Overview of Big Data

CPSC 3620 SP 2015
Linh Ngo

Terminologies

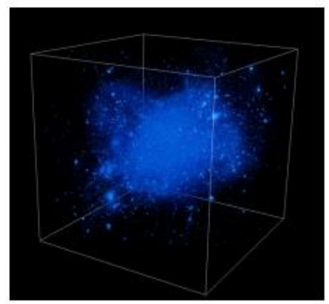
- Big data problems are problems whose not only the processing power, but the size of the data is also the limiting factor in being able to find a timely solution.
- Big Data (Industry, Social Sciences)
  - Input data carrying characteristics of Big Data (the 4V)
  - Computational process is simple and straightforward, with minimal intermediate data being generated
- Data Intensive Computing (Natural Sciences)
  - Input data may or may not be big data
  - Computational process produces massive and complex intermediate data that needs to be analyzed during the process

Motivating Examples in Science

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

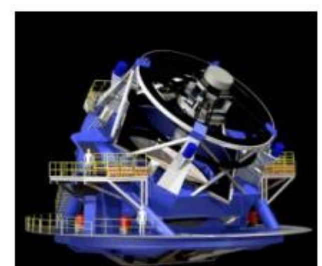
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



LAST FEW DECADES
a computational branch simulating complex phenomena

TODAY
data intensive science, synthesizing theory, experiment and computation with statistics
►new way of thinking required!



Alex Szalay, 2012: Extreme Data-Intensive Scientific Computing: The Fourth Paradigm

Motivating Examples in Industry

- Scientific data is doubling every year, reaching PBs
  - *CERN is at 22PB today, 10K genomes ~5PB*
- Data will never will be at a single location
- Architectures increasingly CPU-heavy, IO-poor
- Scientists need special features (arrays, GPUs)
- Most data analysis done on midsize BeoWulf clusters
- Universities hitting the "power wall"
- Soon we cannot even store the incoming data stream
- **Not scalable, not maintainable…**

Alex Szalay, 2012: Extreme Data-Intensive Scientific Computing: The Fourth Paradigm

Note: CERN's LHC generated 125PB in 2015

CLEMSON UNIVERSITY

Data Generation

- In 2008-2009:
  - Google processed 20PB a day
  - Facebook had 2.5PB of user data + 15TB/day
  - eBay had 6.5PB of user data + 50TB/day
- In 2010-2011:
  - Facebook had 400M users / 125PB of user data
  - eBay had 10PB of user data in 2010, expected to double this number in 2011
- In 2012-2013:
  - Facebook had 900M users
  - Twitter had 400M Tweets/day

