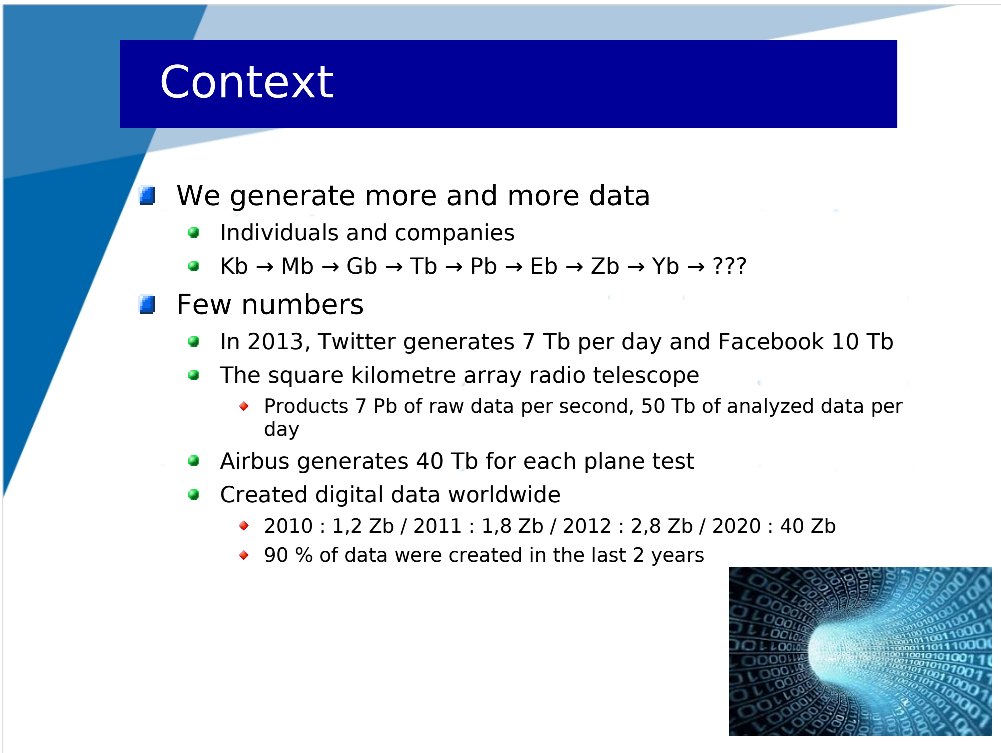
Introduction to Big Data

Daniel Hagimont
hagimont@enseeiht.fr

1

The lecture starts with a general introduction to Big Data.

