CHAPTER 1

TYPES OF DATA & SOURCES

Data is a **foundational resource** that powers strategic insights and actions. But **not all data is created equal** — it varies in **structure**, **format** and **source**, and each type plays a unique role in analysis. This chapter covers **structured, unstructured & semi-structured** data, the **internal / external / primary / secondary** sources, and **ethical** data collection.

CHAPTER ROADMAP

1. Types of data — Qualitative · Quantitative
2. Categories of data — by structure
3. Sources of data — 4 types
4. Importance of data collection — R · A · T
5. Ethical considerations — 4 principles + laws

0 INTRODUCTION TO DATA

the starting point

DEFINITION

Data refers to raw facts, figures, or details collected from observations, measurements, or research.

It can take many forms and is typically used as **input** for processing and analysis to generate **useful insights**.

Numbers | Aa Text | ▣ Images | ··· Other types of information

How data becomes value

Observations Measurements Research → **DATA** raw facts & figures → Processing & Analysis → **Useful Insights**

- Drives decision-making
- Helps understand customer behaviour
- Vital role in developing strategies

1 TYPES OF DATA

classified by nature & how it can be measured

TYPES OF DATA

QUALITATIVE

Non-numerical · describes **qualities or characteristics** · used for **categorical / descriptive** analysis

Nominal

NO ORDER

Categories **without any order or ranking**.

e.g. gender, marital status, blood type

Ordinal

ORDER ✓ · INTERVALS ×

Categories with a **meaningful order or ranking**, but intervals between them are **not defined**.

e.g. satisfaction levels (satisfied, neutral, dissatisfied); education levels (high school, college, postgraduate)

QUANTITATIVE

Numerical · **quantifies** something · allows **mathematical operations & comparisons**

Discrete

COUNTABLE

Countable, often **whole numbers**; values **cannot be subdivided**.

e.g. number of students in a class, number of products sold

Continuous

MEASURABLE

Measurable; can take **any value within a range**, including decimals or fractions.

e.g. height, weight, temperature, time

Quick compare	Nominal	Ordinal	Discrete	Continuous
Belongs to	Qualitative	Qualitative	Quantitative	Quantitative
Numerical?	**No**	**No**	**Yes — counted**	**Yes — measured**
Order / ranking?	**No order**	**Meaningful order**	—	—
Key test	Just labels a category	Ranked, but gaps between ranks not defined	Whole values, cannot be subdivided	Any value in a range, incl. decimals

MEMORY HOOK **N**ominal = **N**ame only · **O**rdinal = **O**rder · **D**iscrete = **D**o a count · **C**ontinuous = **C**an measure. Ask first: *"Is it a number?"* → No = Qualitative, Yes = Quantitative.

2 CATEGORIES OF DATA

three main types based on structure

The structure spectrum

More organised = easier to process & analyse

◀ STRUCTURED · easiest — SEMI-STRUCTURED · in between — UNSTRUCTURED · most challenging ▶

Structured
predefined format



Organised in a **predefined format**, typically stored in **rows and columns**. **Easy to manage, analyse and query** using standard tools like **relational databases (e.g., SQL)**.

Ex 1 · Customer records in a CRM
Names, contact information, purchase history, interactions — well-organised in tabular format → easy to filter & analyse.

Ex 2 · Financial transaction data
Accounting software records date, amount, account numbers → easily processed into **financial statements** or **audit reports**.

Semi-Structured
between the two



Lies **between structured and unstructured** data. **Lacks a rigid format** but contains **tags or markers** that provide a degree of organisation → **easier to process than unstructured** data.

Examples (Sticky Notes)
Data in **JSON**, **XML** · webpages with tags · data from web applications.

LL ILLUSTRATION

```
<customer>
  <name>Ali</name>   ← XML: tags
</customer>
{"name": "Ali"}    ← JSON
```

Unstructured
no predefined format



Does **not have a predefined structure or format → more challenging to process and analyse**. Includes **text, images, videos and audio** files.

