# Big Data Algorithms, Techniques and Platforms
# Lecture 1 : Course introduction

Céline Hudelot, Professor, CentraleSupélec

2017-2018

# What this course is about ?

- 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 on the main concepts.
  - Some references to go deeper.
- Practice to learn.

# Prerequesites

- 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 données
  `https://chewbii.com/is1230/`

# Course Overview

1. Part 1 : Object-oriented programming in JAVA
   C. Hudelot, P. Ballarini, MICS, Centrale Supélec
2. Part 2 : Distributed Computing : Map Reduce - Hadoop
   C. Hudelot
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
   Régis 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`

DEVENEZ

**DATA ARCHITECT**

Relevez le défi du Big Data ! Concevez des infrastructures pour exploiter des données massives.

# Plan

## Can you explain the 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 guys ?

## What is Hadoop ?

What represents this number ?

40000000000000000000000

## What is big data ?

# What is big data ?

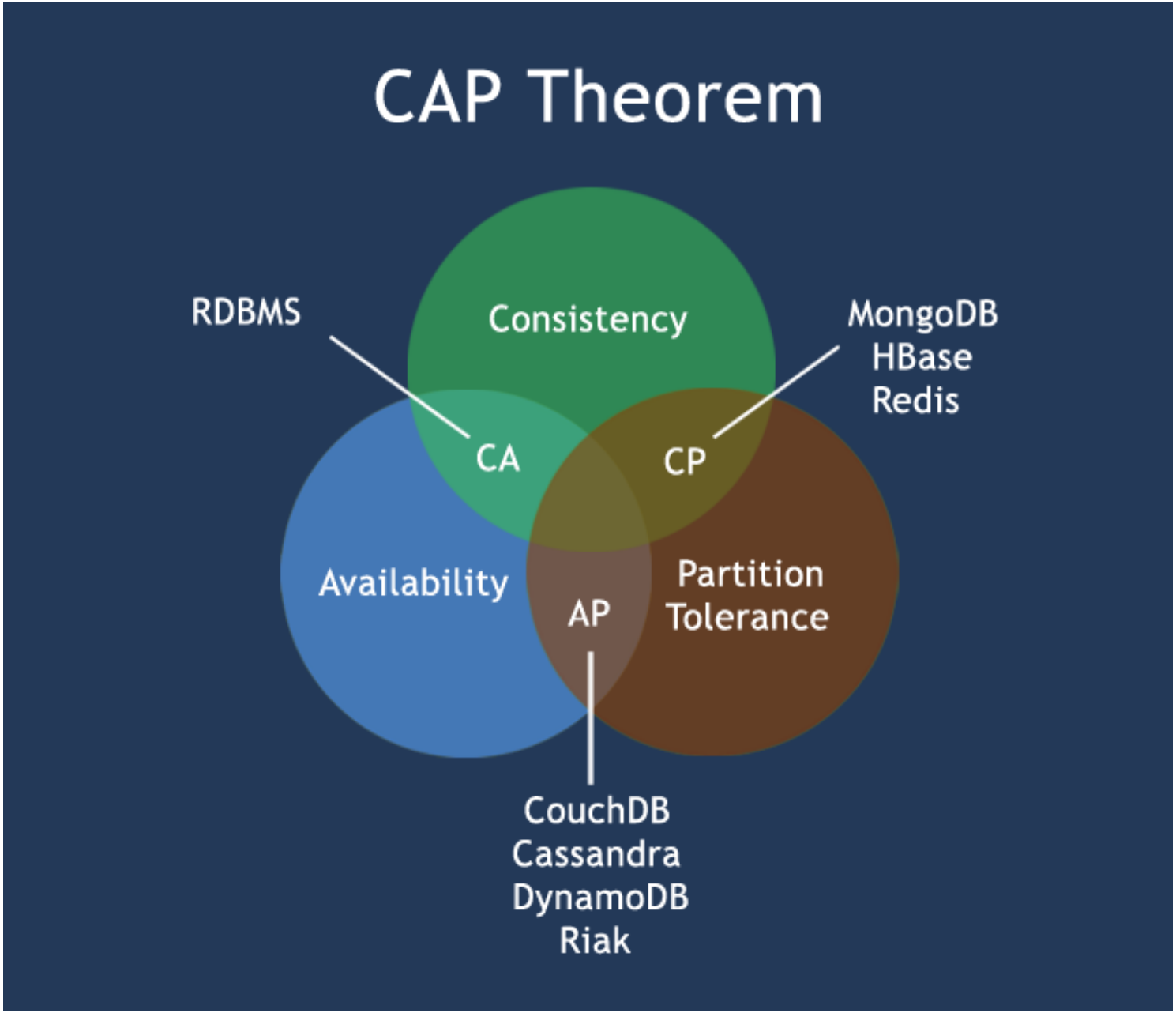

Source : Octo technologies

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

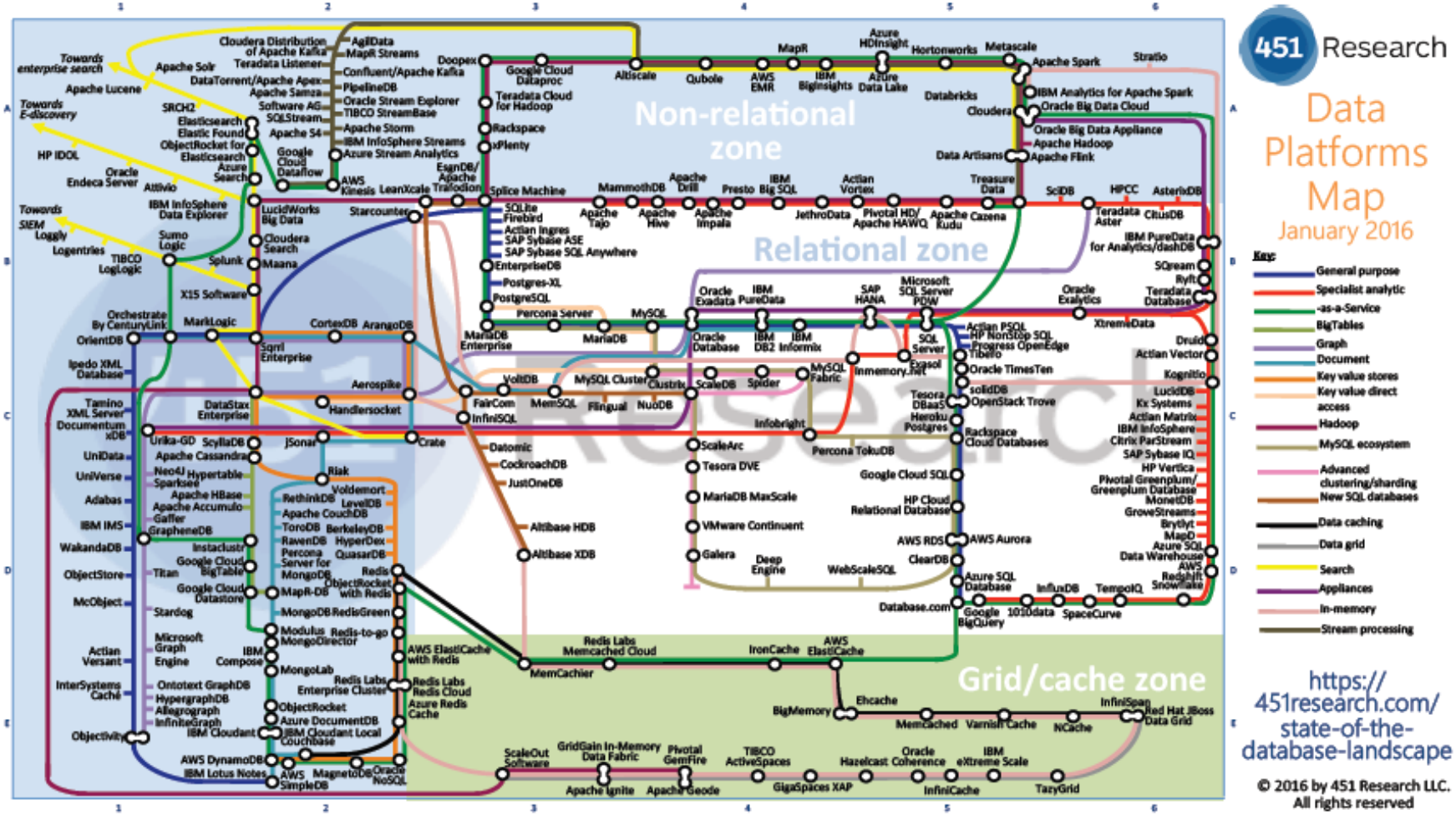


What is the CAP theorem ?

# What is the CAP theorem ?



# Are you able to travel on this map ?

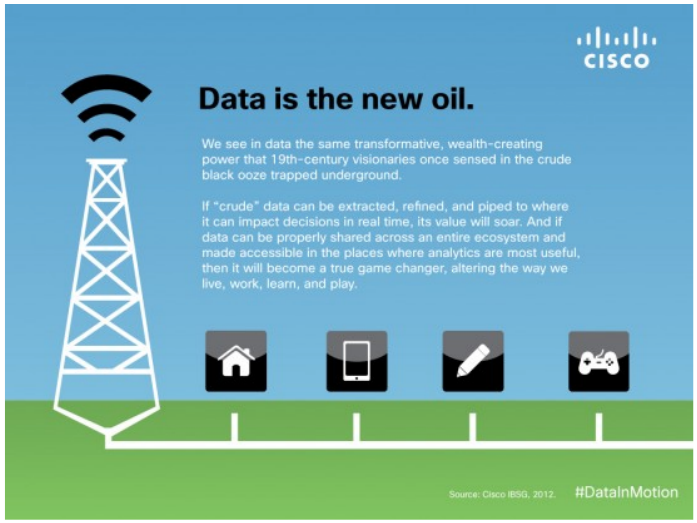


# Plan

# Data Everywhere : *Big Data*

- 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.

