ETL – Extract, Transform, Load

Using Python to shovel data around

What is that?

- Wikipedia

  In computing, extract, transform, load (ETL) is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s). The ETL process became a popular concept in the 1970s and is often used in data warehousing.

Why would I use that?

- Moving spreadsheets into database (for applying queries)
- Migrating databases (old → new)
- File format change (CSV → XLS)
- Data augmentation (extra fields computed from others)
- Merge data sources
- ...

What framework?

- Good question, no idea, depends on problem...
- But let’s look at some of them!
- Let’s start with an overview:

  https://blog.panoply.io/top-9-python-etl-tools-and-when-to-use-them

Gosh, there’s lots!

- Let’s look at these:
  - pandas
  - Bubbles
  - Bonobo
  - petl
- And evaluate them using these criteria:
  - Migrate employees sqlite DB to MySQL
  - Extend a table using computed values (“email” address)
  - Load a CSV file as new table and add index column (“emp_no”)