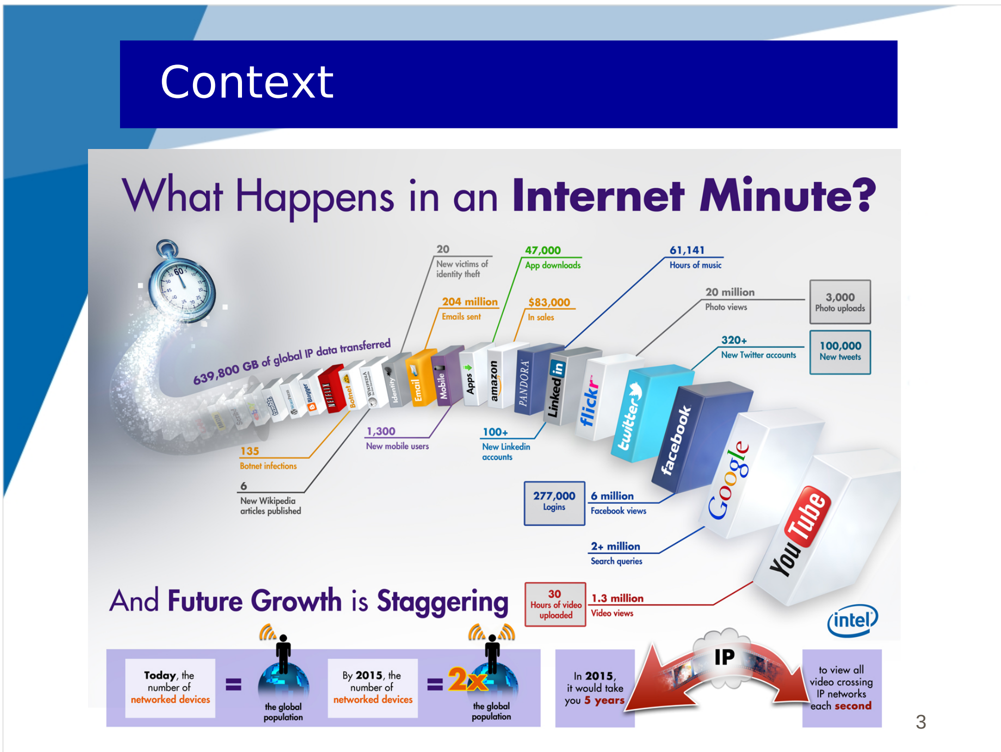


Today's digital applications are generating more and more data. These data may be generated by individuals (through social networks) or companies.

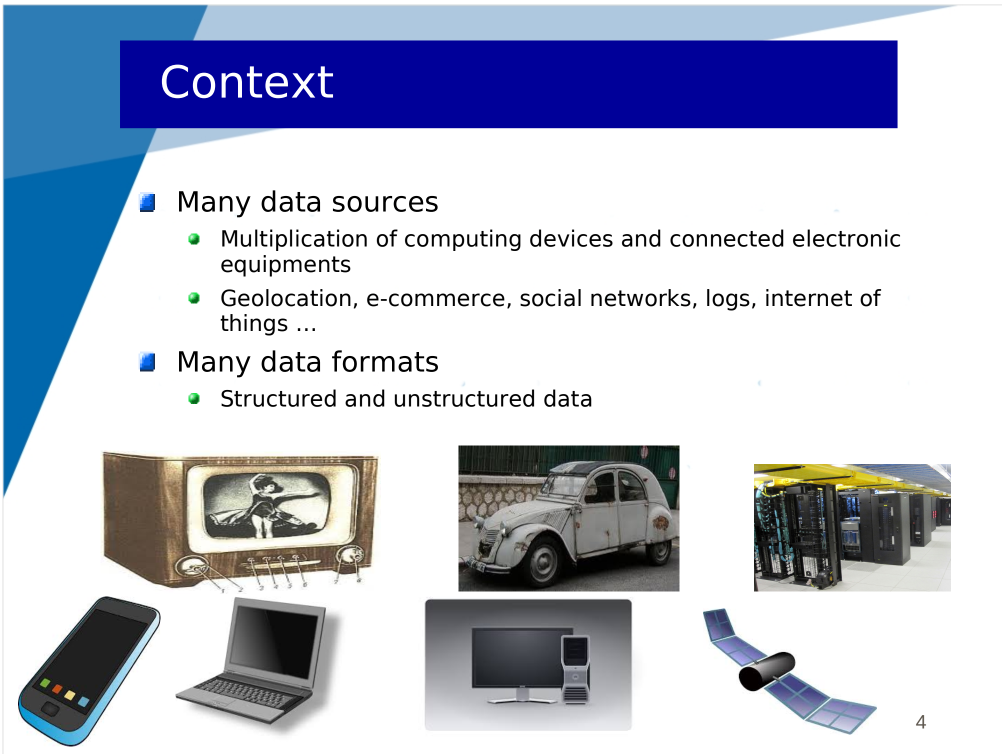
Examples are Twitter and Facebook who analyze the behavior of their users, large scientific equipments like the square kilometre array radio telescope.

Another example I have been working on comes from Airbus. When they have designed and built a new plane, during the test phase, the plane is equipped with a huge set of captors which monitor everything. Each test flight generates about 40 TB of data and they run several test flights per week over a year.

The amount of data generated worldwide becomes so large and it is exponential.