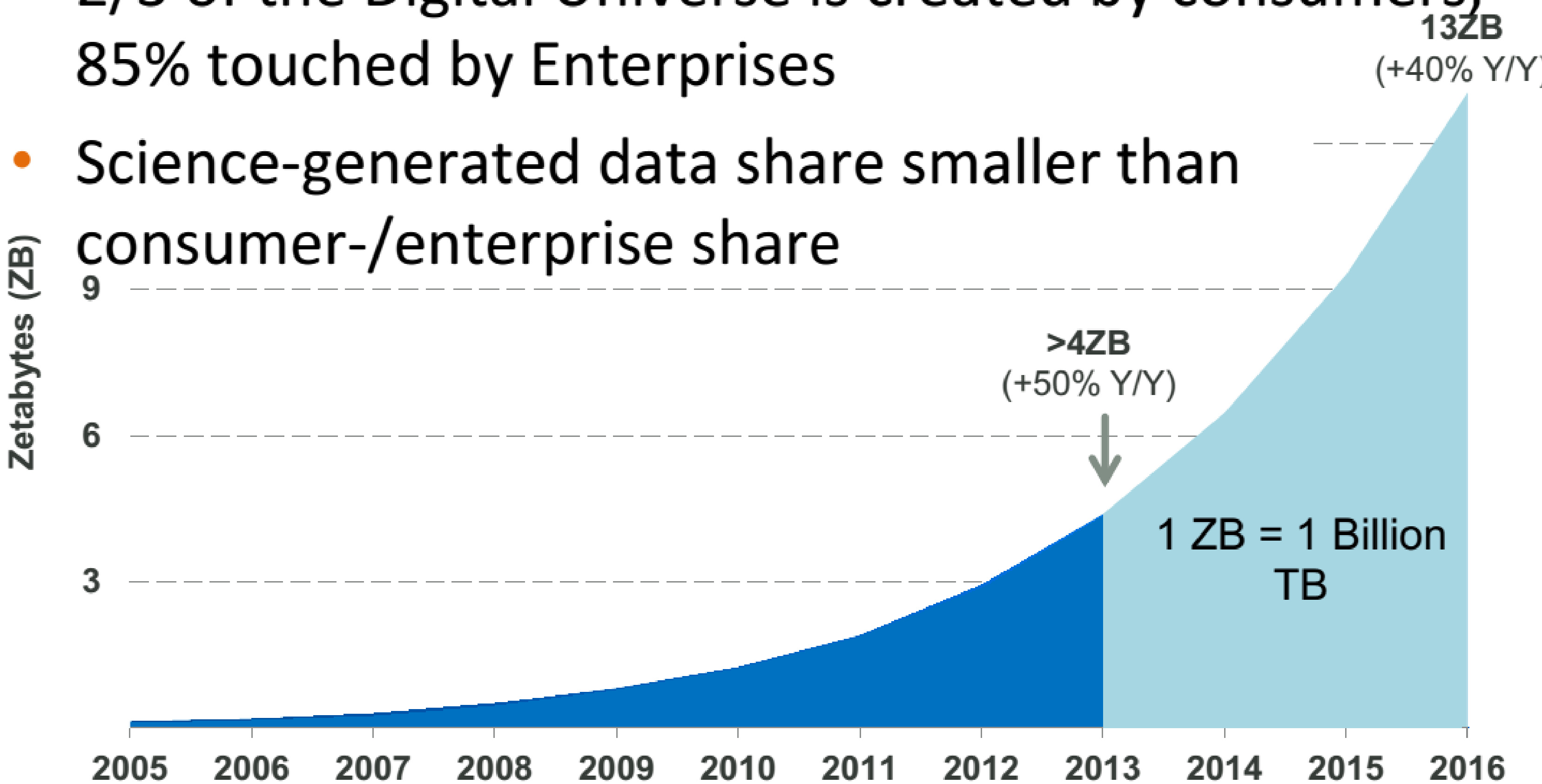
More Data Generations

- GE: In 2020, up to 50 billions devices (many of them industrial machines) will be connected to the internet.

