Instructions

Before we jump into learning data engineering, let's get the basics first! Check out the following—resources in our LMS.

- How I Would Learn Data Engineering in 2025

- DATA ENGINEERING EXPLAINED

- What is a Data Engineer

After watching/ reading the resources, answer the questions below in your own words. Be clear, concise, and **avoid copying** directly from the videos. Save your work as a PDF, and share it with us via a Google Drive link.

I. Write a short paragraph explaining each of the following:

a) What is Data Engineering?

- They are responsible for designing, maintaining, and building infrastructure when collecting data. Based on the articles I've read, they ensure that data is reliable, accessible, and ready for analysis by other teams in an organization

b) What are the main goals of data engineering?

1. **Monitoring and checking of data**
   - A crucial skill for a data engineer is monitoring the data pipeline and databases; it's up to them to ensure everything runs smoothly, since the data fuels business functions, so that the work of data analysts and data scientists has accurate, timely data.
2. **Optimize performance**
   - In terms of optimizing performance, database data processing is crucial for handling large datasets, as it can be time-consuming. It's their job to improve data processing speed and efficiency.
3. **Developing and maintaining**
   - This is where ETL comes in handy, since it's the data engineer's job to ensure the data comes from the right source.
4. **Data cleansing**
   - It's the data engineer's responsibility to ensure that the data they gather is high-quality and in the correct format.
5. **Collaborate with the team.**
   - They meet data access needs and project expectations in collaboration with data scientists, analysts, and other clients.

c) What is a data pipeline?

- It's a series of steps that automatically move data from one place to another/ A source system to a database or analytics platform. It includes as-permenton, cleaning, transforming, and organizing data along the way.

d) What does ETL stand for, and why is it important?

- ETL stands for **Extract**, **Transform**, and **Load**. It's an essential process because it pulls data from its source, converts the raw information into a usable format, and stores it in a database or warehouse so that data analysts and scientists can analyze it effectively.

e) Why is data engineering important for businesses and analytics teams?

- They ensure that the organizations have fast, reliable, and high-quality data. Without it, analysts and data scientists cannot produce accurate insights, and businesses cannot make well-informed decisions — *w/o data malulugi ang kompanya*.

II. Short Answer Questions

(Answer each question in 1–3 sentences.)

a) What is a data pipeline, and why do companies use it?

- It is a system that moves data automatically between different tools or storage systems. Companies use it to ensure that data flows smoothly, accurately, and is ready for analysis.

b) Explain in your own words what ETL means and what each stage does.

- ETL means Extract, Transform, and Load. Data is taken from a source (Extract), cleaned and modified into a usable format (Transform), and then stored in a database or warehouse (Load).

  It's basically a factory where raw materials are stored and organized for use in essential things.

c) Give one real-world example of a data pipeline (e.g., social media analytics) and briefly describe what happens to the data.

Professor Evaluation System (under PUP-SIS Student Module)

- In the professor's evaluation process, a data pipeline collects, processes, and stores the evaluation data submitted by students.

  First, you must complete an online evaluation form (data collection). Next, the system cleans and organizes the data, ensuring missing answers are handled, converting ratings into numerical values, and ensuring the data is consistent (data processing). Finally, the processed data is stored in a database where admins and other department heads can access it to generate reports and analyze faculty performance (data storage and delivery). The pipeline ensures that the evaluation results are accurate, accessible, and ready for use in academic transparency/decisions.