1. What is Data Engineering?

Data engineering is the process of **building systems that collect, ingest, process, store, and deliver data** so it can be used for analytics, dashboards, machine learning, and AI applications.

The main goal is to ensure that:

- Data is reliable
- Data is clean
- Data is available on time
- Data is secure

All analytics and machine learning depend on data engineers doing this correctly.

2. Why Data is Important

Everything starts with **data**:

- Dashboards tell a story using data
- Machine learning predicts future outcomes using data
- AI systems generate results using data

If data quality is poor, results will also be poor.
So data engineering focuses on building **trustworthy data pipelines**.

3. Types of Data

3.1 Structured Data

- Stored in rows and columns
- Fixed schema
- Easy to query using SQL

Examples:

- Customer tables
- Transaction data
- Employee records

Stored in relational databases.

3.2 Semi-Structured Data

- Flexible schema
- Key–value format

Examples:

- JSON
- XML
- API responses

Commonly stored in NoSQL databases.

3.3 Unstructured Data

- No predefined structure
- Hard to analyze directly

Examples:

- Text files
- PDFs
- Images
- Audio
- Videos

Requires ML, NLP, or Deep Learning for processing.

4. Data Sources

Data can come from:

- Databases (PostgreSQL, MySQL, SQL Server)
- APIs (payments, maps, applications)
- Files (CSV, JSON, XML)
- Event streams (logs, transactions, clicks)
- Web scraping (extracting website data)

Understanding the source is the first step in pipeline design.

5. Data Engineering Lifecycle (End-to-End Flow)

1. **Data Collection** – Gather data from sources
2. **Data Ingestion** – Move data to centralized storage
3. **Data Processing** – Clean and transform data
4. **Data Storage** – Store raw and processed data
5. **Data Analysis** – Enable analytics and ML
6. **Data Consumption** – Dashboards, reports, predictions

Every tool and role fits into this lifecycle.

6. Data Collection

- First stage of the lifecycle
- Data is collected from multiple sources
- No transformation is applied
- Focus is accuracy and completeness

7. Data Ingestion

Data ingestion means **bringing data into the system**.

7.1 Batch Ingestion

- Data collected over time
- Runs on a schedule (daily, weekly, monthly)
- Used for historical analysis and reports

7.2 Real-Time (Streaming) Ingestion

- Data processed instantly
- Used in live systems:
  - UPI transactions
  - Ride-hailing apps
  - Live tracking systems
- Requires low latency

8. ETL vs ELT

ETL (Extract → Transform → Load)

- Data transformed before storage

- Raw data is changed
- Used when immediate processed output is required

ELT (Extract → Load → Transform)

- Raw data loaded first
- Transformations done later
- Raw data preserved
- Preferred in cloud systems and analytics

9. Raw Data Preservation Rule

- Raw data should **never be modified**
- Always keep original data
- Helps in:
  - Reprocessing
  - Debugging
  - Audits
  - Governance compliance

10. Medallion Architecture (Bronze–Silver–Gold)

Bronze Layer

- Stores raw data
- No transformations
- Acts as a data lake

Silver Layer

- Cleaned and validated data
- Schema applied
- Ready for analytics

Gold Layer

- Business rules applied
- Aggregated and optimized
- Used by dashboards and ML models

11. Data Processing & Transformation

Includes:

- Cleaning missing or invalid values
- Removing duplicates
- Applying schemas

- Business logic
- Aggregations

This step converts data into usable information.

12. Data Storage Systems

Databases

- Used for transactional data
- Supports CRUD operations

Data Lakes

- Stores raw structured and unstructured data
- Scalable storage

Data Warehouses

- Optimized for analytics
- Stores historical and processed data

Data Marts

- Subset of data warehouse
- Business-specific data

13. Data Consumption Layer

Final goal of the entire pipeline:

- BI dashboards
- Reports
- Machine learning models
- APIs
- Chatbots and GenAI systems

If data is not consumed, it has no business value.

14. Machine Learning Basics

Machine learning:

- Learns patterns from historical data
- Uses those patterns to predict future outcomes

ML Tasks

- **Regression** – Predict continuous values
- **Classification** – Predict categories or labels

Data quality directly affects model accuracy.

15. NLP and Generative AI

Natural Language Processing (NLP)

- Helps machines understand human language
- Used for:
  - Text analysis
  - Search
  - Chatbots
  - Sentiment analysis
  -

Generative AI

- Creates new content using learned patterns
- Generates text, images, code, responses

RAG (Retrieval-Augmented Generation)

- Combines LLMs with enterprise data
- Improves accuracy
- Reduces hallucination

16. Roles in the Data Ecosystem

- **Data Engineer** – Builds pipelines and manages data flow
- **Data Analyst** – Creates dashboards and reports
- **Data Scientist** – Builds ML models
- **ML / GenAI Engineer** – Builds AI pipelines and agents
- **DevOps Engineer** – Manages CI/CD and deployments
- **Data Governance Team** – Ensures security and compliance

17. DevOps in Data Engineering

DevOps ensures reliability and automation:

- CI/CD pipelines
- Git version control

- Automated testing

- Environments:
    - Development
    - QA
    - UAT
    - Production

18. Data Governance

Data governance ensures data is handled securely and legally.

- Data masking
- Synthetic data
- Access control
- Compliance rules

Sensitive data must never be exposed.

19. Security & Compliance

- Never expose real client data
- Never use sensitive data in public tools
- NDA violations can lead to legal consequences