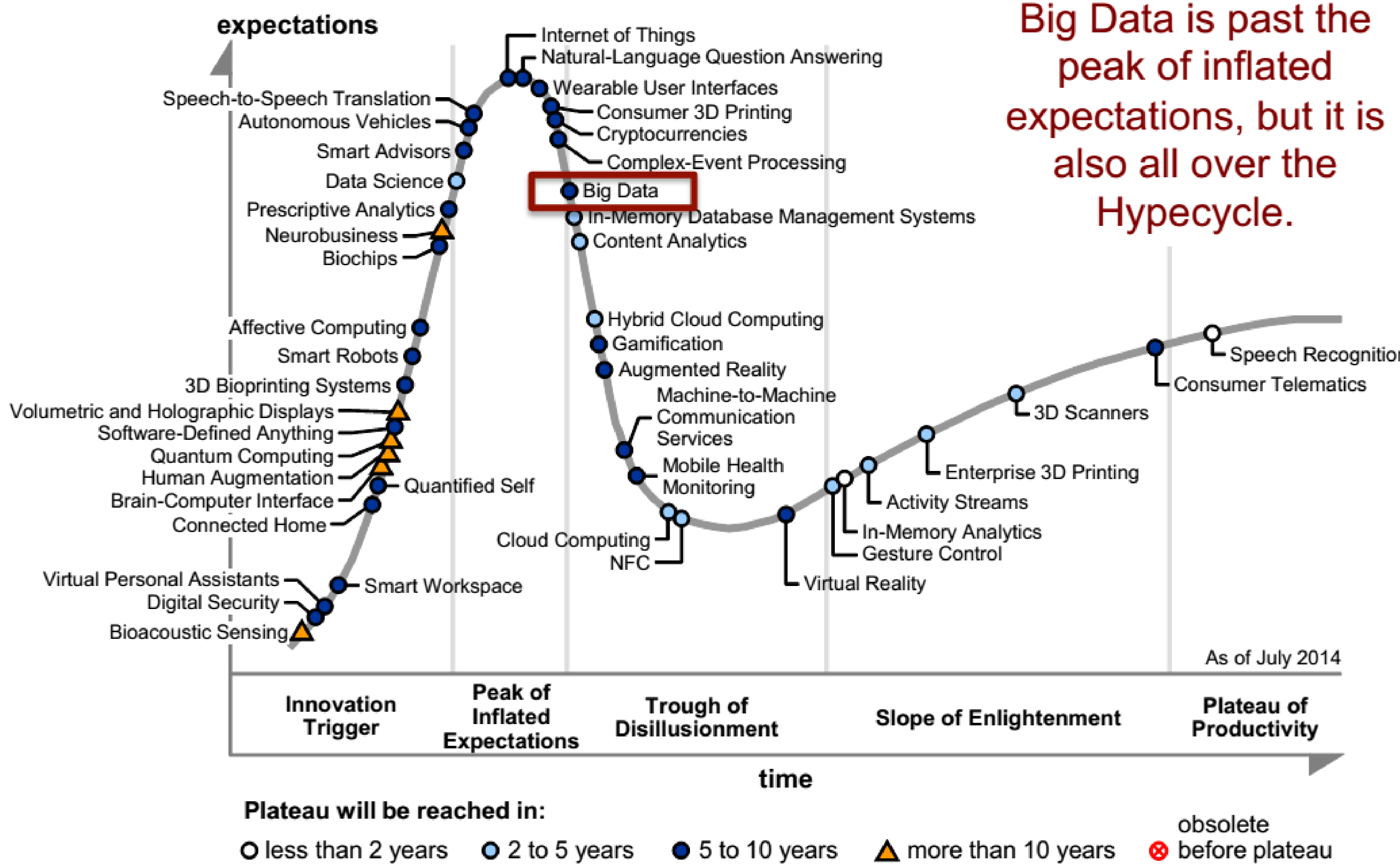
- Amazon: use big data analytic to analyze sale and provide recommendation/suggestion to buyers (“Customers who bought this …”)

- Germany national soccer team: In-depth second-by-second performance analysis of all opposing players (Reduce Germany’s possession time from 3.4s to 1.1s)

- 4.4 ZB digital data created in 2013
- 2/3 of the Digital Universe is created by consumers, 85% touched by Enterprises
- Science-generated data share smaller than consumer-/enterprise share

