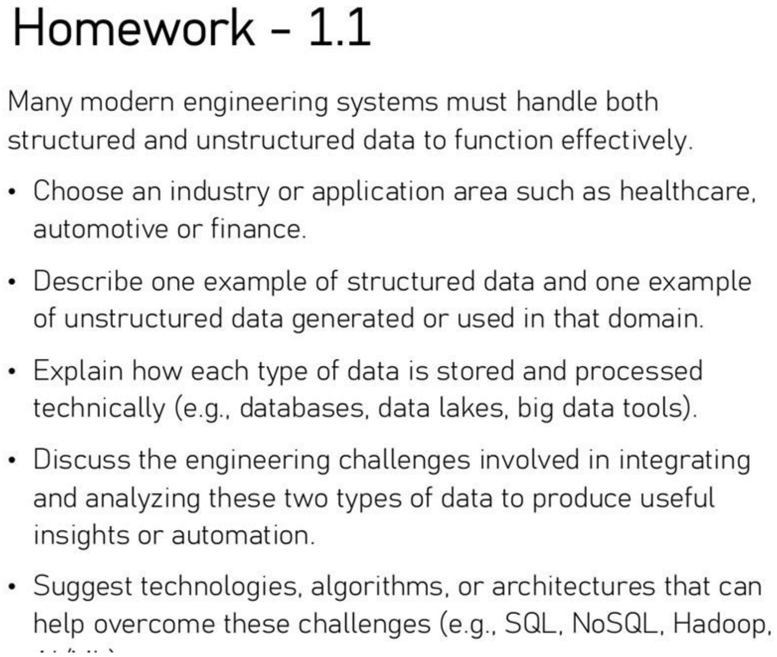
Assignment 2.1 – Homework 1.1



Homework - 1.1

Many modern engineering systems must handle both structured and unstructured data to function effectively.

- Choose an industry or application area such as healthcare, automotive or finance.
- Describe one example of structured data and one example of unstructured data generated or used in that domain.
- Explain how each type of data is stored and processed technically (e.g., databases, data lakes, big data tools).
- Discuss the engineering challenges involved in integrating and analyzing these two types of data to produce useful insights or automation.
- Suggest technologies, algorithms, or architectures that can help overcome these challenges (e.g., SQL, NoSQL, Hadoop,

Industry: Healthcare

Data Example:

a. Structured :
1. Patient Biodata or Information → Includes details such as patient name, age, gender, contact number, address, and ID.
2. No. of patients visited → Represents numerical data like the number of patients visiting daily, weekly, or monthly.
3. Electronic Medical Records →Contains structured entries such as test results, treatment dates, and discharge summaries.
4. Financial Data and Insurance Data

b. Unstructured:
1. MRI Scans or X-ray Images → Image files containing visual information about a patient's internal organs.
2. Doctor's Prescription → Consists of handwritten or typed notes describing medication, dosage, and advice.

How data stored and processed

a. Structured Data : stored in relational databases like MySQL, PostgreSQL. It is processed using SQL queries, ETL pipelines, and bigdata tools like Hadoop , Spark are used to handle large volume of data . For analytics we can use tools like PowerBI, Tableau.

b. Unstructured Data: stored NoSQL databases like MongoDB or storage systems like AWS S3, Azure Blob storage and Apache Kafka , Elastic search to handle large data . It can be processed using NLP for text, OCR for scanned reports and computer vision for images.

Challenges

- Structured and unstructured data comes in different format and sources making it difficult to integrate due to massive amount of data generated continuously.
- Missing values, duplicates formats affect analysis.
- Large files require scalable and distributed storage systems.

Solution

- Use of SQL and NoSQL databases to store structured and unstructured data respectively.
- Hadoop and HDFS for distributed and processing of large volume of data.
- Apache Kafka for real-time streaming from medical devices or hospital systems.
- Data lake can be used to store raw data from multiple sources.
- ETL tools like Apache NiFi , Talend can help extract, transform and load data from different systems into unified platform.