CentraleSupelec 2018-2019 MSC DSBA / DATA SCIENCES

Big Data Algorithms Techniques & Platforms

Hugues Talbot

ESSEC BUSINESS SCHOOL

CentraleSupélec

Introduction

- Course set in collaboration with Pr. Céline Hudelot, department of mathematics, CentraleSupélec.
- 25h, lectures and project.

Objective of the course

- *Big Data* and data-intensive information processing.
- Algorithms that scale on Big data and programming paradigms.
- Distributed computing strategies (e.g. Map Reduce) - Distributed File
- Systems - Distributed Access Structures
- Basic practice on some Big Data platforms (Hadoop, Spark,
- Cassendra, AWS...)

Essence of the course

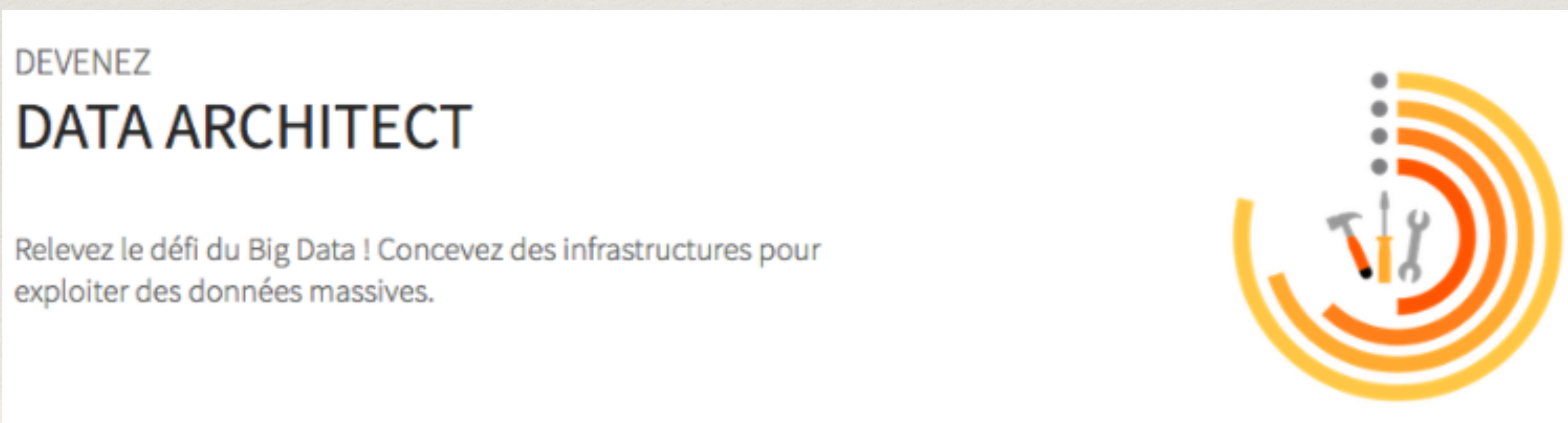
- Small introductions to the main concepts.
- Some references to understand deeper.
- Practice to learn (Confucius: I hear, I forget ; I see, I remember ; I practice, I understand).

Prerequisite

- Knowledge on Programming and Advanced Programming.
- IS1220BC : Object oriented Software design http://cours.etudes.ecp.fr/claroline/course/index.php?cid=TI1220
- Knowledge on Algorithm Design and Data structures.
- Knowledge on Database systems : SQL, relational algebra, ACID
- properties.
- IS1210 : Introduction aux bases de donnees https://chewbii.com/is1230/

Syllabus

1. Part 1 : Object-oriented programming in JAVA : H. Talbot, C. Hudelot, P. Ballarini, MICS, CentraleSupélec
2. Part 2 : Distributed Computing : Map Reduce - Hadoop
3. Part 3 : No SQL Nicolas Travers, Assistant Professor, CNAM, http://chewbii.com/
4. Part 4 : Stream Computing : Real-time Processing of Massive Data ; Spark, Mlib Regis Behmo, Data Architect https://fr.linkedin.com/in/regisb/fr

- Data Architect path in progress in OpenClassRoom with the team of this course (**in French**) https://openclassrooms.com/paths/data-architect
  - Follow it, Become a Mentor : http://jobs.openclassrooms.com/o/mentor--parcours-data-architect
  - 

Tiny quiz: explain code below

- 1. class HumptyDumpty
- 2. {
- 3. void myMethod() {}
- 4. }
- 5.
- 6. class HankyPanky extends HumptyDumpty
- 7. {
- 8. public void myMethod() {}
- 9. }

Who are these two people ?

Have you hear of Hadoop?

What does this number represent?

