Lecture 2 – Big Data

Open Data Management & the Cloud

(Data Science & Scientific Computing / UniTS – DMG)

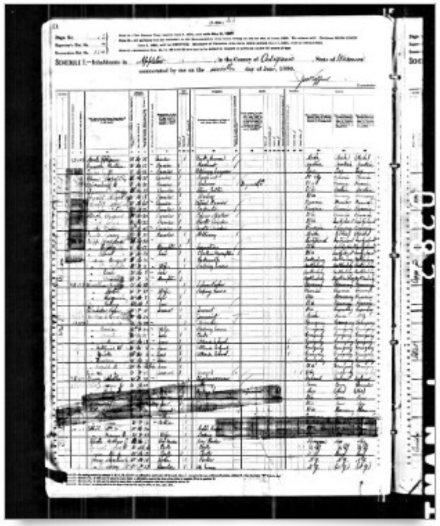
Big Data

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

History

https://www.winshuttle.com/big-data-timeline/

History



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.

‹ 1875 What is Big Data?

› 1881 The Hollerith Tabulating Machine

What is Big Data?

The Start of Information Overload

The Hollerith Tabulating Machine

1870 1874 1878 1880 1882 1886 1890 1894

History

Source: Frguentsch

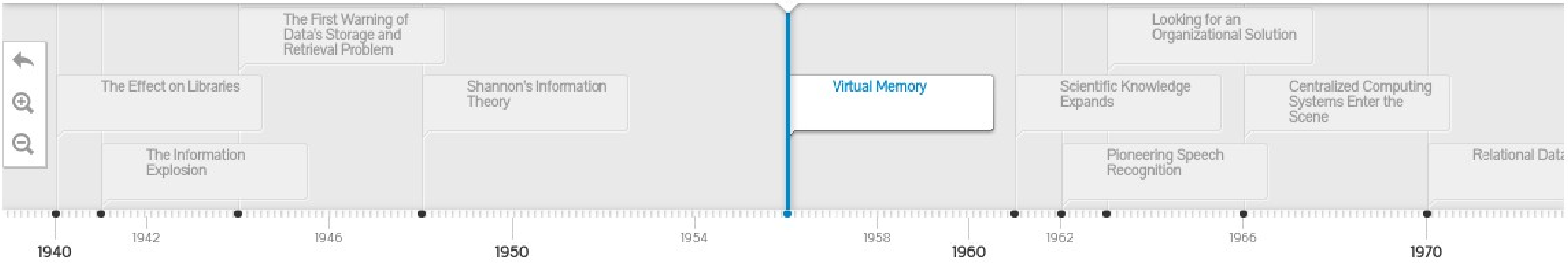
