Extract-Transform-Load

The lecture will discuss ETL (Extract-Transform-Load) process

introduced by the field of Data Warehouses

which is the fore-runner of the Data Wrangling process in Data Analytics.

READ Golfarelli and Rizzi Section 1.4

Introduction to Data Warehousing, Chapter 1 of Data Warehouse Design: Modern Principles and Methodologies, 2009, by Matteo Golfarelli and Stefano Rizzi.

We will discuss these steps again in the lectures on Data Wrangling.

The sample picture will give you an idea of what the operations do

- ▶ **Extract** is data collection, data cleaning
- ▶ **Transform** is filtering, aggregation, and shaping (structuring)
- ▶ **Load** is load, deploy or publish in the data warehouse and data mart

though there is debate about where Extract ends and Transform starts in the process

Operational
and external data

Extraction

Cleansing

Validation

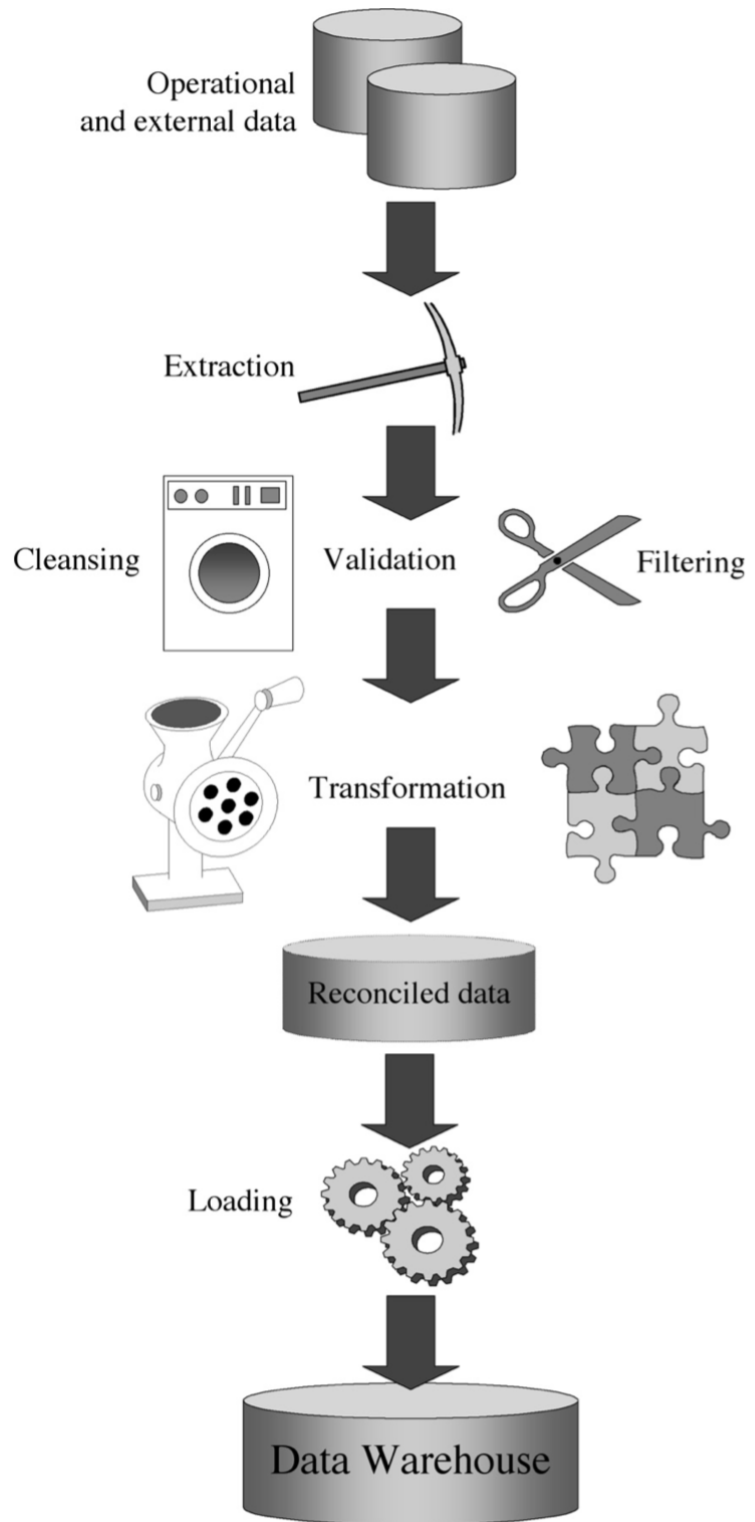Filtering

Transformation

Reconciled data

Loading

Data Warehouse

FIGURE 1-9
Example of
cleansing and
transforming
customer data

John White
Downing St. 10
TW1A 2AA London (UK)

Normalization

```
firstName:John
lastName:  White
address:   Downing St. 10
ZIPCode:   TW1A 2AA
city:      London
country:   UK
```

Standardization

```
firstName:John
lastName:  White
address:   10, Downing Street
ZIPCode:   TW1A 2AA
city:      London
country:   United Kingdom
```

Correction

```
firstName:John
lastName:  White
address:   10, Downing Street
ZIPCode:   SW1A 2AA
city:      London
country:   United Kingdom
```