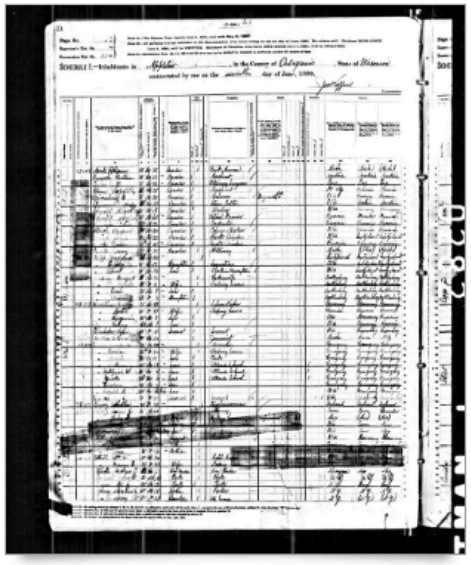
Lecture 2 - Big Data

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

History



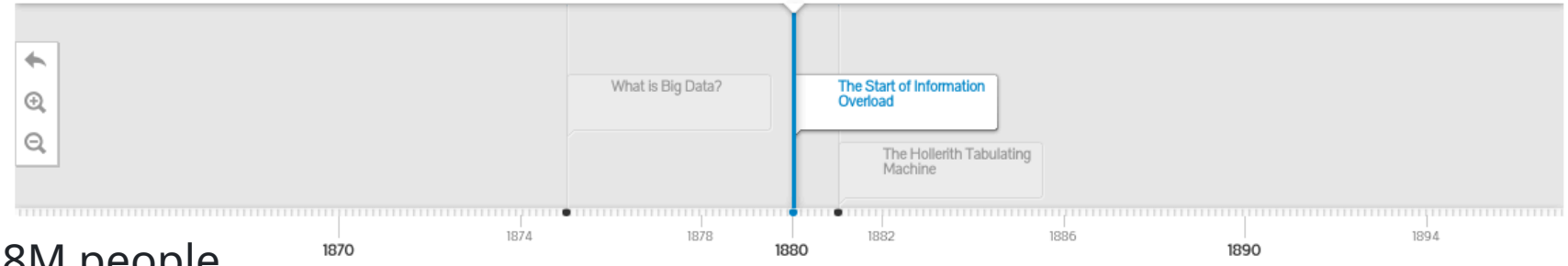
1880

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

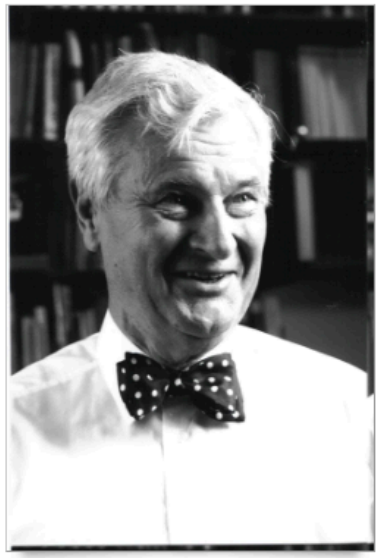


U.S. Census:

- 1870: ~38M people
- 1880: ~50M people
- 1890: ~63M people

History

1948
Shannon's Information Theory

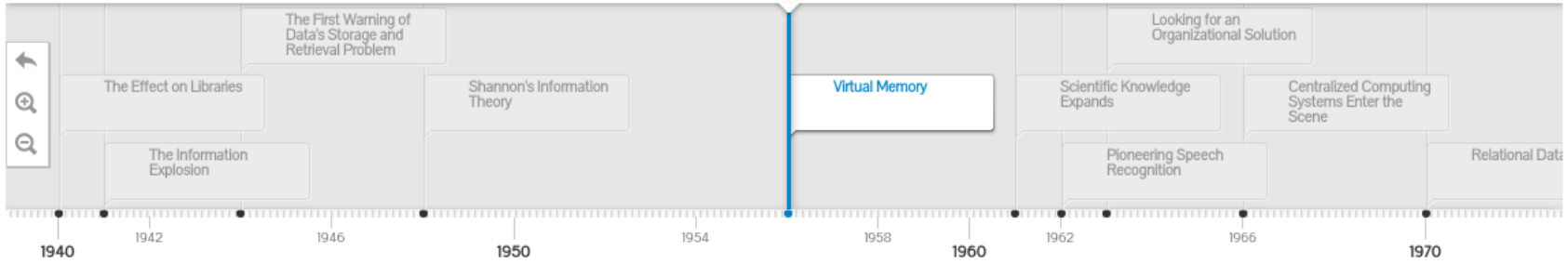


Source: Frguentsch

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



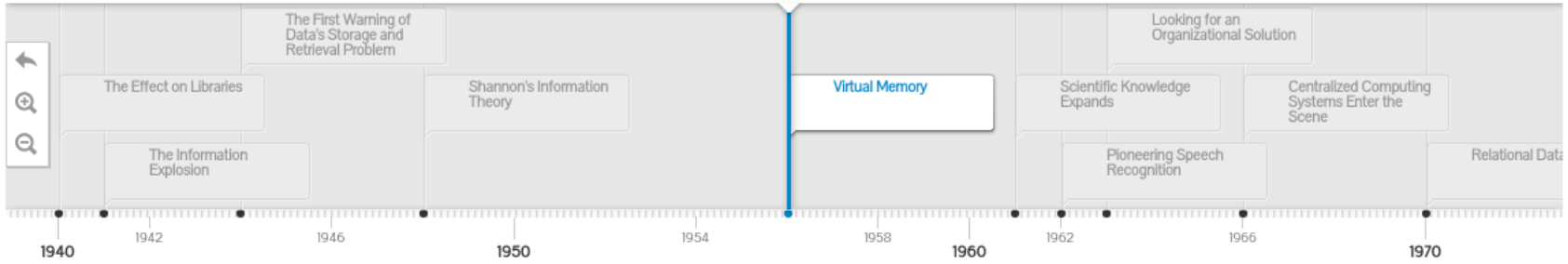
History

1956

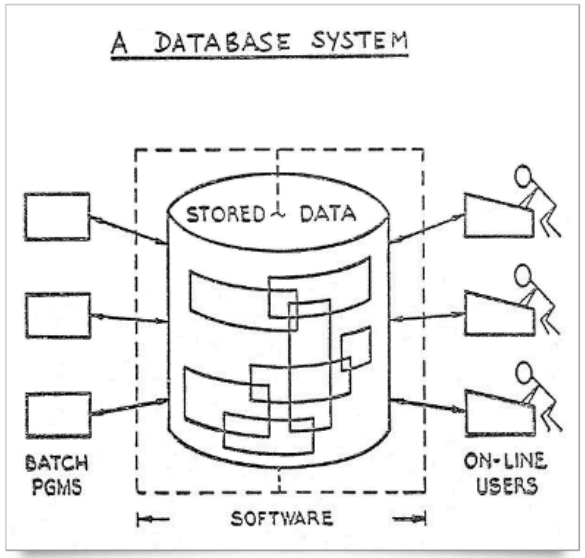
Virtual Memory

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

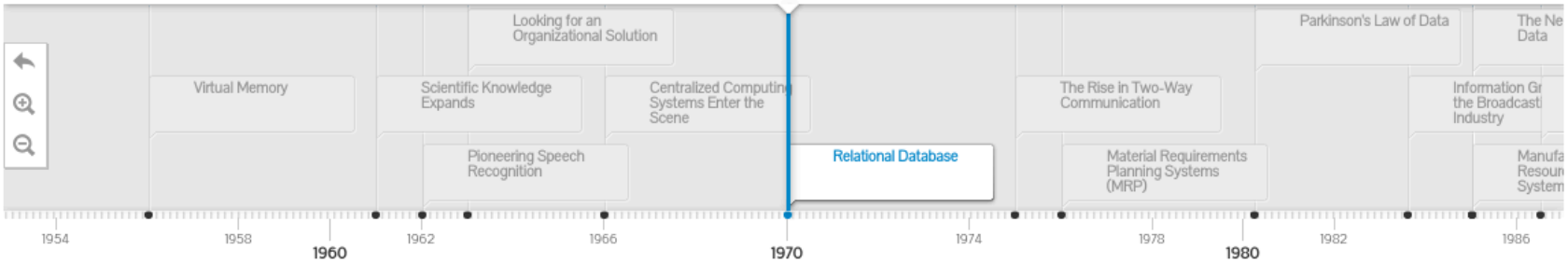
1970

Relational Database

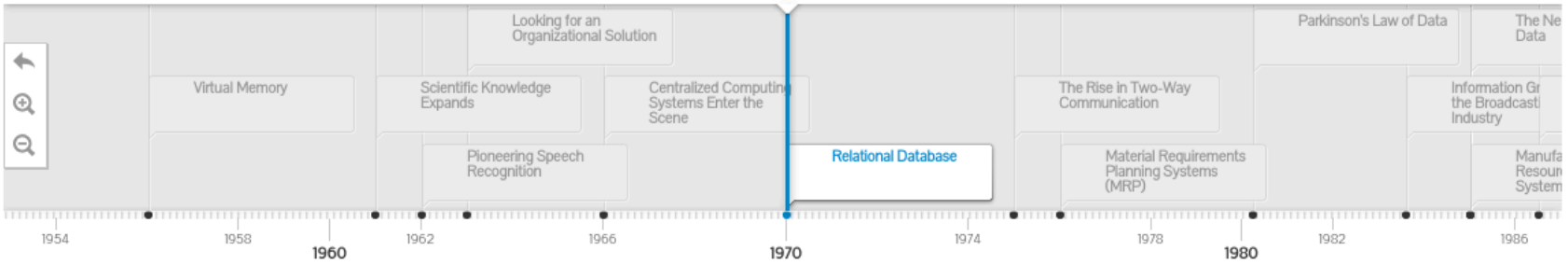
In 1970, Edgar F. Codd, an Oxford-educated mathematician working at the IBM Research Lab, published a paper showing how information stored in large databases could be accessed without knowing how the information was structured or where it resided in the database. Until then, retrieving information required 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

1966
Centralized Computing Systems Enter the Scene

1975
The Rise in Two-Way Communication



History


Looking for an Organizational Solution
Parkinson's Law of Data
Virtual Memory
Scientific Knowledge Expands
Centralized Computing Systems Enter the Scene
The Rise in Two-Way Communication
Pioneering Speech Recognition
Relational Database
Material Requirements Planning Systems (MRP)
1954
1958
1960
1962
1966
1970
1974
1978
1980
1982
1986

History

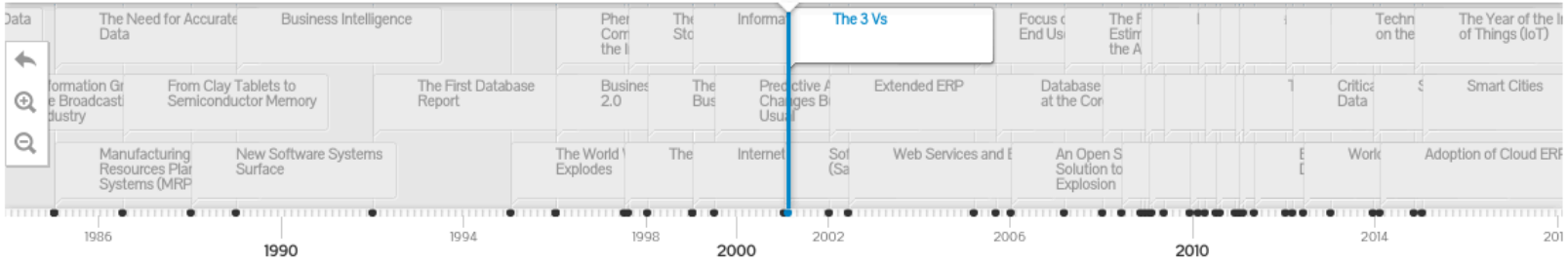
February 2001

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.