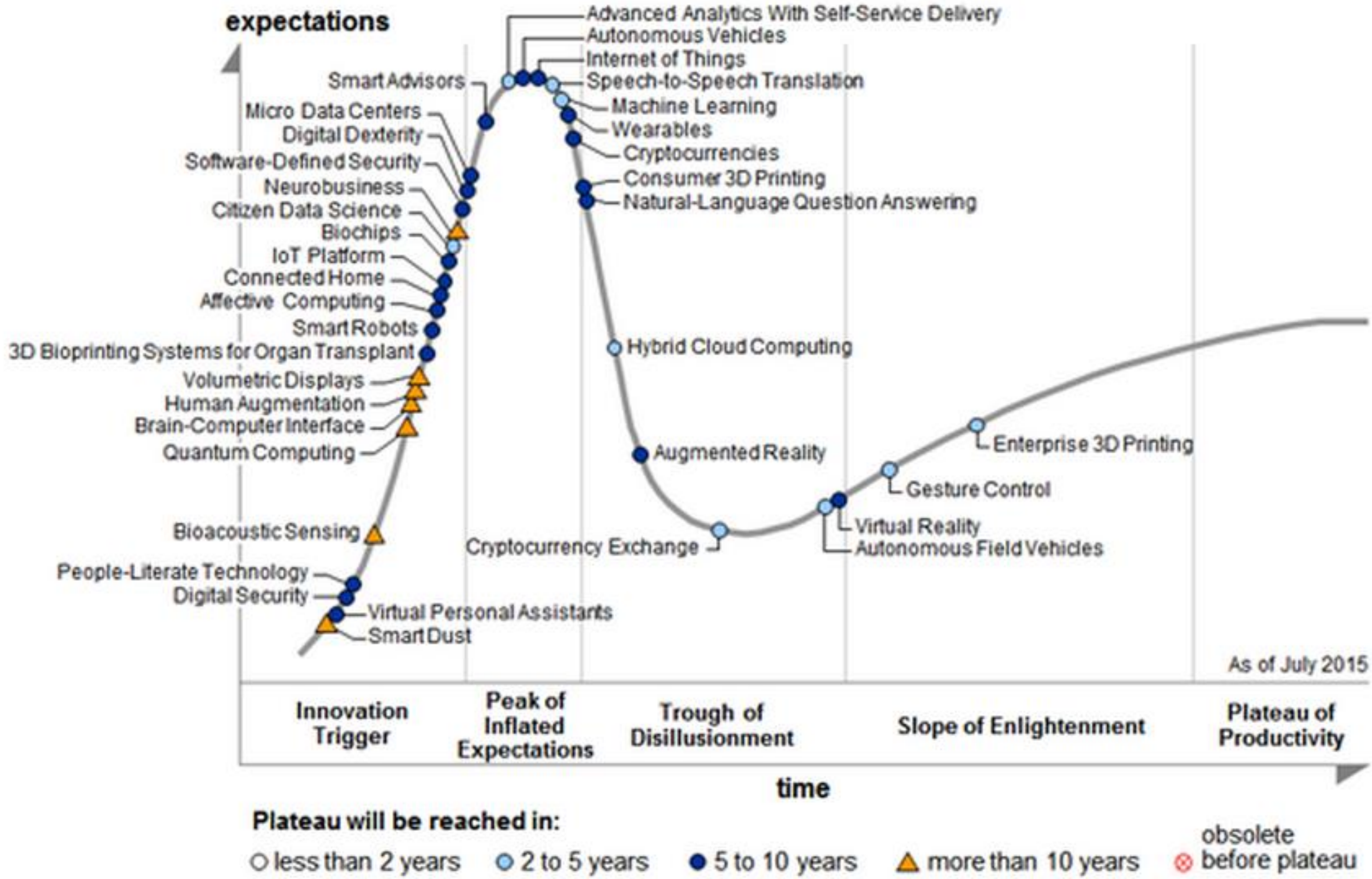


IDC Digital Universe Study, http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf, 2014, KPCB, http://www.kpcb.com/internet-trends, 2014

