Ministry of Higher Education and Scientific Research
University of Mohamed Boudiaf, Msila

Faculty of Mathematics and Computer Science
Department of Computer Science

Mater 1

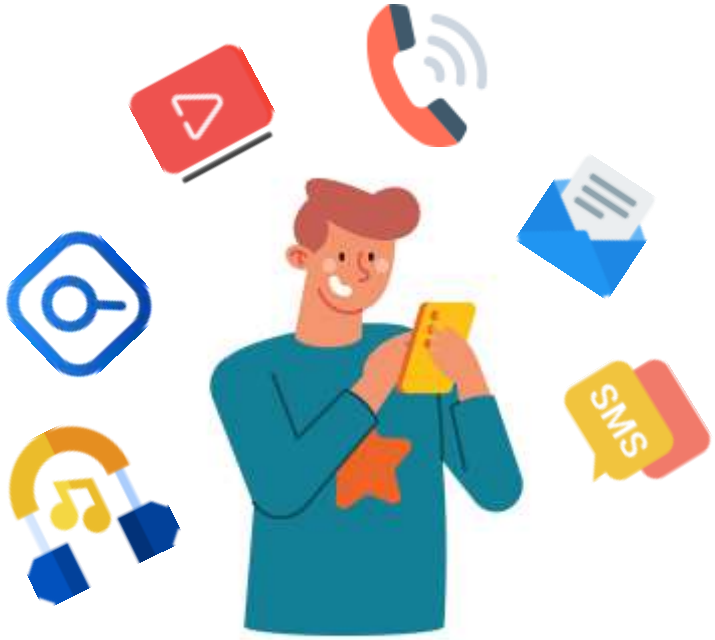
Big Data Analytics

Dr. N.CHALABI

Septembre 2025

Introduction to Big Data

Fundamentals



40 Exabyte of Data /month

=

* 5 billion users

Big Data

“Big data” refers to the **extremely large and complex datasets generated** daily by **individuals** and **machines**

The term ‘’**Big Data**’’ not only represents large volumes of information but also encapsulates the need for new **tools, techniques, and architectures** to **store, process, and extract** value from it.

Big data processing

Industry Examples (2024):

- **Facebook/Meta** handles over 4 petabytes of new data every day from posts, videos, and interactions.

- **self-driving cars** generate around 4 terabytes of sensor data per day.

- **Healthcare sector** produces vast genomic, imaging, and patient data with transformative potential.

IDC projection: Global data volume will reach 175 zettabytes by 2025.

- These examples highlight the scale, complexity, and business potential embedded in Big Data.

Evolution of Big Data

The story of Big Data is closely linked to computing and communication advancements:

- Pre-computer era: Large datasets recorded manually, e.g., the U.S. Census of 1880 took over a decade to process.
- Mechanization (1890): Hollerith Tabulating Machine revolutionized processing with punch cards.
- Early computing (1940s–1950s): Programmable computers handled millions of records.
- Networking age (1969): ARPANET laid the foundation for the Internet.
- World Wide Web (1990s): Explosion in data generation from online content.
- Web 2.0 & Social Media (2000s): User-generated content soared via Facebook, YouTube, Twitter.
- IoT & Cloud (2010s): Sensors, smartphones, and cloud services expanded reach.
- AI-driven data ecosystem (2020s): AI models depend on massive datasets for learning and adaptation.

What is Data?

- Before delving deeper, it is crucial to define what we mean by data. According to Russell L. Ackoff (1989), ***data*** *are raw symbols that represent properties of objects or events, and in this sense, they have no meaning by themselves. Only when data is* ***contextualized*** *does it become information.*

Example:

