unimc
UNIVERSITÀ DI MACERATA
l'umanesimo che innova

vision robotics
artificial
intelligence

Laboratorio di Cloud Computing

Lecture 2 - Big Data & Databases

Docente: Riccardo Rosati

riccardo.rosati@unimc.it

Today's Lesson

- Knowledge assessment
- Introduction to Big Data
- Algorithms for Big Data
- Introduction to Database
- Question Time

Big Data

Data definition

Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.

Data is measured, collected, reported, and analysed, whereupon it can be visualized using graphs, images or other analysis tools.

Data can be..


STRUCTURED DATA
CSV
XLS
@
LOG
DOC
SEMI-STRUCTURED DATA
UN-STRUCTURED DATA
PNG
MP3

Data can be..

Structured: the data that has a structure and is well organized either in the form of tables or in some other way and can be easily operated is known as structured data. Searching and accessing information from such type of data is very easy. For example, data stored in the relational database in the form of tables having multiple rows and columns. The spreadsheet is an another good example of structured data.

Semi-Structured: is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements.

Un-Structured: The data that is unstructured or unorganized Operating such type of data becomes difficult and requires advance tools and softwares to access information. For Example, images and graphics, pdf files, word document, audio, video, emails, powerpoint presentations, webpages and web contents, wikis, streaming data, location coordinates etc.

Big Data definition

The term “big data” refers to data sets so large and complex that traditional tools, like relational databases, are unable to process them in an acceptable time frame or within a reasonable cost range. Problems occur in sourcing, moving, searching, storing, and analyzing the big data

Big Data definition

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Big data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data.

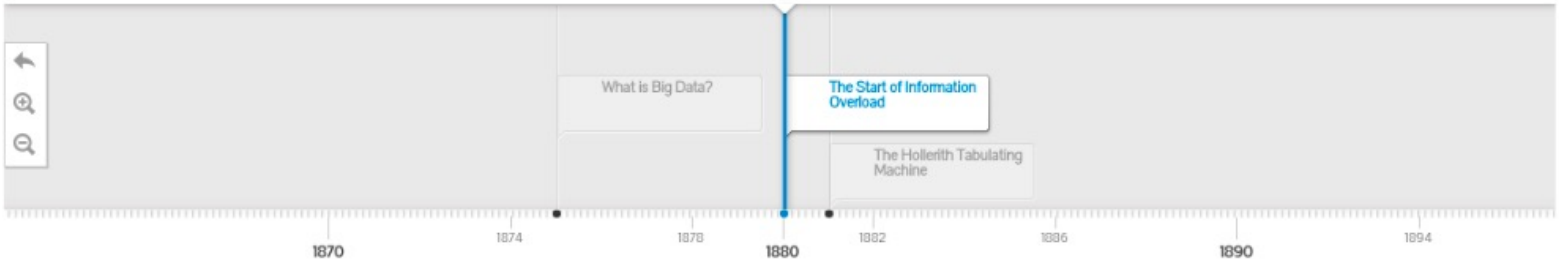
History

- U.S. Census
  - 1870: ~38M people
  - 1880: ~50M people
  - 1890: ~63M people

1880

The Start of Information Overload

The 1880 U.S. Census took eight years to tabulate, and it was estimated that the 1890 census would take more than 10 years using the then-available methods. Without any advancement in methodology, tabulation would not have been complete before the 1900 census had to be taken.



History

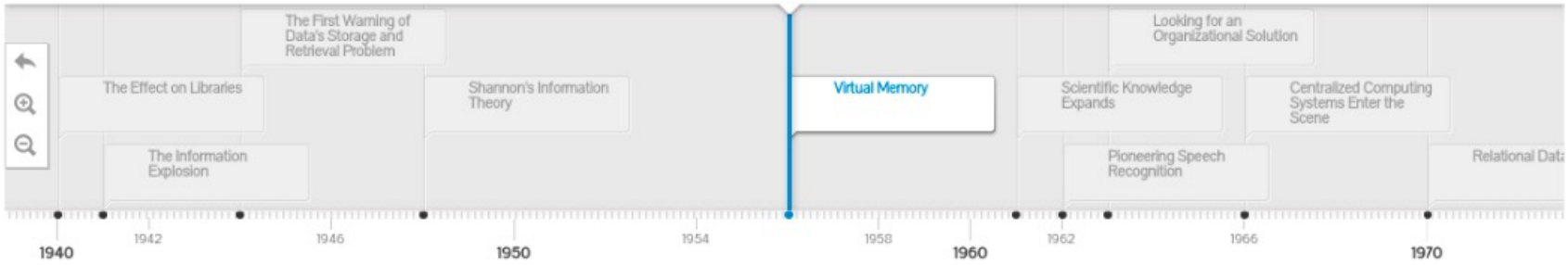
Source: Frguentsch

1956

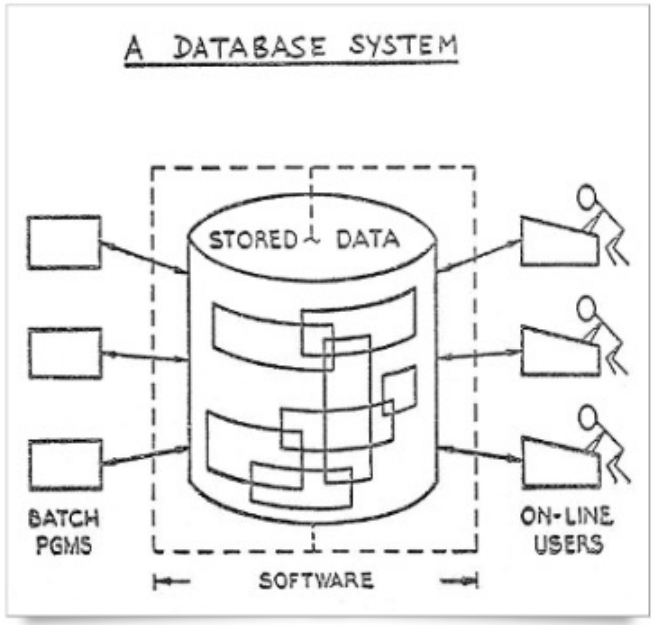
Virtual Memory

The concept of virtual memory was developed by German physicist Fritz-Rudolf Güntsch as an idea that treated finite storage as infinite. Storage, managed by integrated hardware and software to hide the details from the user, permitted us to process data without the hardware memory constraints that previously forced the problem to be partitioned (making the solution a reflection of the hardware architecture, a most unnatural act). With special thanks to @ajbowles