1956

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

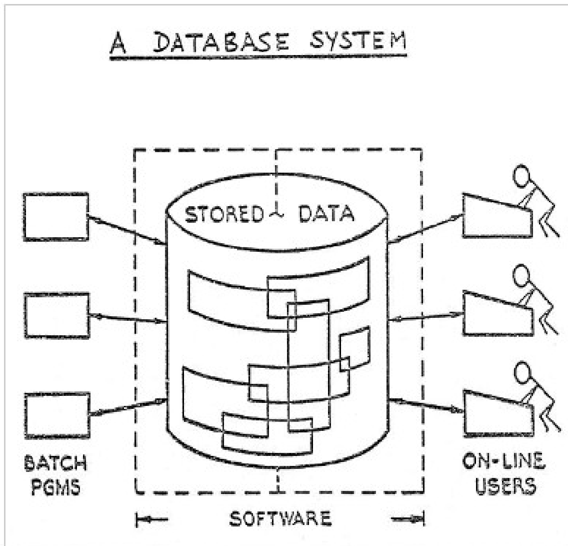
‹ 1948 Shannon's Information Theory

› 1961 Scientific Knowledge Expands



History

1966
Centralized Computing Systems Enter the Scene



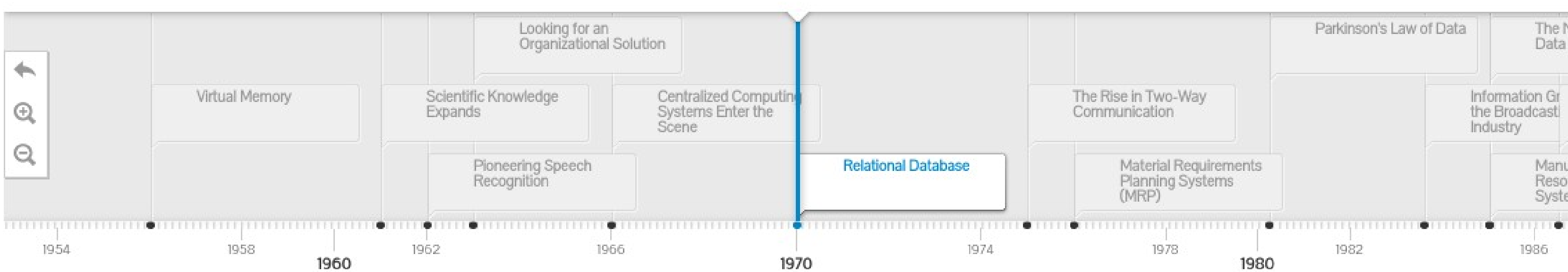
Source: IBM.com

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

1975
The Rise in Two-Way Communication



History

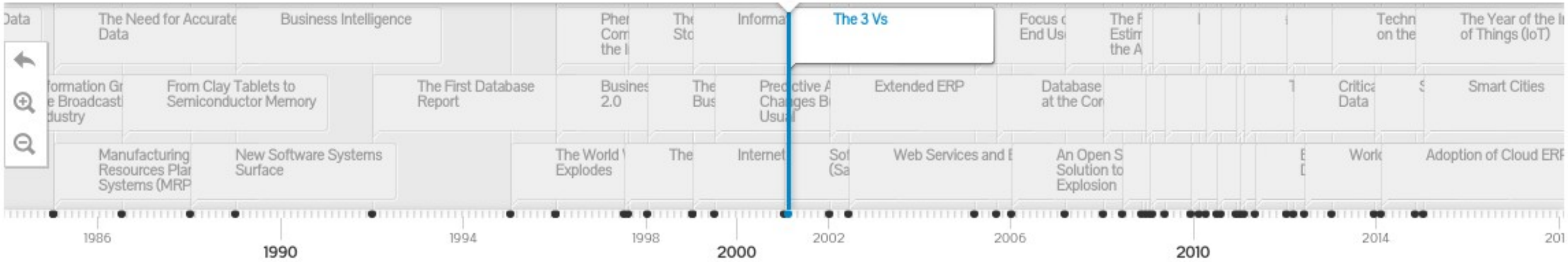
2001
Software as a Service (SaaS)

February 2001

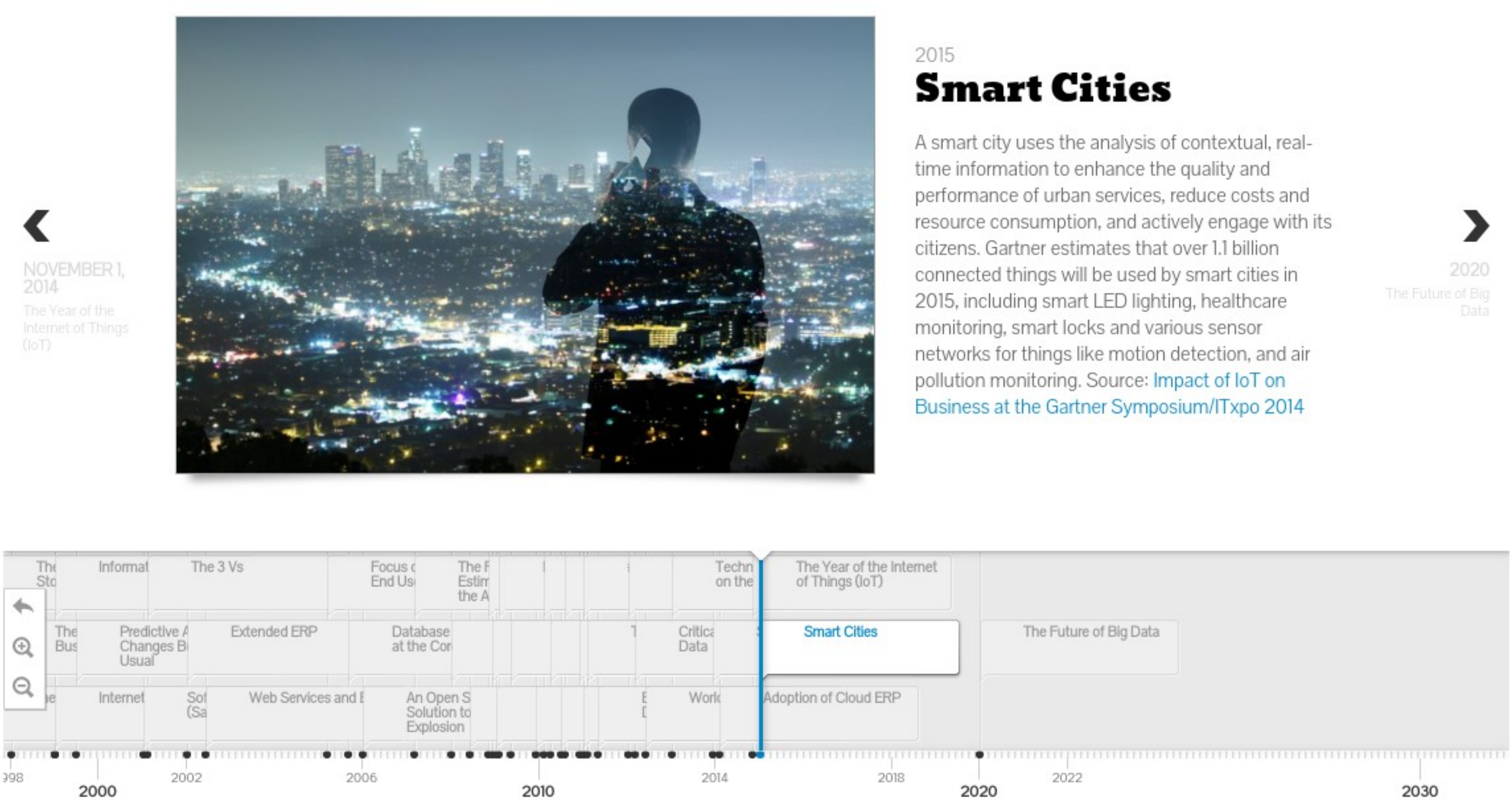
The 3 Vs

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.

