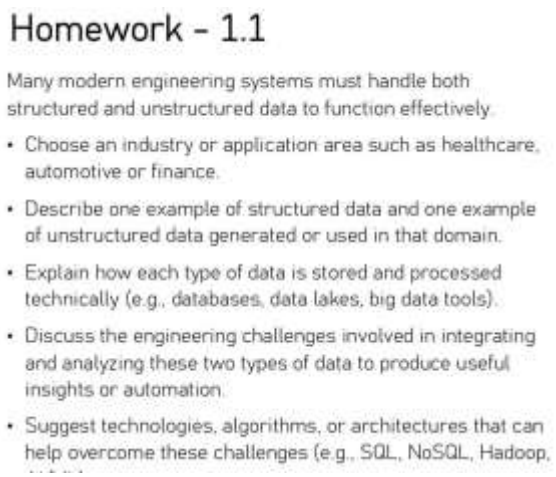
Homework 1.1

Structured and Unstructured Data



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,

1. Industry: AdTech

2. Data examples:
    a. Structured: Data about the advertisers, user profile info, user interest & engagement metrics (like no of mouse clicks, hover, skips, etc).
    b. Unstructured: Ad videos, photos, popup text, gifs, social media posts

3. How it is stored & Processed:
    a. Structured: Relational databases like PostgreSQL, MySQL or data warehouse like Google BigQuery, Snowflake. It can be processed by using data analytics statistical/ML models or using tools like PowerBI, Tableau.
    b. Unstructured: Stored in object storage systems like Amazon S3, Azure Blob storage. Processing can be done using natural language understanding like text parsers, sentiment analysis engines, user feedback analysis, etc.

4. Challenges for data analysis:
    a. Scalability – So many sources and types of data to process which is exponentially increasing. It can be hard to scale up the analysis in terms of resources and accuracy.
    b. Latency – Obtaining the right ad based on user's interests involves analysing the structured user engagement data and unstructured ad data analysis. Combining these across so many permutations in real-time is challenging

5. Possible solutions:
    a. Sharding and microservicing to parallelise the data processing and increase the quantity of data being processed.
    b. Robust ETL pipelines like Kafka, Apache Spark can be used for real-time stream processing and scalability.

- Nithya (1RV22CS099)