Ex 1 · Social media posts
Customer opinions on Twitter & Facebook via text, images, emojis → gauge **brand sentiment** & identify **customer preferences**.

Ex 2 · Emails
Inbox holds communication from customers, vendors & employees. Text body lacks a structured format → needs advanced tools like **NLP**.

Key Terms & Explanations

Sentiment Analysis
Using **computational tools** to identify and classify the **emotional tone** of text — e.g., whether customer feedback is positive, negative or neutral.

Natural Language Processing (NLP)
A field of **artificial intelligence** enabling computers to understand, interpret and manipulate human language — used in sentiment analysis, chatbots and text mining.

Tags & Markers
Organise and identify information. **Tags label** parts of data (like text or images); **markers show the beginning or end** of a section.

XML (Extensible Markup Language)
Markup language that **uses tags to define data**. Flexible; stores data in a **structured, hierarchical form** (nested information) instead of tables like in databases.

JSON (JavaScript Object Notation)
Lightweight format to store and exchange data. Easy for humans to read, easy for computers to process; widely used in web apps and **simpler than XML**.

HTML (HyperText Markup Language)
Standard language to **create webpages**. Uses tags to define structure — <h1> for headings, <p> for paragraphs — to display content in browsers.

Feature	Structured	Semi-Structured	Unstructured
Format	Predefined format	No rigid format, but has tags / markers	No predefined structure or format
How stored	Rows & columns (tables)	Tagged / hierarchical (e.g., XML, JSON)	Text, images, videos, audio files
Processing	**Easy to manage, analyse & query**	Easier than unstructured	**More challenging**
Tools named	Relational databases (SQL)	XML, JSON, HTML (tag-based)	NLP, sentiment analysis
Examples	CRM customer records; financial transactions	JSON/XML data; webpages with tags; web-app data	Social media posts; emails

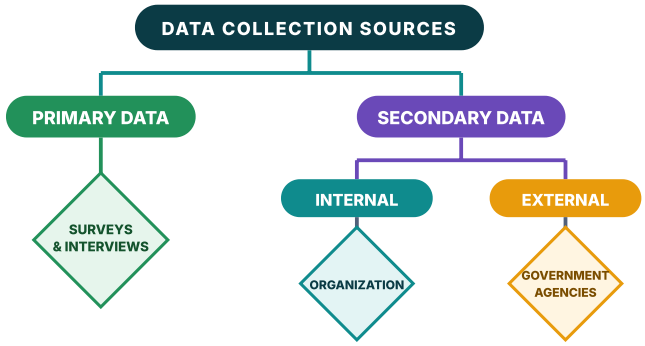
In practice · turning unstructured text into insight

Social media posts & emails (unstructured text) › **NLP** (computer understands & interprets human language) › **Sentiment Analysis** (classifies tone: positive · negative · neutral) › **Brand sentiment & customer preferences** (useful insight)

3 SOURCES OF DATA

quality & relevance depend on how data is collected and managed

Data Collection Sources (study text chart)



NOTE · DATA QUALITY

The degree to which data is **accurate, complete, reliable, relevant and timely** for its intended use. High-quality data ensures better **analysis, decision-making and compliance**, especially in financial and business contexts. Quality data is:

✓ Accurate · ✓ Complete · ✓ Reliable · ✓ Relevant · ✓ Timely

Poor data quality looks like...

Missing values · Outdated information · Inconsistent formats

...and leads to

Errors in reporting · Flawed insights · Regulatory issues

3.1 Internal Data — from WITHIN

Comes from **within an organisation**; usually **specific to its operations**. Already collected & managed through **internal systems**.

Sales records
POS system tracks every transaction — products sold, prices, customer purchases.

Employee data
HR keeps contact info, job roles, salaries, performance reviews.

CASE · MANUFACTURING

Machines generate data: output, downtime, defects ›
Internal database › Analyse → find bottlenecks ›
Corrective action → productivity ↑

3.2 External Data — from OUTSIDE

Comes from **outside the organisation** — via **third-party providers, government agencies** or **public sources**.