- 40 000 000 000 000 000 000 000

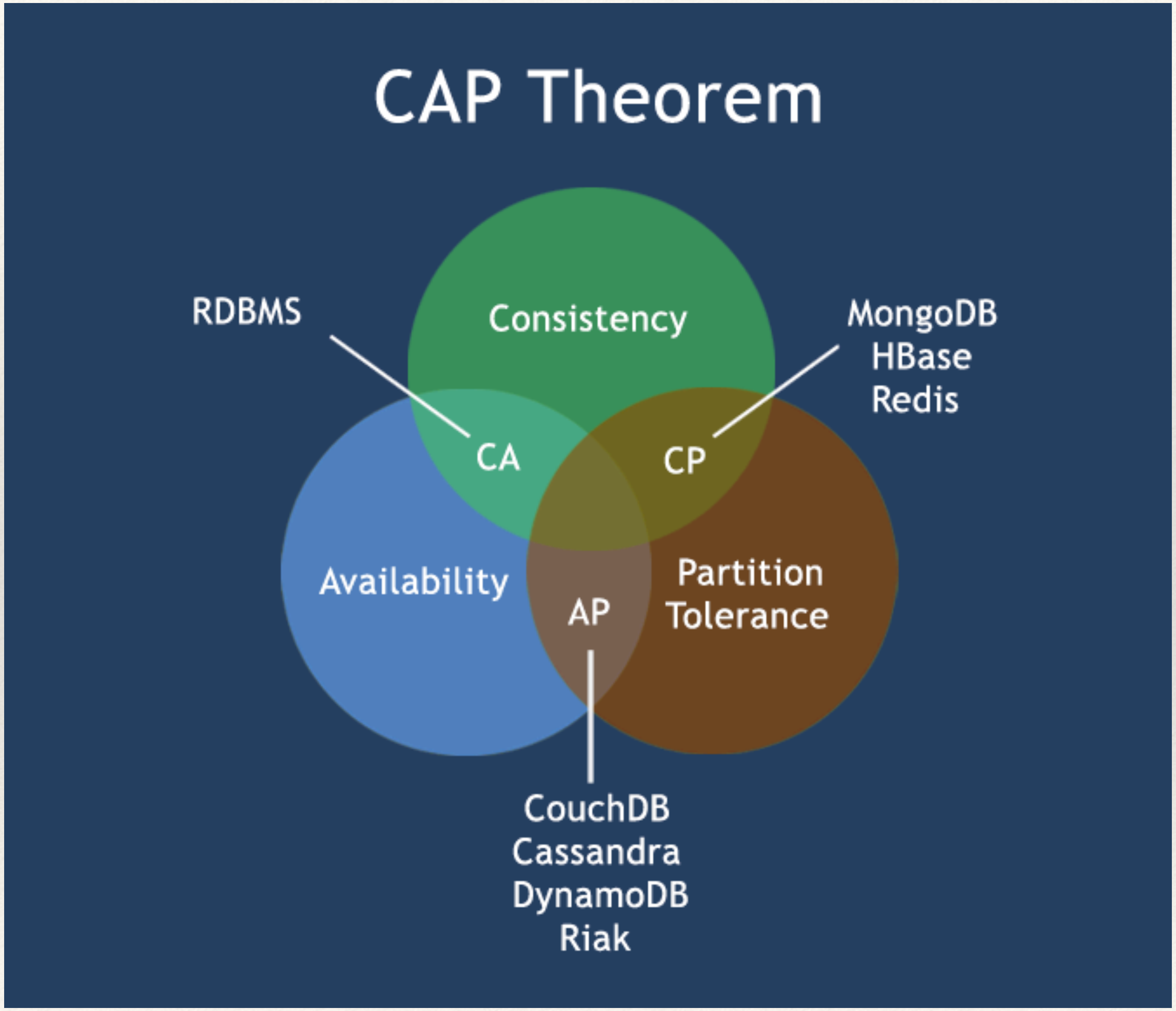
What is big-data ?

What is big-data ?

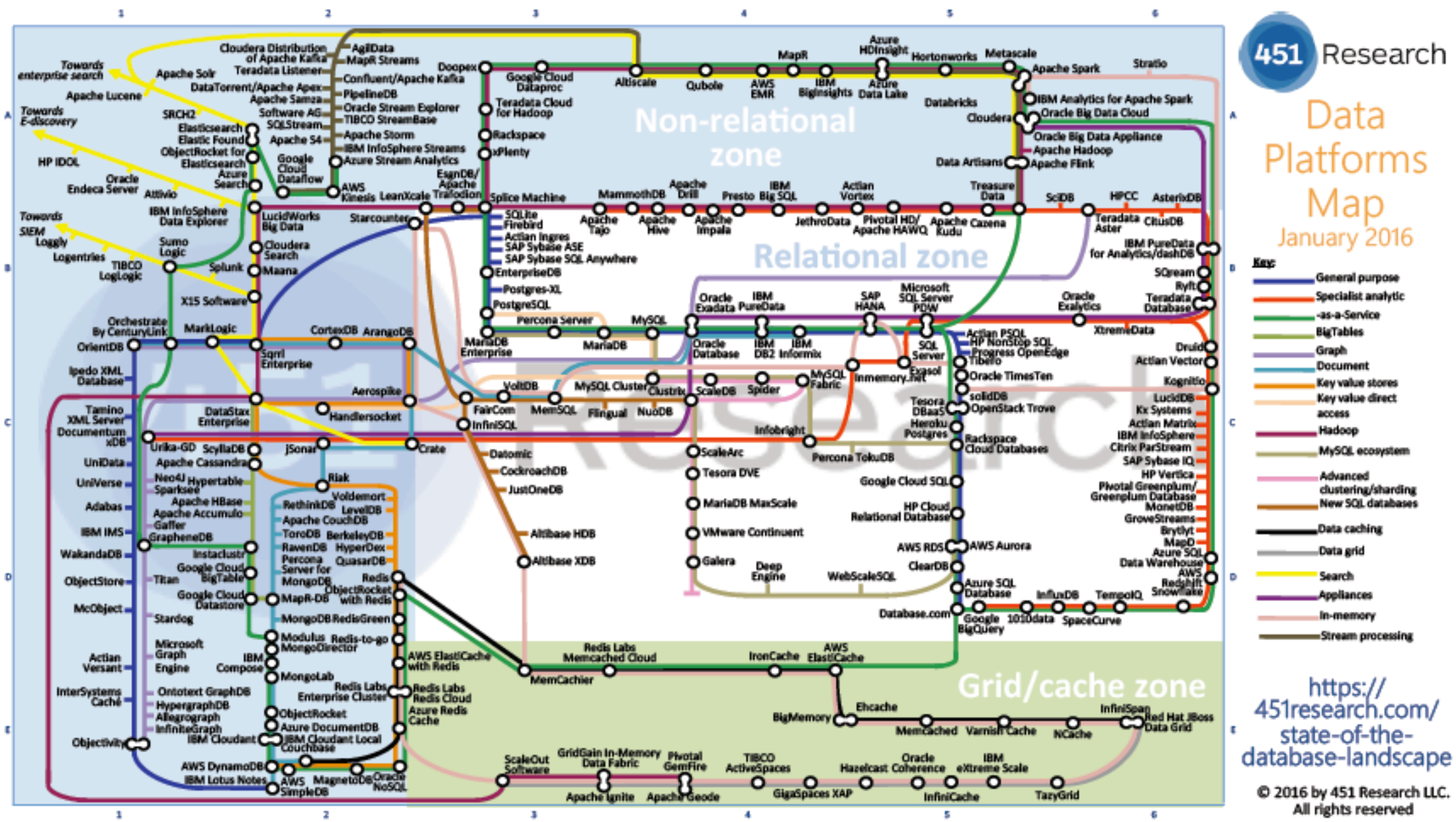
What is the difference between CASSENDRA, Mongo-DB and Neo4J ?

What is the CAP theorem?

What is the CAP theorem?