2002
Extended ERP



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
NOVEMBER 1, 2014
The Year of the Internet of Things (IoT)
2020
The Future of Big Data
The 3 Vs
Extended ERP
The Year of the Internet of Things (IoT)
Smart Cities
The Future of Big Data
Internet
Web Services and
Adoption of Cloud ERP
2000
2002
2006
2010
2014
2018
2020
2022
2030

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

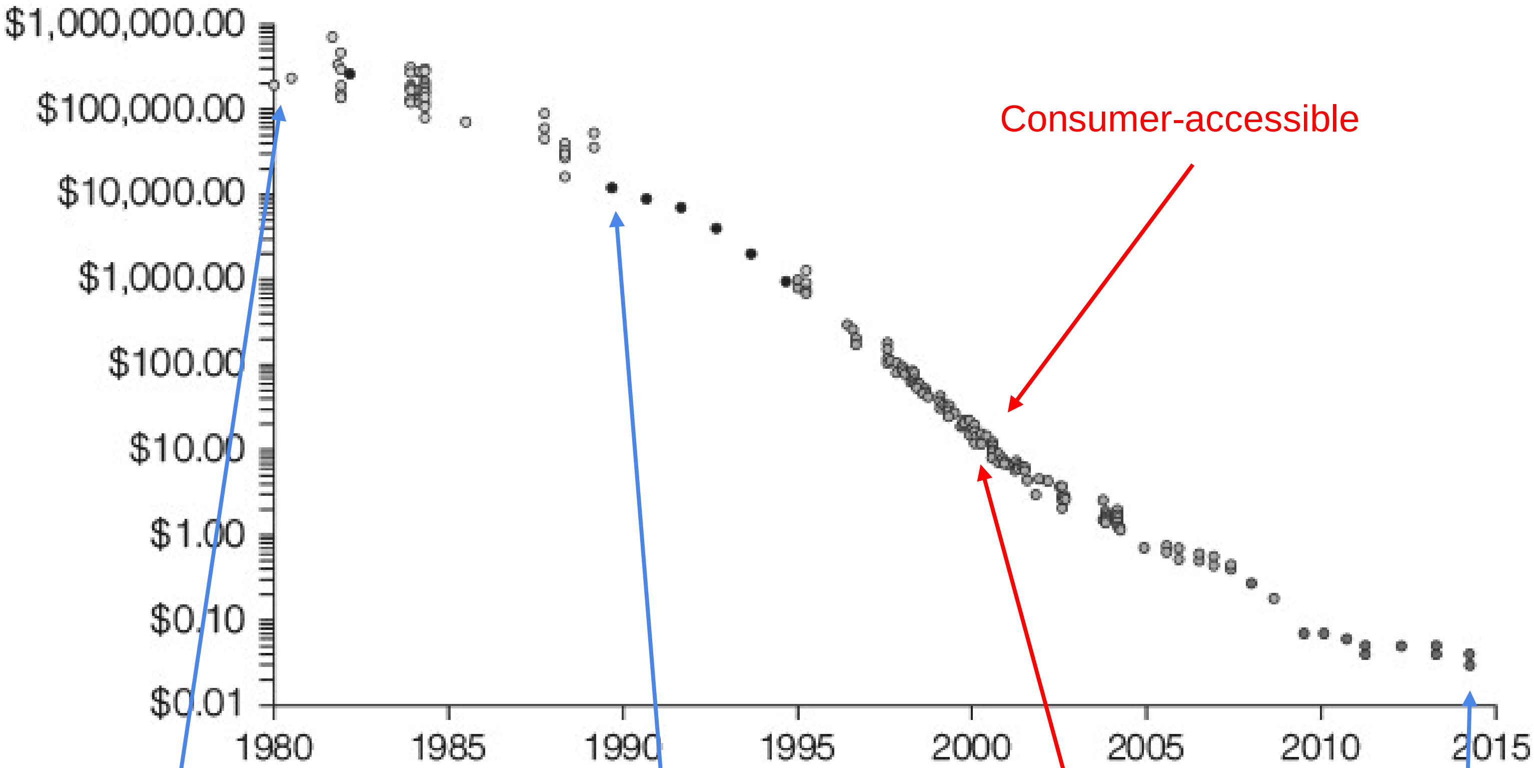
larger and complex experiments, …

Digital Storage cost

Digital storage:

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

Disk storage cost



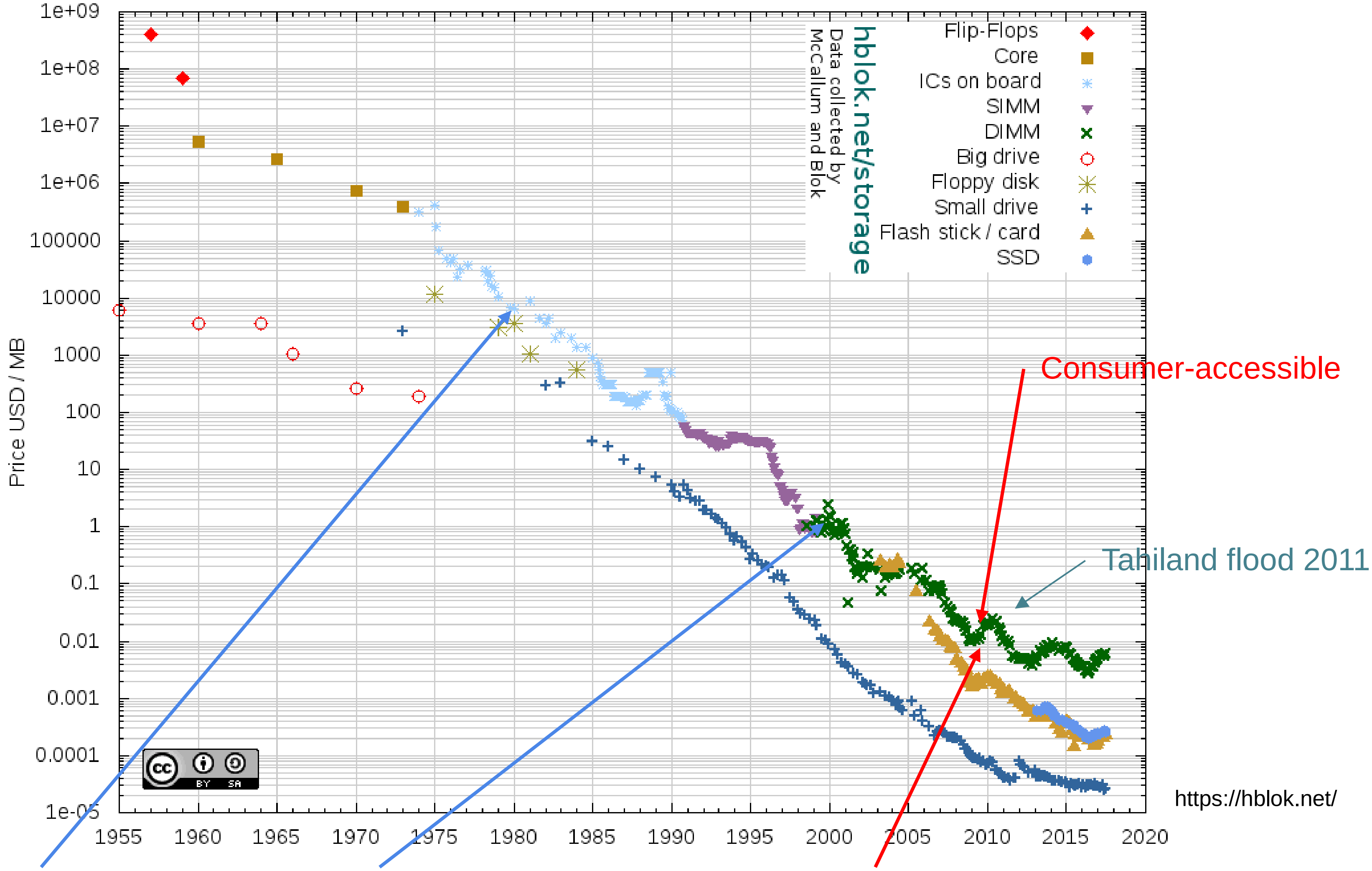
1980: 200 K$/GB

1990: 9 K$/GB

2000: 10 $/GB

2015: 0.3 $/GB

RAM cost

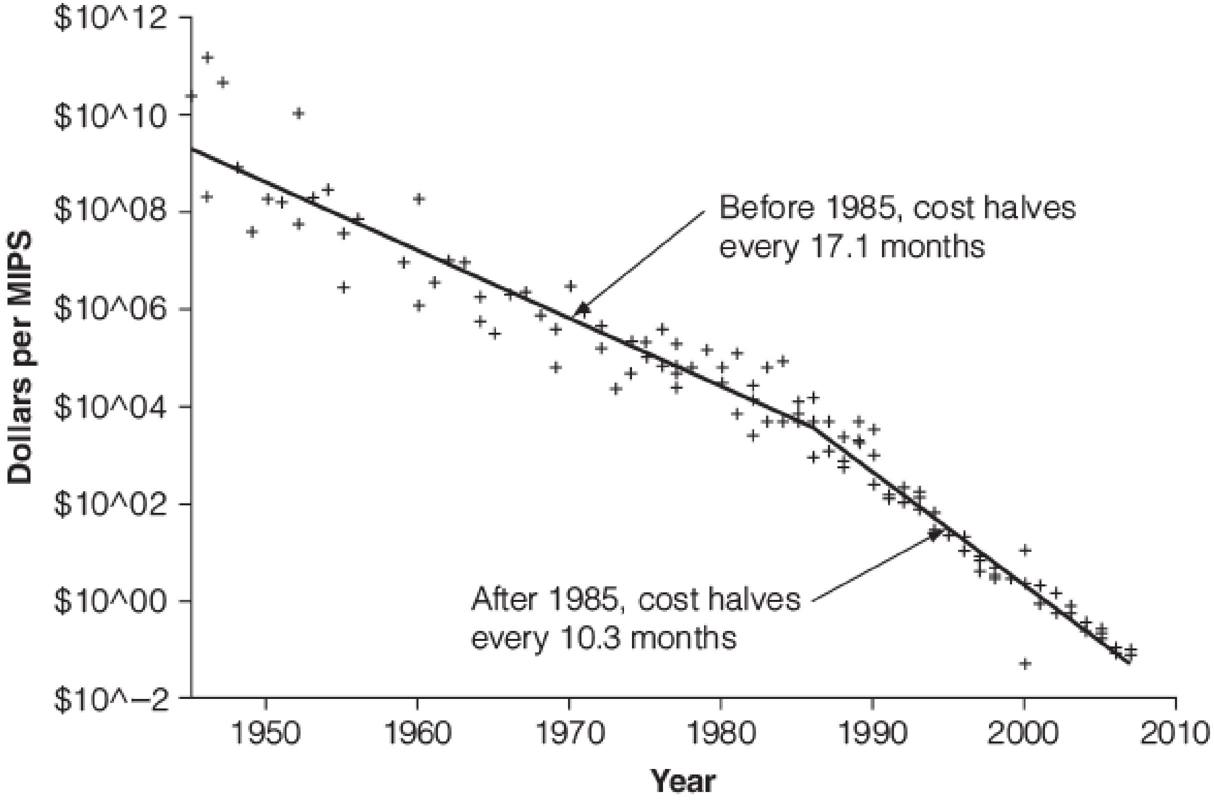


1980: 6 M$/GB

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

Big Data - characteristics

Big Data can be defined in terms of how the data will be manipulated

1) VOLUME