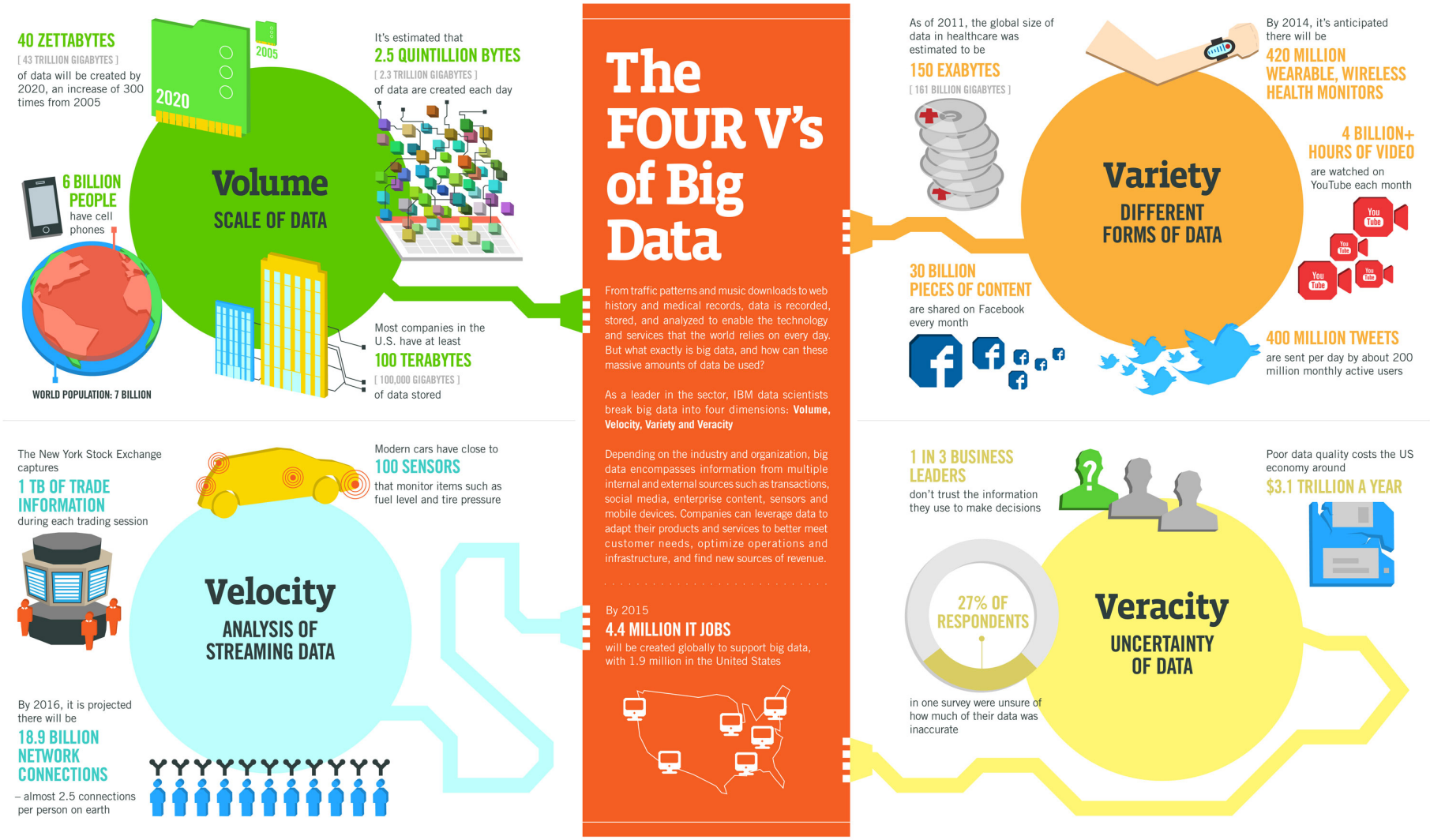
# Data Everywhere : *Big Data*



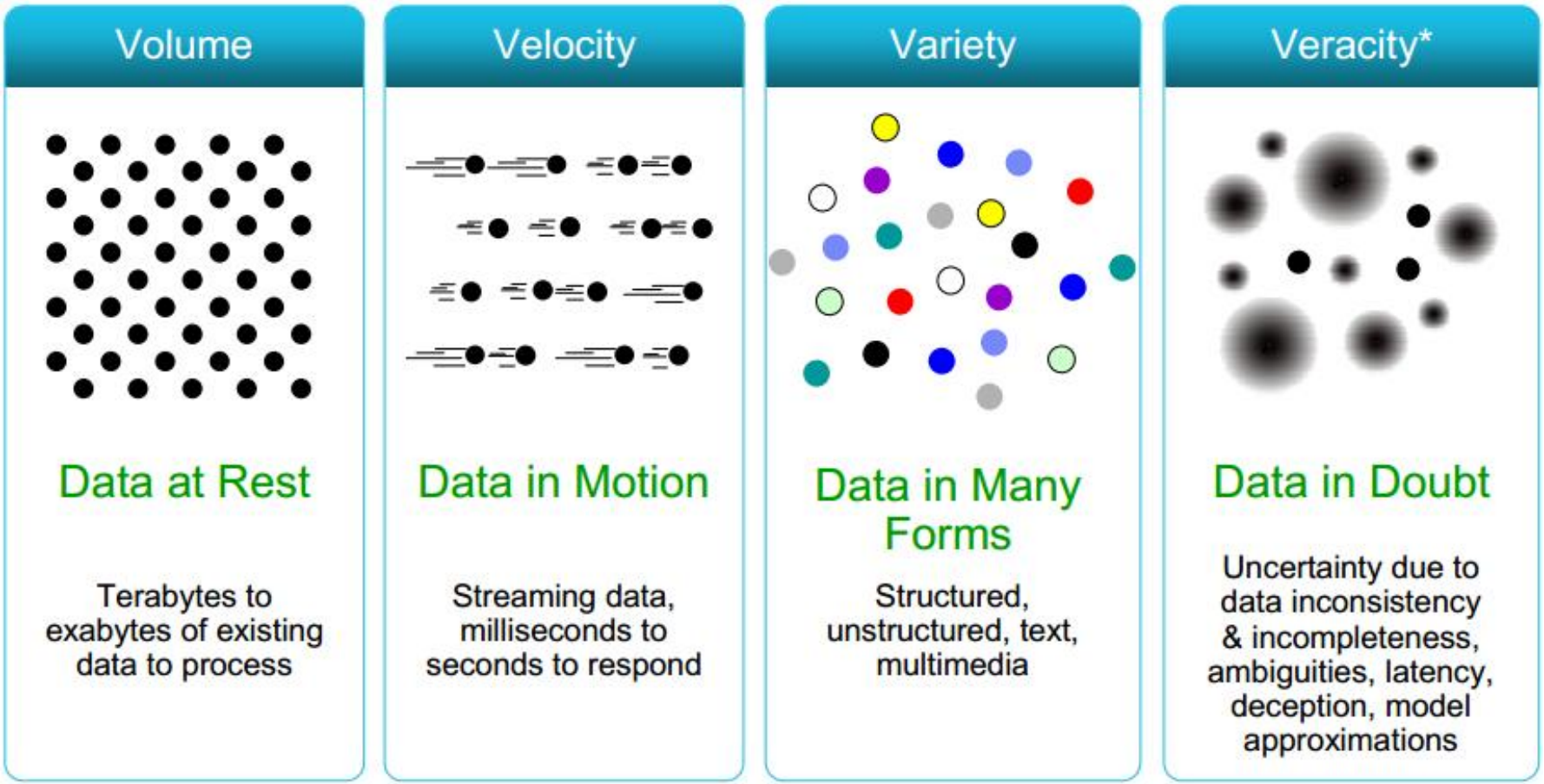
The data power

# Challenges of Big Data : the four V



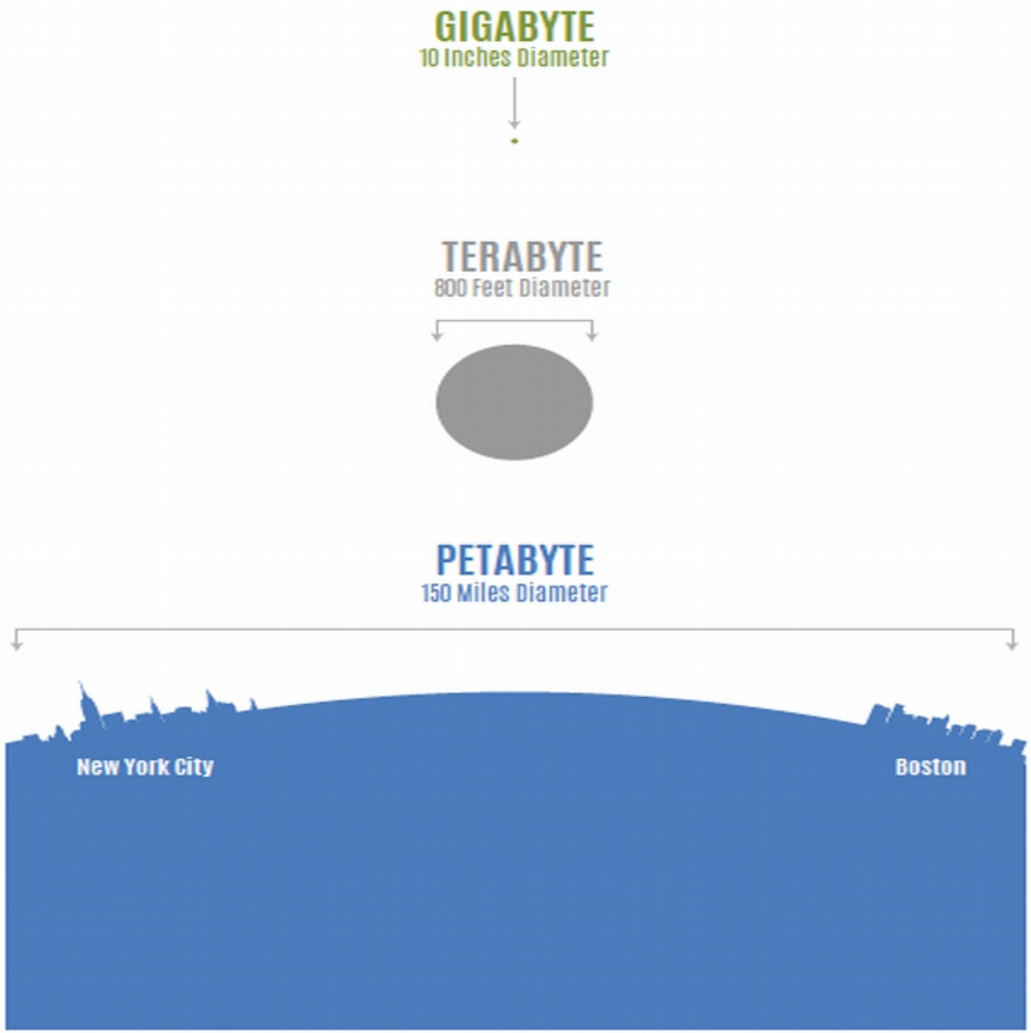

Source : IBM

# Big Data : the four V



# Big Data : Volume

Quantity of generated and stored data