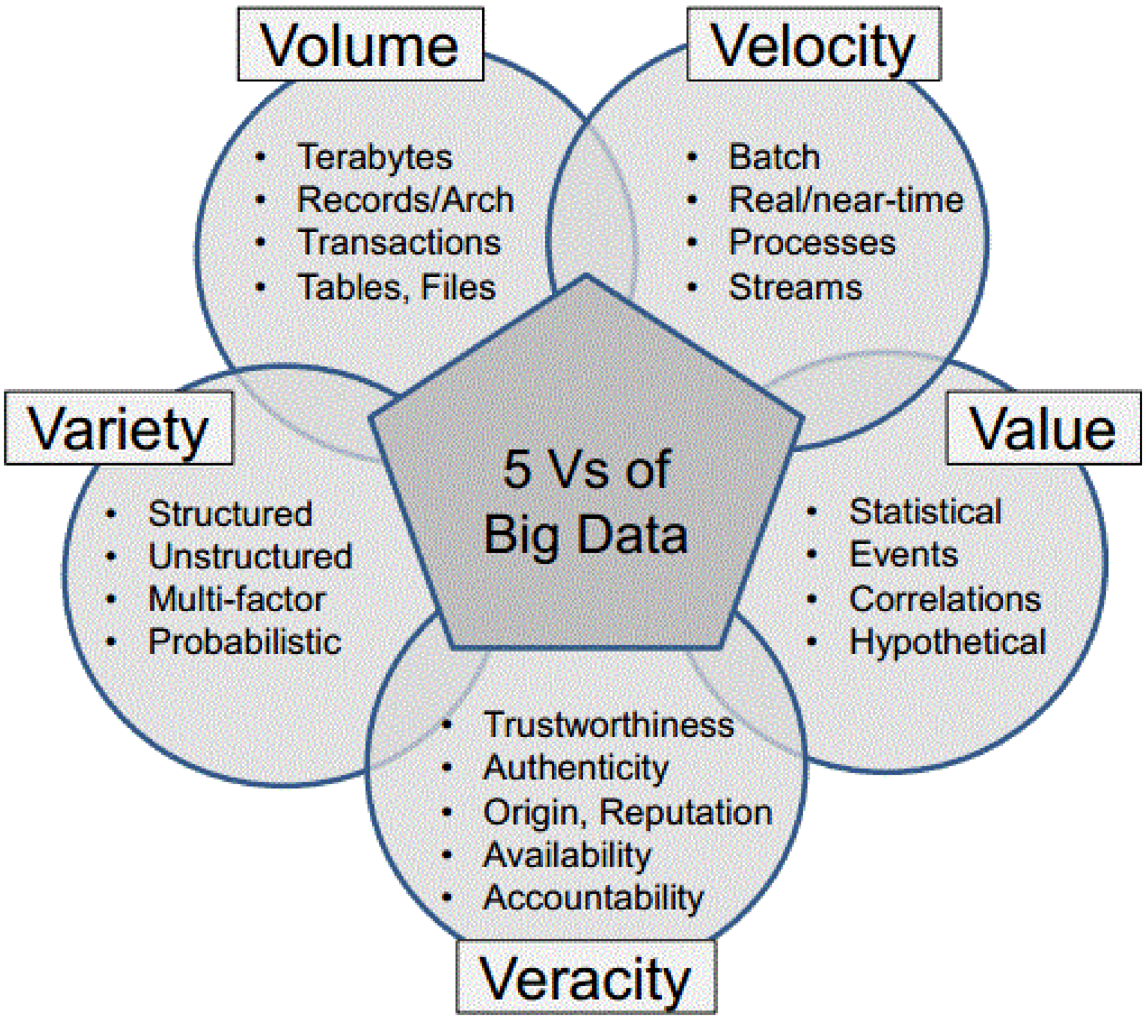


Big Data is past the peak of inflated expectations, but it is also all over the Hypecycle.

Figure 1. Hype Cycle for Emerging Technologies, 2015



The Vs of Big Data



Big Data Applications

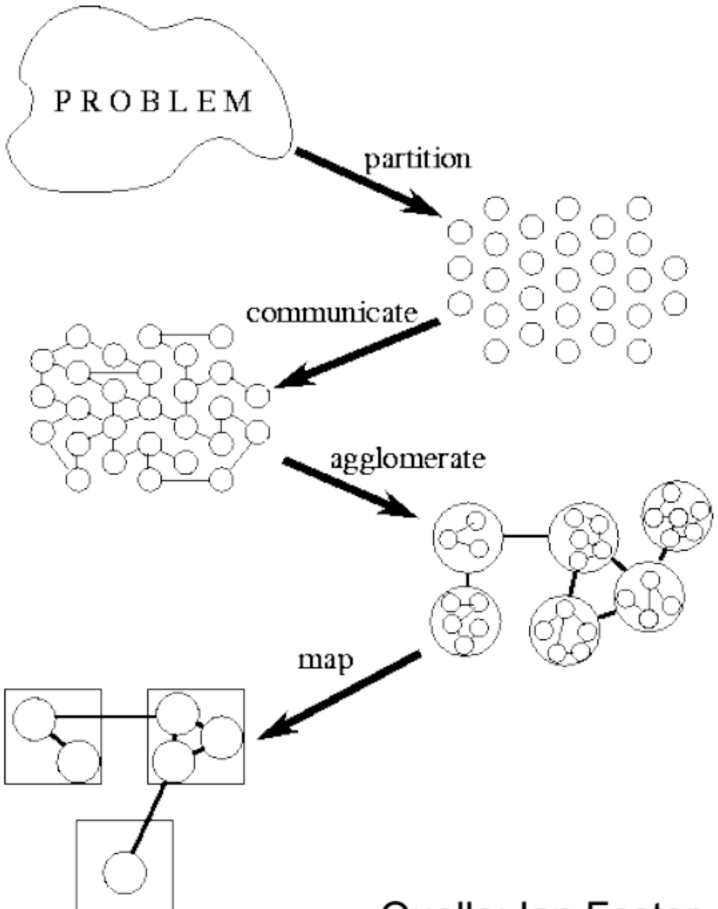
- Consumer Services:
  - Web search, recommendation engines(Amazon, Netflix), Social networks, video analytics (YouTube), Internet of Things (NEST, wearables, fitness tracker, connected vehicles …)
- Industrial Manufacturing:
  - Supply Chain and Logistics, Assembly Quality, Smart Machines
- Government:
  - Census, Archiving, Image Surveillance, Situation Assessment
- Sciences:
  - Genome sequencing, astrophysics, particle physics