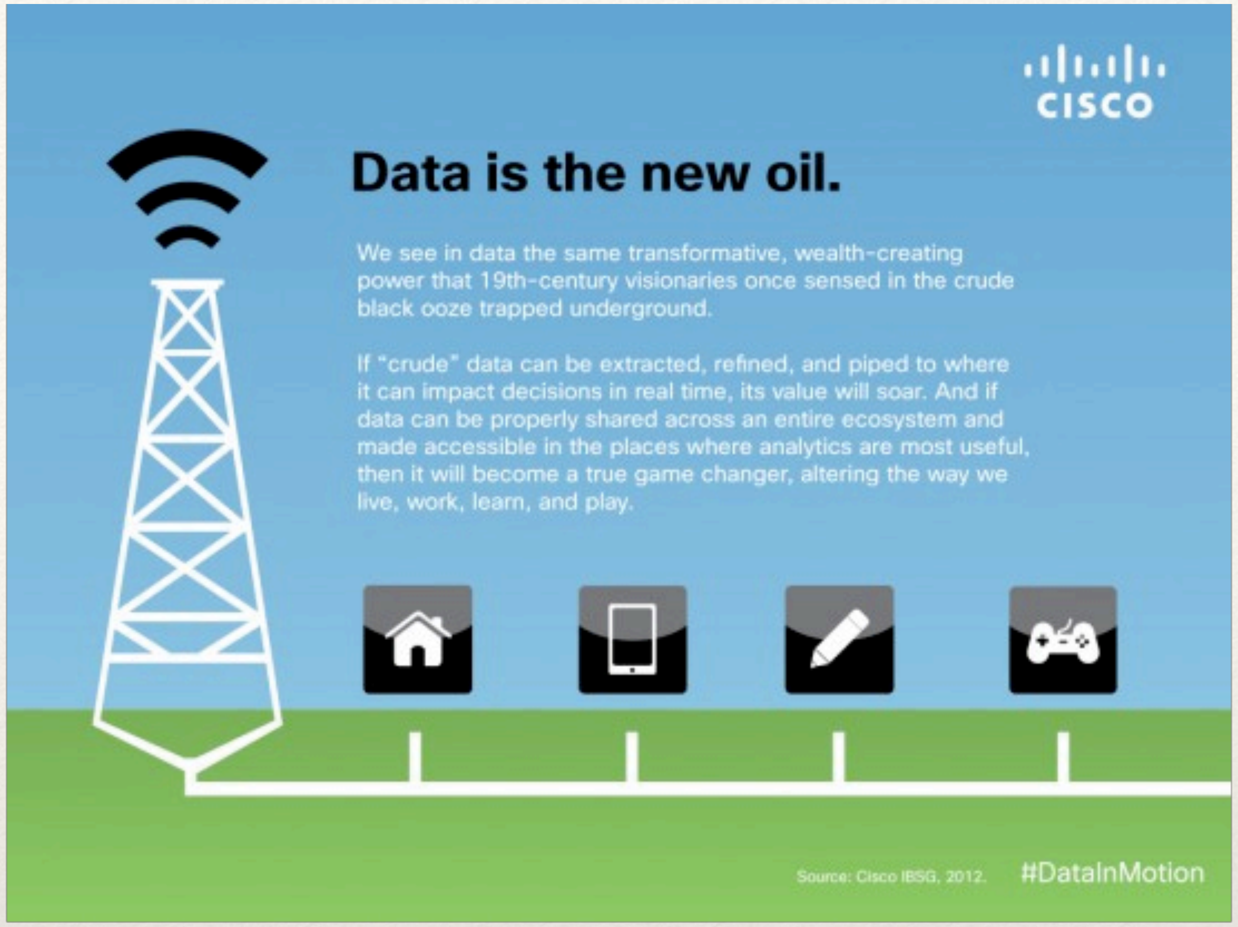
Can you travel on that map?



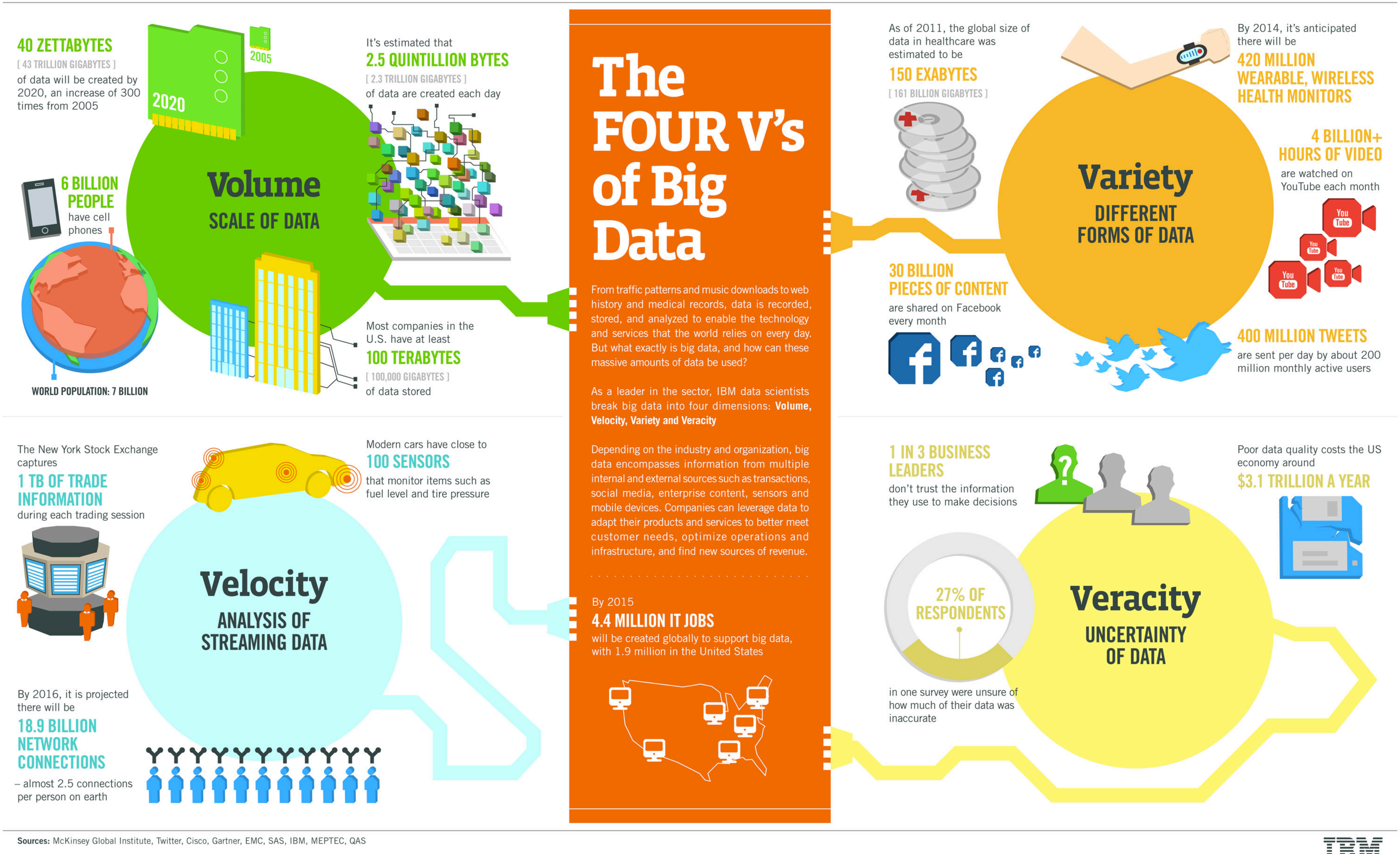
Big Data: data is everywhere

- Massive data are collected and warehoused.
  - Web data, e-commerce
  - Bank/ Credit Card transactions or other card transactions (e.g. navigo pass)
  - Social network.
  - Internet of Things.
  - but also *scientific data* .