   Quantity of data to be stored: affects storage, processing, latency

2) VELOCITY

   How rapidly data accumulates: affects capture, storage

   (SKA completed will reach 750 TB per second)

   How fast the data should be processed: affects processing, latency, storage

   (velocity is not necessary a volume challenge → real time)

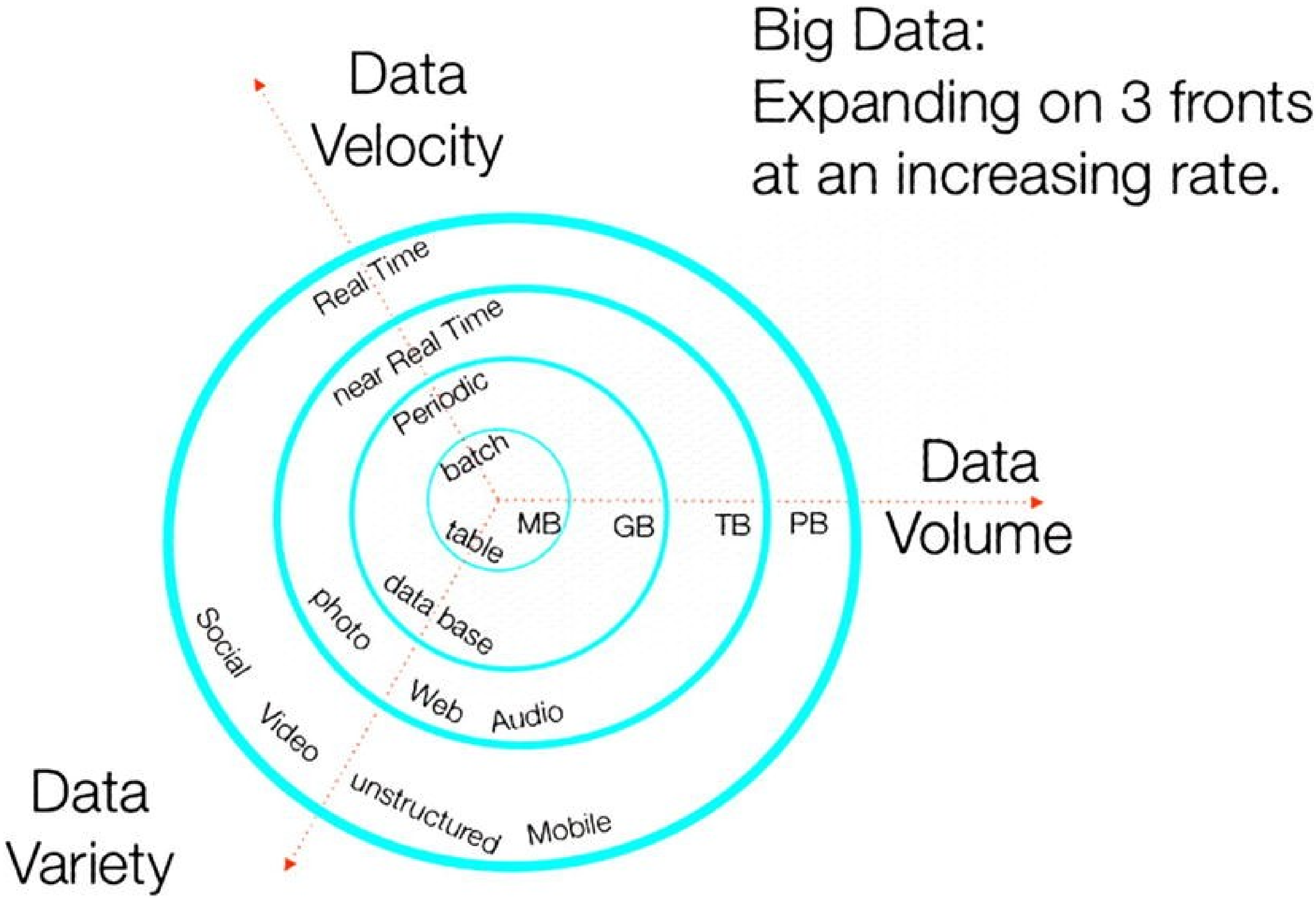
3) VARIETY

   Wide range of different datasets (logs, photo, video,..)