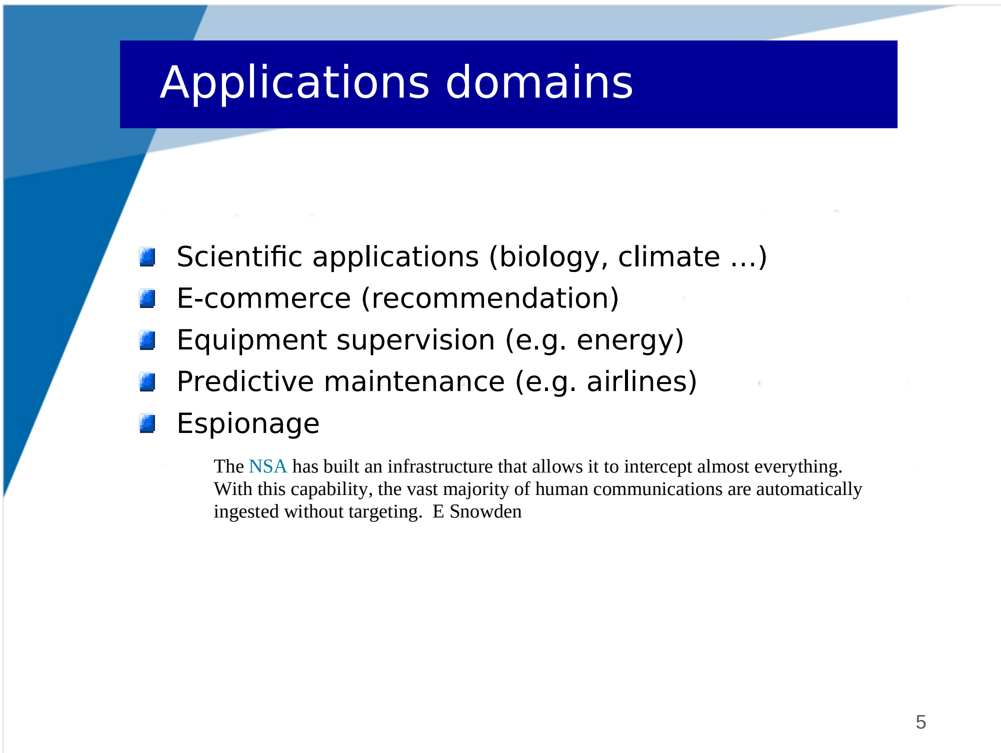


This is a picture I grabbed on the net which illustrates this deluge of data.



It is important to remember that these data are coming from many sources, with the multiplication of devices (personal computers, phones, tablets) and the development of cloud applications.

These diverse sources generate data in many different formats. Such format can be structured (with XML, Json ..) or unstructured (like textual data).



These data are generated and used (analyzed) in many application domains

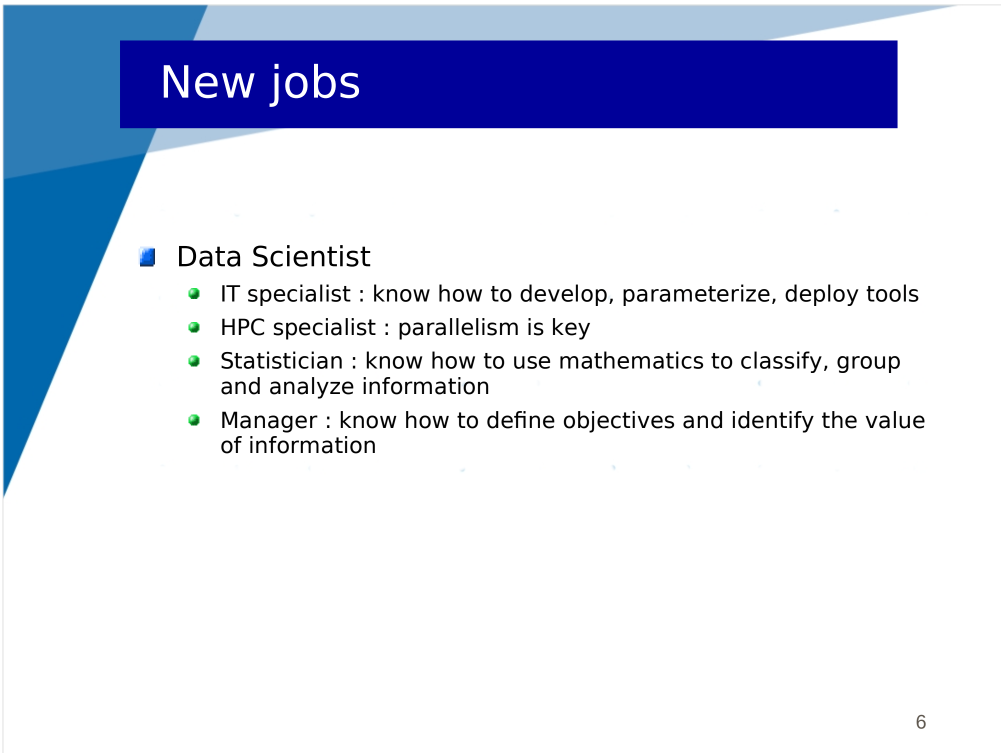
Scientific application : biology (genomic) or climate (weather prediction)