2001
Software as a Service (SaaS)

2002
Extended ERP



History

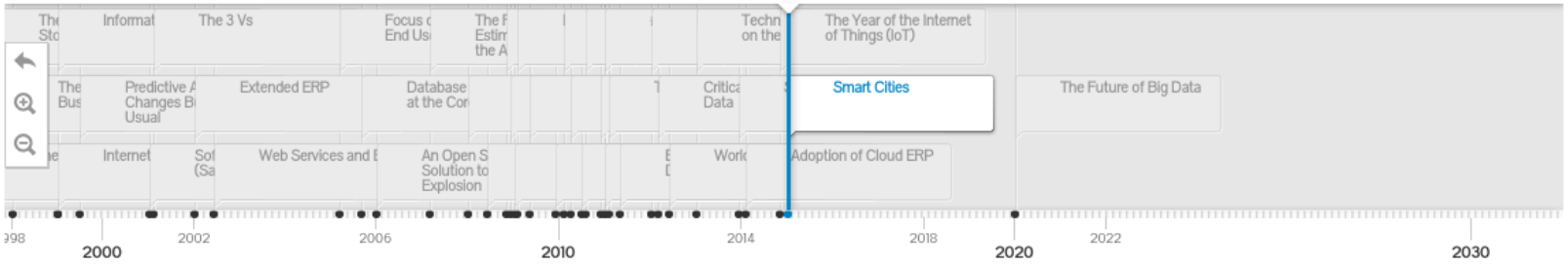
2015

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



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
  - Internet of Things (IoT)

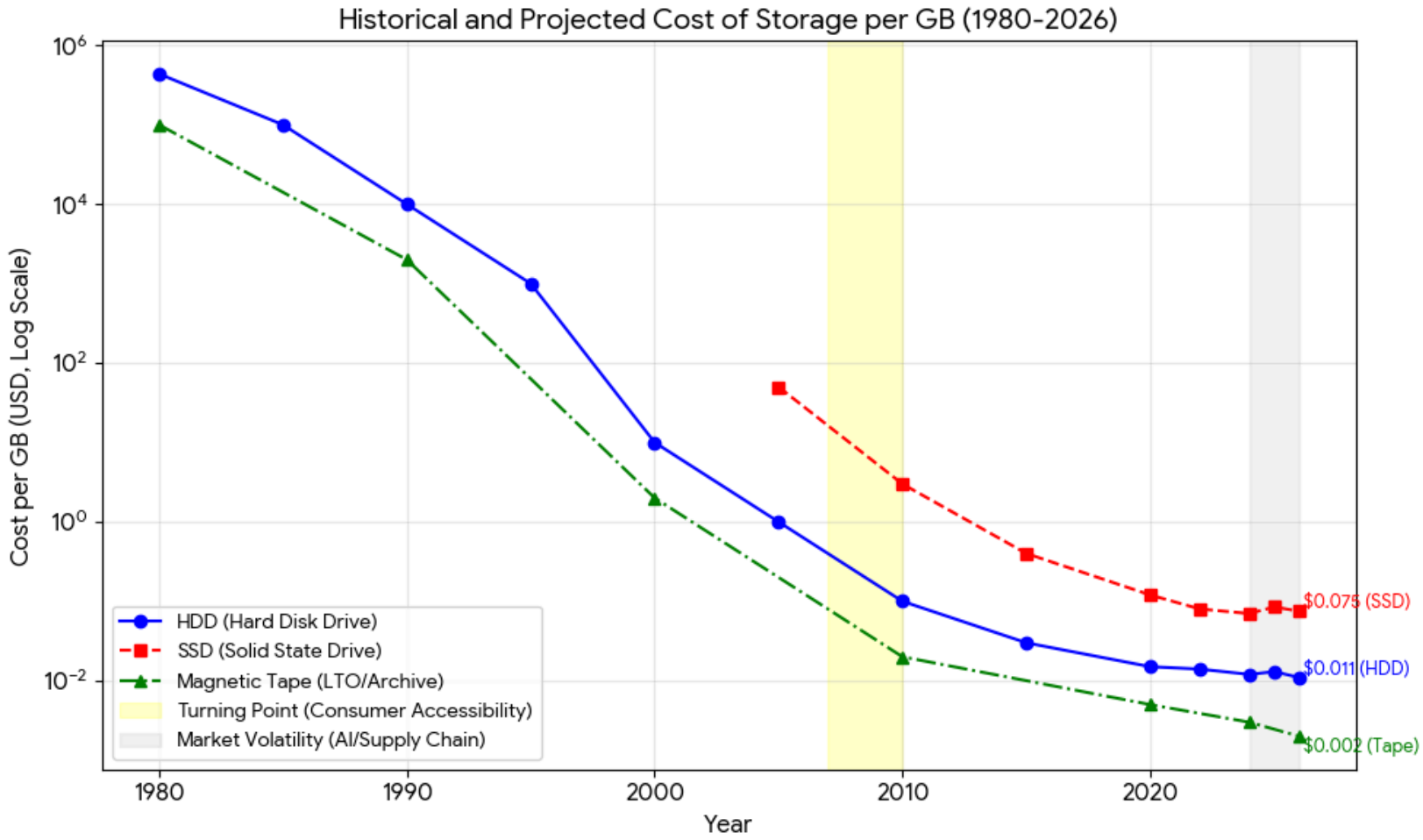
Generating Digital Data

- **Self-published content**: FB, Blogs, YouTube, Instagram, etc.
  - technology completely changed and facilitated publishing: massive growth in human-generated content
- **Consumer Activity**: business and marketing
  - digital footprint, tracking, insights, security cameras, etc.
- **Machine data and IoT**
  - devices exchanging data, integration of physical world into computer-based systems, connectivity, etc.
- **Science**
  - larger and complex experiments

