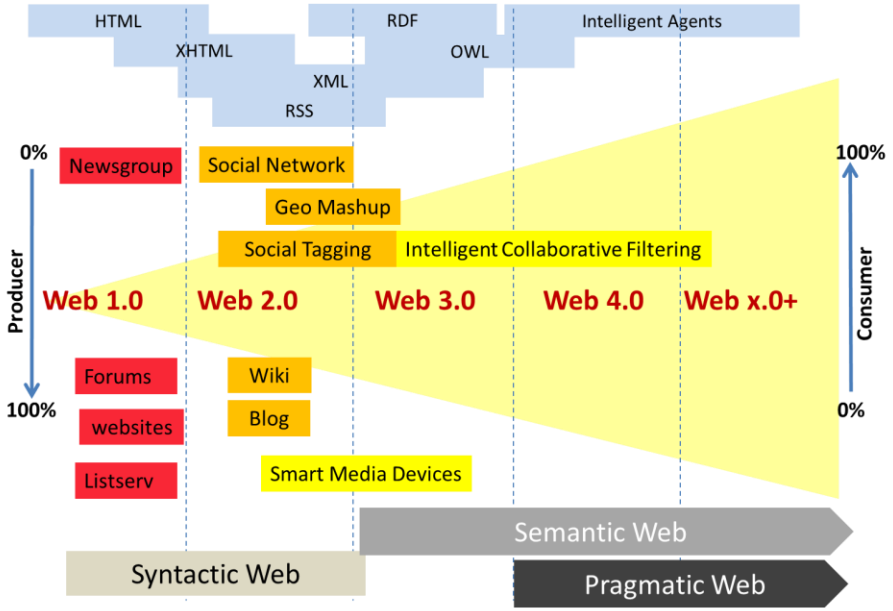
http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/

Internet of Everything – Web X.0

Physical world

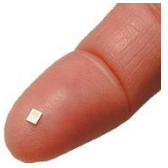


Virtual world



Digital breadcrumbs

Macy's announced plans to begin its RFID rollout this year. Meanwhile, **J.C. Penney's** CEO, Ron Johnson, told a conference audience that he expects his firm to start affixing RFID tags to 100 percent of its merchandise, and to begin using the technology to enable self-checkout (**Walmart**, **American Apparel**)



http://www.rfidjournal.com/articles/view?10263 Jan 2013

Everything, Anytime, Anywhere at Internet Speeds

The paradigm shift we are seeing in data management is more about giving customers the technologies they need to **store** and **analyze**:

- any data set
- any type of data (Variety),
- any size of data (Volume), for
- any type of user, and
- in any timeframe (Velocity), and

Morgan Stanley

- anywhere.

The Challenges

- Acquisition and Storage of Data
- Data Processing
- Data Management (Governance)

Acquisition and Storage

- Physical medium
  - Disks, SSD, Tape, Cloud storage
- Acquisition
  - Data format (metadata)
  - Data cleaning
  - Data Integration
  - Data access

"The value of any piece of information is only known when you can connect it with something else that arrives at a future point in time," Hunt said. "Since you can't connect dots you don't have, it drives us into a mode of, we fundamentally try to collect everything and hang on to it forever." CIA's CTO Gus Hunt –- $600M contract Amazon