   Unstructured

   Incomplete

Big Data: 3V increasing



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)

Approaches and examples

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

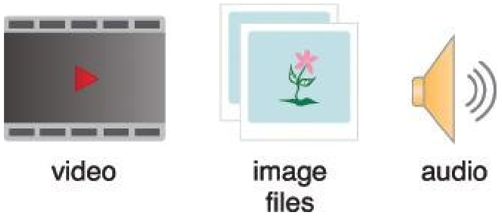
System capable to deal with Big Data require:

- 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

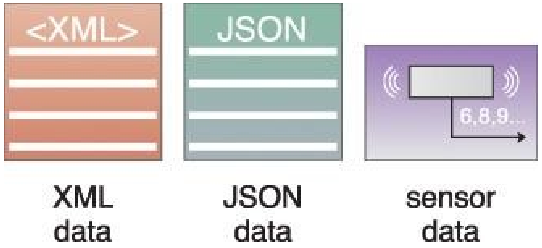
Big Data: type

- Structured: conforms to a data model or a schema
  Express relations between entities, generally stored in relational database

- Unstructured: not conforming to fixed data model o schema
  Special purpose logic required to process (i.e codecs for video) cannot be directly processed or queried using SQL: stored as a Binary Large Object (BLOB) or NoSQL database



- Semi-structured: hierarchical or graph-based structure
  have some level of structure, self describing



Metadata: information about a dataset characteristics and structure crucial to Big Data processing, storage and analysis because it provides information about the data

Big Data Storing Concepts

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)

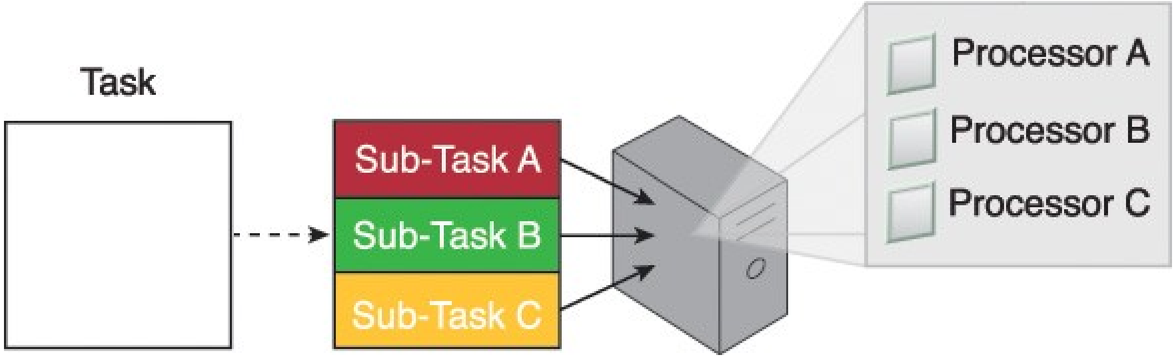
Need to store multiple copies of Big Data datasets: technologies and strategies

- clusters: tightly coupled collection of servers (nodes) to work as a single unit
  - distributed files systems: store large files spread across the nodes of a cluster (GFS, HDFS)
  - databases: RDBMS, NoSQL (structured storage)
  - Distribution models to access data: Sharding, replication

Big Data Processing Concepts - partitioning

Speed up the processing of large amounts of data require partitioning

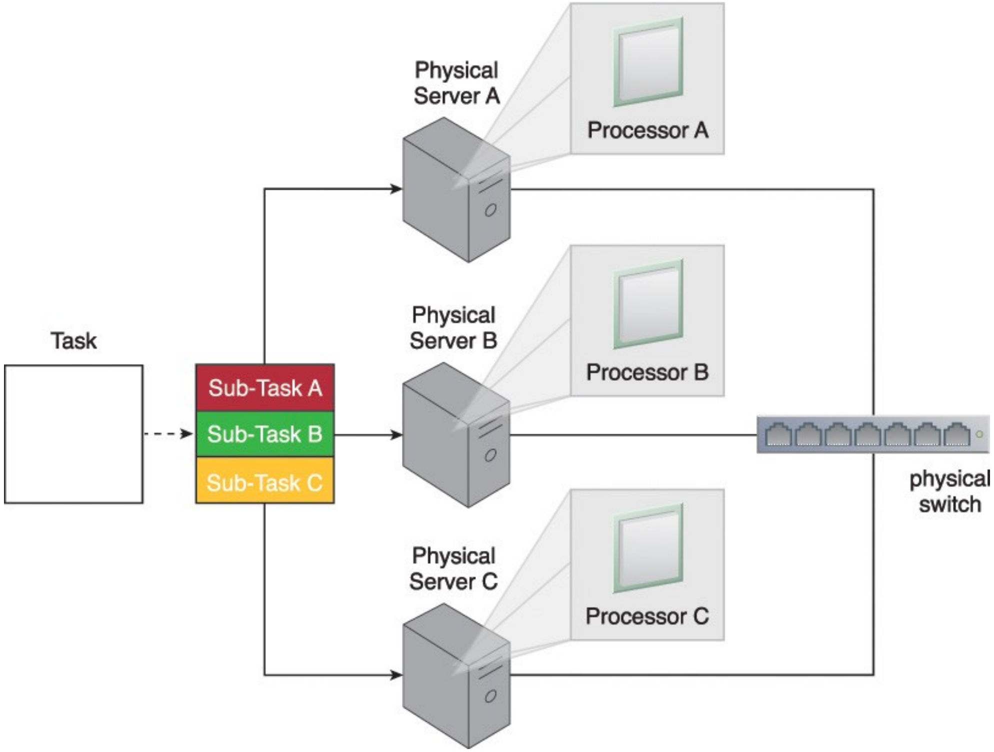
Parallel processing: reducing time by dividing large task into small sub-tasks