Market reports
Purchased research on industry trends, customer demographics, competitive landscapes.

Social media data
Platforms provide **APIs** to gather customer interactions, brand mentions, public opinions.

CASE · FINANCE

Stock indices, economic forecasts, industry reports ›
Bank analyses GDP growth, interest rates, inflation ›
Forecast impact on loan portfolios

3.3 Primary Data — FIRSTHAND

Collected **firsthand for a specific purpose** — through **surveys, interviews, experiments** and **direct observation**.

Customer satisfaction surveys
After a product launch — data is **unique** & directly applicable to the company's objectives.

Focus group discussions
Marketing: direct feedback on new products/campaigns → targeted improvements.

CASE · HEALTHCARE

Pharma company runs clinical trials › Collects data from patients ›
Conclusions on drug's safety & efficacy

3.4 Secondary Data — REPURPOSED

Collected **by someone else for a different purpose**, but considered relevant and **repurposed** for current analysis.

Government statistics
Census data gives demographic information to target specific customer segments.

Consulting-firm industry reports
Trends & forecasts used to inform strategies.

CASE · MARKETING

Retailer uses consulting report on e-commerce consumer behaviour ›
Not collected for it, but gives insight › Refines digital marketing

API — Application Programming Interface

A **set of rules and tools** that allows **different software applications to communicate** with each other. Social media APIs (Twitter, Facebook, Instagram) let companies access posts, likes, hashtags or mentions **without manually browsing** the sites — e.g., a retailer auto-collects customer comments about its brand from Twitter. APIs make it **easier and faster to integrate external data** into business systems or dashboards.

Two questions to classify any source

INTERNAL — originates within the organisation **vs** **EXTERNAL** — comes from outside the organisation

↑ WHERE does it come from? | WHO collected it & for WHAT purpose? ↓

PRIMARY — firsthand, for your specific purpose **vs** **SECONDARY** — collected by others for a different purpose

4 THE IMPORTANCE OF DATA COLLECTION

Accurate and relevant data collection is **essential for decision-making**. Poorly collected or irrelevant data can lead to **incorrect conclusions**, which may **negatively affect business outcomes**.

KEY CONSIDERATIONS FOR DATA COLLECTION

Relevance R
Ensuring the data collected **aligns with the organisation's objectives**.

Accuracy A
Gathering data from **reliable and valid sources**.

Timeliness T
Using data that **reflects the current state of affairs** to make informed decisions.

Case Study · Data Collection in the Insurance Industry

Customer applications | Claims history | External databases › **analysed by insurer**

✓ **Good data** → assess risk **more accurately** → offer **personalised policies**.

× **Outdated / inaccurate data** → **mispricing** of insurance products or **higher claim losses**.

5 ETHICAL CONSIDERATIONS IN DATA COLLECTION

As organisations collect more data, it is crucial to consider the **ethical implications** of data collection. Key ethical principles:

Privacy
Respecting individuals' **rights to control their personal data**.

Transparency
Ensuring data collection practices are **clear to individuals**.

Security
Protecting collected data from **unauthorised access or misuse**.

Fairness
Avoiding bias that could lead to discrimination or unethical outcomes.

Global Data Privacy Regulations

Establish **stringent rules** for protecting individuals' personal data:

- **GDPR** — **General Data Protection Regulation** — European Union
- **CCPA** — **California Consumer Privacy Act** — U.S.

Pakistan's Framework

A **comprehensive data protection law is still under development**; several regulations provide a **foundational framework**:

- **PECA 2016** — **Prevention of Electronic Crimes Act** — criminalises **unauthorised access to personal data** and **cyber offences**.
- **NSCP** — **National Cyber Security Policy** — emphasises **secure digital infrastructure** and **data protection**.
- **ETO 2002** — **Electronic Transactions Ordinance** — recognises **legal validity of electronic documents and transactions** → basis for digital trust.