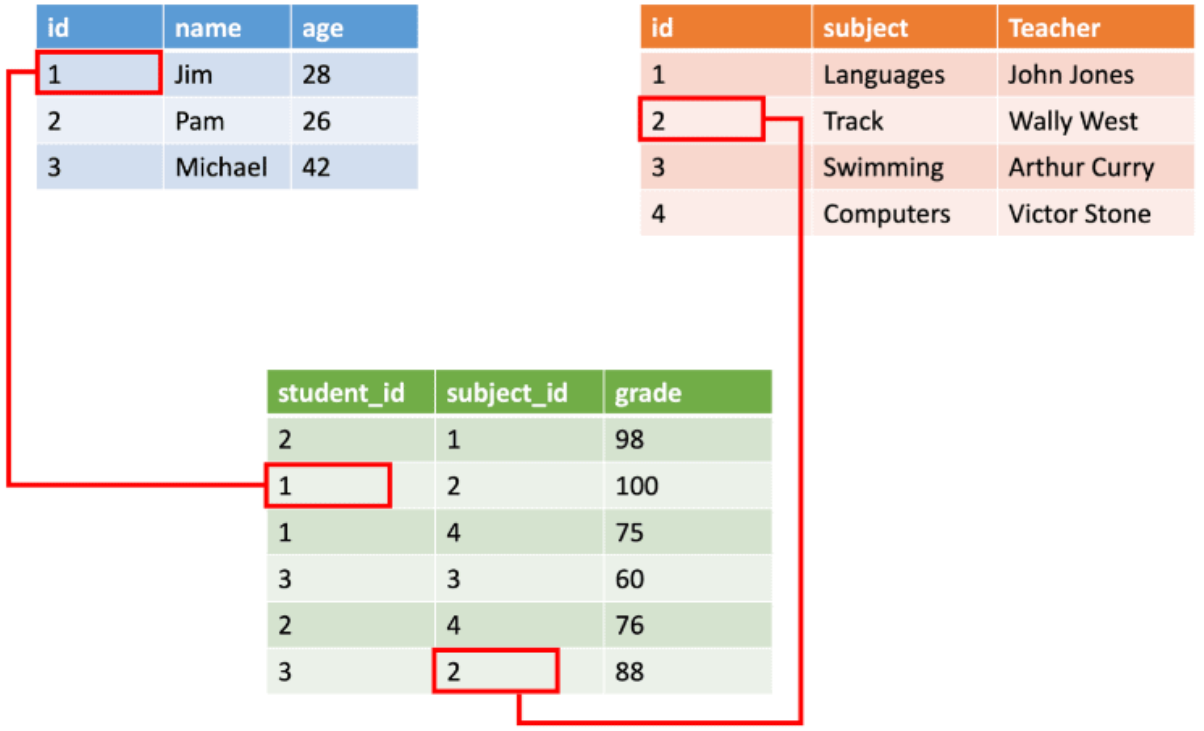
- Raw data: "John", "Smith", "30000"
- Information: ("First name", "John"), ("Last name", "Smith"), ("Salary", 30000)

A **dataset** is a **structured collection** of such data points, **organized** for efficient **processing**, **analysis**, and **visualization**.

Types of Data

1. Structured Data

- Organized in tables with rows and columns.
- Easy to store, query, and analyze.
- Examples: Excel sheets, SQL databases.



