CHAPTER 4

BIG DATA

Data is being generated and collected at an **unprecedented rate**. Big Data has emerged as a **critical asset** for businesses across industries. This chapter explores its **characteristics**, **sources**, **methods of collection**, **applications** in different sectors, and the **challenges** organisations face with such vast datasets.

CHAPTER ROADMAP

1. Characteristics of big data — 5 Vs
2. Sources of big data — 4
3. Methods of big data collection — 4
4. Applications of big data — 4 sectors
5. Challenges of big data — 4 + solutions

DEFINITION

Big Data refers to the massive volumes of structured and unstructured data generated at high velocity from various sources.

Social media · Sensors · IoT devices · Transaction records · ... and more

Why it's special

- Characterised by its **complexity**
- Needs **advanced techniques & technologies** to **capture, store, process and analyse**
- With **digital transformation**, an asset for gaining **insights** and improving **decision-making**

1 CHARACTERISTICS OF BIG DATA — THE 5 VS

unique challenges & opportunities



V VOLUME
enormous & growing

Amount of data is **enormous** and grows **exponentially**. Organisations process **terabytes or even petabytes** in real-time or near-real-time.

Sources: social media, websites, IoT devices, sensors, transactions, videos, emails. More devices & people online → **more volume**.

V VELOCITY
high speed

Data generated at **high speed**; must be processed in **real-time / near-real-time** to extract value. IoT gives a **constant** data stream.

e.g. Financial markets make **split-second trading decisions** on real-time market changes.

V VARIETY
multiple formats

Structured (databases, spreadsheets), **unstructured** (social posts, videos), **semi-structured** (XML, JSON).

e.g. E-commerce: transactions (**structured**), reviews (**unstructured**), clickstream (**semi-structured**).

V VERACITY
accuracy & trust

Accuracy and trustworthiness of data. Large volumes may contain **inconsistencies, biases or errors** to fix for reliable analysis.

e.g. Misinformation on social media can skew user-behaviour insights.

V VALUE
insight & benefit

Potential **insights and business value** extracted. The ability to derive **actionable insights** is what makes Big Data valuable.

e.g. Retailer analyses customer behaviour → **tailored marketing**, more engagement & sales.

MEMORY HOOK **"Very Vast, Very Varied, Very Valuable"** → **Vo**lume (size) · **Ve**locity (speed) · **Va**riety (formats) · **Ver**acity (truth) · **Va**lue (worth).

2 SOURCES OF BIG DATA

internal & external

Big Data is collected from a wide range of sources, **both internal and external** to an organisation.

2.1 Social Media

user data

Facebook, Twitter, LinkedIn, Instagram generate **posts, likes, comments, shares**. Used to understand **customer sentiment**, improve **brand awareness**, **personalise marketing**.

EXAMPLE

Consumer goods company tracks **keywords & hashtags** during a product launch → quickly spots issues → adjusts marketing campaigns.

2.2 IoT Devices

continuous

Sensors, smart appliances, connected vehicles continuously generate data — critical for **manufacturing, logistics, healthcare**: monitoring, efficiency, **predicting failures**.

EXAMPLE

Smart thermostat collects temperature data → energy companies optimise **energy consumption predictions**.

2.3 Transactional Data

sales & payments

POS systems, financial transactions, online payments, e-commerce. Helps **track sales, detect fraud, optimise pricing**.

EXAMPLE

Online retailer analyses **purchasing patterns** → customer preferences → optimised **inventory management**.

2.4 Machine-Generated Data

automated

From **sensors, industrial machinery, automated systems**. Used in **predictive maintenance**, logistics optimisation, **industrial automation**.

EXAMPLE

Vehicle sensors record **engine performance, fuel consumption, wear-and-tear** → predict maintenance, avoid breakdowns.

3 METHODS OF BIG DATA COLLECTION

specialised tools & techniques

Given the **variety** of sources, organisations need **specialised tools and techniques** to collect data effectively.

Web Scraping

Extracting data from websites and storing it for analysis — product prices, customer reviews, **competitor analysis**.

e.g. Price-comparison site collects prices from many e-commerce platforms.

Data from Sensors

IoT devices & sensors **continuously** gather data; e.g., plant sensors send **real-time** machinery data to central systems.

e.g. Smart farming sensors measure soil moisture, temperature, nutrients.

Log Files

Generated by **servers, applications, networks** — user interactions, errors, system performance, **security breaches**.

e.g. Cybersecurity firm analyses firewall & IDS logs for suspicious patterns.

Surveys & User-Generated

Surveys, polls, user content from a **large audience**; platforms aggregate feedback in **real-time**.

e.g. Post-purchase satisfaction survey → improve products & services.

Don't confuse	Source (where data comes from)	Collection method (how it is gathered)
Study-text lists	Social media · IoT devices · Transactional data · Machine-generated data	Web scraping · Data from sensors · Log files · Surveys & user-generated data
Overlap	IoT devices & machine-generated data both involve **sensors**	"Data from sensors" is the **method** of gathering that IoT data

4 APPLICATIONS OF BIG DATA

efficiency · decisions · innovation

4.1 Healthcare

Early disease detection, improved patient care, **personalised treatment plans** — patterns in large medical data sets → better outcomes.