E-commerce : profiling users for product recommendation

Equipment supervision : energy where they profile consumption

Predictive maintenance : for instance in the airline industry, they monitor each flight with captors embedded in planes. When a failure occurs, they analyze the monitored data before the failure to identify symptoms that preceded the failure. Therefore, by analyzing data from each flight, they can detect such symptoms before a failure occurs and anticipate a maintenance to prevent the failure.

Espionage : E. Snowden revealed that the NSA was analyzing data from so many sources in a wish to control everybody.

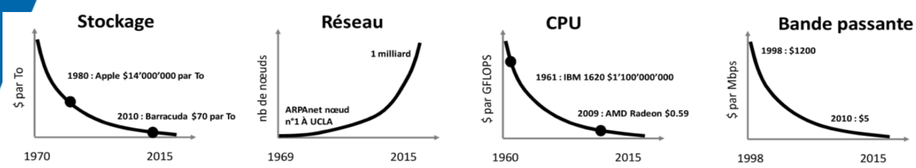


The field of big data is generating many jobs:

- manager who are dealing with high level objectives
- mathematicians who are analyzing and modeling data
- HPC specialists who know how to improve performance
- IT specialists who master the infrastructure (hardware and software)

Computing infrastructures

- The reduced cost of infrastructures



- Main actors (Google, Facebook, Yahoo, Amazon …) developed frameworks for storing and processing data
- We generally consider that we enter the Big Data world when processing cannot be performed with a single computer

7

Decades ago, data treatments were performed by a single computer, a multiprocessor machine if computing power was required.

Two main evolutions modified that:

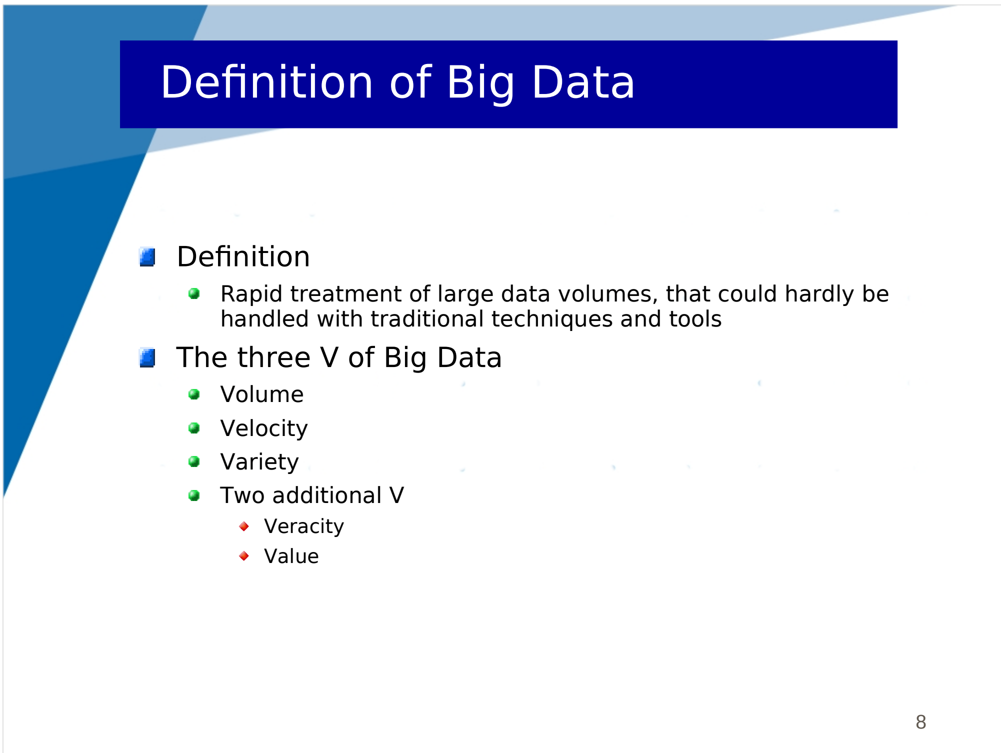
- new applications which require huge computing power