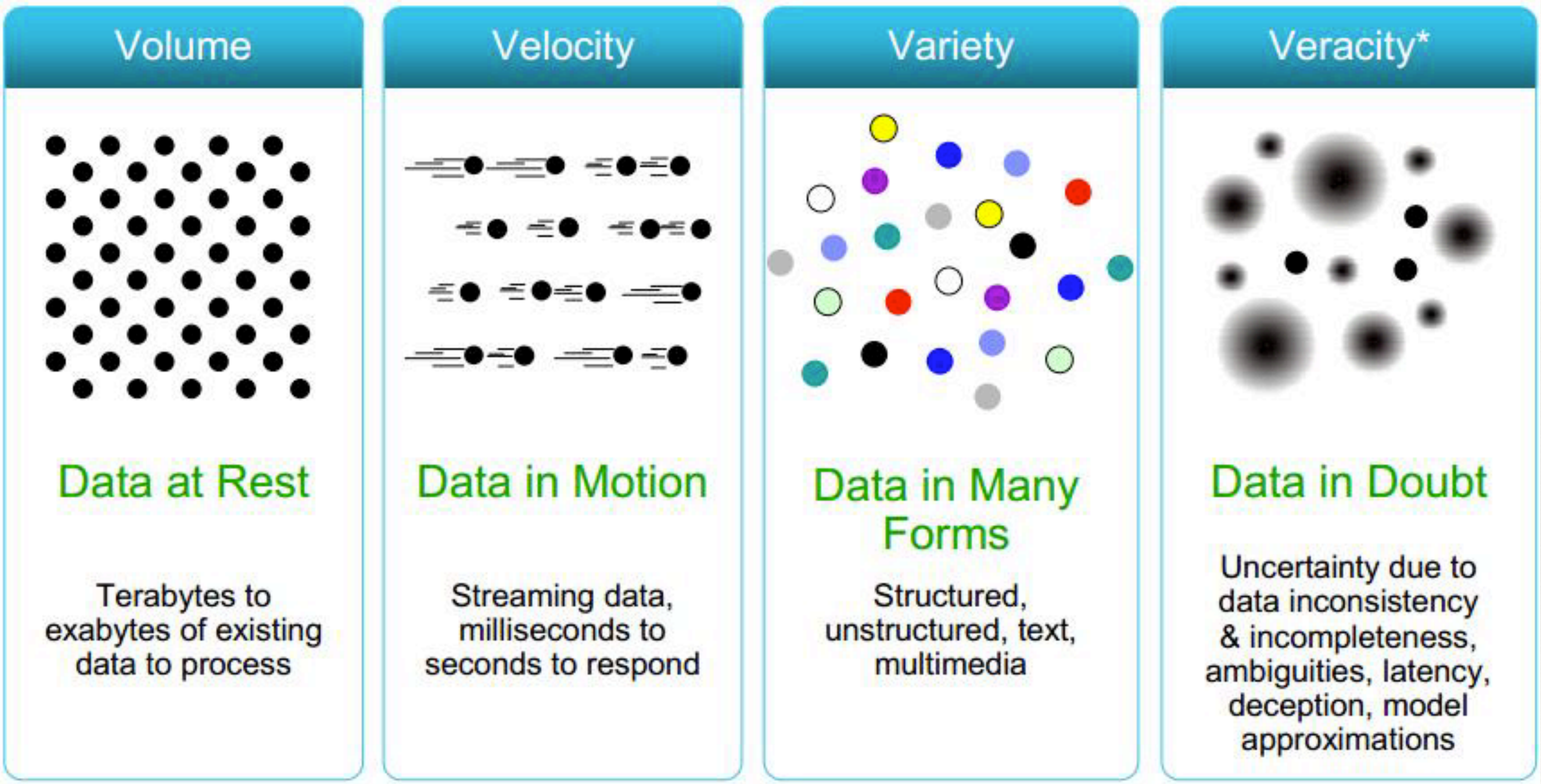

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate p.52
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins? p.38
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics p.69
PLUS
Juan Enriquez Reprograms Life p.31
James Gleick Unsplits the Bit p.58
AND Lawrence Weschler Questions the Cloud p.76
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE


CISCO
Data is the new oil.
We see in data the same transformative, wealth-creating power that 19th-century visionaries once sensed in the crude black ooze trapped underground.
If "crude" data can be extracted, refined, and piped to where it can impact decisions in real time, its value will soar. And if data can be properly shared across an entire ecosystem and made accessible in the places where analytics are most useful, then it will become a true game changer, altering the way we live, work, learn, and play.
Source: Cisco IBSG, 2012.
#DataInMotion