Big Data Processing Concepts - partitioning/2

Speed up the processing of large amounts of data require partitioning

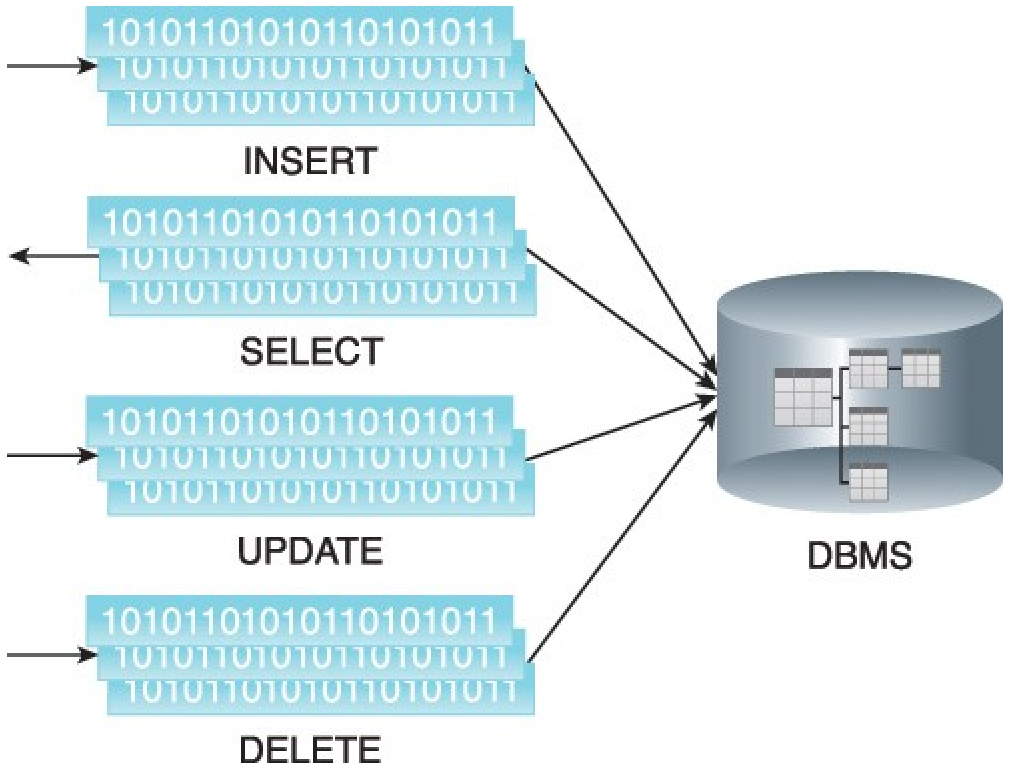
Distributed processing: reducing time by executing sub-tasks in different machines



Big Data Processing Concepts - processing

Batch Processing:

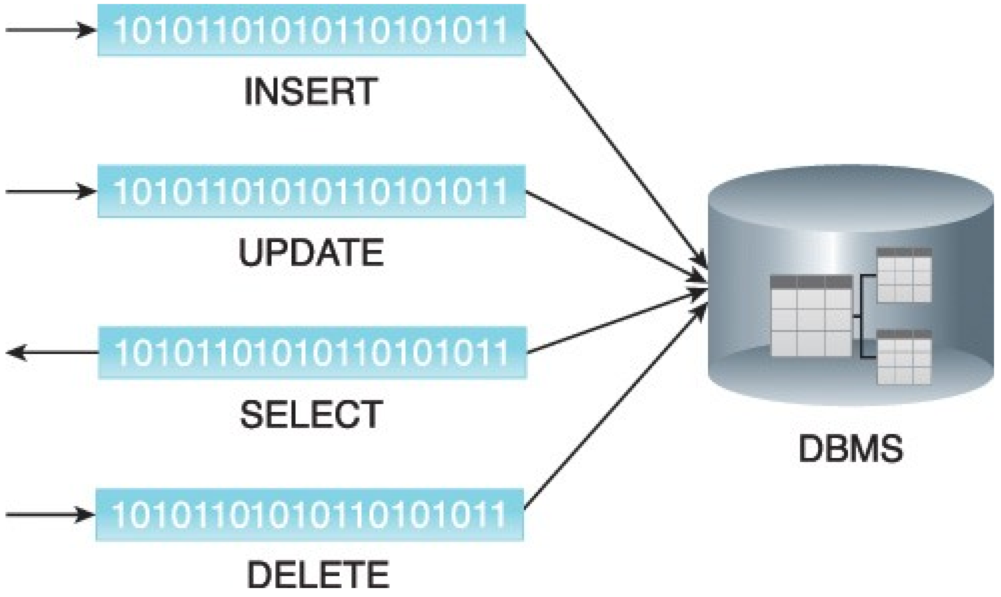
- offline processing
- large amounts of data querying - reading - writing.
- data stored on disk
- high latency - min to hours
- easy to set up and low-cost



