Data and Database Systems

From raw data to database management systems

By

Dr. Zohaib Ijaz

What is Data?

- Raw facts (numbers, text, signals, media)
- Needs processing to gain meaning
- Forms the foundation of information & knowledge

Activity 1 (Discussion)

- Not more than 20 minutes
- Please jot down your reason without exploring over internet

Data is or Data are ?

Activity 2

Reading:

https://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular

Data vs. Information

- Data: raw facts (e.g., '1200, 22/09/25, Alex')
- Information: meaningful context (Alex purchased goods worth €1200 on Sept 22, 2025)
- Knowledge: using information for decisions

Types of Data

- Structured: Tables, rows, columns (databases, spreadsheets)
- Semi-Structured: Flexible schema (XML, JSON, NoSQL)
- Unstructured: Text, images, video, logs, social media

Structured	Unstructured
tables, organized, observations	No hierarchy or arrangement
Row is instance, Column is attributes Examples: company records scientific observations	Raw signals that need processing Examples: tweets & social media posts server logs media (images, video, etc)
Easier for Machine Learning to work with (kind of)	More challenging to work with. How to turn into “Structured”?

Semi-Structured

- Flexible schema: not as rigid as relational tables
- Self-describing: metadata embedded in the data
- Example: API responses, configuration files, emails

Breakdown

- In semester 1, we will mostly deals with structured data.
- In semester 2, we will deals with unstructured/ semi-structured data.

Files, Streams, and Databases

- Files: CSV, TXT, images — persistent storage
- Streams: IoT, video, logs — continuous real-time input
- Databases: Structured repositories for efficient queries

The Four Vs of Data

- **Volume** – massive scale (TBs → ZBs)
- **Velocity** – high-speed generation & ingestion
- **Variety** – multiple formats and sources
- **Veracity** – ensuring quality and trustworthiness

Contemporary Global Data Trends

- Data doubling every 2 years (IDC, 2020s)
- 80% of enterprise data is unstructured
- Cloud-first strategies dominate storage
- AI/ML demand large high-quality datasets
- Data regulation: GDPR, localization laws

Data Modeling Considerations

- Representing real-world entities and relationships
- Relational (tables), Document (JSON/XML), Graph (nodes/edges), Key-value
- Balancing flexibility, scalability, and performance

Data Acquisition & Ingestion

- Batch: ETL, bulk uploads (CSV, database dumps)
- Real-time: Kafka, Flink, MQTT for streaming
- APIs/Web scraping for external sources
- Sensors/IoT continuous data generation

Storage & Retrieval Patterns

- OLTP – transactions, small queries (e.g., ATM)
- OLAP – analytics, aggregation (e.g., dashboards)
- Data lakes vs. warehouses
- Indexing & caching for performance

Distributing & Scaling Data

- Horizontal scaling – add servers (NoSQL, Hadoop, Spark)
- Vertical scaling – more CPU/RAM in one machine
- Sharding – partitioning data across nodes
- Replication – redundancy, fault tolerance

File & Stream Formats

- Files: CSV, Parquet, ORC, Avro, JSON, XML
- Streams: Kafka Avro/Protobuf, JSON streams, MP4, MP3
- Choice affects efficiency, compression, portability

Data Compression

- Lossless: GZIP, Snappy, LZ4 (perfect recovery)
- Lossy: JPEG, MP3, MP4 (smaller, but some info lost)
- Trade-offs: Size vs. speed vs. fidelity

Database

- A **database** is an organized collection of data that is stored electronically and managed in a way that makes it easy to **store, retrieve, update, and share** information. Instead of keeping data scattered in files or spreadsheets, a database brings everything together in a structured way.

What is a Database?

- Shared, integrated structure containing:
- End-user data (facts of interest)
- Metadata (data about data, structure, constraints)
- Databases are self-describing

Why Databases?

- Businesses need structured access to: Customers, employees, products, transactions
- Databases provide fast storage & retrieval
- Ensure data integrity & accuracy
- Support decision-making

Types of Databases

- Single-user vs. Multi-user
- Centralized vs. Distributed
- Operational vs. Data warehouse
- Enterprise vs. Workgroup
- Examples: MySQL, PostgreSQL, Oracle, MongoDB

From File Systems to Databases

- Manual files → paper-based
- Computerized files → digital but isolated
- Modern databases → integrated, scalable, queryable

Problems with File Systems

- Structural & data dependence
- Data redundancy
- Inconsistent formats
- Lack of modeling and scalability

Importance of Database Design

- Good design → accuracy, efficiency, scalability
- Poor design → redundancy, errors, inefficiency
- Transactional DBs: optimize for speed
- Data warehouses: optimize for insights

Database System Components

- Hardware – servers, storage, networks
- Software – DBMS (MySQL, Oracle, SQL Server)
- People – users, administrators, designers
- Procedures – policies, standards
- Data – raw facts + metadata

Database management Systems

- A **Database Management System (DBMS)** is software that allows you to define, create, maintain, and control access to databases. It acts as an interface between users/applications and the database itself. Its main purpose is to manage data efficiently and ensure data integrity, security, and consistency.

DBMS Functions

- Data dictionary management
- Storage & retrieval optimization
- Multi-user access control
- Security enforcement
- Backup & recovery
- Communication interfaces

Role and Advantages of DBMS

- Centralized data control
- Reduces redundancy and inconsistency
- Improves sharing, security, and integrity
- Supports transactions and analytics

Managing Database Systems

- Shift from program-focused to data-focused view
- Challenges: cost, complexity, security
- Importance of skilled DBAs and designers

Summary & Takeaways

- Data is diverse: structured, semi-structured, unstructured
- Modern data = Big Data, Four Vs, cloud-driven
- Databases solve file system problems
- DBMS provides efficient, secure, and scalable data management
- Good design is the foundation of reliable systems