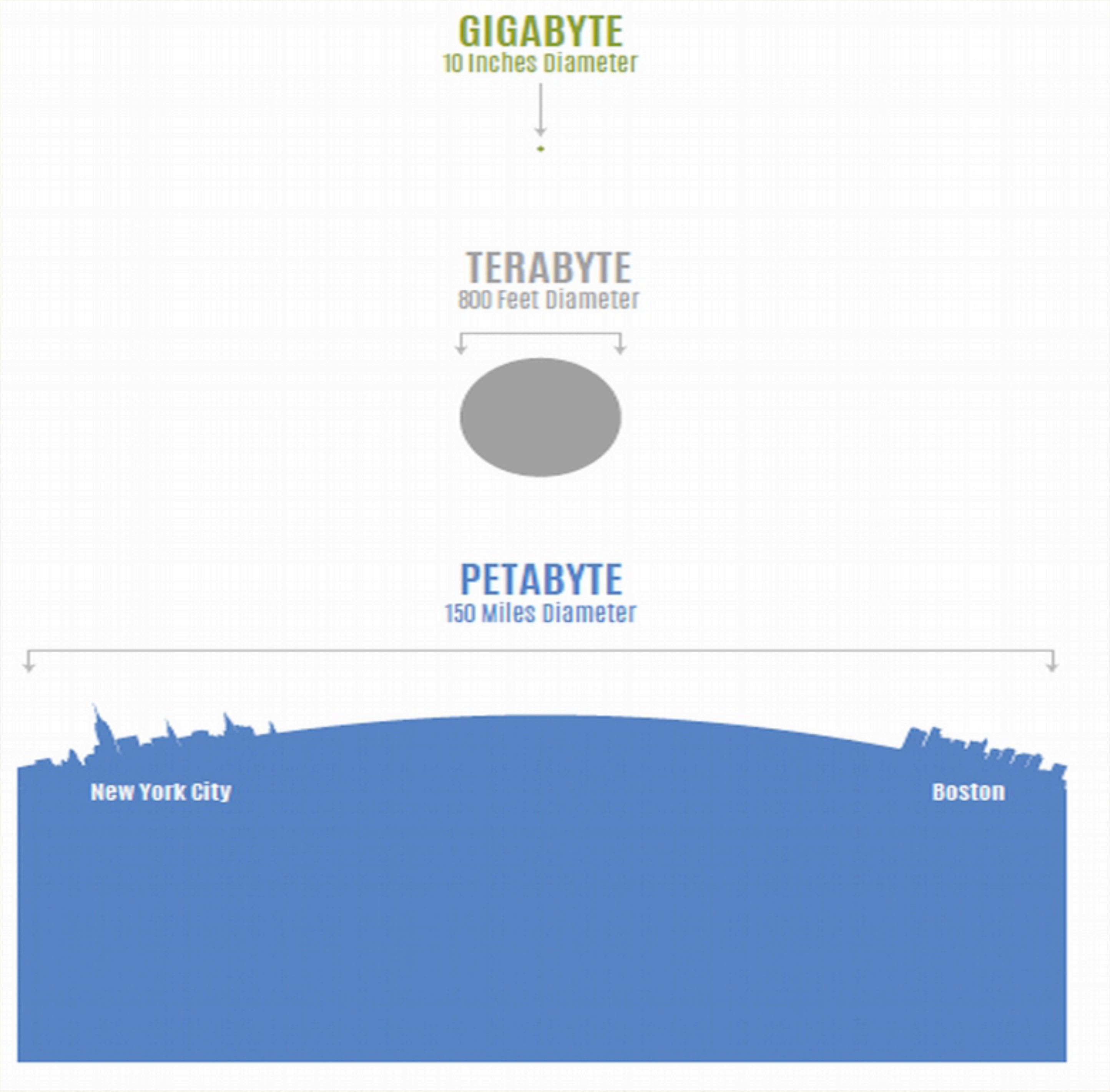

The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it is going to be a hugely important skill in the next decades.
Hal Varian, Chief Economist, Google
@kyleplacy
#NSD2013



The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
Volume
SCALE OF DATA
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

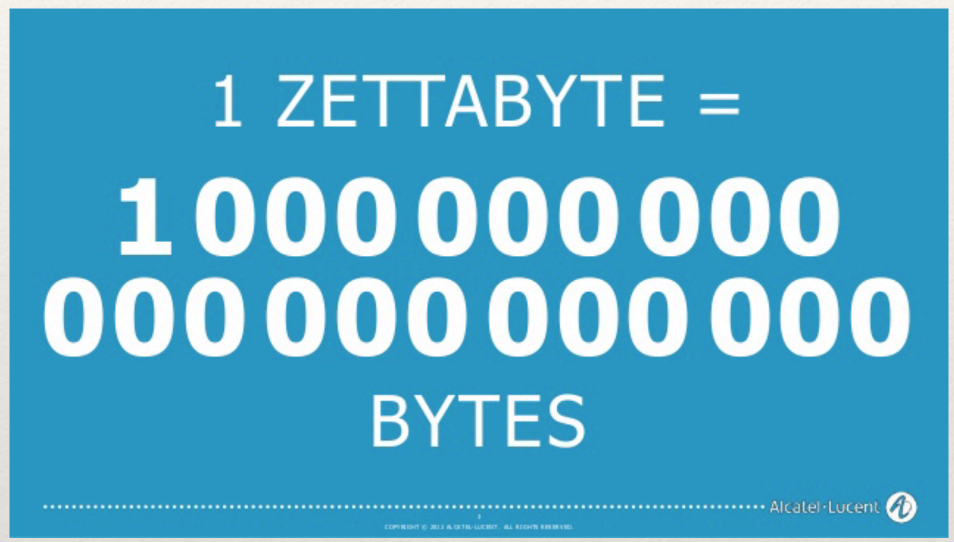

Volume
Velocity
Variety
Veracity*
Data at Rest
Data in Motion
Data in Many Forms
Data in Doubt
Terabytes to exabytes of existing data to process
Streaming data, milliseconds to seconds to respond
Structured, unstructured, text, multimedia
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Volume


GIGABYTE
10 Inches Diameter
TERABYTE
800 Feet Diameter
PETABYTE
150 Miles Diameter
New York City
Boston

Volume

- In 2020, a total of 40 zettabytes of data on the web produced per year



- Scientific installations :
  - The radiotelescope Square Kilometre Array will generate 50 TB of *reduced* data per day ; with 7 000 TB of raw data per second
  - The LSST will produce 100 Petabyte of reduced data over a 10-year period starting in 2020.

Volume

- Today:
  - 150 millions emails every minute
  - Facebook : <u>4000 TB / day</u> (i.e, 4 PB)
  - CERN, LHC : 15 PB / year (Source : Wikipedia)