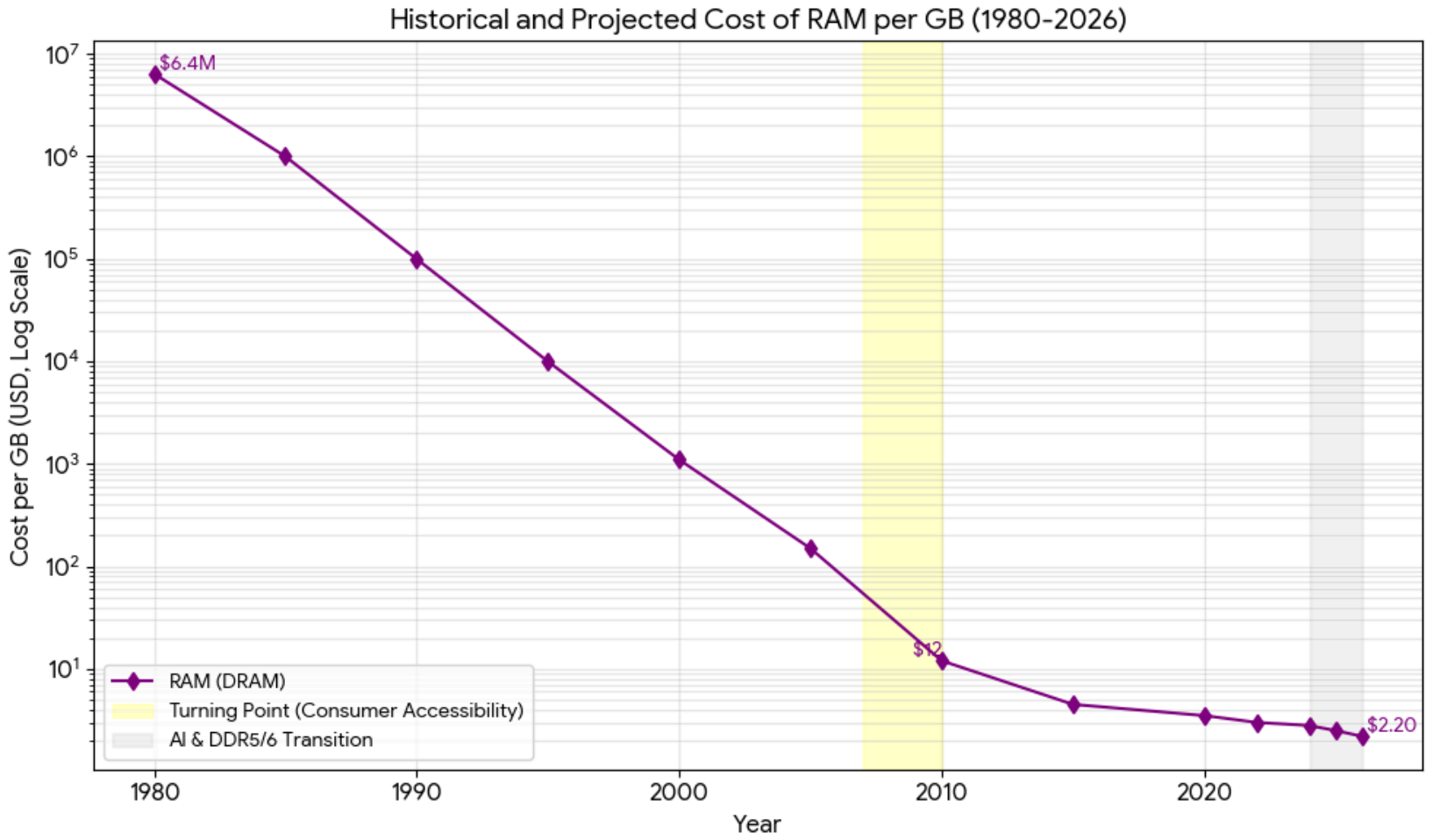
Digital Storage Cost

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

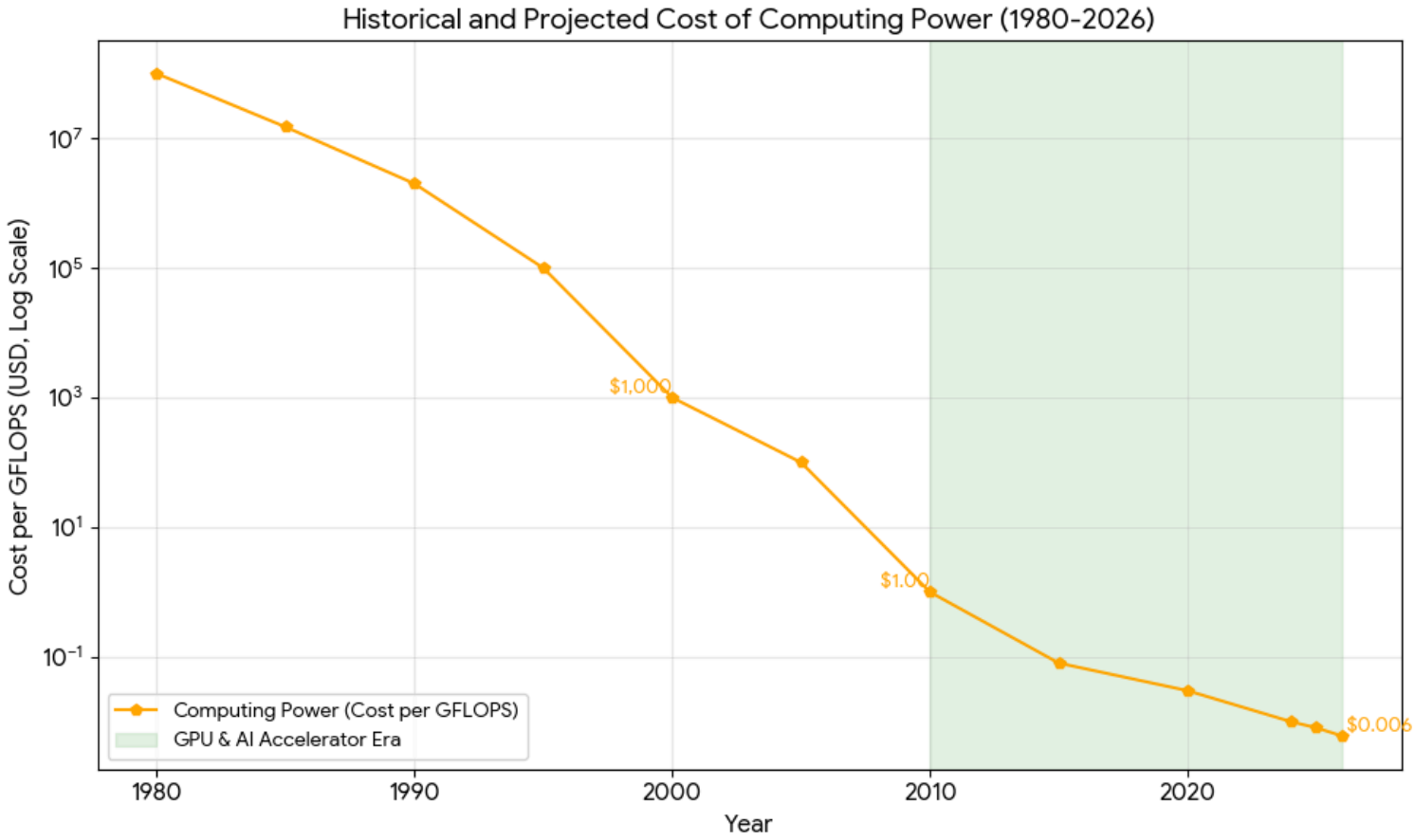
Disk Storage Cost



RAM Cost


Historical and Projected Cost of RAM per GB (1980-2026)
Cost per GB (USD, Log Scale)
10^7
10^6
10^5
10^4
10^3
10^2
10^1
$6.4M
$12
$2.20
RAM (DRAM)
Turning Point (Consumer Accessibility)
AI & DDR5/6 Transition
1980
1990
2000
2010
2020
Year