1961
Scientific Knowledge Expands



History

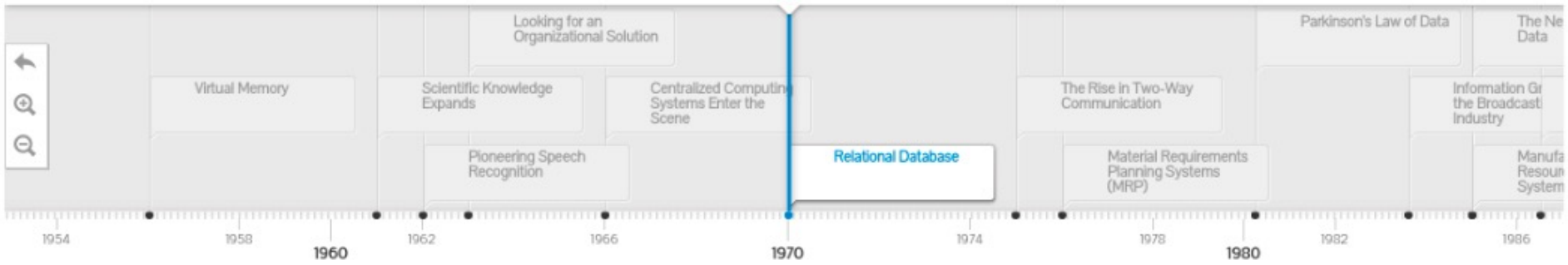


Source: IBM.com

1970

Relational Database

In 1970, Edgar F. Codd, an Oxford-educated mathematician working at the IBM Research Lab, published a paper showing how information stored in large databases could be accessed without knowing how the information was structured or where it resided in the database. Until then, retrieving information required relatively sophisticated computer knowledge, or even the services of specialists —a time-consuming and expensive task. Today, most routine data transactions—accessing bank accounts, using credit cards, trading stocks, making travel reservations, buying things online—all use structures based on relational database theory. Source and special thanks to @TheSocialPitt

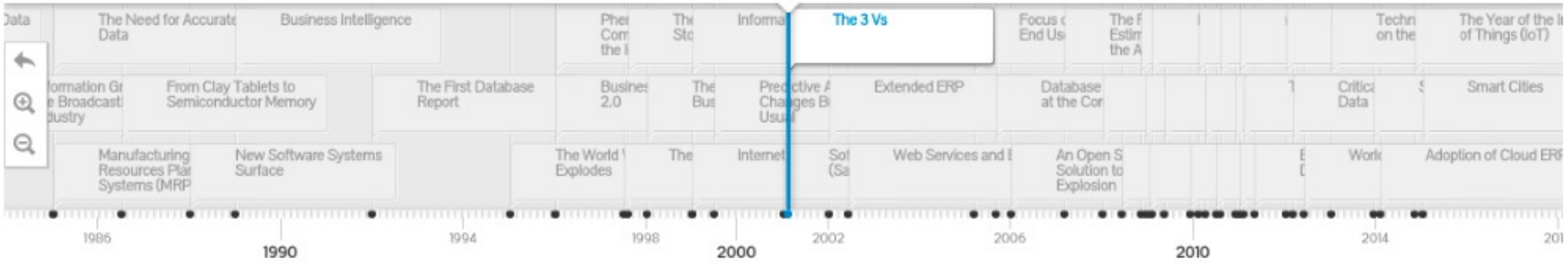


History

February 2001

The 3 Vs

Gartner Analytst, Doug Laney, published a research paper titled 3D Data Management: Controlling Data Volume, Velocity, and Variety. Even today, the "3Vs" are the generally-accepted dimensions of big data.



History

2015

Smart Cities

A smart city uses the analysis of contextual, real-time information to enhance the quality and performance of urban services, reduce costs and resource consumption, and actively engage with its citizens. Gartner estimates that over 1.1 billion connected things will be used by smart cities in 2015, including smart LED lighting, healthcare monitoring, smart locks and various sensor networks for things like motion detection, and air pollution monitoring. Source: Impact of IoT on Business at the Gartner Symposium/ITxpo 2014

Why so many data?

- Drop of digital Storage cost
- Increase of computing power
- Proliferation of devices that generate digital data (consumer accessible technology) (computers, smartphones, cameras, RFID systems, IoT)

Generating digital Data

- Self-published content: FB, Blogs, YouTube, Instagram, ...

technology completely changed and facilitated publishing: massive growth in human- generated content

- Consumer Activity: business and marketing

digital footprint, tracking, insights, security cameras, ...

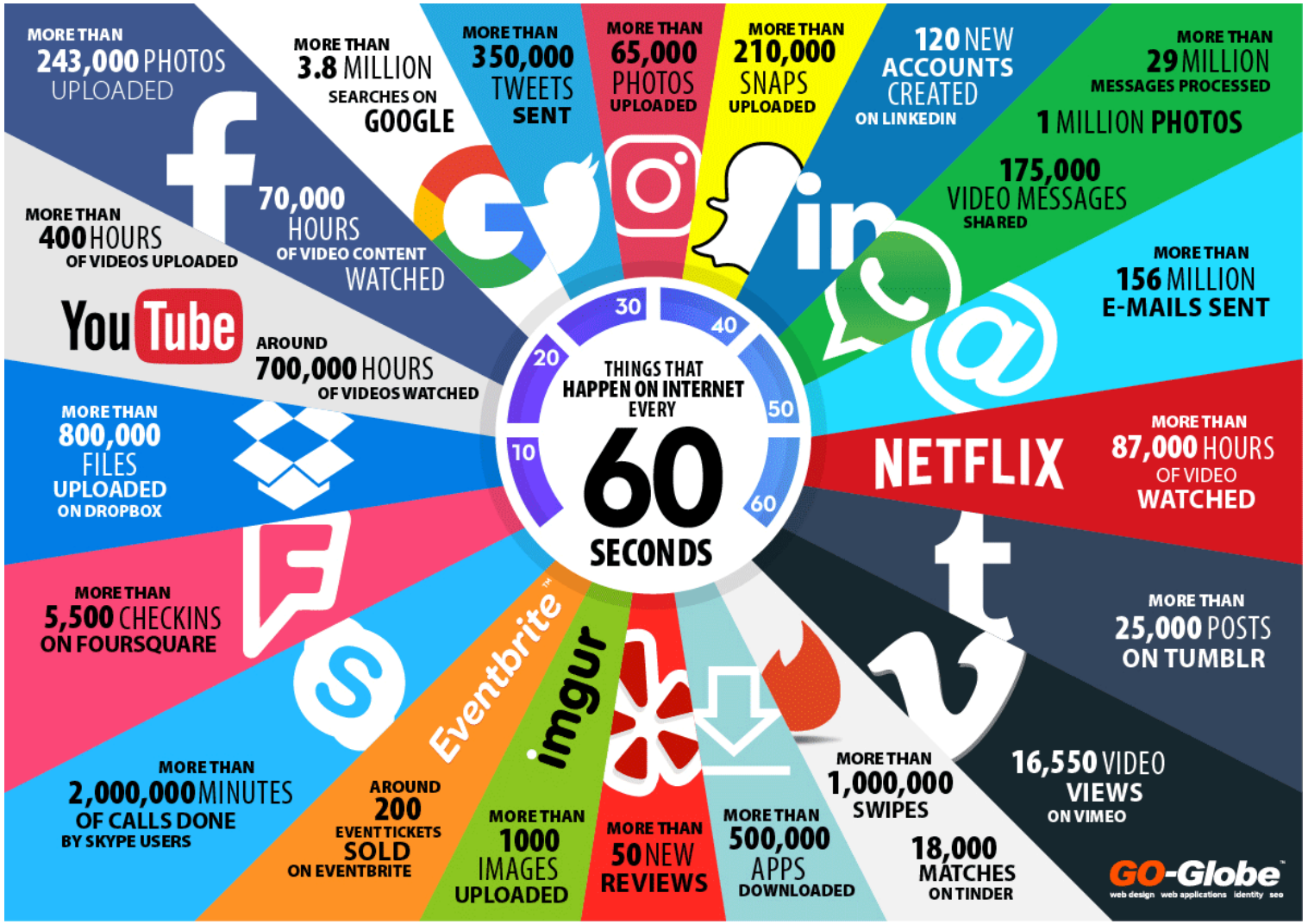
- Machine data and IoT

devices exchanging data, integration of physical world into computer-based systems, connettivity, ...