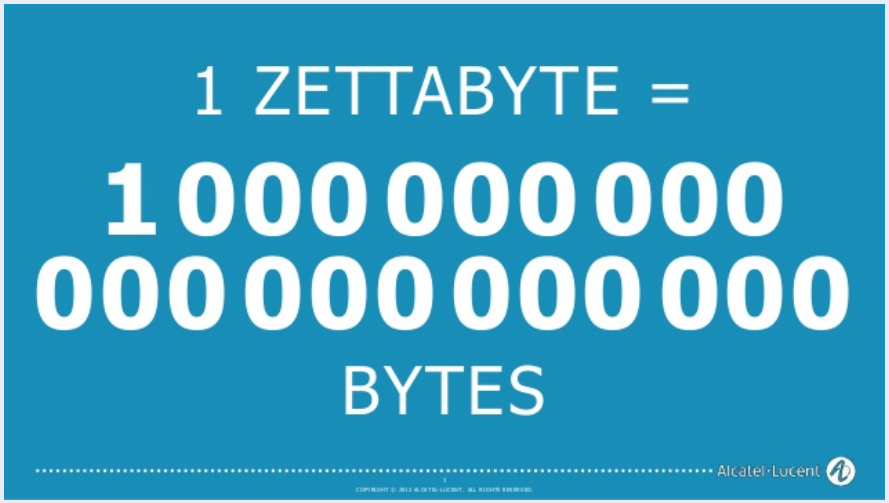



Source : The Executive's guide to big data & Apache Hadoop

# Big Data : Volume

## Some orders of magnitude (Source Wikipédia)

- In 2020, 40 zettabytes of data by the web per year




- Scientific installations :
  - The radiotelescope *Square Kilometre Array* will generate 50 TB of analyzed data per day ;
  - with 7 000 TB of raw data per second

# Big Data : Volume

## Today

- 150 milions emails every minute
- Facebook : 4000 TB / day
- CERN, LHC : 15 PB / year

