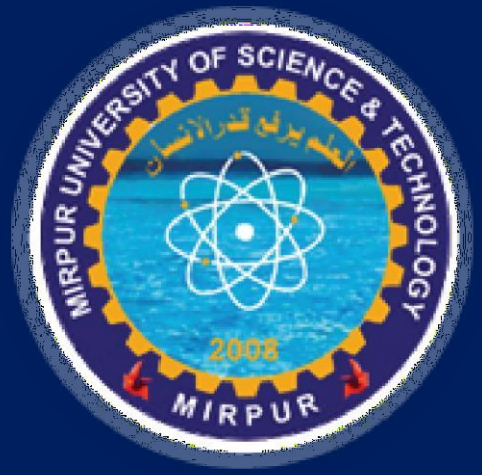


MUST

Wisdom & Virtue

MIRPUR UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF SOFTWARE ENGINEERING

Data Cleaning/ETL (Hands-on)

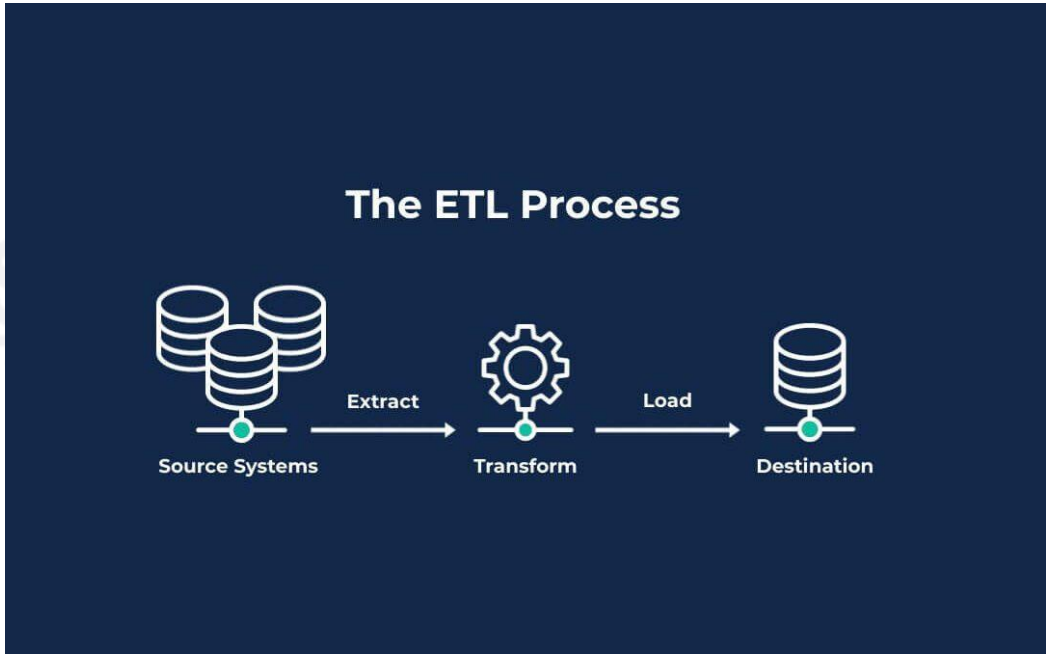
Saman
Fatima
(Lecturer)

What is Data Cleaning?

The process of identifying and correcting errors, in consistencies, and anomalies in data to improve its quality and usability for analysis and decision-making.

What is ETL?

ETL = Extract → Transform → Load
Extract: Pull data from sources
Transform: Clean, fix, and standardize
Load: Save into a final destination



Why is Data Cleaning Important?

- Improves data quality
  - ✓ Reduces errors
  - ✓ Boosts decision-making accuracy
  - ✓ Essential for Machine Learning models

Reasons for Dirty Data

Human Errors (Typos)

- System Errors (Migration failures)
- Outdated Information
- Measurement Errors (Sensor faults)

Example of Dirty Data (Table)

Name	Age	Country	Email Address
John Smith	28	USA	john@email.com
John Smith	28	USA	john@email.com
Maria Anders	300	Germany	maria@website
Lee Wong	_	China	lee.wong@email.com
Ayesha Khan	24	Pakstian	ayesha@domain.com

Types of Missing Data

MCAR: Missing Completely at Random
MAR: Missing At Random
MNAR: Missing Not At Random

MCAR: Missing Completely at Random

The missing data **has no relationship** with any other variable or the missing value itself.
It happens **purely by chance**.
Example:
A survey page was accidentally skipped by a participant.

MAR: Missing At Random

MAR: Missing At Random

The missing data is **related to other observed variables**, but **not to the missing value itself**.

Example:

In a survey, younger people are less likely to disclose their income, but age is recorded.

MNAR: Missing Not At Random

The missing data is **related to the value of the missing data itself**.
Example:
People with very high incomes are more likely to leave the income field blank because it's sensitive.

THANKS