- the reduced code of hardware (mainly servers and networks), leading to the development of clusters (a huge set of interconnected servers).

A cluster is a set of interconnected servers which are exploited by a distributed operating system in order to give the illusion of a giant computer.

The main actors of the field developed such distributed operating system frameworks to implement a sort of giant computer for storing and processing data.

We generally consider that we enter the Big Data world when processing cannot be performed with a single computer (even a large one) and when a cluster is required.



One (among many) definition of Big Data is the rapid treatment of large data volumes that could hardly be handled with traditional techniques and tools (e.g. a database system on a single computer).

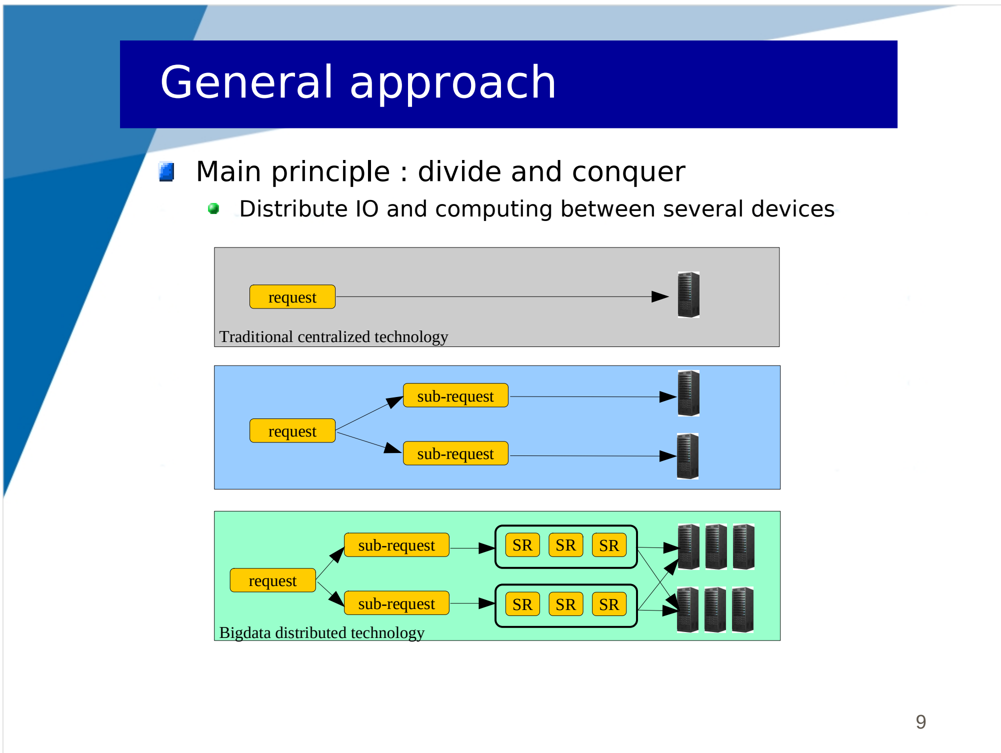
In the field, people are talking about the 3 V which characterize Big Data:

- Volume : dealing with very large data volumes
- Velocity : treatment should be fast enough
- Variety : data may be of very different types

Sometimes, they add 2 additional V :

- Veracity : big data may also consider the trust we have into the handled data
- Value : it may also consider the value of the data

And may be at the time of today, they added again other Vs.