id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88

Types of Data

2. Semi-structured Data

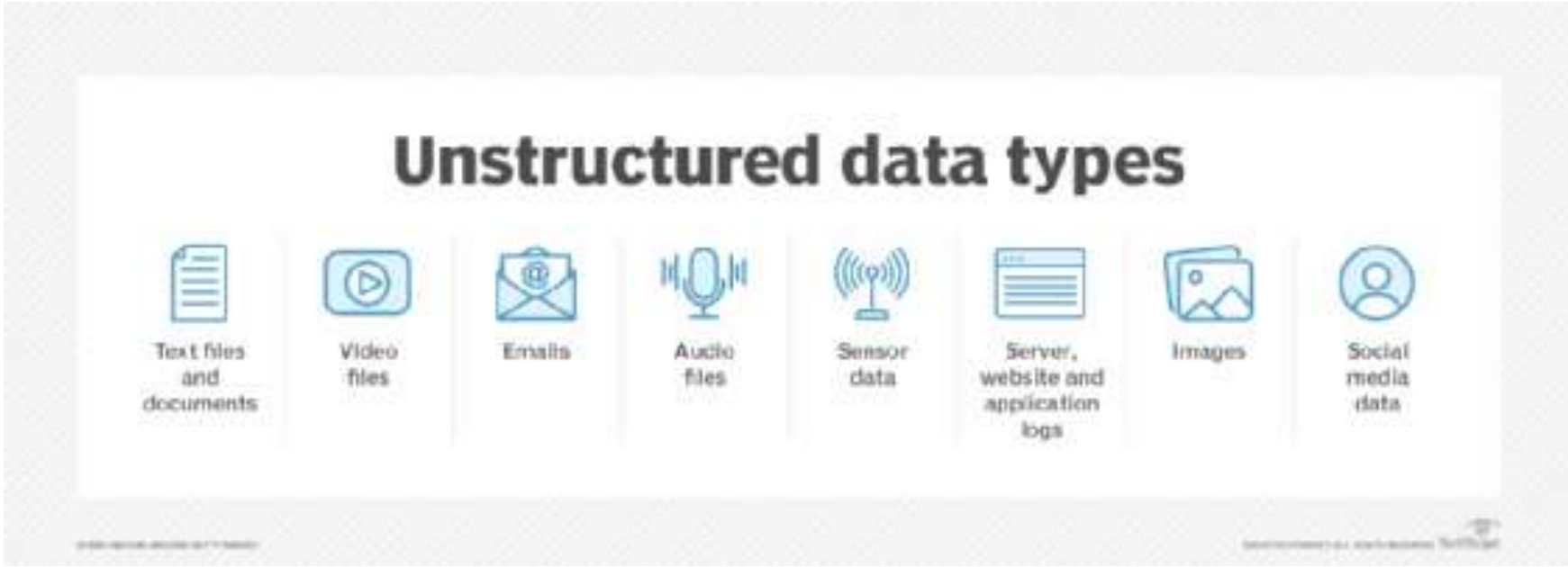
- Partially organized; uses tags or markers for structure.
- Flexible and adaptable, but not rigid like tables.
- Examples: JSON, XML, CSV with varying fields, NoSQL databases.

```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Types of Data

3. Unstructured Data

- No predefined format or organization.
- Requires advanced tools (like NLP, computer vision) for analysis.
- Examples: Images, videos, audio files, social media posts, emails.



let have a look at the data generated per minute on the internet ⏱️

Over **231 million emails** are sent every minute.

Over **500 hours** of new videos uploaded every minute on YouTube

~ **5.9 million Google searches** happen per minute

~ **510,000 comments / minute**; ~ **136,000 photos uploaded / minute**; ~ **4 million likes / minute**.

~ **95 million** photos & videos shared *per day* on Instagram

THAT A LOT OF DATA...!

How to classify any data as Big Data

The classification is possible with the concepts of

5Vs

- Volume
- Velocity
- Variety
- Veracity
- Value

Example: 5Vs in Medical Care

Volume

Hospitals generate **petabytes** of data from electronic health records (EHRs), MRI/CT scans, lab results, and patient histories.

Velocity

Real-time monitoring from **ICU sensors, pacemakers, or wearable devices (like smartwatches tracking heart rate and blood oxygen)** streams data every second. means collecting data at high speed

Variety

Data comes in many formats: structured (lab test results), semi-structured (doctor's notes, prescriptions), and unstructured (X-ray images, genomic sequences).

Veracity

Patient data may have errors, duplicates, or missing information. Ensuring accuracy is critical for diagnosis and treatment.

Value

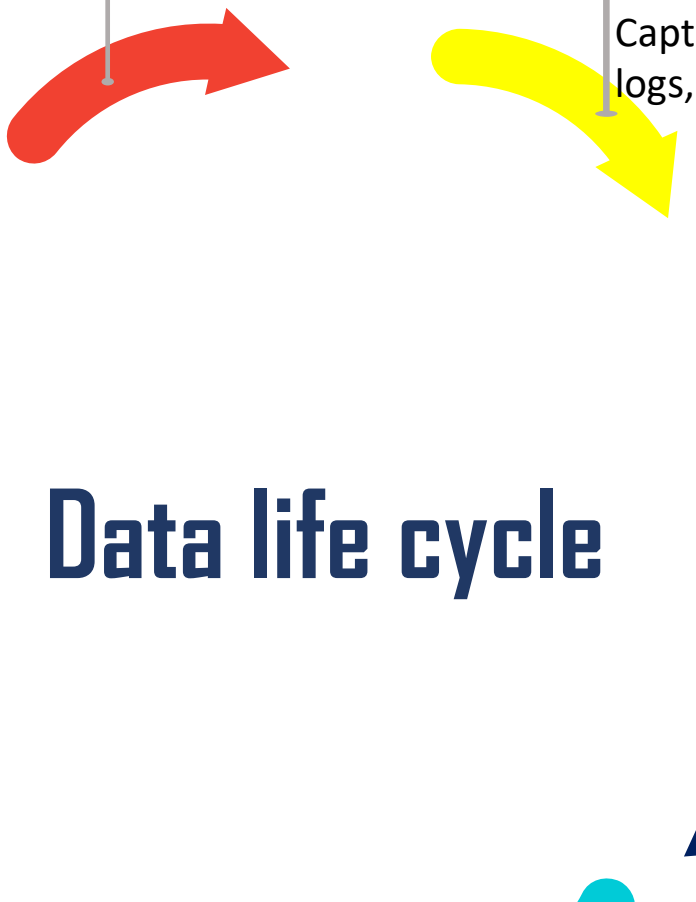
Analyzing medical big data helps detect diseases early (e.g., predicting heart attacks), personalize treatments, optimize hospital operations, and even support **AI-powered drug discovery**.

How Big Data is managed (Data)

Big Data doesn’t just “exist” — it moves through stages. Managing it well ensures security, efficiency, and useful insights.

Data life cycle

Data life cycle



1. Data Generation
Created by sensors, apps, transactions, or online actions.

2. Data Collection
