🎓 Video 17 Summary: The Load Stage in Data Pipelines

🔁 ETL Recap

A typical data pipeline includes three stages:

1. **Extract** – Collect data from various sources
2. **Transform** – Clean, manipulate, and enrich the data
3. **Load** – Move the data to its final storage destination

🚚 Analogy: Donating Items to Charity

- **Extract**: Collect items to donate
- **Transform**: Clean and prepare the items
- **Load**: Deliver them to the charity truck
  This mirrors how data is handled in a pipeline.

📦 What Happens in the Load Stage

- Data is moved from a **staging area** to **destination storage** (e.g., database, data warehouse, data lake).
- The **goal** is to make data available for analysis.

🔄 ETL vs. ELT

- **ETL**: Extract → Transform → Load
- **ELT**: Extract → Load → Transform
  - ELT is used when transformation happens **after** loading, often in cloud environments.

🛠️ Preparing for Loading

Before loading:

- Create or configure **tables**, **directories**, or **schemas** in the destination system.
- Ensure the destination is ready to accept the data.

📥 Loading Methods

1. Batch Loading

- Moves data in **groups (batches)** at scheduled times.
- Efficient for large datasets.
- Risk: Can **overload** the system if data volume is too high.

2. Streaming Loading

- Moves data **continuously** as it becomes available.
- Ideal for **time-sensitive** data.
- Prevents overload by processing data **record-by-record**.

3. Incremental Loading

- Loads **only new or changed data** since the last load.
- Saves **time and resources**, especially for frequently updated datasets.
- Frequency depends on:
  - Dataset size
  - Update rate
  - System performance needs

✅ Post-Load Check

- Always **verify data integrity and accuracy** after loading.
- Ensures data is ready for reliable analysis.

🤖 Automation Tools

- Many teams use **automated tools** to:
  - Load data efficiently
  - Prevent errors and data loss
- Even with automation, understanding the **loading process** is essential for maintaining **data quality**.

🧠 Takeaway

The **load stage** is critical for making data usable.
Whether using **batch**, **streaming**, or **incremental** loading, the method should align with your **data volume**, **timing needs**, and **system capabilities**.