Processing the Data

- Distributed system design
  - Scalability
  - Performance
  - Cost

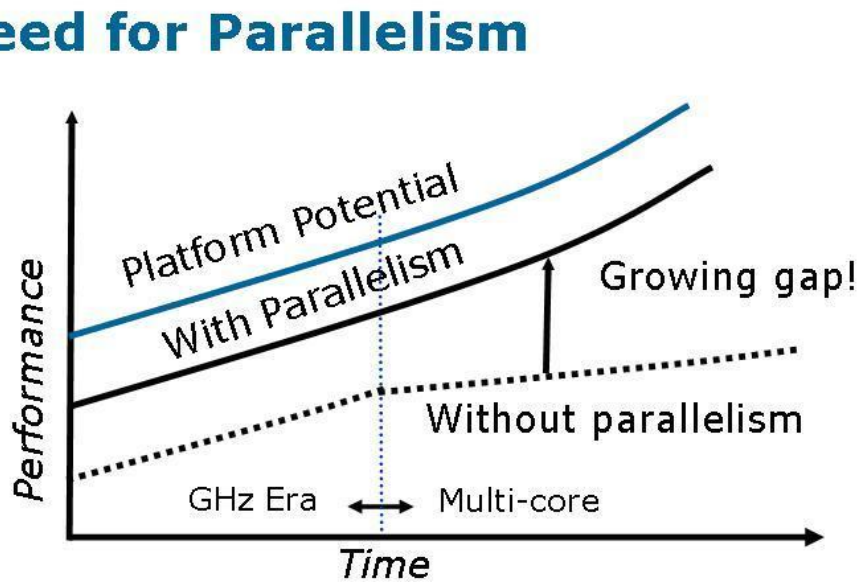
“the impulse to store lots of data because it can be cheap can lead to storing too much and make answering simple questions harder. "You want to have some sort of control over what data you push into an application, … Otherwise, your juice isn't really worth the squeeze."

Rob Bearden, CEO of Hortonworks

Paradigm Shift in Computing

- Engineers are unable to make faster processors
  → **"Multi (many)-core" is the solution**

- The problem is most software runs sequentially on single processors and will not speed up on multi-core systems
  → **Need for parallelism**



Why is scalability so hard? Because scalability cannot be an after-thought. It requires applications and platforms to be designed with scaling in mind …
Werner Vogels CTO Amazon

http://www.allthingsdistributed.com/2006/03/a_word_on_scalability.html

Distributed Computing

Essentially everyone, when they first build a distributed application, makes the following eight assumptions. All prove to be false in the long run and all cause *big* trouble and *painful* learning experiences.

1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. The network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is zero
8. The network is homogeneous

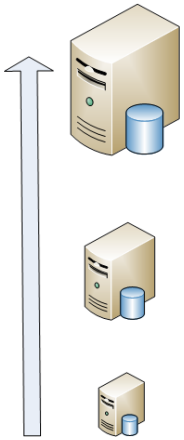
The Eight Fallacies of Distributed Computing
Peter Deutsch

- L1 cache reference 0.5 ns
- Branch mis-predict 5 ns
- L2 cache reference 7 ns
- Mutex lock/unlock 25 ns
- Main memory reference 100 ns
- Compress 1K bytes with Zippy 3,000 ns
- Send 2K bytes over 1 Gbps network 20,000 ns
- Read 1 MB sequentially from memory 250,000 ns
- Round trip within same datacenter 500,000 ns
- Disk seek 10,000,000 ns
- Read 1 MB sequentially from disk 20,000,000 ns
- Send packet CA->Netherlands->CA 150,000,000 ns

http://perspectives.mvdirona.com/2009/10/17/JeffDeanDesignLessonsAndAdviceFromBuildingLargeScaleDistributedSystems.aspx

