🎓 Video 16 Summary: Data Manipulation Techniques in Flexible Pipelines

🔄 Data Pipelines Aren’t Always Linear

While data pipelines often follow a **collect** → **process** → **store** → **analyze** sequence, in practice they can be **flexible and adaptive** depending on:

- The **project’s needs**
- The **data’s condition**
- The **desired outcomes**

🧠 Why Order Can Vary

- Typically, **cleaning** comes before **manipulation**.
- But if data is full of errors, **manipulating it first** (e.g., removing obvious issues) can **speed up cleaning**and improve accuracy.

🔧 Three Common Data Manipulation Techniques

1. Data Standardization

- Ensures all data follows a **common format**.
- Improves **consistency** and **reliability**.
- **Example**: Kyle standardizes product names to lowercase to unify formatting across entries.

2. Data Enrichment

- Adds **additional information** to existing data.
- Can involve **joining datasets** or adding new fields.
- **Example**: Kyle enriches product data by adding SKU numbers from a catalog.

3. Data Conversion

- Changes the **format** of data for compatibility, readability, or efficiency.
- **Example**: Kyle converts CSV files to **Parquet**, a columnar format optimized for storage and analysis.

🧠 Takeaway

As a cloud data analyst, your choice of **techniques** and their **order** depends on:

- The **structure and quality** of your data
- The **tools** available
- The **business goals** you're supporting

Understanding and applying **standardization**, **enrichment**, and **conversion** effectively ensures your data is **ready for storage and analysis**.