Computing Power Cost



Why Is Big Data Useful?

We should change our perspective and look at Big Data more as a challenge than as a problem

New ways to use data:

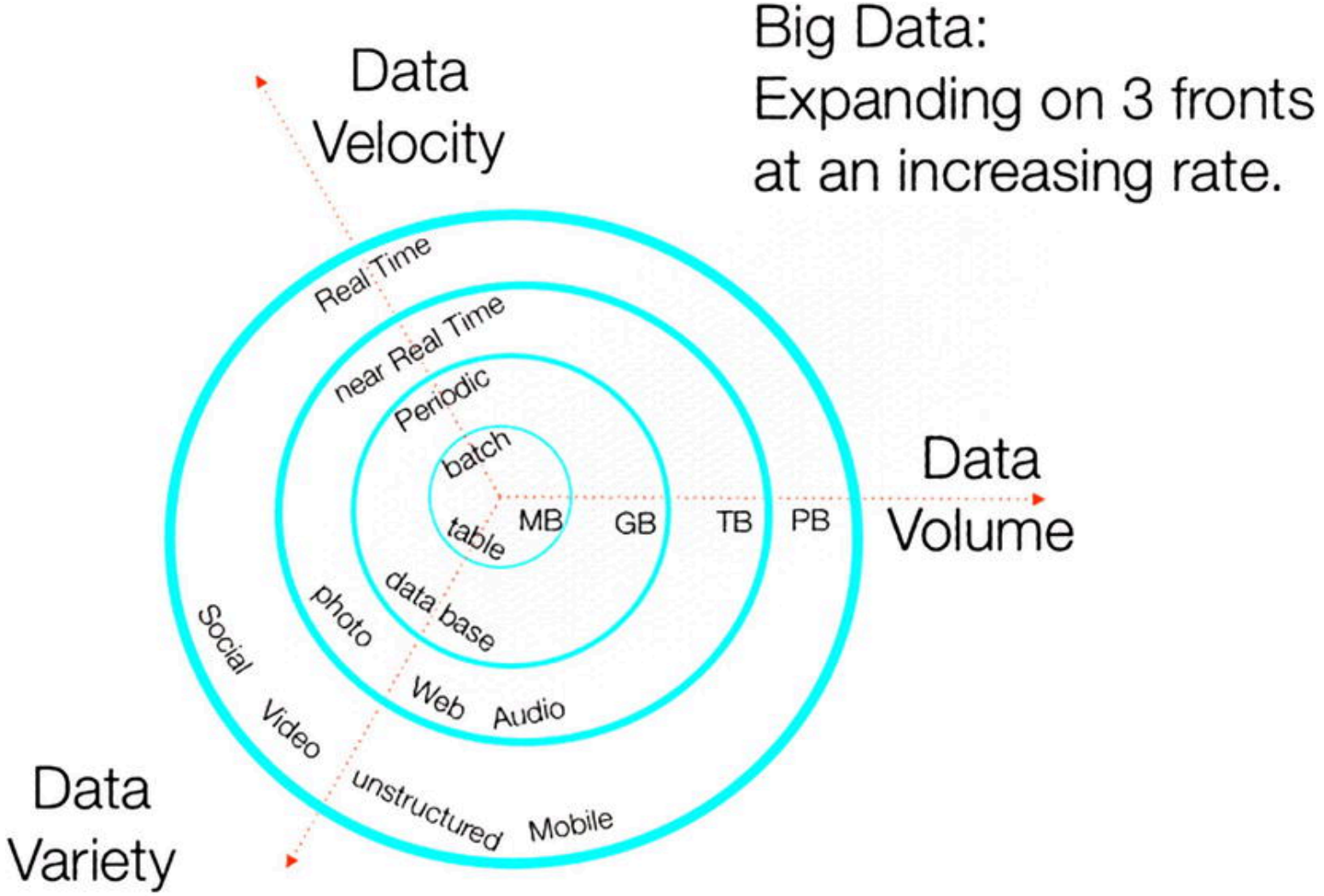
- rationing storage and selecting the "valuable" data
- storing raw data in "data lakes" for future questions and application (>100Gbps) where data is located is not important
- heavy "data driven" approach
- data insights: analytics VS analysis