- Science

larger and complex experiments, ...

The impact of Big Data



The impact of Big Data

“There are some things that are so big that they have implications for everyone, whether we want it or not.
Big Data is one of those things, and is completely transforming the way we do business and is impacting most other parts of our lives.”

The impact of Big Data

The US National Security Agency has built a huge data centre in Bluffdale, Utah - codenamed Bumblehive - capable of storing a yottabyte of data - that's one thousand trillion gigabytes.

Yottabyte??

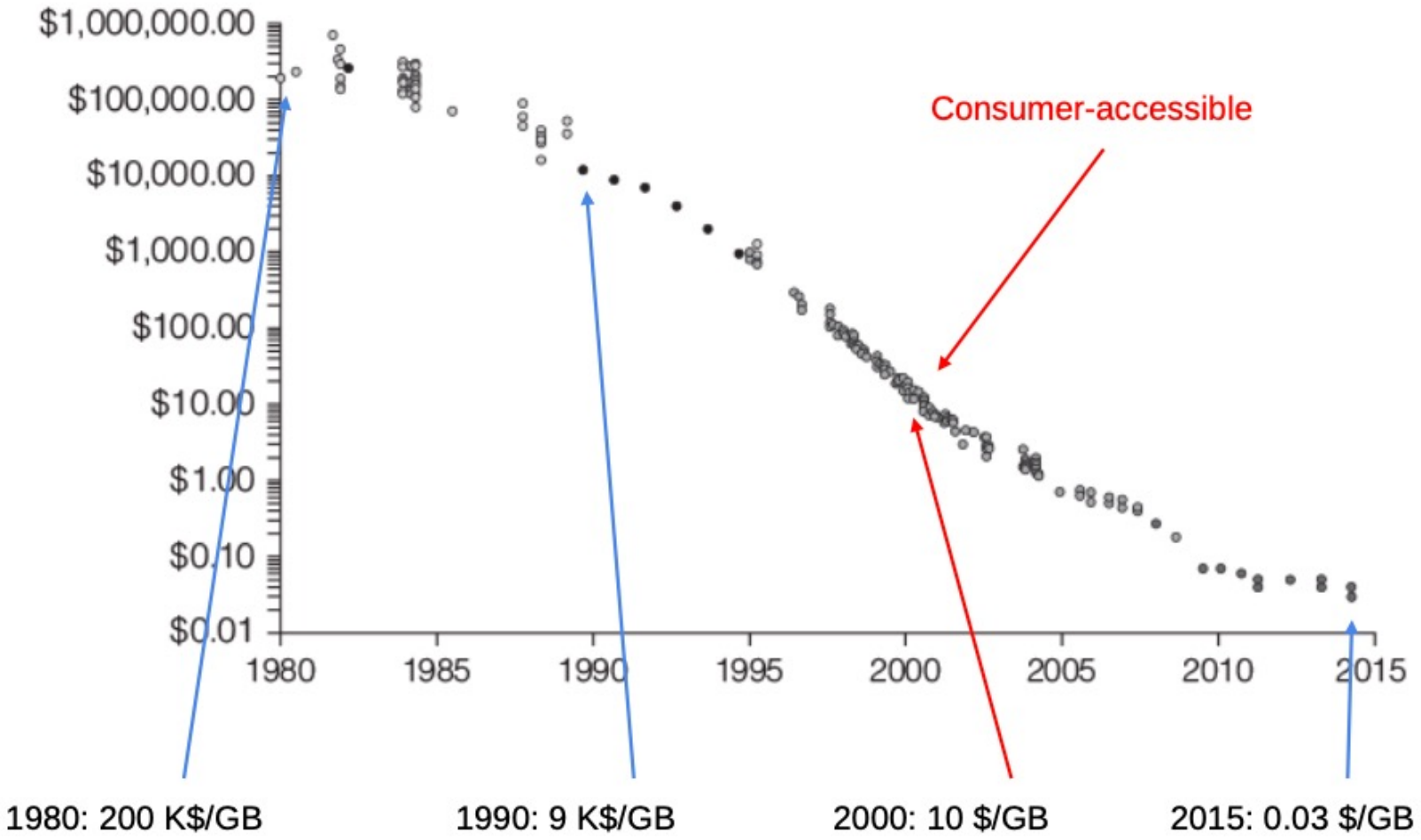
Unit	Numeric Representation	Exponential Representation
Megabyte	1,000,000 bytes	10^6 bytes
Gigabyte	1,000,000,000 bytes	10^9 bytes
Terabyte	1,000,000,000,000 bytes	10^{12} bytes
Petabyte	1,000,000,000,000,000 bytes	10^{15} bytes
Exabyte	1,000,000,000,000,000,000 bytes	10^{18} bytes
Zettabyte	1,000,000,000,000,000,000,000 bytes	10^{21} bytes
Yottabyte	1,000,000,000,000,000,000,000,000 bytes	10^{24} bytes

Digital Storage cost

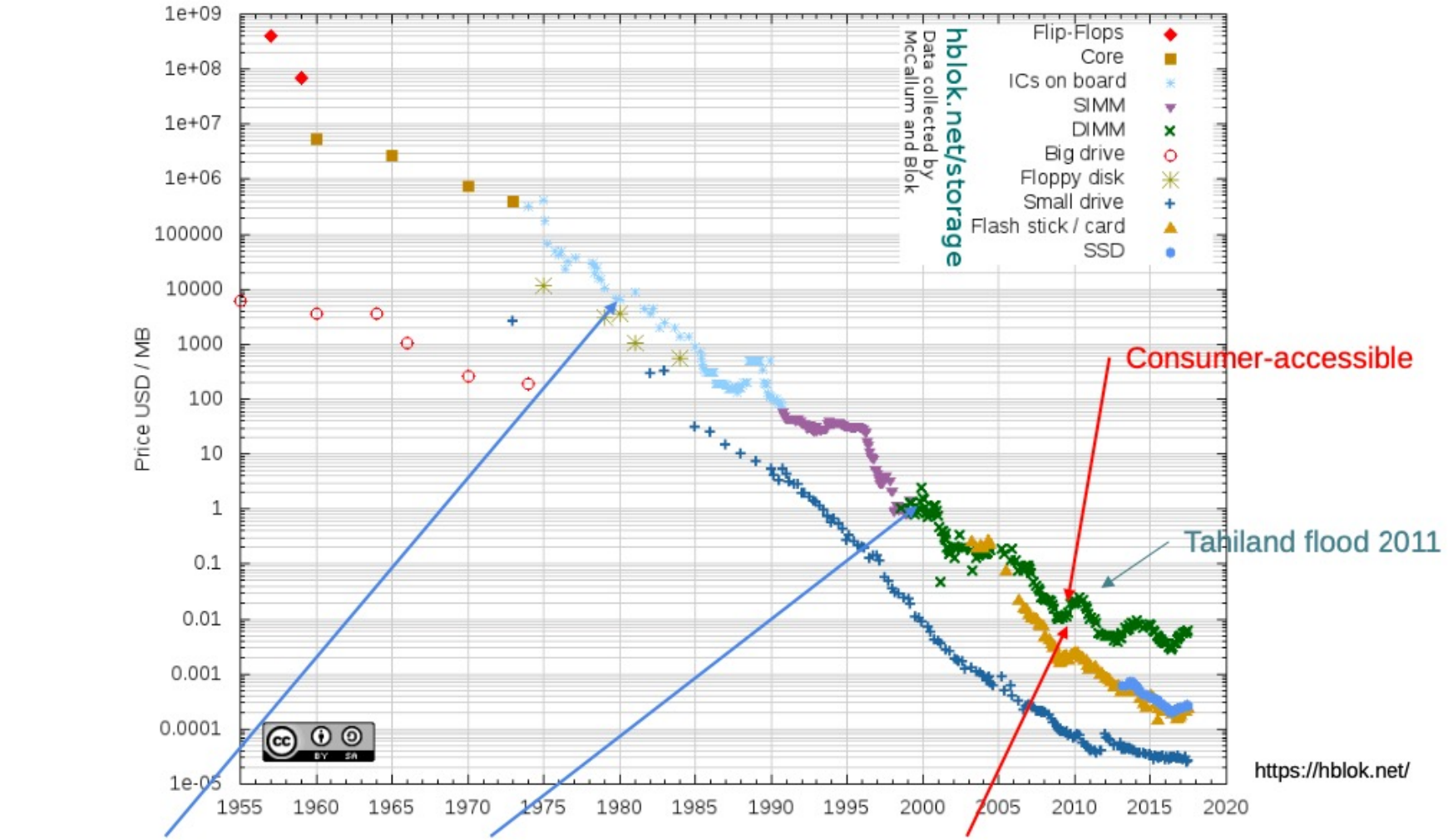
Digital storage:

- Disk: low cost, high capacity, slow access
- RAM: high cost, “small” capacity, fast access

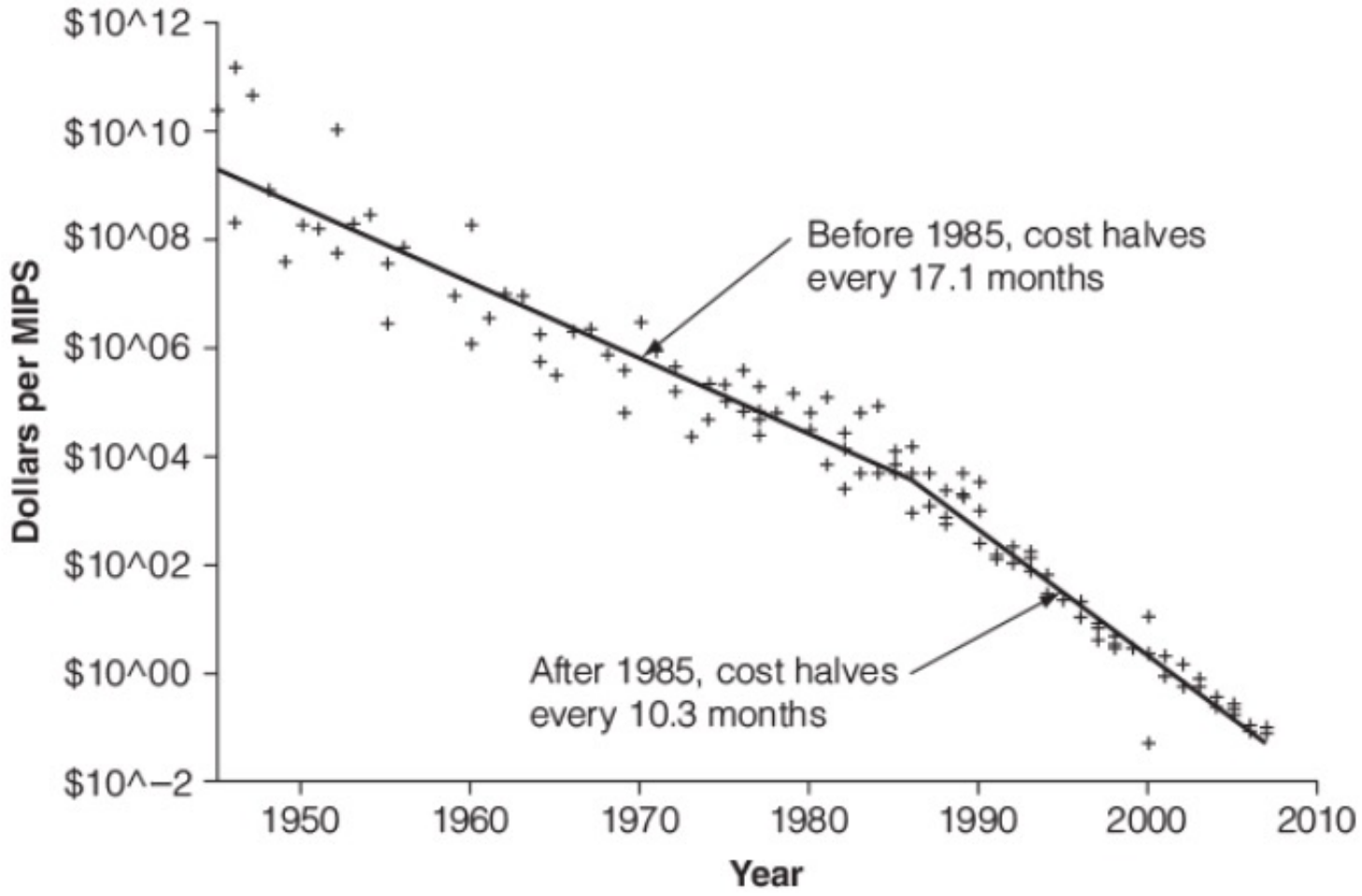
Digital Storage cost



RAM cost


Price USD / MB
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
1e-05
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
hblok.net/storage
Data collected by McCallum and Blok
Flip-Flops
Core
ICs on board
SIMM
DIMM
Big drive
Floppy disk
Small drive
Flash stick / card
SSD
Consumer-accessible
Tahiland flood 2011
https://hblok.net/
1980: 6 M$/GB
2000: 1 K$/GB
2010: 10 $/GB

Computing power cost



Why is big data useful

New ways to use data:

- no rationing storage and selecting the “valuable” data
- storing raw data in “data lakes” for future questions and application (>100Gbps) where data is located is not important
- heavy “data driven” approach
- data insights: analytics VS analysis

The 4-V model for Big Data

Volume

The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not.

Variety

The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through *data fusion*.

The 4-V model for Big Data

Velocity

In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data are produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.

Veracity

It is the extended definition for big data, which refers to the data quality and the data value. The data quality of captured data can vary greatly, affecting the accurate analysis.

Volume Challenge

Traditional tools quickly can become overwhelmed by the large volume of data

- disk space
- latency in retrieving data

Common approach:

- discard data (filtering)
- increase device storage (until the device limit is reached)
- distribute the storage in different devices working together

Velocity Challenge

Big Data analysis can be performed:

- realtime (immediate response)
- near-realtime (fast response)
- batch (huge datasets)
- custom (on-call activity)
- analytical (reports)

<u>Approaches and examples</u>

- Real time data analysis (e.g adaptive optics: deforming real time a mirror to compensate for atmospheric distortion over 0.1- 0.01s)
- Data lakes: store data without structuring (import any amount of raw data saving time by avoiding structure)
- Speed up storage using multiple disks (RAID) and distributed storage

Variety Challenge

Diversity of data acquired by different sources:

- different format
- different structure
- incomplete datasets
- complex datasets

Common approach:

- NoSQL and structured storage: embedding, referencing
- Metadata

Big Data System

System capable to deal with Big Data require:

- A method of collecting/categorizing data
- A method to transfer data
- A storage distributed, scalable, redundant
- A parallel data processing
- System monitoring tools
- Scheduling tools
- Local processing tools to reduce network bandwidth

Components and ecosystem of big data

Acquired data can’t be directly processed (variety): filtering, cleanse,...

