🎓 Video 14 Summary: Data Mapping in the Data Pipeline

🧭 What Is Data Mapping?

Data mapping is the process of **matching fields from one data source to another** to ensure consistency and usability in analysis.

🔭 Analogy: Bird Watching

Just like identifying a bird by comparing its features to a field guide, data mapping involves:

- Observing data fields
- Comparing them to a known schema
- Matching them to create a unified dataset

🔄 Why Data Mapping Matters

- It ensures **standardization** and **consistency** across datasets.
- It's a **critical step** after data ingestion in the pipeline.
- It enables **accurate analysis** and **better decision-making**.

🏛️ Example: Public Library System

A library wants to combine two datasets:

1. **Library catalog**: ISBN, title, author, publisher, publication date
2. **Circulation database**: Barcode, title (as "Book Title"), author, due date

Steps in Data Mapping:

1. **Identify fields to map** (e.g., title, author)
2. **Standardize naming conventions** (e.g., unify "Book Title" and "Title")
3. **Create mapping rules** (e.g., convert barcode to ISBN)
4. **Test rules** on a small data sample
5. **Create a map** showing relationships between fields
6. **Combine datasets** into one unified dataset

⚠️ Manual vs. Automated Mapping

- **Manual mapping** can be **time-consuming** and **error-prone**.
- **Automated tools** help match fields efficiently, especially for large or complex datasets.

Factors to Consider:

- Structure of the data
- Size of the project
- Available tools

🧠 Takeaway

Data mapping is essential for building reliable data pipelines.
It improves **data quality**, supports **standardization**, and enables **effective analysis**.