Case Study · Ethical Data Collection in Social Media

Platforms collect browsing habits, personal preferences, location › Data-privacy concerns → stricter regulations › Scrutiny for not adequately informing users › **Be TRANSPARENT → maintain trust & comply with law**

! EXAM FOCUS · MCQ TRAPS

commonly confused points

1 Nominal vs Ordinal
Both are categories. The **only** difference is **order**. Satisfaction / education levels = **Ordinal**; blood type / gender = **Nominal**.

2 Ordinal is not "numeric"
It has a ranking, but the **intervals between ranks are not defined**.

3 Discrete vs Continuous
Discrete = **counted**, cannot be subdivided. Continuous = **measured**, can take decimals — **time** & **temperature** are continuous.

4 JSON / XML / HTML ≠ structured
Tag-based data is **semi-structured**. Structured = **rows & columns** (relational DB / SQL).

5 Emails = Unstructured
The study text treats emails as **unstructured** — the text body lacks a structured format, so **NLP** is needed.

6 Semi-structured position
Lacks a rigid format, yet is **easier to process than unstructured** due to tags/markers. JSON is **simpler than XML**.

7 Two different lenses
Internal/External = **where** data comes from. Primary/Secondary = **who collected it & why**. In the study-text chart, internal & external sit **under Secondary**.

8 Don't mix the lists
Data **quality** = 5 (accurate, complete, reliable, relevant, timely). Collection **considerations** = 3 (relevance, accuracy, timeliness). **Ethical principles** = 4.

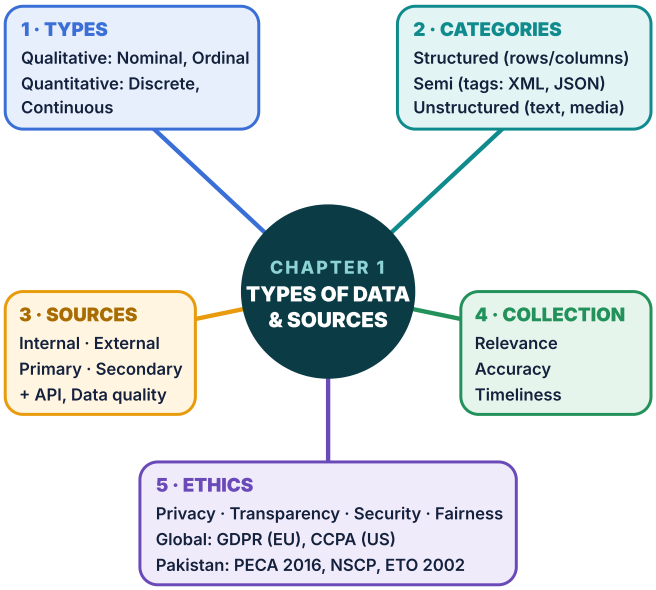
9 Pakistan's law position
No comprehensive data protection law yet (**under development**). PECA **2016** = crimes; ETO **2002** = legal validity of e-documents.

10 Definitions to know exactly
API = rules & tools letting software applications communicate. **NLP** = AI field for human language. **Fairness** = avoiding bias.

CLASSIFY IT · PRACTICE

1	Blood type of patients	Qualitative · Nominal
2	Education level (high school, college, postgraduate)	Qualitative · Ordinal
3	Number of products sold	Quantitative · Discrete
4	Temperature readings	Quantitative · Continuous
5	Transaction records in accounting software	Structured
6	Customer opinions posted on Facebook	Unstructured
7	Web-application data stored in JSON	Semi-structured
8	POS sales records of the company	Internal source
9	Purchased market research report	External source
10	Clinical-trial data gathered by the drug company	Primary source
11	Census data used to target customers	Secondary source

REVISION MAP



Data
Raw facts, figures or details → input for analysis → insights.

Structure decides ease
Structured easiest, unstructured hardest, semi in between.

Source decides quality
Quality & relevance depend on how data is collected & managed.

Collect right
Relevant, accurate, timely — or risk wrong conclusions.

Collect ethically
Uphold privacy, transparency, security & fairness → trust.