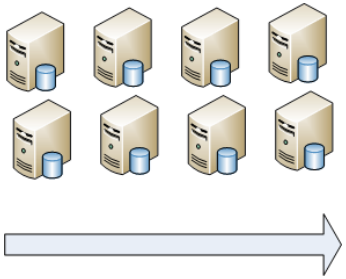
Increasing Performance

Vertical Scaling

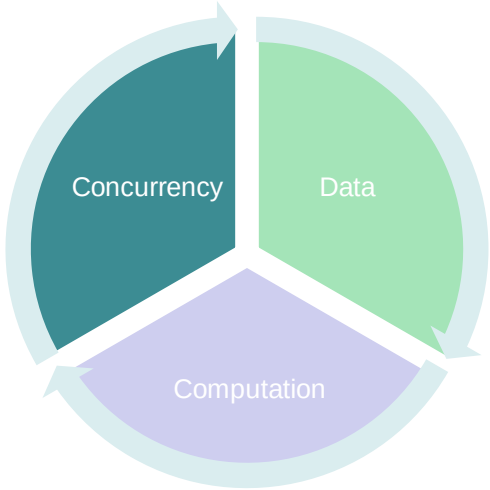


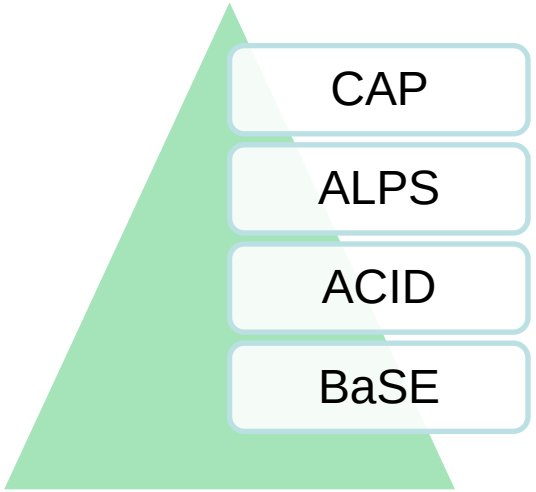
Physical limitations

Horizontal Scaling



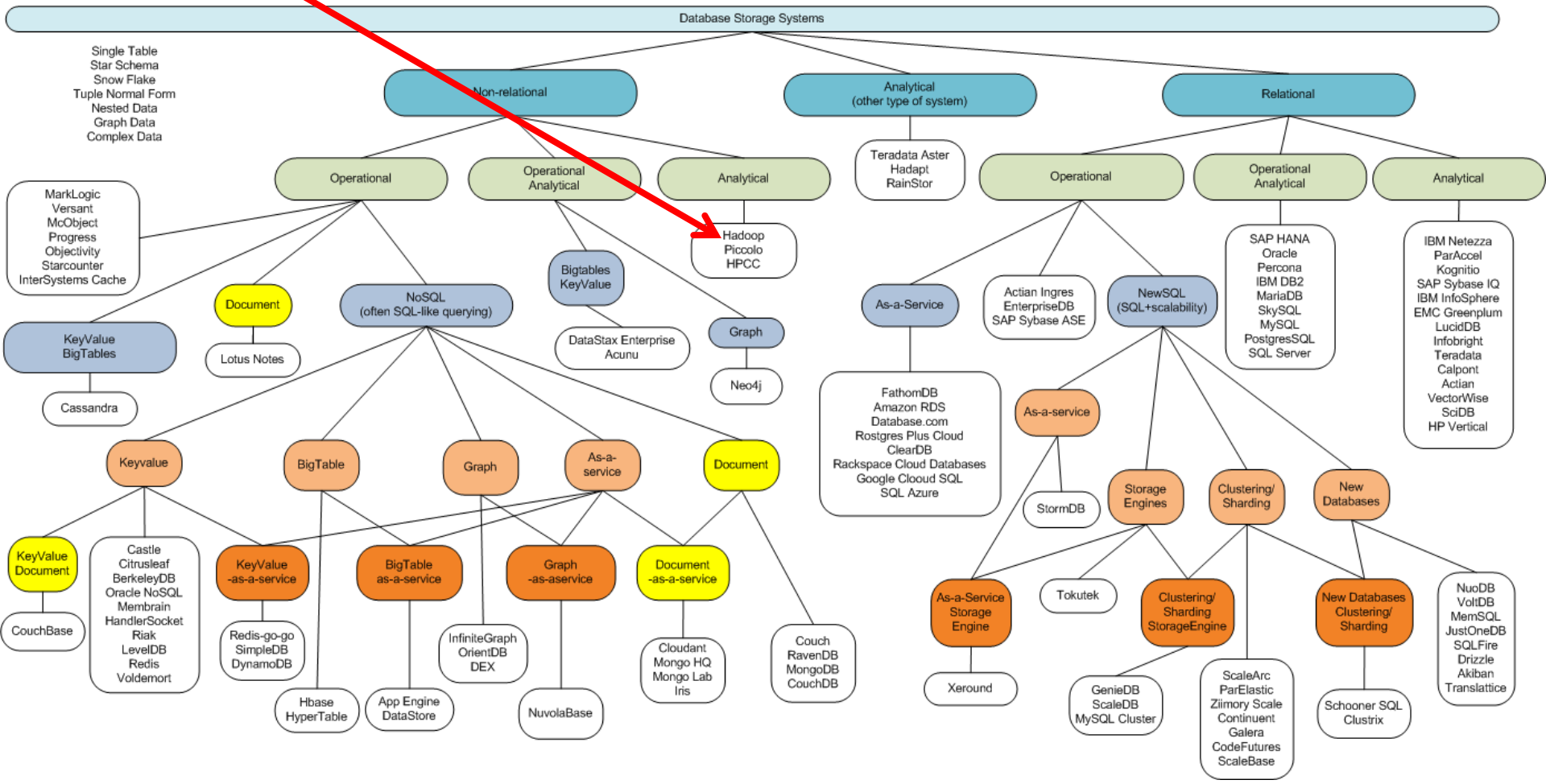
Communication overheads





Slower disks, faster networks, more computation

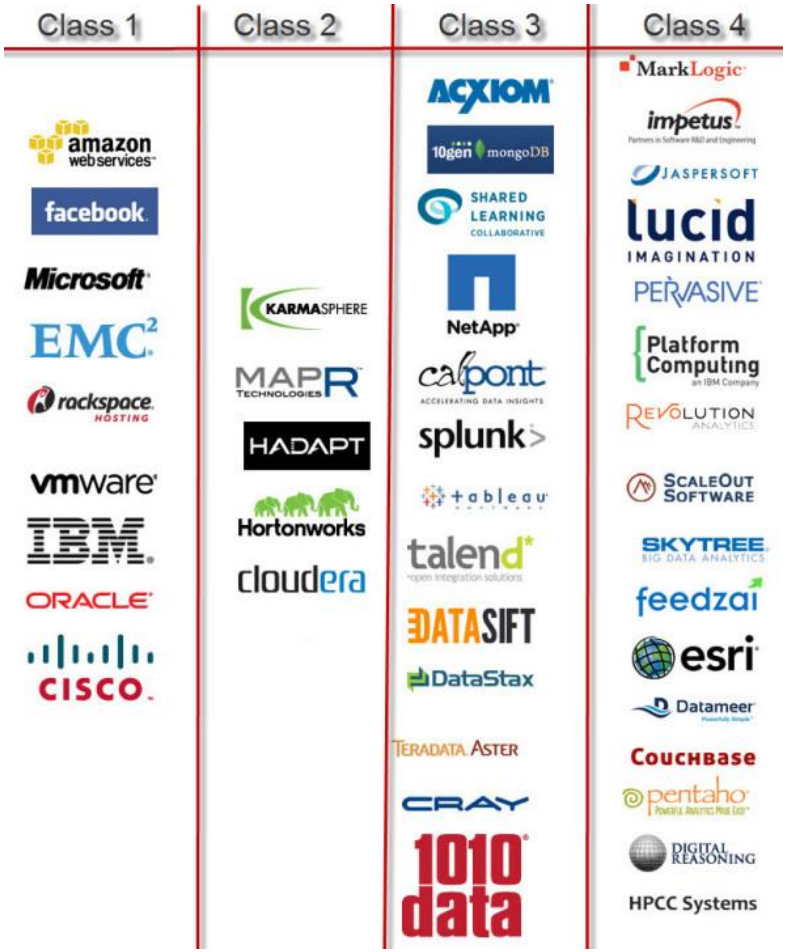
Landscape



Data source the451group.com updated-database-landscape-graphic Matthew Aslett

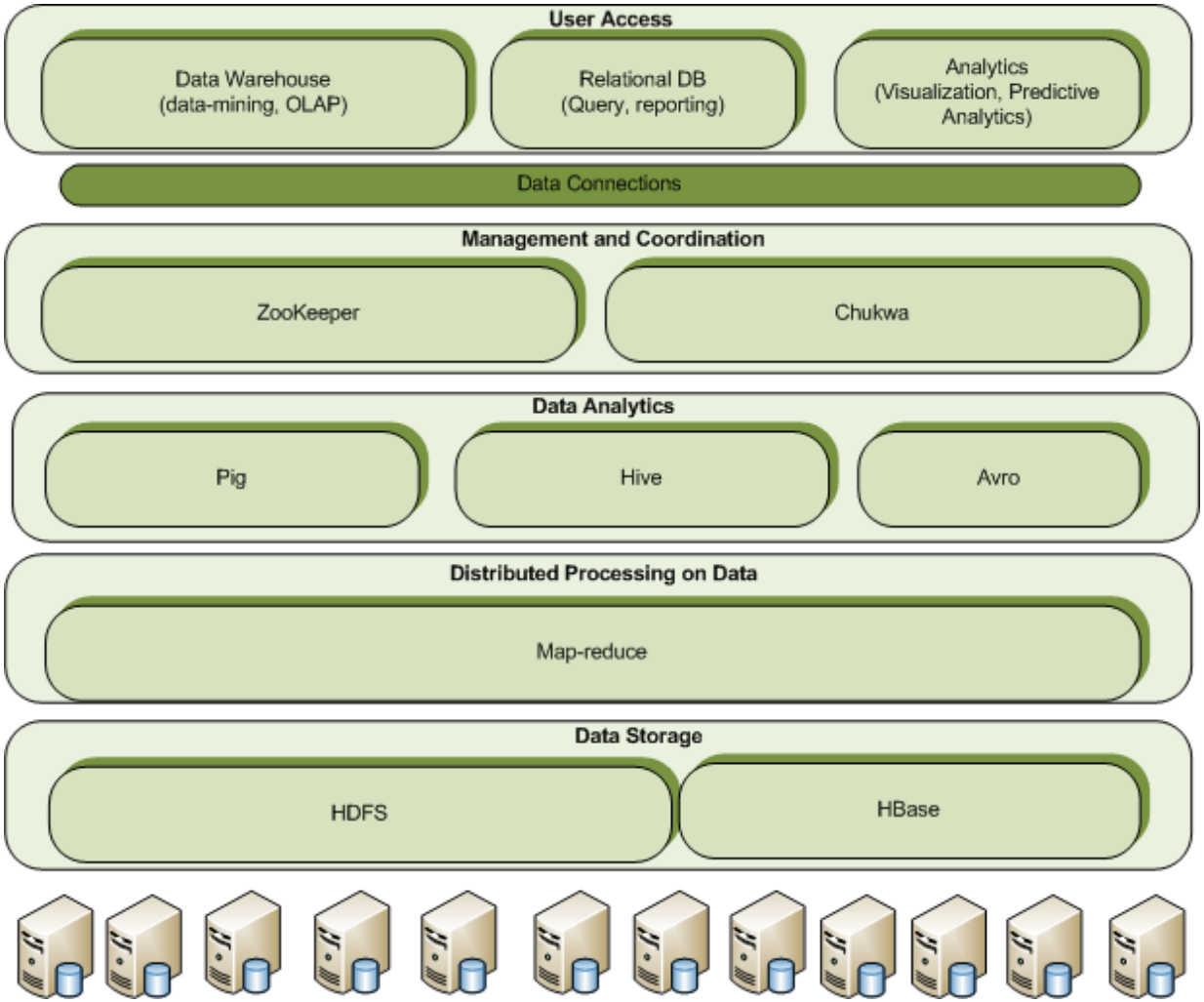
Hadoop Ecosystem

1. “Big data” vendors
2. Hadoop providers
3. Hadoop add-ons
4. Hadoop services



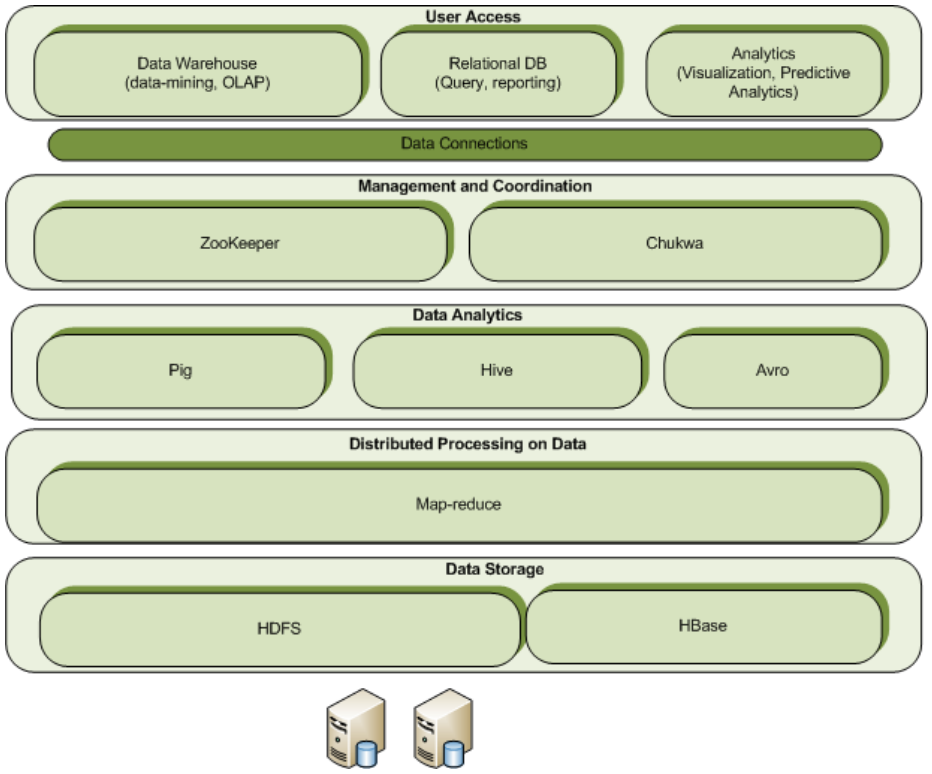