Big Data Processing Concepts - processing/2

Transactional Processing:

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



Big Data in Science - LHC CERN

Large Hadron Collider uses detector to analyze particles produced by collisions in the accelerator

27 km ring of superconducting magnets

Collision energy of 14 TeV

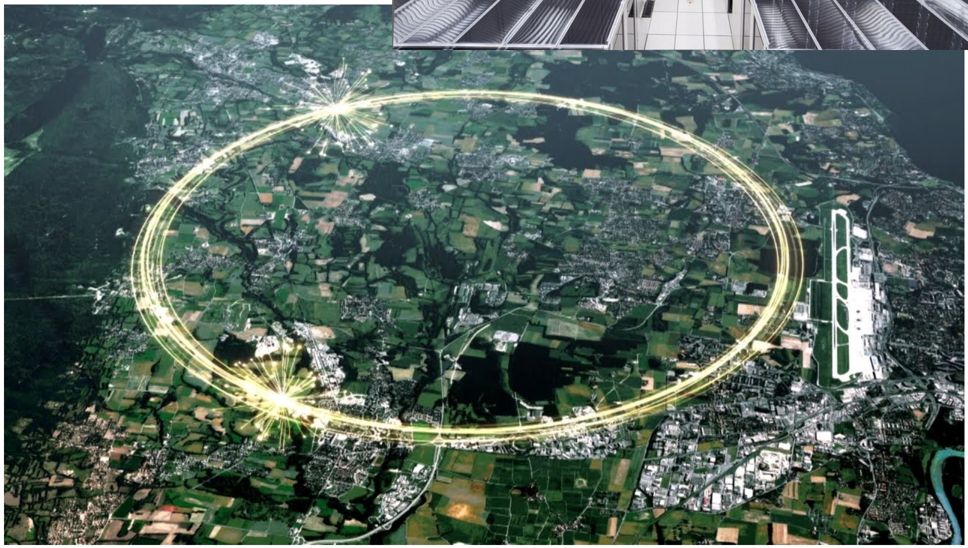
10^9 collisions per second

3.5 MW for computing

45 PB storage, 1 PB/day processed

100.000 cores

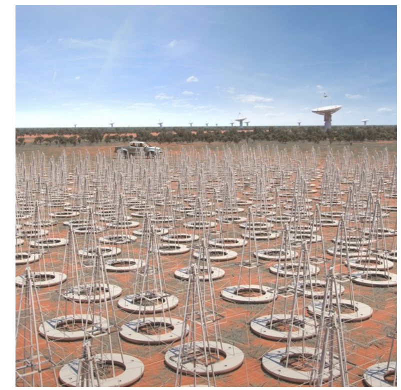
200 PB of permanent tape storage



https://home.cern

Big Data in Science - SKA

Square Kilometre Array is the largest international radio telescope



Australia - low freq: 512 stations with 250 antennas

South Africa - mid freq: 133 antennas of 64m

Data transfer antenna -->processing

2020: 20000 PB/day

2028: 200000 PB/day

Imaging:

2020: 100 PBytes/day

2028: 10000 PBytes/day

Processing power:

2020: 300 PFlop

2028: 30 EFlop

Computing & data for the SKA Peter Braam Oct 2016

Big Data in Science - EUCLID

ESA cosmology mission to map the evolution of cosmic structures - 4 yr mission

2 instruments VISible imager, Near-InfraRed Spectrometer

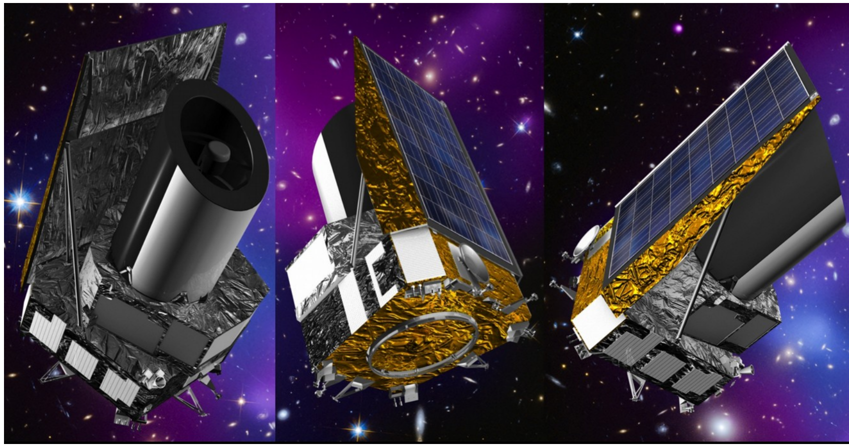
850 Gbit of raw data (compressed) in 4h download

Final data: 1Pbit/year processed

12 Science Data Centres (1 per country)

20 fields (images) per day ~ 30PB images tot

10^{10} galaxies observed



The Euclid mission design, 1610.05508