Programming Paradigm for Big Data

- Multi-faceted challenges:
  - Require not only parallel computation but also parallel data processing
- New computational tools and strategies
- New data intensive scalable architectures
- Science is moving increasingly from hypothesis-driven to data-driven discoveries
- Industry is at a stage where big data infrastructures are integrated and big data sets are beginning to be analyzed to produce business insights

Alex Szalay, 2012: Extreme Data-Intensive Scientific Computing: The Fourth Paradigm

Parallel Programming:



Partition: Decompose work into tasks that can be executed concurrently

Communicate: Design communication structure depending on which data needs to be exchange how often.

Agglomeration: Aggregate tasks to optimize performance.

Map: Map tasks to processes.

Data Parallelism: partitions data and maps chunks to different processors. Each processor performs the same task on different data.

Quelle: Ian Foster, Designing and Bulding Parallel Programs

Distributed Computing

L1 cache reference	0.5 ns
Branch mispredict	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 USA->Germany->USA	150,000,000 ns

Why is Distributed Computing hard?

8 **Fallacies** of Distributed Computing:

- The network is reliable
- Latency is zero
- Bandwidth is infinite
- The network is secure
- Topology doesn't change
- There is one administrator
- Transport cost is zero
- The network is homogeneous

Data Access is Hitting a Wall

FTP and GREP are not adequate

- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years

- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (= 1 $/GB)
- … 2 days and 1K$
- … 3 years and 1M$

- Oh!, and 1PB ~4,000 disks

- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where **databases** can help

Slide from Jim Gray (2005)

Alex Szalay, 2012: Extreme Data-Intensive Scientific Computing: The Fourth Paradigm

Current Parallel Programming Paradigms

- It is difficult to write parallel programs
  - Difficult in converting algorithms from serial to parallel
  - Difficult in identifying different ways that the program can fail
  - No reliable way to detect failure of a process
- It is even more difficult to write parallel programs at large scale
  - Same set of errors, but scale up with size
- It is even more difficult to debug large scale parallel programs
  - What if the program doesn't fail but only produce incorrect results?

Data-Intensive Approach

- Scale “out”, not “up”
  - It is easier and cheaper to add nodes to an existing cluster than to build a faster cluster.
- Move computation to the data
  - Reduce data movement.
- Sequential processing, avoid random access
  - Reduce seek movement on disks.
- Seamless scalability

Common Analysis Pattern

- Huge amounts of data, aggregates needed
  - *But also need to keep raw data*
  - *Need for parallelism*
  - *Heavy use of structured data, multi-D arrays*
- Requests enormously benefit from indexing
- Computations must be close to the data!
- Very few predefined query patterns
  - *Everything goes….*
  - *Rapidly extract small subsets of large data sets*
  - *Geospatial/locality based searches everywhere*
- Data will never be in one place
  - *Remote joins will not go away*
- No need for transactions
- Data scrubbing is crucial

DB!

Alex Szalay, 2012: Extreme Data-Intensive Scientific Computing: The Fourth Paradigm

Increased Diversification

One shoe does not fit all!

