### ****Objectives****

* Understand the importance of **data quality** and its impact on decision-making.
* Learn about **data standards** and governance principles for ensuring **data integrity**.
* Identify **key data quality metrics** such as accuracy, completeness, consistency, and timeliness.
* Explore **common file types** (CSV, JSON, XML) and their associated challenges.
* Learn how to apply the **FAIR principles** (Findability, Accessibility, Interoperability, Reusability) to data management.
* Understand **data validation techniques** to ensure reliability and consistency.

### ****Key Concepts****

#### ****1. Data Quality & Governance****

* **Garbage In, Garbage Out (GIGO)**: Poor-quality data leads to inaccurate insights.
* **Role of Data Engineers**: Gatekeepers of data quality; responsible for cleaning and maintaining structured datasets.
* **Key Data Quality Metrics**:
  + **Accuracy**: Reflects real-world values correctly.
  + **Completeness**: No missing or null data where information is required.
  + **Consistency**: Ensures uniformity across different datasets.
  + **Timeliness**: Data is up-to-date and refreshed as required.
  + **Reliability**: Ensures data remains trustworthy over time.

#### ****2. Data Standards & FAIR Principles****

* **FAIR Principles**:
  + **Findability**: Using metadata, indexing, and unique identifiers.
  + **Accessibility**: Secure role-based access controls.
  + **Interoperability**: Data formats (CSV, JSON, XML) and consistent schema structures.
  + **Reusability**: Standardized metadata descriptions and data governance.
* **Data Catalogs & Dictionaries**: Used for organizing and documenting metadata.

#### ****3. Common Data Quality Issues & Fixes****

* **Inconsistent Formatting**: Different date formats, missing headers, or incorrect delimiters.
* **Encoding Errors**: Issues with special characters (UTF-8 standard recommended).
* **Data Duplication**: Implementing validation checks and unique identifiers.
* **Schema Validation**: Ensuring data follows predefined structures for compatibility. Schema - structure of the data/table, data dictionary provides info about the data

#### 

#### ****4. File Types & Considerations****

* **CSV (Comma-Separated Values)**
  + Simple, portable, widely used.
  + Issues: Lacks structure, inconsistent delimiters, data type mismatches.
* **JSON (JavaScript Object Notation)**
  + Human-readable, key-value format.
  + Issues: Complex nesting, large files affect performance.
* **XML (Extensible Markup Language)**
  + Structured, widely used in data exchange.
  + Issues: Parsing errors, improper tag structures.

### ****Questions****

1. What are the best practices for handling large JSON datasets without impacting performance?
2. How can organizations automate **data validation** to ensure real-time data quality?
3. Are there industry-specific **data governance frameworks** recommended for compliance?
4. What tools are commonly used for creating and managing **data catalogs**?
5. How do companies handle **data silos** when integrating multiple data sources?
6. Can you provide examples of **real-world implementations** of FAIR principles improving data quality?