EXAMPLE · PREDICTING DISEASE OUTBREAKS

During **COVID-19**, agencies analysed **mobile tracking, social media, health records** → understood spread → deployed resources.

4.2 Retail & E-Commerce

Understand **customer preferences**, **optimise pricing**, **manage inventory** using sales data, feedback and **clickstream** data.

EXAMPLE · PERSONALISED MARKETING

Platform tracks **browsing & purchase history** → recommends personalised products → higher **satisfaction & conversion rates**.

4.3 Financial Services

Detect fraud, **assess credit risk**, enhance customer service. **Predictive analytics** identifies fraud patterns & improves risk assessment.

EXAMPLE · FRAUD DETECTION

Global bank's **ML models trained on Big Data** flag suspicious transactions **in real-time** → faster response, fewer losses.

4.4 Manufacturing & Supply Chain

Optimise production, enhance **product quality**, improve **supply chain efficiency**; sensor data predicts maintenance → less **downtime**.

EXAMPLE · PREDICTIVE MAINTENANCE

Car maker's production-line sensors + predictive analytics → detect **wear-and-tear** → maintain **before failure**.

5 CHALLENGES OF BIG DATA

problem → solution

Challenge	Solution
5.1 Data Quality Volume, velocity & variety make quality hard; **inaccurate, incomplete or inconsistent** data → flawed insights, poor decisions.	**Solution** Rigorous **data validation** processes and **cleansing** techniques.
5.2 Storage & Processing Needs specialised infrastructure; **traditional databases are often inadequate**.	**Solution** **Cloud storage** & **distributed computing** platforms — **Apache Hadoop** and **Spark**.
5.3 Data Privacy & Security Must protect sensitive info, **comply with data protection regulations** and guard against **cyber threats**.	**Solution** Strong **encryption**, **access controls** and **monitoring systems**.
5.4 Data Integration Combining structured & unstructured data from many sources; **data silos** & **compatibility issues**.	**Solution** **Data integration platforms** & **middleware** → consolidate into a **unified system**.

EXAM FOCUS · MCQ TRAPS

commonly confused points

1 Velocity ≠ Volume
Volume = **how much** (TB/PB). Velocity = **how fast** it's generated & must be processed (real-time).

2 Veracity = trust
Accuracy & trustworthiness — biases, errors, **misinformation**. Not "variety".

3 Variety formats
Clickstream & XML/JSON = **semi-structured**; reviews & videos = **unstructured**; transactions = **structured**.

4 Value
The ability to derive **actionable insights** — what makes Big Data valuable.

5 Source vs method
Social media is a **source**; web scraping is a **collection method**. Log files are a **method** (from servers, apps, networks).

6 Hadoop & Spark
Solution to **storage & processing** — distributed computing, because traditional databases are often inadequate.

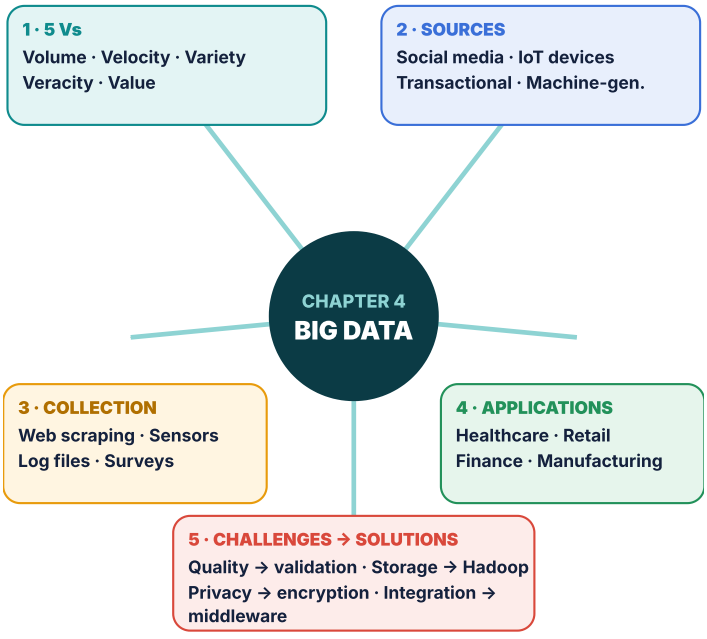
7 Middleware
Solution to **data integration** (silos & compatibility) — not to security.

8 Predictive maintenance
Appears in **manufacturing** (production-line sensors) and machine-generated data (vehicle sensors).

CLASSIFY IT

1	Petabytes processed daily	Volume
2	Split-second trading decisions	Velocity
3	Reviews + transactions + clickstream	Variety
4	Misinformation skews insights	Veracity
5	Tailored marketing lifts sales	Value
6	Price-comparison site	Web scraping
7	Firewall & IDS records	Log files
8	Data silos problem	Integration

REVISION MAP



Big Data
Massive, fast, structured + unstructured data needing advanced tech.

5 Vs
Volume, Velocity, Variety, Veracity, Value.

Get it
Social, IoT, transactions, machines — via scraping, sensors, logs, surveys.

Use it
Healthcare, retail, finance, manufacturing.

Tame it
Quality, storage, privacy, integration challenges.