Source : Wikipedia

# Big Data : Volume

## Capacity of a big server

- Memory : 256 GB
- Disk storage : 24 TB
- Disk speed : 100 MB /s

# Big Data : 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, ...

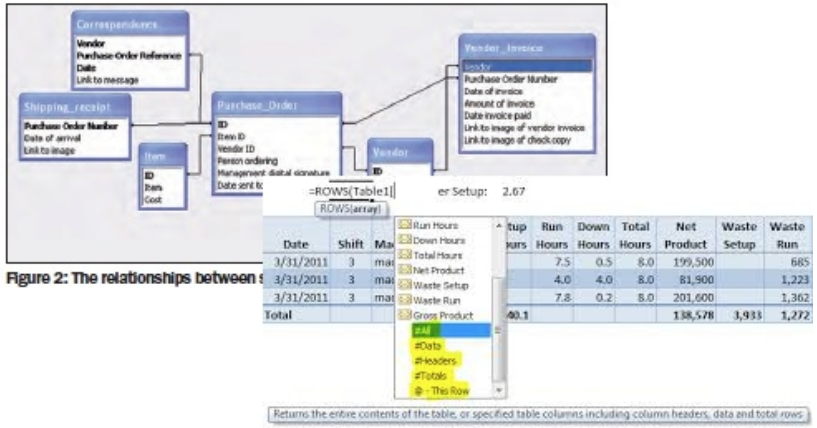



Figure 2: The relationships between

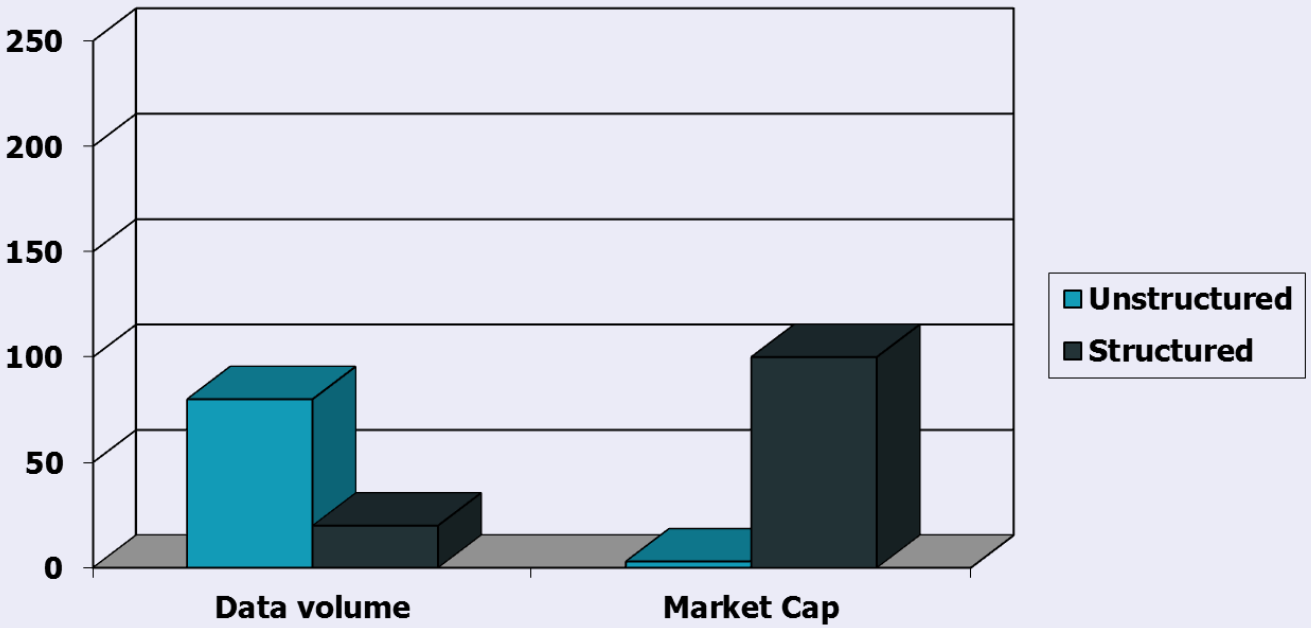
## Non structured data

Without strong structuration

- e.g. : emails, documents, images, social network data...

# Big Data : Variety

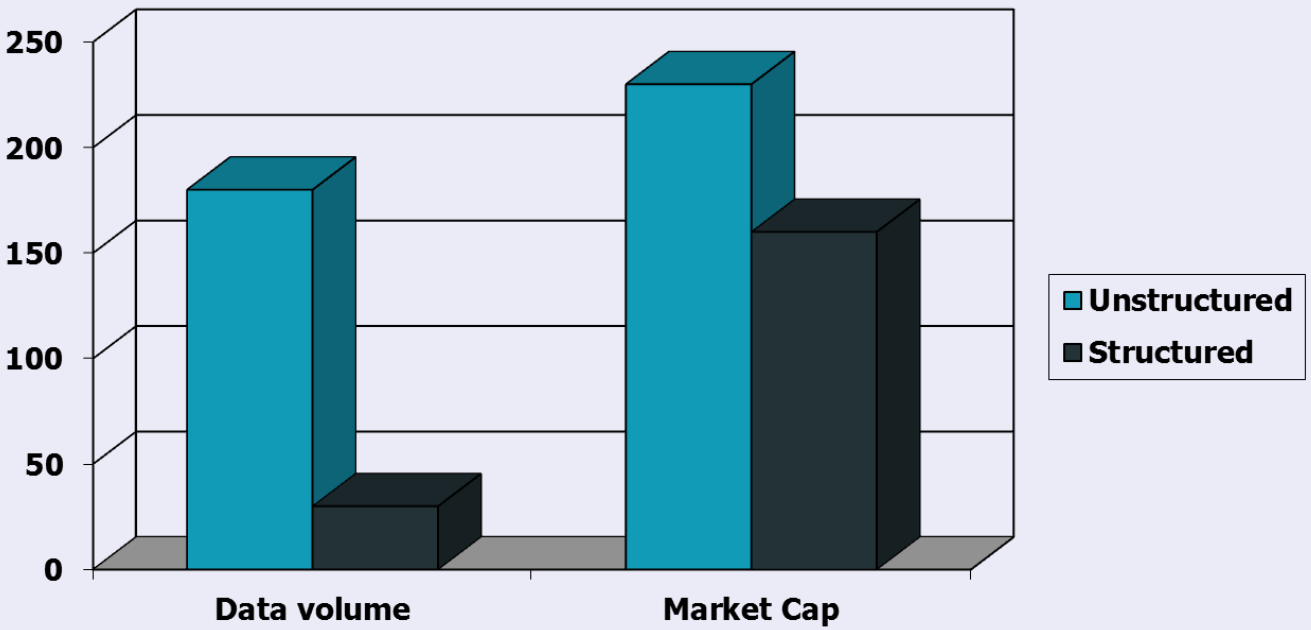
## Structured vs non-structured data in the 90's