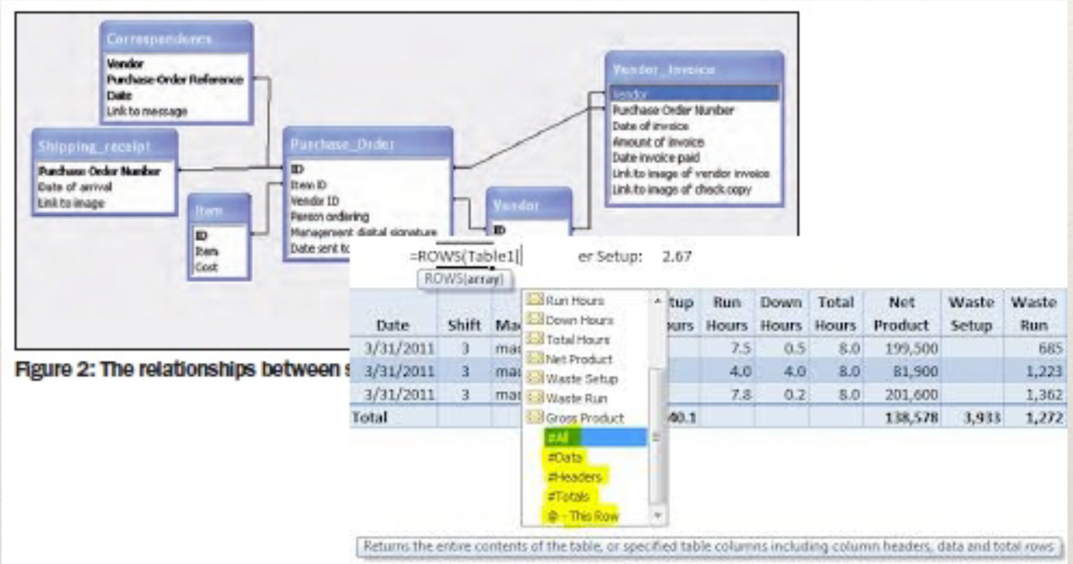
Dealing with volume

- Capacity of a big server
  - Memory : 256 GB
  - Disk storage : 24 TB
  - Disk speed : 100 MB / s

-

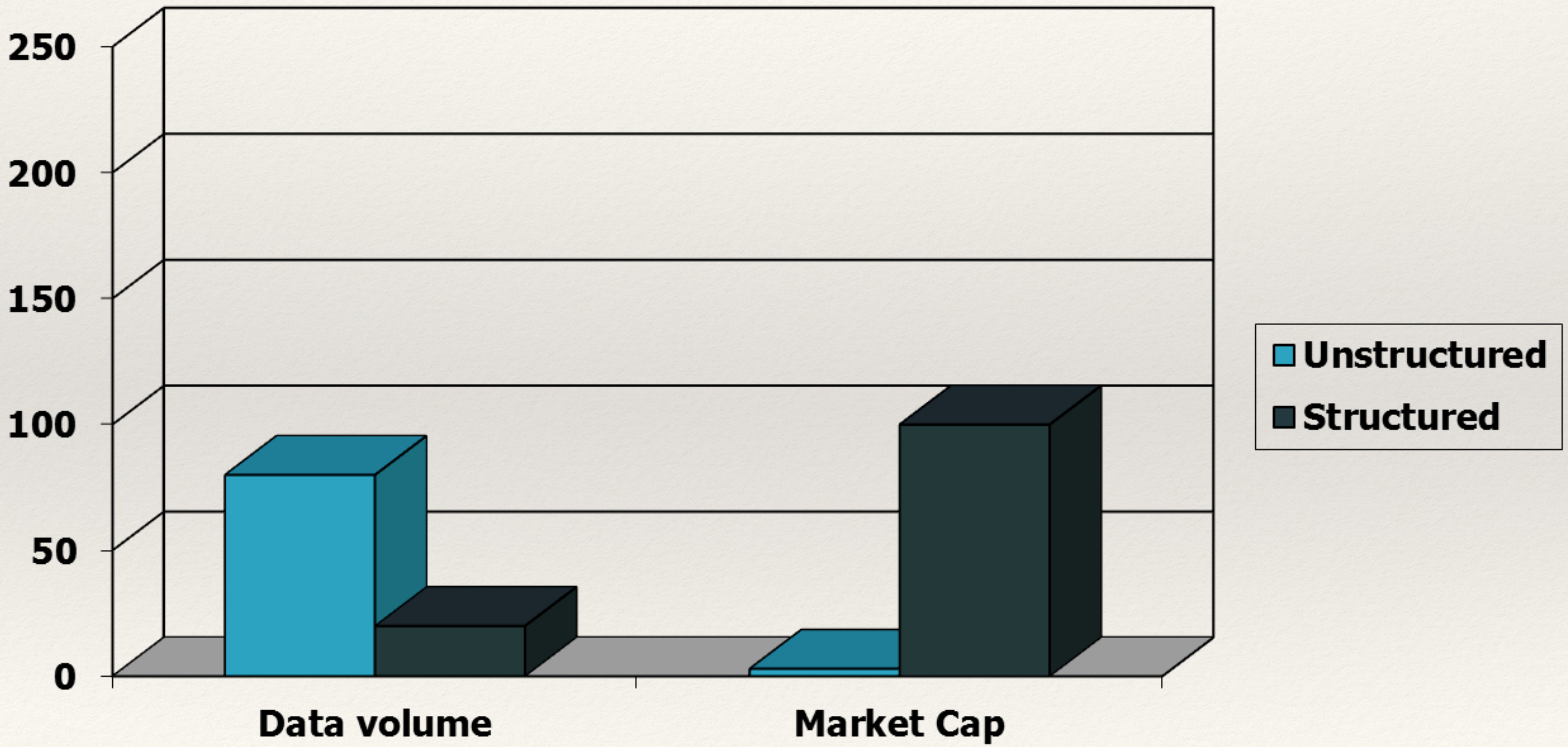
Variety

- The type and nature of the data : structured and unstructured data
- *Structured data*
  - Data with a level of organization.
  - e.g. : databases, excel sheets, ...