- Storage of raw datasets (acquisition)
- Storage of (pre)processed datasets (manipulation)
- Storage of processed data/results (analysis)

Technologies and strategies:

- Techniques for analysing data, such as **machine learning**
- **Big data technologies**, like business intelligence, cloud computing and databases
- **Visualization**, such as charts, graphs and other displays of the data
- **Clusters**: tightly coupled collection of servers (nodes) to work as a single unit

What is Machine Learning ?

Machine Learning is a field of computer science that gives the ability to the computer for self-learn without being explicitly programmed. We move the computation from instructions to the data. With Big data we have now a huge amount of data available, we can construct labeled dataset and give them to machine learning algorithms. They will learn the patterns we are a looking for looking at the data, looking at the examples.

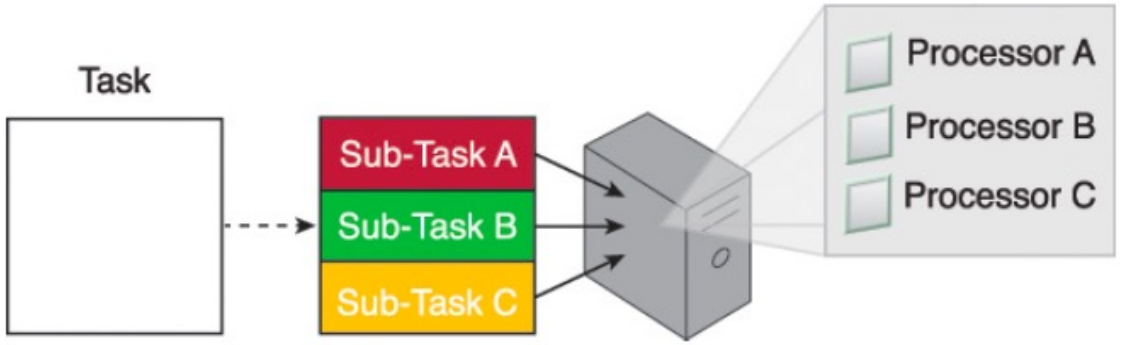
Just to make an example, we want an algorithm to detect cats in pictures. We can feed this kind of algorithm with thousand of labelled images with cats (images where a bounding box is drawn around the cat, if present). The algorithm will learn itself what are the features of a cat and will be able to detect a cat on a image he never seen before.

There are now several public labelled datasets out there, see for example imagenet.
This is just an example! Machine Learning is a huge topic!

Partitioning

Speed up the processing of large amounts of data require partitioning

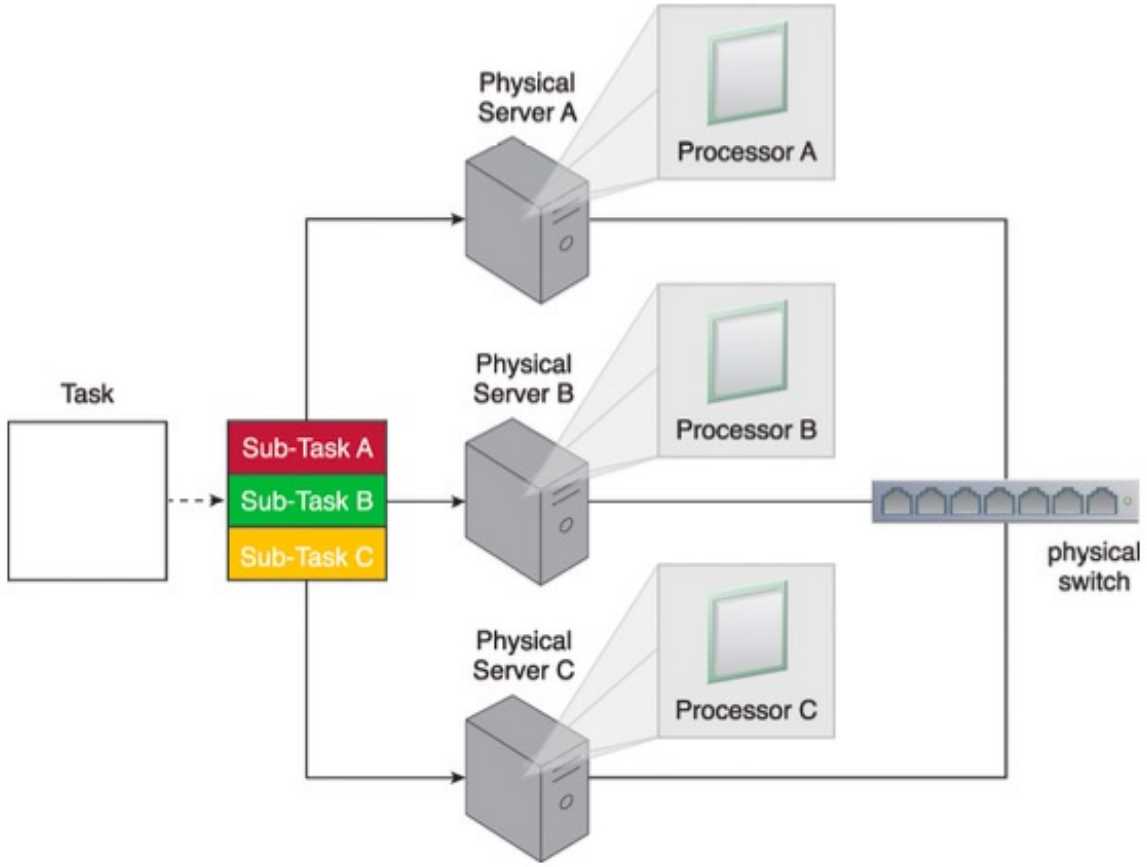
Parallel processing: reducing time by dividing large task into small sub-tasks



Partitioning

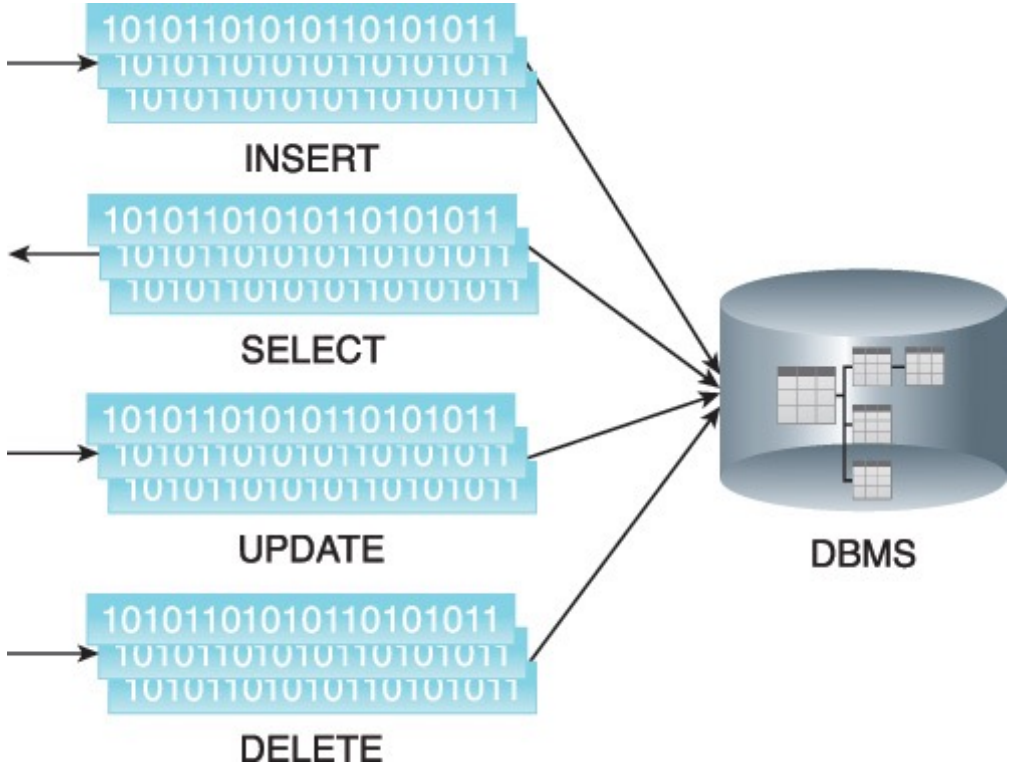
Speed up the processing of large amounts of data require partitioning

