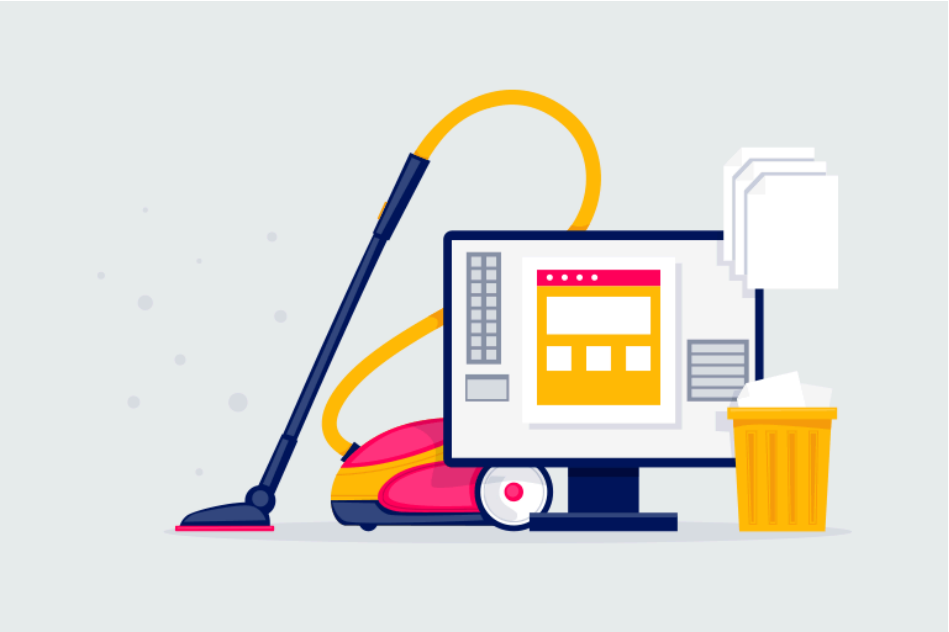
Extract Transform Load



You are a Data Engineer and your team is being tasked with **extracting** the data from a company's flat file data storage, **transforming** that data into a normalized structure for storage in an SQL database, then **loading** it into a data dashboard for displaying insights.
(ETL)

The company generally works with very large datasets, you must demonstrate your ETL techniques on a sample dataset. This dataset should:

- **Contain at least 1000 rows**, but the more rows the better!
- **Contain around 10 Columns**, with columns for
  - Categorical Data - Data that separates rows into discrete states
  - Numerical Data - Data that represents scalar values
  - (optional) Ordinal Data - Data that ranks rows on non-numerical categories
  - (optional) Binary Data - Data that is in either of 2 distinct states

Example Topics:

You are free to choose any topic of interest, but below you can find one we recommend

Spotify Data set: https://spotify-public-dataset.s3.amazonaws.com/spotify_songs.csv

The above data set is scraped from the Spotify API. There are several columns with ambiguous purposes, which you can read more on in the **Spotify Documentation**

Some examples of other data sets you can look for are

Cognixia

1. **Sales over periods of time** - Tracking sales in a given industry or of a single company
2. **Medical Records** - Spread of a disease, effectiveness of a treatment, health trends of a given population
3. **Scientific Measurements** - Recordings of physical phenomena, measurements and results of experiments
4. **Sports Statistics** - Tracking performance of a single team, the results of individual players, or an entire league

Data Cleaning

Import your chosen dataset into a Python Notebook. This notebook should clearly demonstrate the steps taken to **clean** and **transform** your data.

- **Clean out any missing data** - Empty cells can be removed or left to be counted in other measurements, depending on context
- **Replace or remove any incorrect data** - Incorrect data can be either removed entirely, or replaced with logically consistent data
- **Separate data into any number of appropriate data frames** -
    - To write your data to a SQL database, it should be normalized
    - Separating data can also reveal insights.
- **Write data to the relational database of your choice** - Table data must be normalized before being written to the database

Data Visualization

- **Import your clean data to a new Jupyter Notebook** - This notebook can act as your proof of concept dashboard
- **Create plots that display your insights in individual cells** - Clearly state your intended analysis in markdown cells
- **Display any relevant rows or data** - Display notable rows of data, as well as data points like column counts, minimums, or maximums
- **Use at least three kinds of visualizations** - Include at least one Visualization type that we did not go over in class.