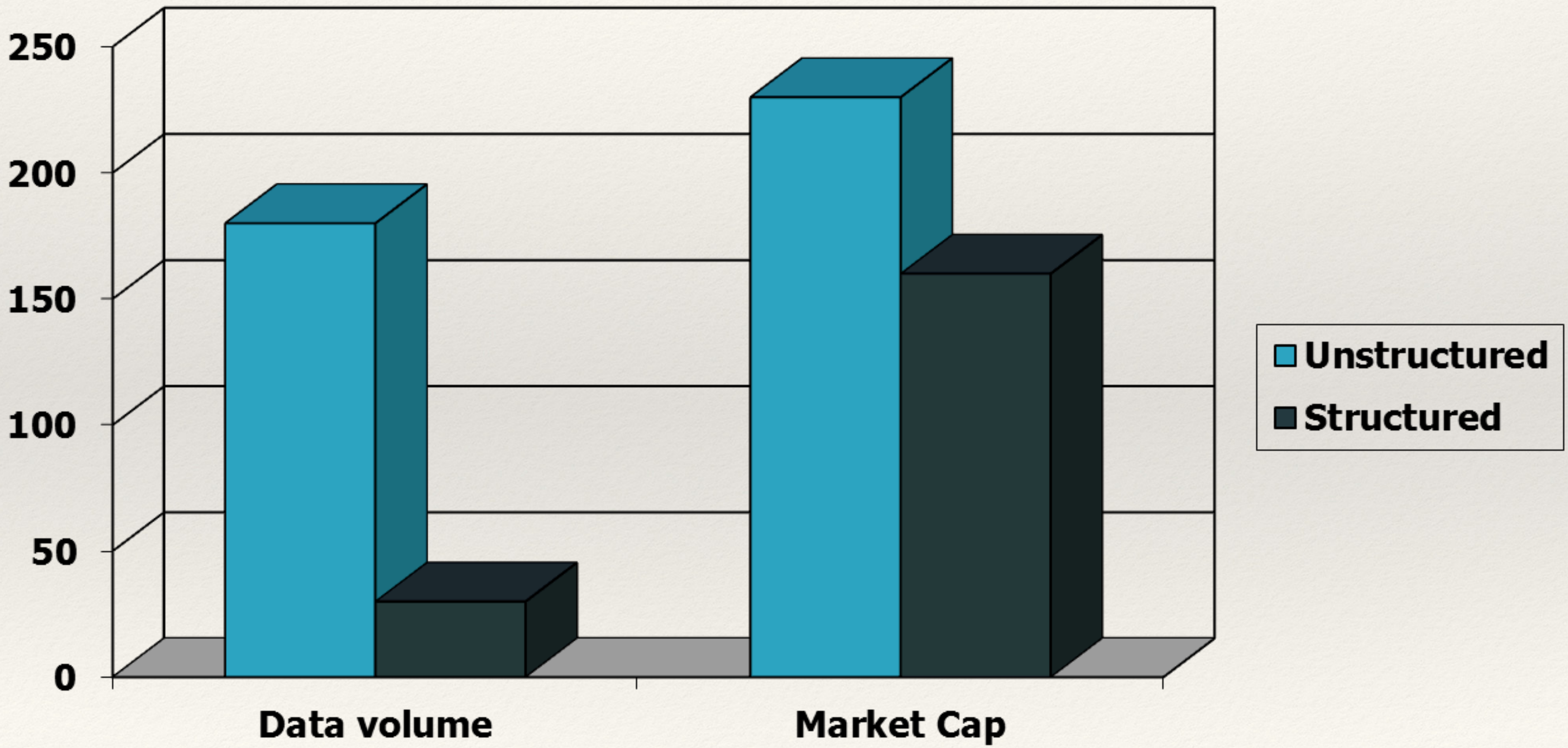
- *Unstructured data*
  - Without strong structuration
  - e.g. : emails, documents,
  - images, social network data...

In 1990



250
200
150
100
50
0
Data volume
Market Cap
Unstructured
Structured

In 2005



250
200
150
100
50
0
Data volume
Market Cap
Unstructured
Structured

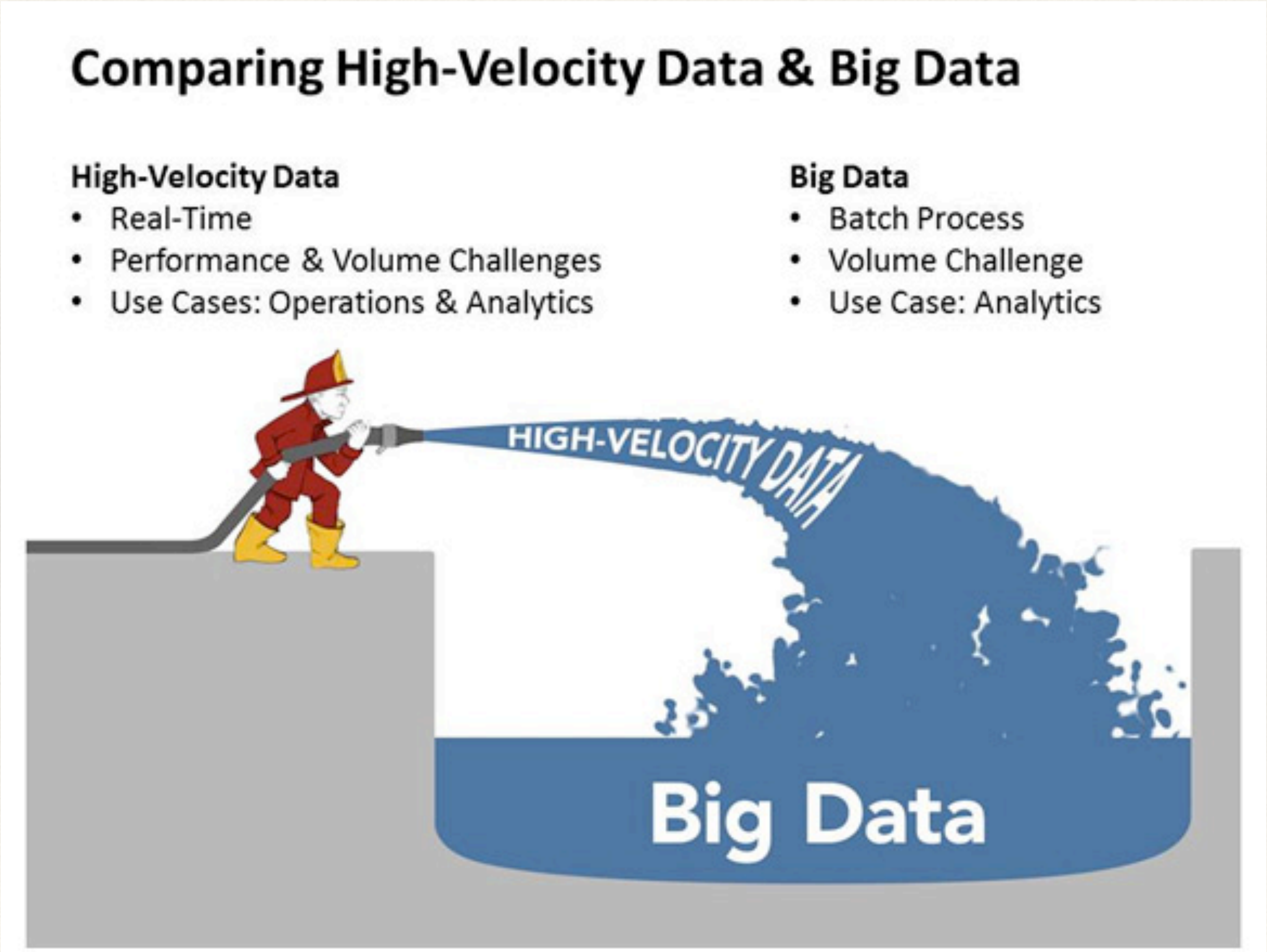
Variety

- Non-structured data
  - 175 millions tweets per day.
  - 571 new websites every minute.
  - 2.5 exabytes of data per day.
  - Source