http://cloudcelebrity.wordpress.com/2012/02/29/classification-of-companies-in-big-data-space/

hadoop System



Parallel Disk Access

hadoop System



Simplified Distributed Computing

Data Management

…the vast majority of Big Data is either duplicated data or synthesized data.

• Let's take a look at a leading medical research facility that generates 100 terabytes of data from various instruments.

•This data is then copied by 18 different research departments that further process the data and add 5 terabytes of additional synthesized data each.

•Now they must manage a total of over a petabyte of data, of which less than 150 terabytes is unique.

• Yet, the entire petabyte of data is backed up, moved to a disaster recovery site, consuming additional power and space used to store it all.

•So now, the medical center has used over 10 petabytes of storage to manage less than 150 terabytes of real unique data.

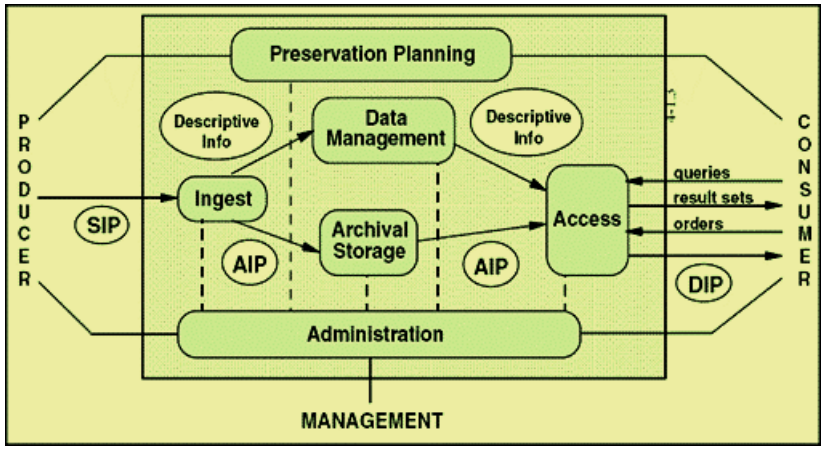
Ash Ashutosh is CEO of Actifio, "Best practices for management big data" Forbes 7/5/2012

The Ontario Public Service (OPS) spends $15.3 million every year backing up redundant information. The OPP cost of backing up redundant information is $1.99 million every year. (OPP Strategic plan 2011-2013)

De-duplication

Data Governance

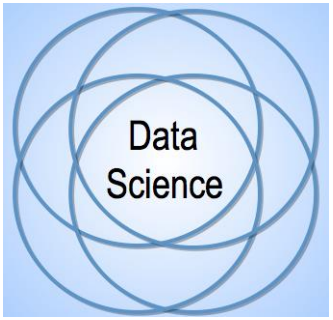
- Physical medium
  - Disks, SSD, Tape, Cloud storage
- Planning
- Archival
  - Retention rules, Disaster recovery, formats/technologies
- Backup (operational data), Processes
- Data Management
  - Acquisition, veracity – data cleaning
  - Format, metadata
- Administration
  - Integrity, security
- Access and Availability
  - Transport, access control, privacy



OAIS Model for Archival Storage

Total Cost of Ownership

Data Science



just gather huge amounts of information, observe the patterns and estimate probabilities about how people will act in the future.

