🎓 Video 18 Summary: Data Validation in the Data Pipeline

🧪 Analogy: Quality Control Inspector

Just like a **quality control inspector** checks products on an assembly line for defects and compliance with standards, a **cloud data analyst** performs **data validation** to ensure data is:

- **Complete**
- **Accurate**
- **Consistent**
- **Secure**

🔁 Where Validation Happens

- Validation can occur during **Extract**, **Transform**, or **Load** stages.
- It's **especially critical during the Load stage**, as it's the **last chance** to catch errors before analysis.

✅ Common Data Validation Techniques

1. **Type Validation**
   Ensures data is of the correct type (e.g., zip code should be a number, not a string).
2. **Format Validation**
   Checks that data follows a consistent format (e.g., dates formatted as `YYYY-MM-DD`).
3. **Uniqueness Validation (Duplicate Check)**
   Ensures no duplicate records exist (e.g., same email or name appears twice).
4. **Range Validation**
   Verifies values fall within acceptable limits (e.g., age between 0 and 130).
5. **Null Validation**
   Detects missing or empty values that could affect analysis.

⚠️ Handling Invalid Data

Depending on the **validation rules**, invalid data can be:

- **Discarded** (e.g., age outside valid range)
- **Flagged** for manual review (e.g., misspelled email)
- **Automatically corrected** (e.g., zip code matched to a valid address)

🧠 Takeaway

Data validation is like quality control for your data.
It ensures that the data used for analysis is **reliable**, which leads to **trustworthy insights** and **better decisions**.