Source : C. Manning

# Big Data : Variety

## Structured vs non-structured data today (2005)




Source : C. Manning

# Big Data : Variety

## Non-structured data

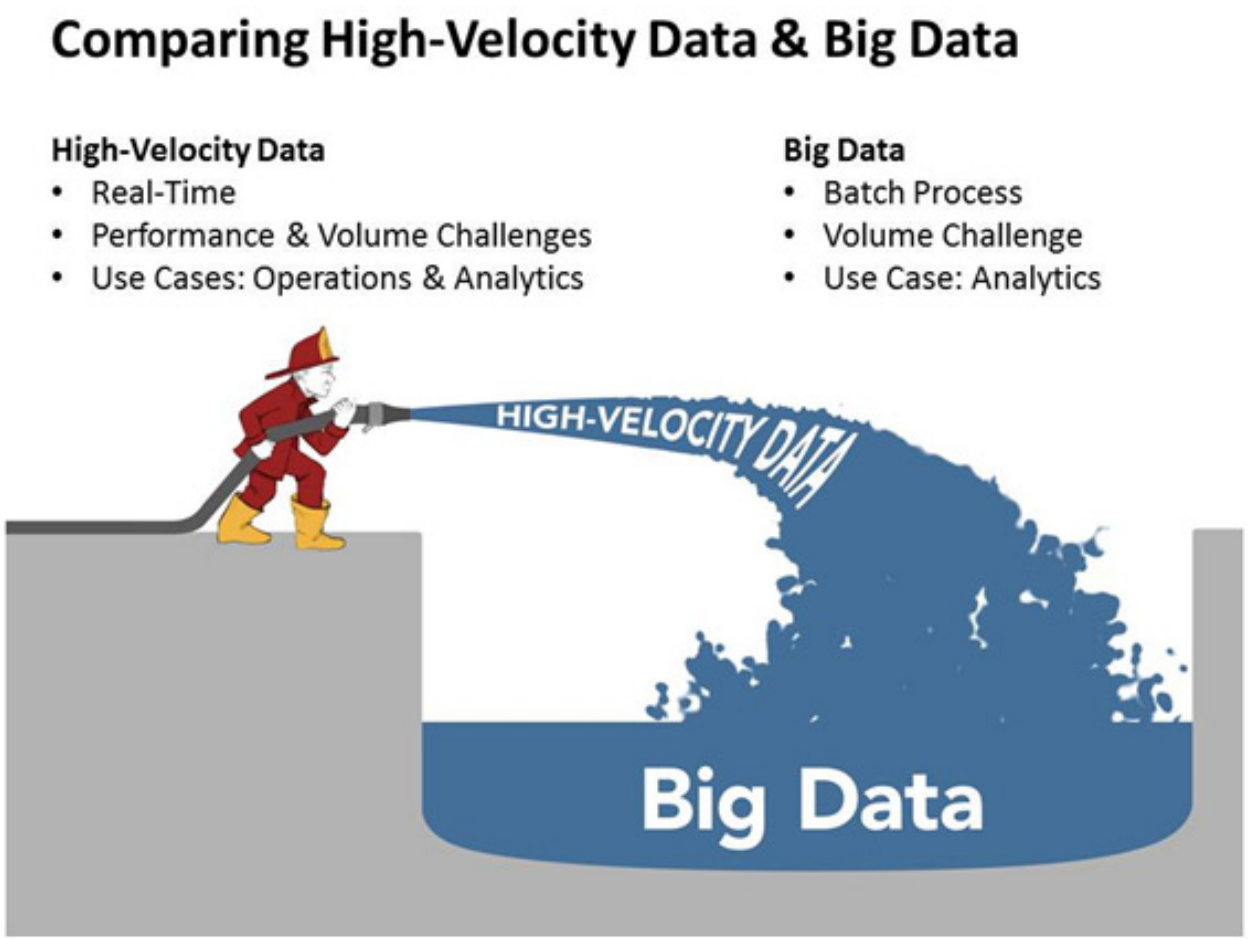
- 175 millions tweets per day.
- 571 new websites every minute.
- 2.5 quintillions bytes of data per day.