Captured through surveys, APIs, logs, or scraping.

3. Data Processing
Cleaned, transformed, and formatted for use.

4. Data Storage
Saved in databases, data lakes, or the cloud.

5. Data Management
Organized with security, compliance, and tracking.

6. Data Analysis
Insights gained using statistical and AI methods.

7. Data Visualization
Results shown in charts, dashboards, or reports.

8. Data Interpretation
Insights applied to decisions, strategies, or actions.

Where does all this Big Data actually come from

Every click, sensor reading, or transaction leaves a digital footprint. By grouping these footprints, we can better understand the main **sources of Big Data**.

Human-Generated Data

- Social media, emails, blogs, reviews, clickstreams
- Mostly unstructured→ shows behavior & preferences

Machine-Generated Data

- IoT devices, sensors, GPS,server logs
- Semi/unstructured, very fast streams of data

Transactional Data

- Sales, banking, insurance, e-commerce records
- Structured → backbone of operations & analytics

Public & Government Data

- Census, health, transport, environment datasets
- Open data → supports research & policymaking

Can traditional systems handle Big Data

Data today is bigger, faster, and more varied than ever before. Traditional processing systems were designed for smaller, structured datasets, but modern data demands new approaches. Let’s compare traditional systems with Big Data systems to understand why updated architectures are essential for handling today’s information challenges.

Traditional Systems (RDBMS)

Big Data Systems

	Traditional Systems (RDBMS)	Big Data Systems
Data Size	MB–GB	TB– PB+
Data Type	Structured	Structured, Semi-structured, Unstructured
Processing	Batch	Batch & Real-time
Scaling	Vertical (bigger machines)	Horizontal (add more machines)
Cost	High at scale	Lower with cloud/commodity hardware

Big Data Challenges for Business

Big Data offers great opportunities, but organizations face several challenges in implementing and leveraging it effectively.

Key Challenges:

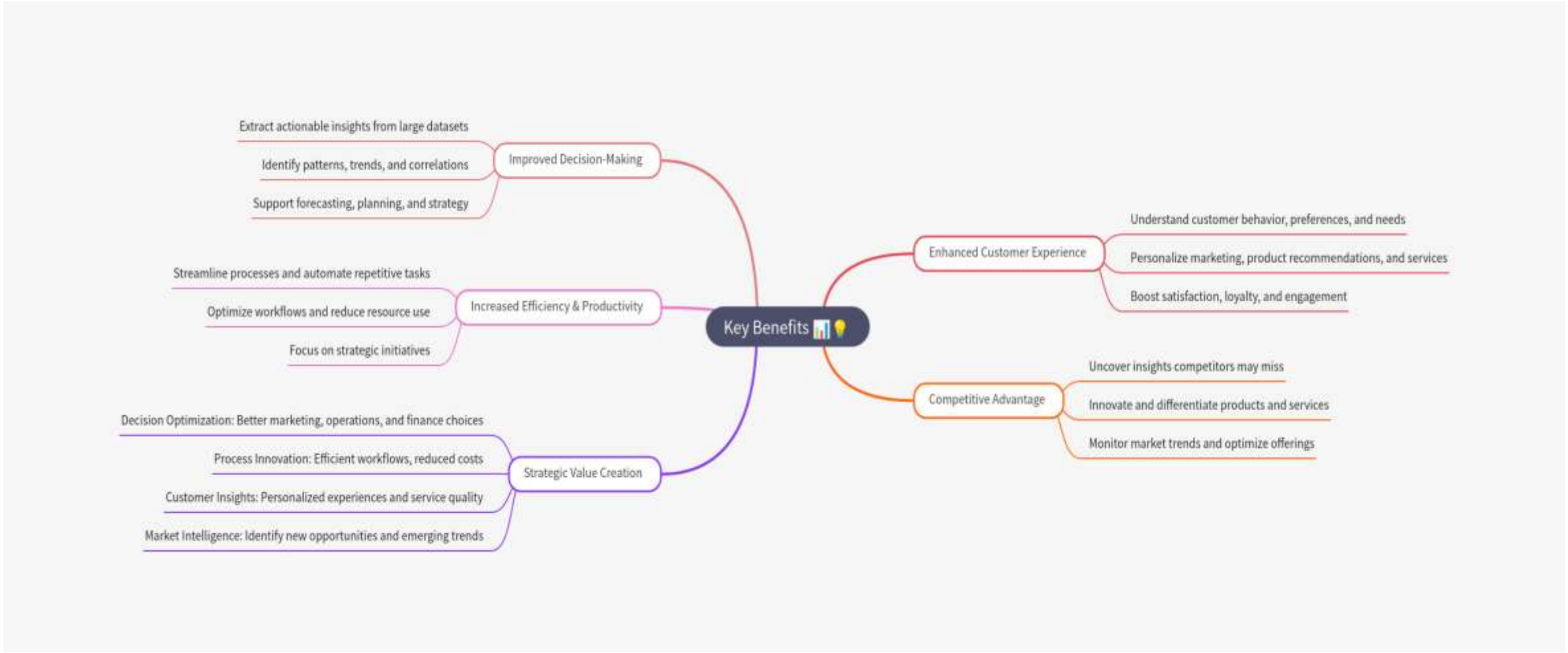
- **Data Quality** – Ensuring accuracy, completeness, and consistency; poor quality leads to unreliable insights.
- **Storage & Infrastructure** – Handling massive datasets requires scalable storage and high-performance computing.
- **Data Privacy & Security** – Protecting sensitive information with encryption, access control, and compliance measures.