The general approach used to efficiently treat very large datasets is to apply a well known principle called "divide and conquer".

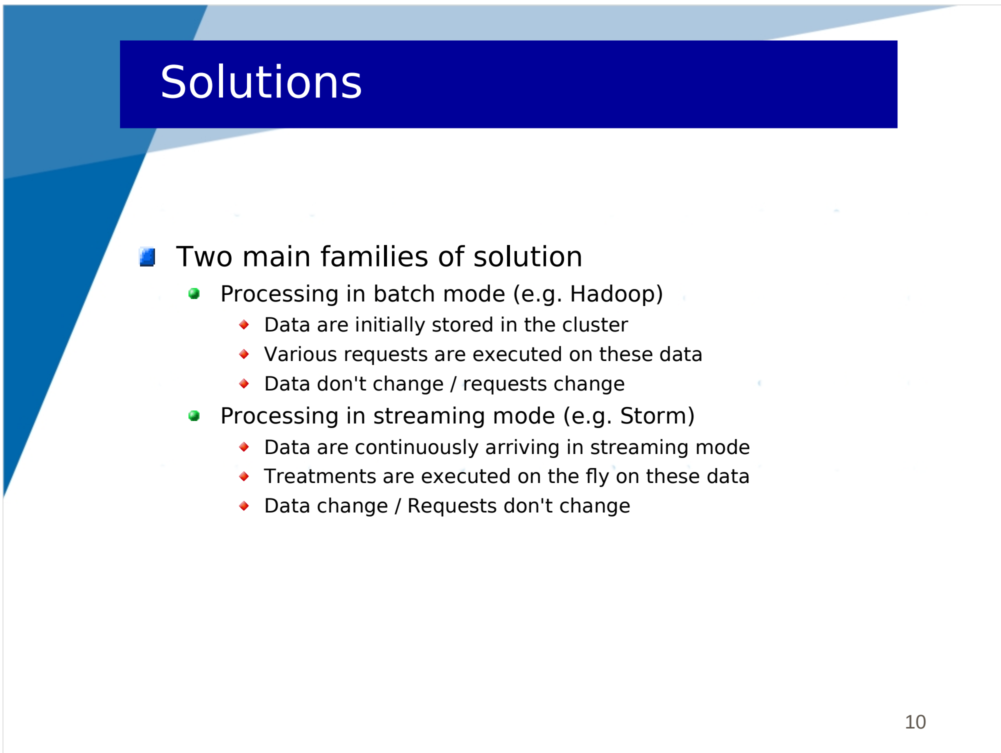
The principle is to divide a task into several sub-tasks which can be distributed on distributed computers, therefore benefiting from parallel IO (reads from different disks) and/or parallel computation (on different processors).

Figure top : in a traditional centralized setting, a request or task is executed on one computer.

Figure middle : a request can be divided into 2 sub requests executed in parallel on 2 separate computers.

Figure bottom : the request can be divided into many sub-requests scheduled on many computers. This can be done if the initial request is sufficiently large.

Notice here that in this general principle, we don't precise how data are moved to the computers where sub-requests are executed.



The previous general principle leads to 2 main families of solution.

Processing in batch mode (Hadoop is an example). Data are initially stored on the computers' disks. For instance a very large dataset is divided into blocks and the blocks are distributed on the machines. Then, various requests can be issued to analyze the data. Such a request is divided into sub-requests which will handle the blocks on the different machines (in parallel). What is important here is that the data are installed in the cluster (installed means here that data don't change, they are here to be read and analyzed, not modified) and that many requests can be issued on the same dataset.

Processing in streaming mode (Storm is an example). Data are continuously arriving in streaming mode. Treatments are decomposed in tasks that are deployed in the cluster