David Brookes NYT 4-16-13

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
title of an 2008 article by former *Wired* editor-in-chief Chris Anderson, who also stated “with enough data, the numbers speak for themselves”

DEBATE: Noam Chomsky pioneer in the field of linguistics; and Peter Norvig, Senior Research Director at Google

http://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/
http://norvig.com/chomsky.html

The Hidden Biases in Big Data
Harvard Business Review, Kate Crawford April 1, 2013

Experiences

“Big Data” is here, and opportunities for processing existing data and acquiring new varieties of data is growing.

A. There is considerable risk
    1. Cost of software/hardware
    2. Skills needed
    3. Achieving the goals

B. Focus …
    1. Specific opportunities
    2. Veracity and pedigree of data

C. Existing Practices

Financial Markets

MarketView:

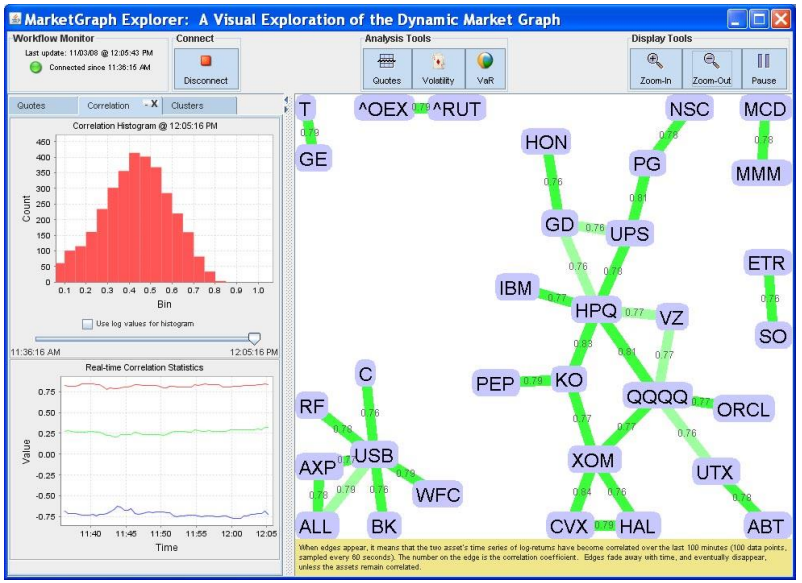
Real-time discovery of correlation patterns across entire markets

RiskView:

Real-time value-at-risk analysis for large portfolios

SentimentView:

Gauge market sentiment with our proprietary indicators



Back-testing Hedge fund

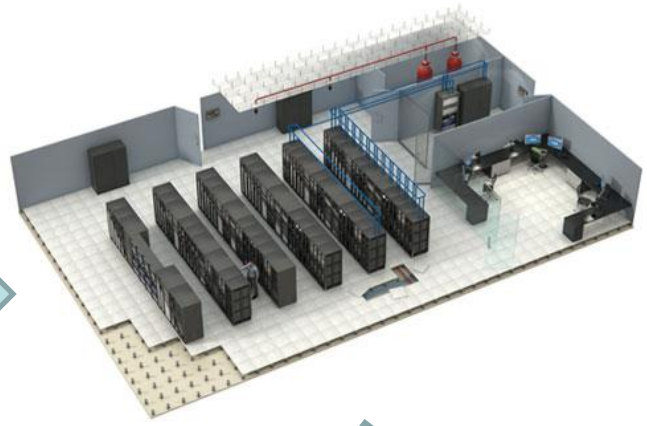
end of year processing 2 days for 6 months worth of computing

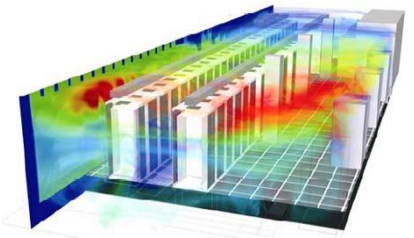
Lesson: Need to fit into customers workflow, value of data

Datacenters

CRAC Units

Cooling Strategies





Thermal Characteristics

Component Level

Sensors

Load and Scheduling

External variables affecting operation

Lesson: The challenge to scale, is it feasible?