Big Data Characteristics

Big Data can be defined in terms of how the data will be manipulated, the so called **3V**

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)
  - Near Real Time (e.g space weather: monitoring conditions within the Solar System that may condition space and ground activities)
  - 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

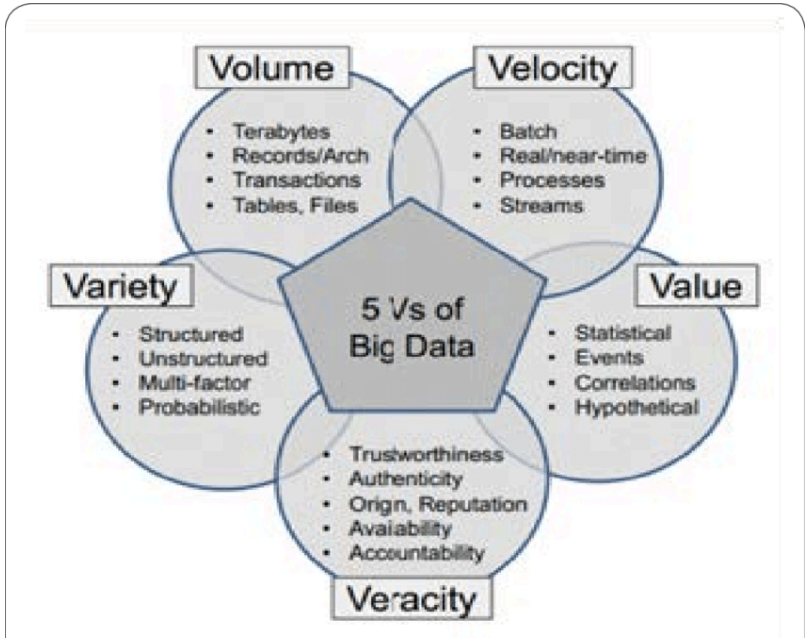
Variety Challenge

- 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 - 5V

In addition to the standard 3V (Volume, Velocity and Variety), in the last years two more Vs were added

4. *Value*:
    - How beneficial the data to be analyzed?
    - Is it worth to dig in the data?
    - How much it costs in terms of time and money to analyze the data?
5. *Veracity*:
    - Data quality referred to the data noise and accuracy.



DOI:10.15344/2456-4451/2017/125

The Extended Big Data Model: The 10Vs

The concept of Big Data has evolved over time to describe the growing complexity of data, arriving at more complex frameworks such as the 10 Vs.

6. *Variability*
   - The data's meaning is constantly changing.
   - Inconsistent speed at which data points are loaded into the system, often depending on context or seasonal trends.
7. *Volatility*
   - How long is the data relevant? This defines the retention policy. In advanced systems, we must decide when data becomes "stale" and can be archived or deleted to save costs.
8. *Validity*
   - Focuses on the accuracy and correctness of the data for its intended use. Even if data is "clean" (veracity), it might not be valid for a specific analytical model.
9. *Visualization*
   - The challenge of making billions of data points understandable to humans. Advanced management requires specialized tools to represent complex patterns without losing detail.
10. *Vulnerability*
    - Big Data equals big risk. This covers the security challenges specific to large-scale architectures, including data breaches and the protection of sensitive information in distributed environments.

Big Data System

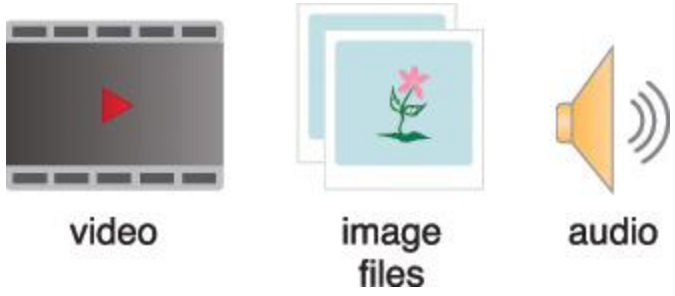
System capable to deal with Big Data require:

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

Big Data Types

- 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,...

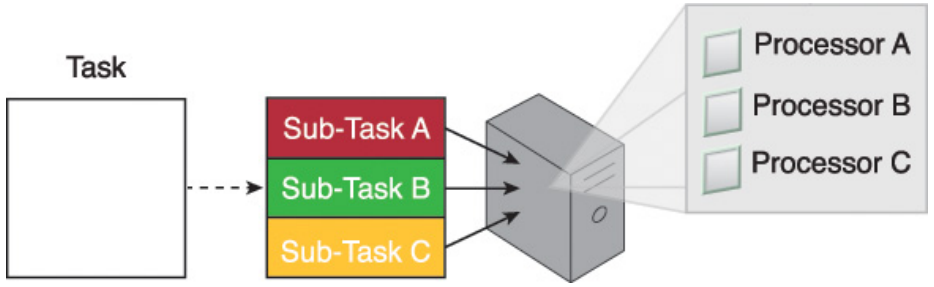
- Storage of raw datasets (acquisition)
- Storage of (pre)processed datasets (manipulation)
- Storage of processed data/results (analysis)