- Diversity grows naturally, no matter what
- Evolutionary pressures help
- Individual groups want specializations

At the same time

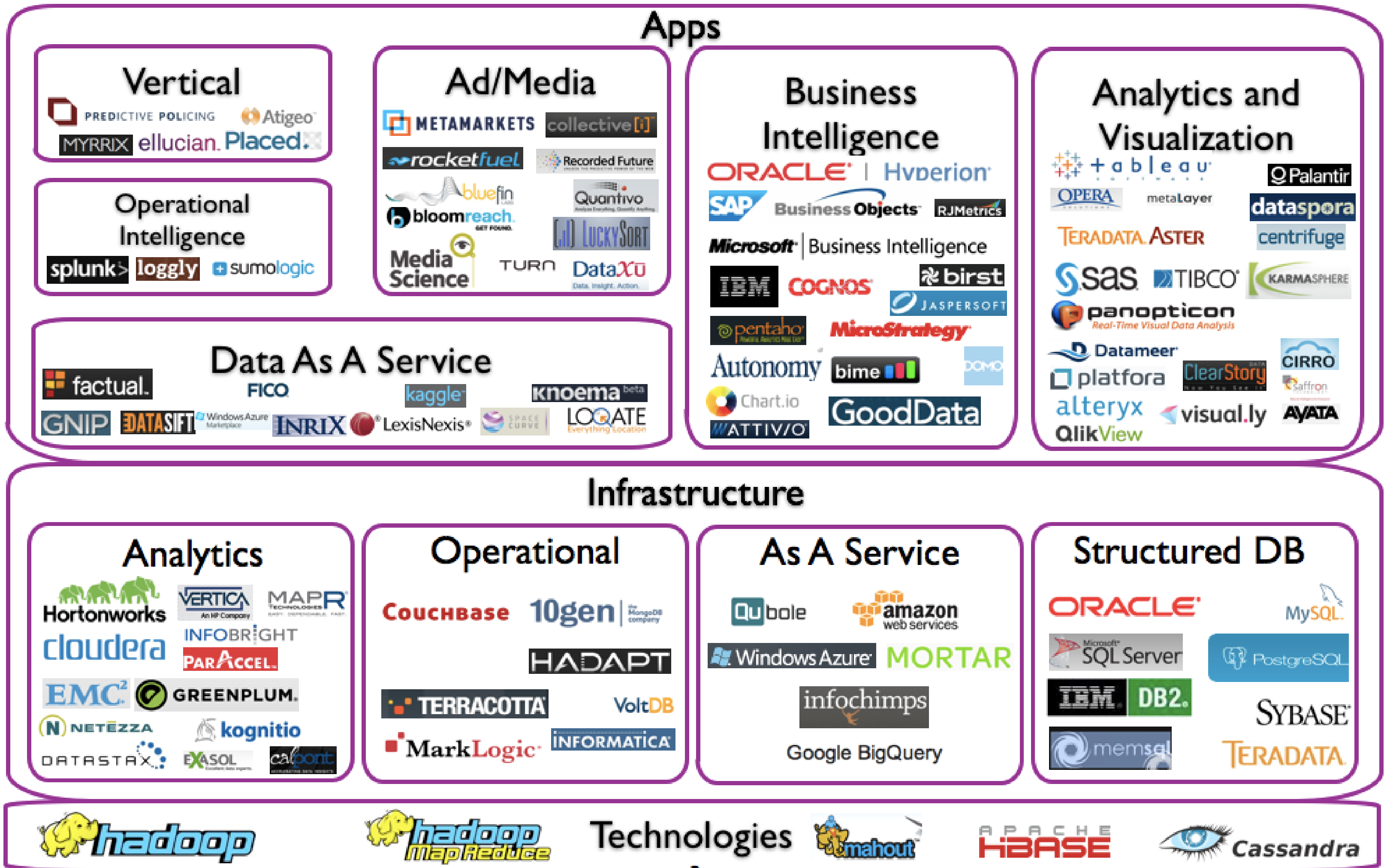
- What remains in the middle?
  - Common denominator is Big Data
- Data management
  - Everybody needs it, nobody enjoys doing it
- We are still building our own… over and over

- Large floating point calculations move to GPUs
- Big data moves into the cloud (private or public)
- RandomIO moves to Solid State Disks
- High-Speed stream processing emerging
- noSQL vs databases vs column store vs SciDB …

Alex Szalay, 2012: Extreme Data-Intensive Scientific Computing: The Fourth Paradigm

CLEMSON UNIVERSITY

The Big Data Landscape



dave@vcdave.com

bigdatalandcsape.com

CLEMSON UNIVERSITY