Distributed processing:
reducing time by executing
sub-tasks in different machines



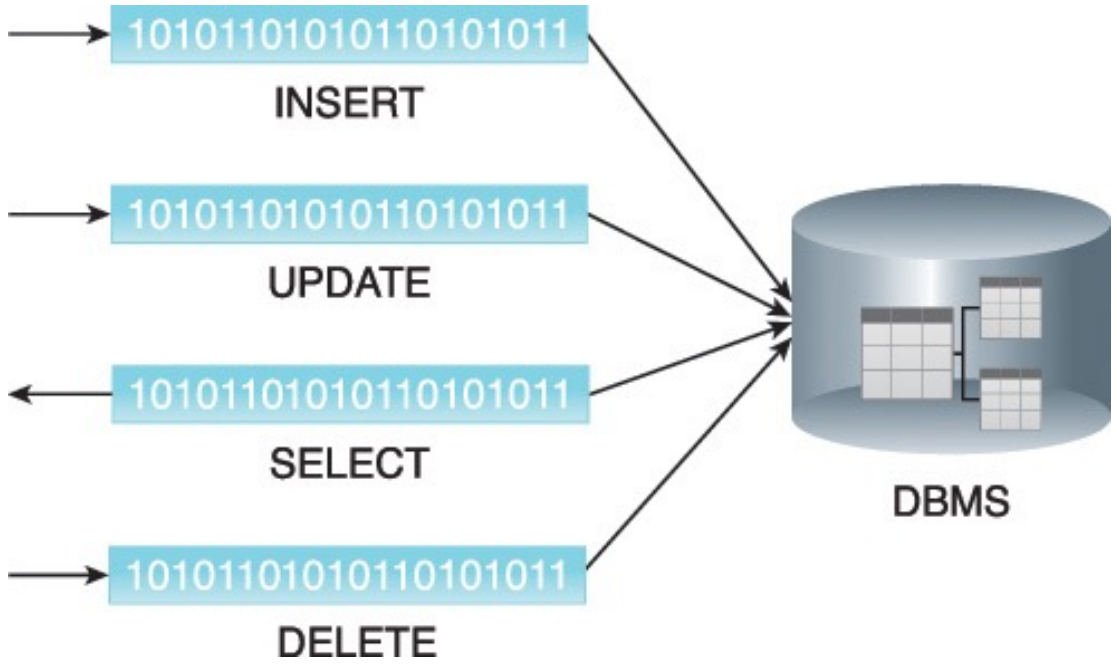
Batch Processing:

- offline processing
- large amounts of data querying - reading - writing
- data stored on disk
- high latency - min to hours
- easy to set up and low-cost

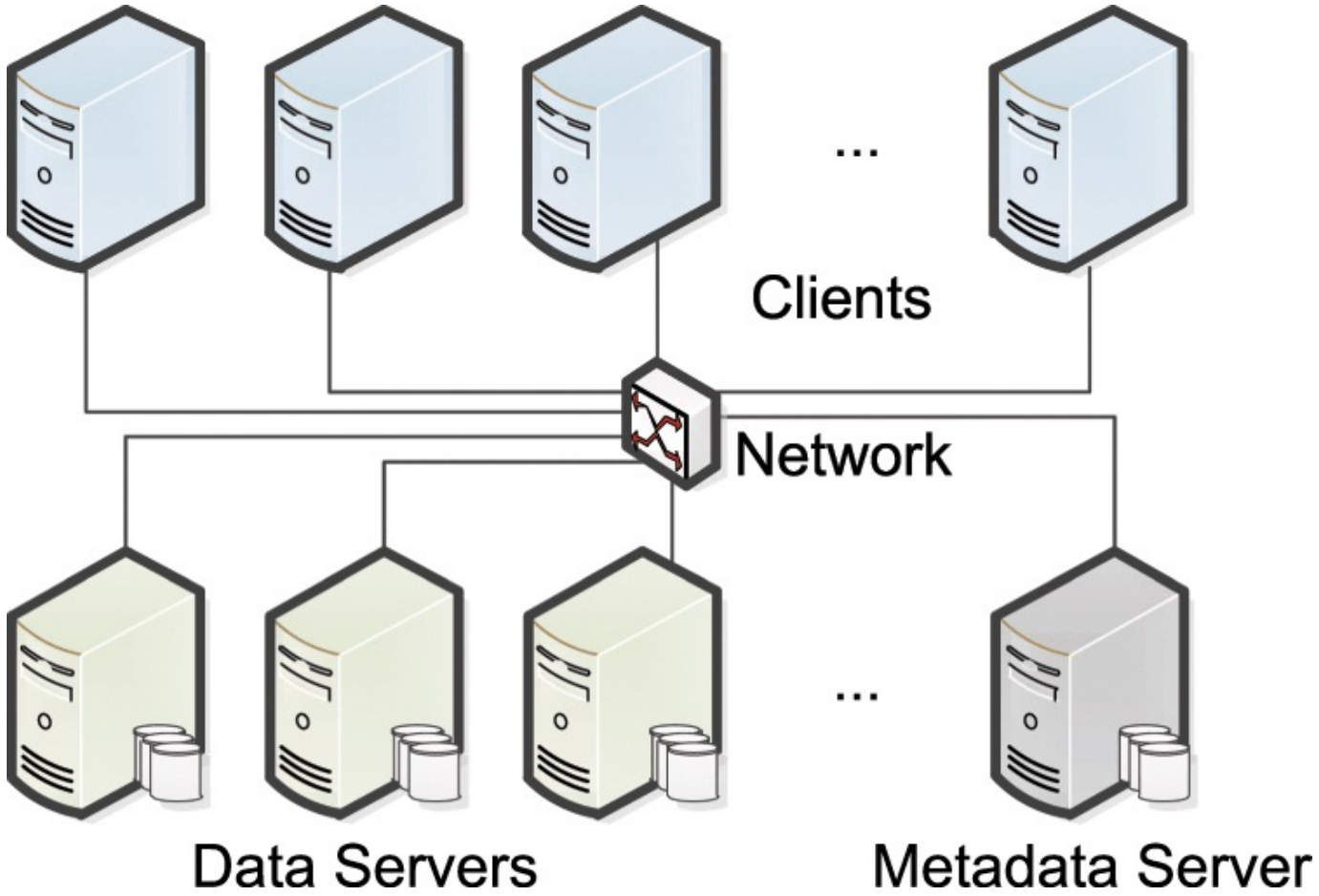


Transactional Processing:

- online processing (realtime)
- processing in-memory, then storage
- low latency (< 1min)
- small amounts of data but continuous



New tools for Big Data



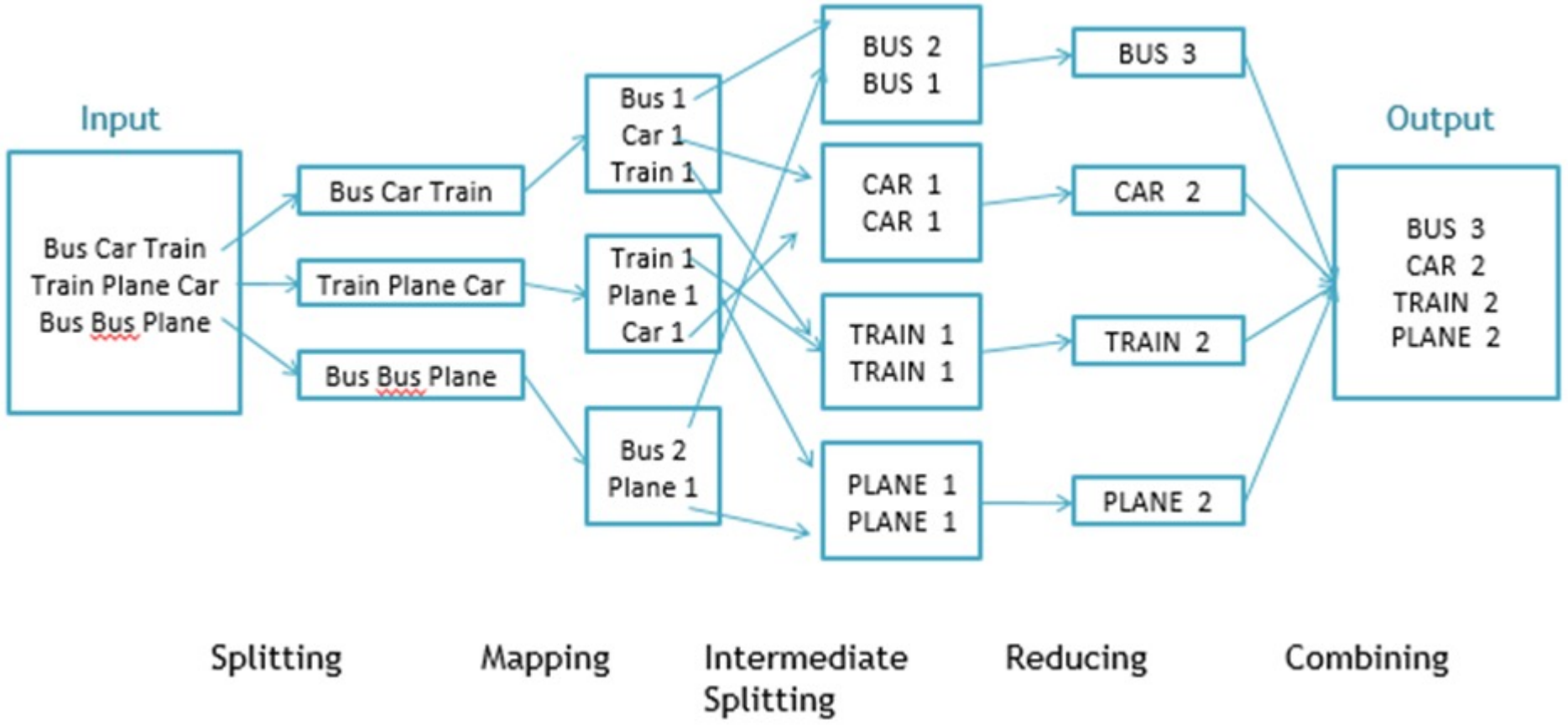
- Distributed File System
- Pallelism

Algorithm for Big Data: MapReduce

With Big Data we need new tools in order to manage data in a reasonable time we need to use **parallelism**. Data can also be stored in a distributed storage.

MapReduce is a computation that decomposes large manipulation jobs into individual tasks that can be executed in parallel across a cluster of servers. The results of tasks can be joined together to compute final results.

Algorithm for Big Data: MapReduce



Algorithm for Big Data: MapReduce

MapReduce consists of 2 steps:

Map Function – It takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (Key-Value pair).

Input	Set of data	Bus, Car, bus, car, train, car, bus, car, train, bus, TRAIN,BUS, buS, caR, CAR, car, BUS, TRAIN
Output	Convert into another set of data (Key,Value)	(Bus,1), (Car,1), (bus,1), (car,1), (train,1), (car,1), (bus,1), (car,1), (train,1), (bus,1), (TRAIN,1),(BUS,1), (buS,1), (caR,1), (CAR,1), (car,1), (BUS,1), (TRAIN,1)

Algorithm for Big Data: MapReduce

Reduce Function – Takes the output from Map as an input and combines those data tuples into a smaller set of tuples

Input (output of Map function)	Set of Tuples	(Bus,1), (Car,1), (bus,1), (car,1), (train,1), (car,1), (bus,1), (car,1), (train,1), (bus,1), (TRAIN,1),(BUS,1), (buS,1), (caR,1), (CAR,1), (car,1), (BUS,1), (TRAIN,1)
Output	Converts into smaller set of tuples	(BUS,7), (CAR,7), (TRAIN,4)

Algorithm for Big Data: MapReduce

Workflow of MapReduce consists of 5 steps:

1. **Splitting** – The splitting parameter can be anything, e.g. splitting by space, comma, semicolon, or even by a new line.
2. **Mapping** – as explained above.
3. **Intermediate splitting** – the entire process in parallel on different clusters. In order to group them in "Reduce Phase" the similar KEY data should be on the same cluster.
4. **Reduce** – it is nothing but mostly group by phase.
5. **Combining** – The last phase where all the data (individual result set from each cluster) is combined together to form a result.

MapReduce: a real example

Facebook has a list of friends (note that friends are a bi-directional thing on Facebook. If I'm your friend, you're mine). They also have lots of disk space and they serve hundreds of millions of requests everyday.

They've decided to pre-compute calculations when they can to reduce the processing time of requests. One common processing request is the "You and … have xxx friends in common" feature.

When you visit someone's profile, you see a list of friends that you have in common. This list doesn't change frequently so it'd be wasteful to recalculate it every time you visited the profile.

We're going to use mapreduce so that we can calculate everyone's common friends once a day and store those results. Later on it's just a quick lookup. We've got lots of disk, it's cheap.