Source : `http://www.digitalreasoning.com/resources/Holistic-Analytics.pdf`

# Big Data : Velocity

Speed at which the data is generated and processed.

Important flows - Short processing times.




Source : ScaledB

# Big Data : Velocity

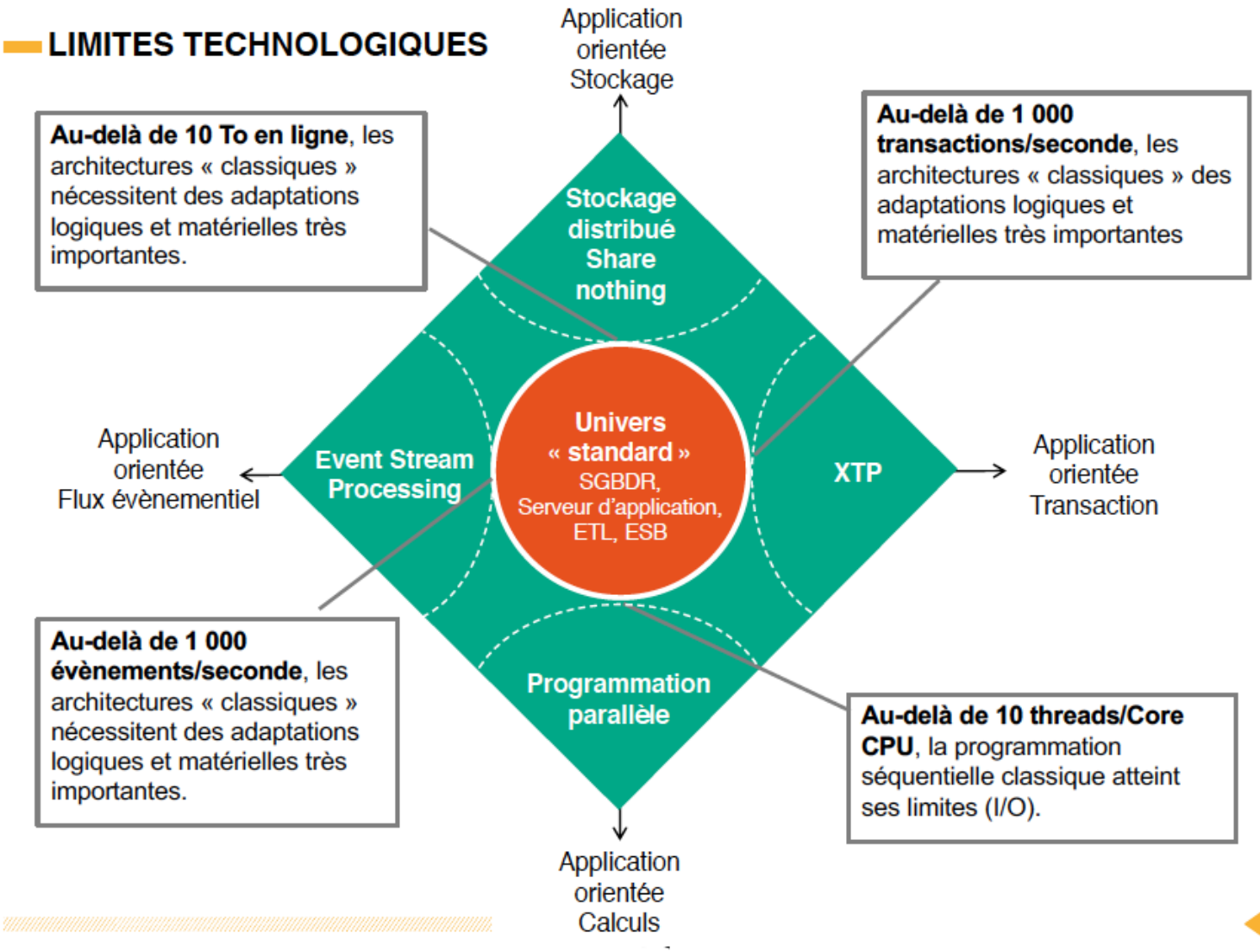

Source : Go Globe

# Big Data : Velocity

## The measure of the eye blink (User experience)

- Amazon : increase of more than 100 ms of the latency $\Rightarrow$ -1 % of the sales
- Google : more than 500 ms at the loading $\Rightarrow$ ? 20 % of the trafic (seen pages)
- Yahoo : more than 400 ms at the loading $\Rightarrow$ + 5 to 9 % of cancelation (rebound)
- Bing :more than 1 second at the loading $\Rightarrow$ -2,8 % of ad revenue.

# Big Data : technical limitations




Source : OCTO

# Processing Big Data

## Solution : parallelism

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

# Processing Big Data

## Some problems

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

# Processing Big Data

## 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 from 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.

# Processing Big Data

**How do we design algorithms for distributed computing ?**

Generic programming models : design patterns.

MapReduce