Mining

MineSense provides a proven platform for the sensing and sorting of low-grade ore to a level of sensitivity unprecedented in the industry.

Lesson: Challenges of real-time, is it feasible

TeraPeak and Benevity

benevity

"Seamlessly Power Any Initiative That Has Charitable Giving as Part of Its Goal"

TERAPEAK
THE LEADING RESEARCH TOOL FOR ONLINE SELLERS

"Terapeak is an eBay Certified Solutions Provider and member of the eBay Developers Program. The Terapeak research tool helps sellers build better auction titles and increase eBay profits."

Lesson: Need to ask "How can we benefit the business model"

The Future

- Cloud Services
- Virtualization, Democratization of Teaching and Delivery
- More and more mobile devices
- Data Management
- Training

The Future

- Cloud Services
- Virtualization of Teaching
- More and more mobile devices
- Data Management
- Developing Expertise

Be a service provider rather than a consumer

Social Media

"Classmates Online, the Renton company that built a business around uniting high-school alumni, is being acquired by Internet service provider United Online for $100 million in cash. It reported sales of $54 million for the first nine months of this year and has 1.4 million paying subscribers."

(No. 2 Internet service provider buying Classmates Online Tuesday, October 26, 2004 , Seattle Times)

United Online, the holding company for a school-themed portfolio of websites, has just acquired schoolFeed, a Facebook app that's been growing like a weed over the past several months.
SchoolFeed acts as a connection-finding tool for current and former high school students. As of today, the app has around 19 million members, with 100,000 new registrations added daily.

(http://venturebeat.com/2012/06/11/schoolfeed/)

University Related Companies

Tutoring is a leading educational services company for university and college students in Canada and the U.S

Content Management Software, universities starting to provide their own (eClass-moodle, OWL-Sakai)

Analytic and collaboration tools for schools (Grades)

Course Evaluation

LinkedIn for researchers

Business Case

- More Successful Graduates
  - Stature of university
  - Alumni fund-raising
  - More research funding
- Information Provider
  - Directed advertising
  - Jobs marketing
  - Directed services (jobs, tutoring…)
- Operational Efficiencies

Sources of Data

HS students → Ugrads → Grads → Alumni

Admissions :
- School grades
- References
- Schools

Classes/Grades

Course Evaluations (TA/Professor)

Course Material

Teaching Artifacts

Admissions :
- School grades
- References
- Schools

Classes/Grades

Course Evaluations (TA/Professor)

Course Material

Teaching Artifacts

Research

Facebook

LinkedIn

Avoid being creepy

How Companies Learn Your Secrets, NY-TIMES Feb 16,2012, Charles Duhigg

Statistician in the Target was asked by a colleague “If we wanted to figure out if a customer is pregnant, even if she didn’t want us to know, can you do that? ”

“We knew that if we could identify them in their second trimester, there’s a good chance we could capture them for years,” Pole told me. “As soon as we get them buying diapers from us, they’re going to start buying everything else too. If you’re rushing through the store, looking for bottles, and you pass orange juice, you’ll grab a carton. Oh, and there’s that new DVD I want. Soon, you’ll be buying cereal and paper towels from us, and keep coming back.”

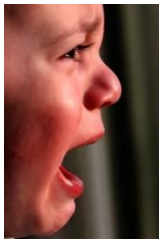
he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Avoid being creepy

How Companies Learn Your Secrets, NY-TIMES Feb 16,2012, Charles Duhigg

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."



"With the pregnancy products, though, we learned that some women react badly," the executive said. "Then we started mixing in all these ads for things we knew pregnant women would never buy, so the baby ads looked random. We'd put an ad for a lawn mower next to diapers. We'd put a coupon for wineglasses next to infant clothes. That way, it looked like all the products were chosen by chance.

"And we found out that as long as a pregnant woman thinks she hasn't been spied on, she'll use the coupons. …. As long as we don't spook her, it works."

THANK YOU

Dr. Alan Wagner
Department of Computer Science
University of British Columbia
wagner@cs.ubc.ca