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

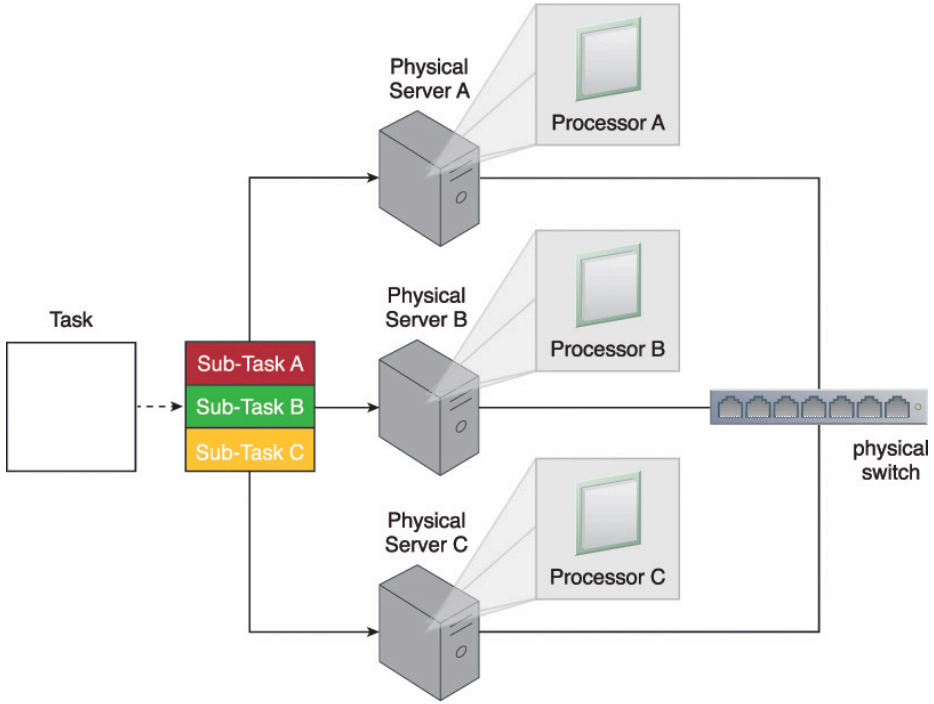
Big Data Partitioning Concepts

- 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 Partitioning Concepts

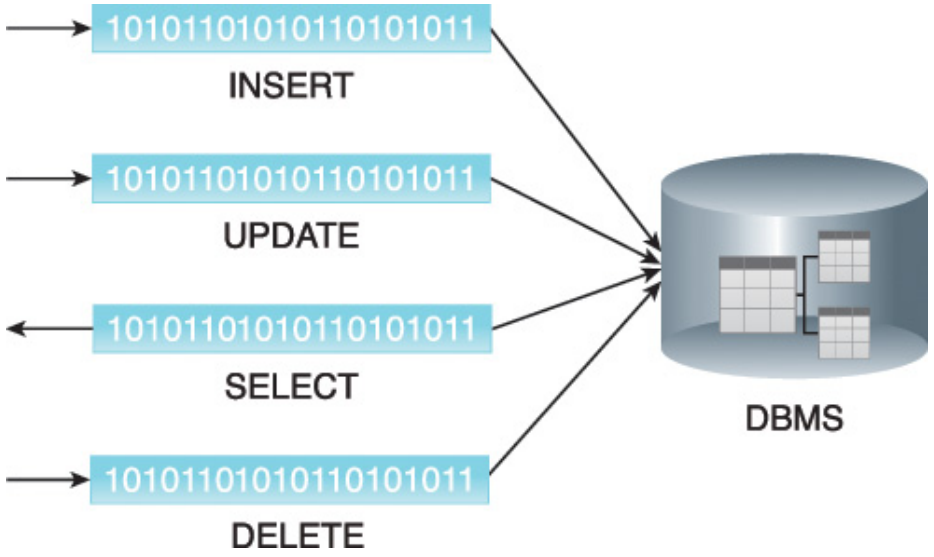
- Speed up the processing of large amounts of data require partitioning
- Distributed processing: reducing time by executing sub-tasks in different machines



Big Data Processing Concepts

Transactional Processing:

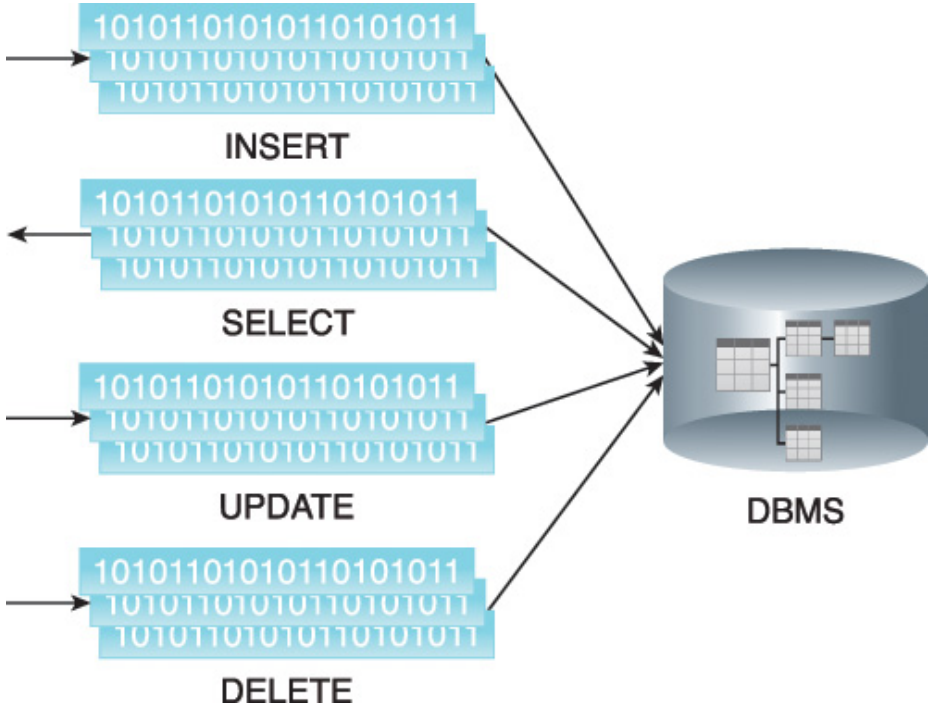
- online processing (realtime)
- processing in-memory, then storage
- low latency (< 1min)
- small amounts of data but continuous



Big Data Processing Concepts

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



Unified Streaming & Batch Processing

- **Breaking the Wall**: *The Lambda Architecture* vs. *The Unified Paradigm*
  - Traditional systems used two separate paths: **Batch** for historical precision and **Streaming** for real-time speed.
  - The modern standard (Unified) treats Batch as a special case of Streaming where the data happens to have a start and an end.
- **Velocity without Compromise**
  - High-velocity data needs immediate processing (Streaming), but long-term analysis needs consistent results (Batch).
  - Unified engines allow developers to write the logic once and apply it to both real-time events and archived petabytes.
- **Key Technologies**
  - **Apache Flink:** Known for true event-time processing and state management, handling millions of events per second with millisecond latency.
  - **Apache Beam:** A unified programming model that provides a "write once, run anywhere" abstraction for various execution engines.
- **Advanced Management Impact**
  - Reduces maintenance overhead by eliminating redundant codebases.
  - Ensures "Exactly-Once" processing semantics, crucial for financial and scientific data integrity.