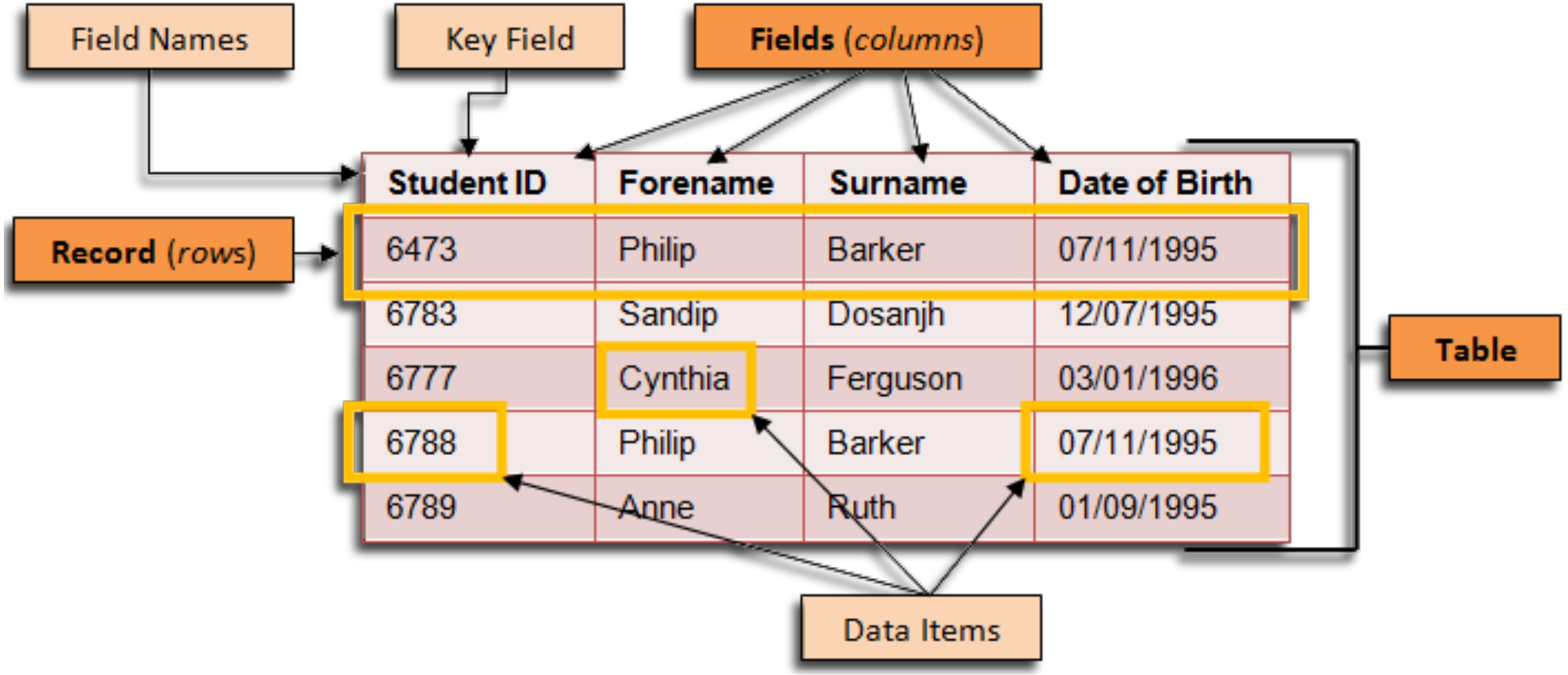
Database

Database and Database management systems (DBMS)

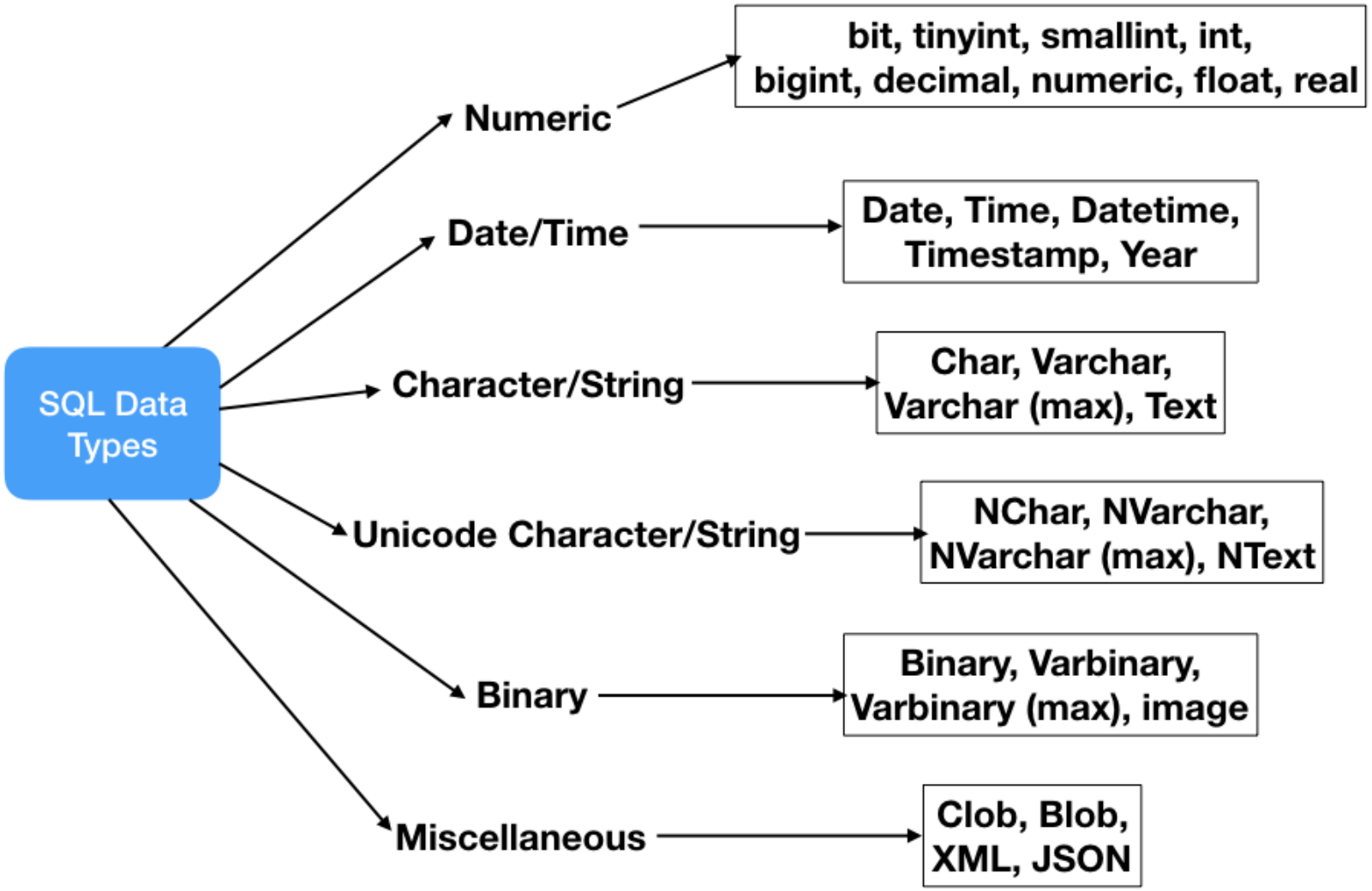
A Database is an organized collection of data, generally stored and accessed electronically from a computer system. A database is essentially a collection of tables. For example:

ID	First_name	Last_name	Age
1	John	Red	25
2	Mark	Red	60
3	Alice	Smith	35

Database main concepts



Datatypes



For each field (column) we have a specific datatype we define during table creation.

Database and Database management systems (DBMS)

Database Management System (DBMS) is software to maintain and utilize the collections of data (MySQL, Oracle, Micorsoft Access, Libreoffice Base). Is the software that interacts with end users, applications, and the database itself to capture and analyze the data.

Query

A precise request for information retrieval with database and information systems

Structured Query Language (SQL)

Structured Query Language (SQL) is a standard computer language for relational database management and data manipulation. SQL is used to query, insert, update and modify data. Most relational databases support SQL. Some slightly difference in syntax can occur between different implementations (e.g. Microsoft Access, Libreoffice Base), but the concept we are going to learn are basic for every DBMS.