- **Data Governance** – Maintaining consistency, transparency, and accountability across departments.
- **Data Integration** – Combining structured, semi-structured, and unstructured data from multiple sources.
- **No One-Size-Fits-All Solution** – Each industry and use case may require tailored tools and approaches.
- **Tool Diversity & Selection** – Choosing scalable, compatible, and future-ready analytics platforms.
- **Volume of Data** – Managing exponential growth in data with distributed computing and parallel processing.

Use case of Big Data

Big Data powers innovation across multiple industries by turning raw data into actionable insights.

- **Healthcare:** Early diagnosis, patient monitoring, epidemic prediction; Electronic Health Records + wearable sensor data.
- **Finance:** Fraud detection, credit risk assessment, algorithmic trading; real-time analysis of transactions.
- **Transportation:** Route optimization, predictive maintenance, fuel reduction; e.g., UPS logistics.
- **Smart Cities:** Traffic monitoring, utilities management, air quality tracking, public safety.
- **Education:** Personalized learning, performance tracking, dropout prevention through analytics.
- **Media & Entertainment:** Audience insights, content personalization; e.g., Netflix recommends 80% of viewed content algorithmically.

How does Big Data create real business value



Best Practices for Big Data Analytics

To succeed with Big Data, organizations should follow these best practices:

- **Start Small, Scale Big:** Begin with pilot projects before expanding.
- **Design for Scalability:** Prepare systems for exponential growth.
- **Choose the Right Tools:** Avoid a one-size-fits-all approach; select tools for the task.
- **Integrate Feedback Loops:** Continuously refine and improve analytics.
- **Maintain Governance:** Ensure compliance, privacy, and security.
- **Use In-Memory Processing:** Tools like Apache Spark speed up computations.

Case Study: Netflix

Netflix collects and analyzes data on viewing patterns, search queries, and device usage:

- Predicts preferences to enhance recommendations.
- Selects new productions based on viewer interest.
- Optimizes streaming quality by network conditions.

This case illustrates data-driven decision-making in content and operations.

Big Data Analytics : Syllabus overview