How big is a
Yottabyte?
TERABYTE
Will fit 200,000 photos or mp3 songs on a single 1 terabyte hard drive.
PETABYTE
Will fit on 16 Backblaze storage pods racked in two datacenter cabinets.
EXABYTE
Will fit in 2,000 cabinets and fill a 4 story datacenter that takes up a city block.
ZETTABYTE
Will fill 1,000 datacenters or about 20% of Manhattan, New York.
YOTTABYTE
Will fill the states of Delaware and Rhode Island with a million datacenters.
The Cost
The cost of buying a 1 terabyte hard drive today is $100. It would cost $100 Trillion dollars to buy a yottabyte of storage for just the hard drives.
YOTTABYTE
$14 Trillion United States GDP in 2008
$18 Trillion European GDP in 2008
$61 Trillion World's GDP
$100 Trillion for a Yottabyte
BACKBLAZE

Velocity

- Speed at which data is generated and processed.

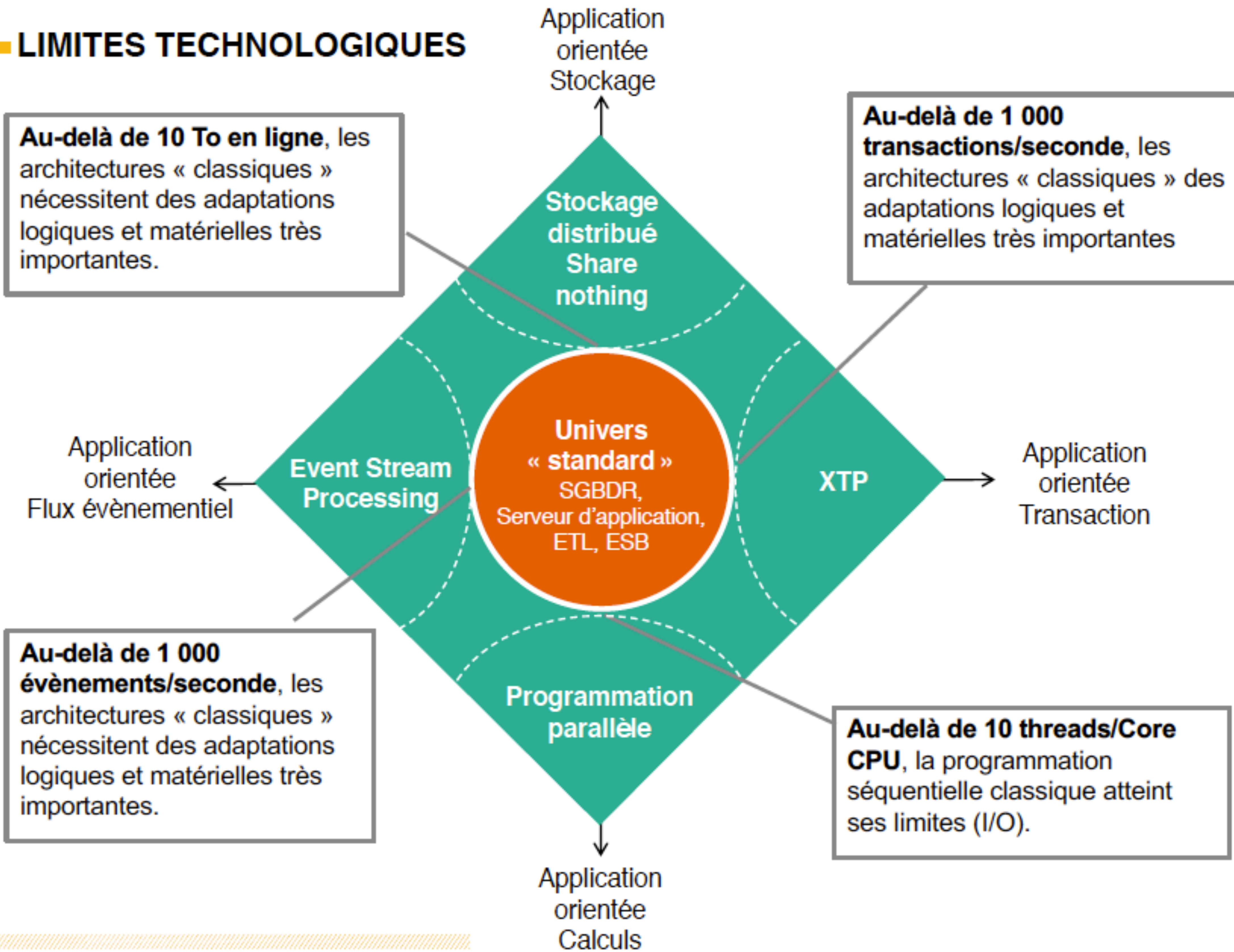


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban dictionary
1,600+ READS ON Scribd.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
You Tube
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
GO-Globe.com web technologies

Cost of velocity

- The measure of the eye blink (User experience)
  - Amazon : increase of more than 100 ms of the latency => -1 % in sales
  - Google : more than 500 ms at the loading => 20 % loss in traffic
  - Yahoo : more than 400 ms at the loading => + 5 to 9 % of cancelations (rebound)
  - Bing :more than 1 second at the loading => -2,8 % of ad revenue.

LIMITES TECHNOLOGIQUES



Processing big data

- Solution : parallelism
  - 1 server
    - 8 disks
    - Read the web : 230 days
  - Cluster Hadoop Yahoo
    - 4000 servers with 8 disks
    - Read the web : 1h20

Problems with this approach:

- Some problems
  - Synchronization.
  - Programming models (share memory, message passing (MPI))
  - Scalability and elasticity (arbitrary numbers of nodes)
  - Fault Tolerance.

Solutions:

- How do we get data for computation ?
  - Solution 1 : Move data to computation ?
  - Solution 2 : Move computation to the data ?
- Solution 2 : not enough RAM to hold all the data in memory and prevent slow disk access
  - Data is stored on the local disks of nodes in the cluster.
  - The programs are started up on the node that has the data local.
- Distributed File Systems : GFS, HDFS.

Distributed computing models

- How do we design algorithms for distributed computing ?
  - Generic programming models : design patterns.
  - MapReduce