# **PRÀCTICA 3: EXTRACCIÓ DE DADES DES DE DIFERENTS FONTS D’INFORMACIÓ**

En l’àmbit de la gestió i l’anàlisi de dades, el procés **ETL** (Extract, Transform, Load) és fonamental per consolidar informació provinent de diverses fonts i preparar-la per a la seva anàlisi en entorns com ara magatzems de dades (*data warehouses*).

La **primera fase del procés ETL** és **l’extracció de dades**, i és una de les més crítiques. Aquesta etapa consisteix a **obtenir les dades originals** des dels seus sistemes d’origen: bases de dades, fitxers pla (com CSV o Excel), serveis web (APIs), sistemes ERP (com Odoo), entre d’altres. Sense una extracció correcta, la resta del procés ETL es veu compromès, ja que les dades poden arribar incompletes, obsoletes o amb errors.

### Objectiu

### L’objectiu d’aquesta pràctica és **obtenir i carregar informació des de tres fonts de dades diferents** (fitxer CSV, Api Odooi base de dades MySQL) mitjançant la llibreria **Pandas** de Python, per tal de preparar les dades per a la seva posterior anàlisi o integració en un procés ETL.

### Requisits previs

* Fonts d’informació disponibles i amb dades
* Entorn Python preparat (Docker ETL)
* Coneixements bàsics de Python

### Procediment pas a pas

Per a la realització de les pràctiques relacionades amb el procés ETL (Extracció, Transformació i Càrrega de dades), s’utilitzarà la llibreria **Pandas** de Python, ja que és una eina àmpliament utilitzada en l’àmbit de l’anàlisi i la manipulació de dades.

Pandas ofereix una sintaxi clara, eficient i altament expressiva per treballar amb dades estructurades provinents de diverses fonts, com ara fitxers CSV i bases de dades relacionals. A més, permet dur a terme operacions comunes del procés ETL com la neteja de dades, la transformació de tipus, la filtració, l’agregació i la combinació de conjunts de dades.

Podeu trobar la documentació de la llibreria a:

<https://pandas.pydata.org/docs/user_guide/index.html>

El alumne per a la realització de la pràctica disposarà de l’estructura del fitxer **main.py** i el fitxer **extraer.py**

El fitxer **main.py** serà on anem a estructura tot el procés ETL , començant per les funcions d’extracció de dades que s’implementaran durant aquesta pràctica.

El fitxer **extraer.py** servirà per organitzar totes aquelles funcions necessàries en la extracció de les dades de les diferents fonts d’informació que van a ser utilitzades en l’execució del projecte.

La funció extraer\_odoo amb la funcionalitat d’extraure información mitjançant l’API d’Odoo ve ja programada per a que l’alumnat puga fer ús d’ella.

### Notes Importants

* **Llegir la documentació insertada en els fitxers main.py i extraer.py**
* El alumnat es lliure de modificar les plantilles i crear les funcions que siguen necessaries per al desenvolupament del seu projecte.
* Els alumnes poden utilizar tantes fonts d’informació com siguen necessaries i creguen pertinents:

Repositori APIs llliure: [GitHub - public-apis/public-apis: A collective list of free APIs](https://github.com/public-apis/public-apis)

# ENTREGA

La **prova d’entrega de la Tasca 3**:

* L’alumne haurà d’entregar el codi font complet, ja siga en format .py on es mostre de forma clara i comentada el procés d’extracció de cada una de les fonts.
* A més, s’hauran d’adjuntar captures de pantalla o mostres de les primeres files de les dades obtingudes de cada font (per exemple, amb df.head()), que evidencien que s’ha realitzat la lectura correcta i que els datasets s’han carregat amb èxit dins de l’entorn de Python.