From Big Data to AI Readiness: The Paradigm Shift

- **Data as Fuel**
  - Volume is no longer the goal; *High-Quality Tokens* are the new gold standard.
  - Big Data is the raw material used to train Large Language Models (LLMs) and Foundation Models.
- **The Shift in Challenges**
  - From *"How do we store this?"* to *"How do we feed this to a model?"*
  - Focus moves from raw ingestion to curation, deduplication, and safety filtering.
- **AI-Ready Infrastructure**
  - *Vector Databases:* Managing embeddings for semantic search and RAG (Retrieval-Augmented Generation).
  - *Data Lineage:* Tracking the origin of data to ensure ethical AI and bias mitigation.
- **The New Bottleneck**
  - Storage is cheap, but *clean, labeled, and governed data* is scarce.
  - Advanced Data Management is now about building the *Data Pipeline* for the AI lifecycle.

Edge Computing & Data Gravity

- **The Problem: Moving Mountains of Data**
  - As datasets reach Petabyte scale, the cost and latency of moving them to a central Cloud become prohibitive.
- **Data Gravity**
  - The concept that data has "mass." As data grows, it attracts applications, services, and even other data toward it.
  - Instead of moving data to the code, we must move the **Code to the Data**.
- **Edge Computing: Processing at the Source**
  - Performing initial analysis, filtering, and reduction directly on the devices (sensors, telescopes, satellites).
  - Reducing the "Velocity" challenge by sending only meaningful insights or compressed summaries to the core.
- **Advanced Management Implications**
  - Architecting **Distributed Workflows** that span from the sensor to the archive.
  - Managing "Smart Storage" that can execute local compute tasks.

The Sustainability Challenge: Green Data

- **The Environmental Cost of "Always-on"**
  - Data centers currently account for ~1-2% of global electricity consumption.
  - The explosion of Generative AI training has significantly increased the power density required per rack.
- **Carbon Footprint of Storage**
  - Keeping "Dark Data" (unused/unclassified data) on high-performance SSDs is ecologically and financially unsustainable.
  - Strategic use of **Tape Storage** and **Cold Tiers** can reduce energy consumption by up to 90% compared to "always-on" disk arrays.
- **Energy-Aware Data Management**
  - **Computational Efficiency:** Optimizing algorithms to reduce CPU/GPU cycles is now a "Green" requirement.
  - **Data Minimization:** Collecting only what is necessary (Quality over Quantity) to reduce the storage energy footprint.
- **The Role of the Advanced Data Manager**
  - Moving beyond "Performance at all costs" toward **Carbon-Intelligent Computing**.
  - Selecting providers and architectures based on Power Usage Effectiveness (PUE) and renewable energy sourcing.

Big Data Governance & Ethics

- **Ownership vs. Stewardship**
  - Data is no longer "owned" by the collector but "stewarded".
  - Respecting the rights of data subjects to control and benefit from their data.
- **GDPR: The Core Foundation**
  - Mandatory adherence to **7 Principles**: Lawfulness, Purpose Limitation, Data Minimization, Accuracy, Storage Limitation, Integrity, and Accountability.
  - New 2026 focus: Speeding up cross-border enforcement and clarifying "personal data" in aggregated datasets.
- **The EU AI Act (Binding from Aug 2, 2026)**
  - **Risk-Based Logic:** High-risk systems (e.g., in healthcare or law enforcement) face stringent data governance and risk management requirements.
  - **Transparency & Safety:** Mandatory labeling for deepfakes/AI content and strict bans on harmful manipulative AI.
- **Ethical Challenges in Big Data**
  - **Algorithmic Bias:** Active monitoring is required to prevent discriminatory outcomes from training data.
  - **Informed Consent:** Moving beyond complex legal jargon to ensure users truly understand how their data is aggregated.

Big Data in Science - LHC CERN

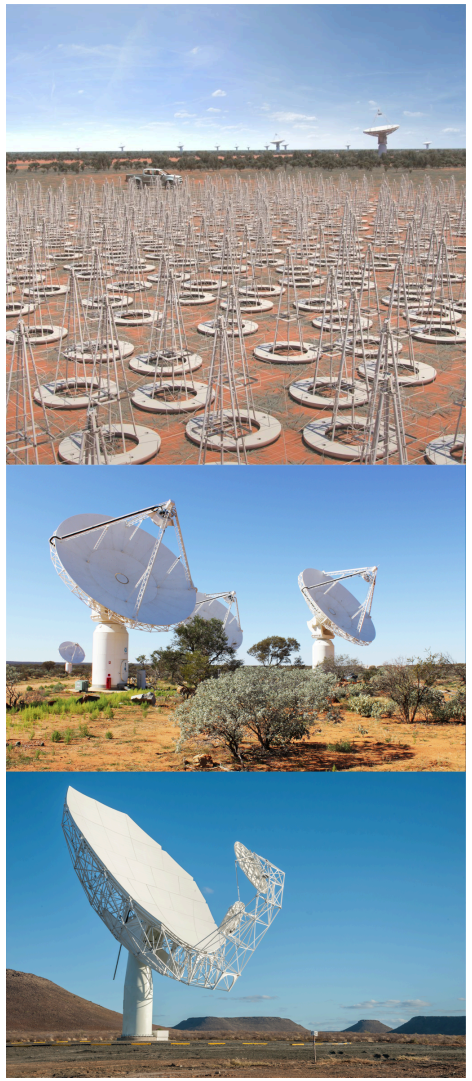
Large Hadron Collider uses detector to analyze particles produced by collisions in the accelerator

- 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

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
  - 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



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