1. **Data warehousing vs Data warehouse vs Database**

* **A data warehouse** is a centralized repository that stores historical data from multiple sources, specifically designed for analysis and reporting.
* **Data warehousing** refers to the process of collecting, transforming, and loading data into a data warehouse for analytical.
* **A database** is for day-to-day transactions, while a data warehouse is for complex data analysis.

1. **ETL (Extract, Transform, Load)**

ETL is a fundamental process in data warehousing that involves extracting data from various sources, transforming it into a standardized format, and then loading it into a central repository like a data warehouse, enabling efficient analysis and reporting by cleaning, organizing, and merging data from disparate systems for business intelligence purposes.

**Key aspects of the ETL process:**

**Extract**: Retrieves raw data from multiple sources like databases, flat files, APIs, web services, and applications.

**Transform**: Cleans and standardizes the extracted data by handling missing values, correcting inconsistencies, and applying data type conversions.

Performs data manipulations like joining tables, calculating derived values, and data aggregation.

**Load**: Transfers the transformed data into the data warehouse, ensuring proper structure and organization for querying and analysis.

Can be done in batch mode (processing large amounts of data at once) or incremental mode (loading only new or updated data).

**Benefits of using ETL in data warehousing**

**Improved data quality**: By cleaning and standardizing data, ETL ensures consistent and reliable information for analysis.

**Enhanced data integration**: ETL combines data from multiple sources, creating a unified view for comprehensive analysis.

**Faster data access**: Data stored in a data warehouse is readily accessible for reporting and analytics.

1. **With python how to extract multiple files from different sources?**
2. **Multiple Sheet Excel Files:** Extract of data from multiple sheet files where there are 3 sheets( Sheet1, Sheet2, Sheet3).

*# import Pandas library import pandas as pd*

*# Read our file. Here sheet\_name=1*

*# means we are reading the 2nd sheet or Sheet2 df = pd.read\_excel('Sample1.xlsx', sheet\_name = 1) df.head()*

**Extracting text from Doc File:** Here we will extract text from the doc file using docx module. For installation:

*pip install python-docx*

*# Importing our library and reading the doc file*

*import docxdoc = docx.Document(r'D:/My\_Stuff/notes/Data Mining/pactice/Task.docx’) # Printing the title*

*print(doc.paragraphs[2].text)*

1. **Extract Text from Images:** we would be making use of pytesseract OCR library.

*pip install pytesseract*

*from PIL import Imageimport pytesseract # Read the image*

*img = Image.open(r'D:\My\_Stuff\notes\Data Mining\pactice\tumba.png’) # Set the Tesseract OCR path (update this if necessary)pytesseract.pytesseract.tesseract\_cmd = r'C:\Program*

*Files\Tesseract-OCR\tesseract.exe’*

*# Extract text*

*text = pytesseract.image\_to\_string(img)*

*print(text)*

1. **Extracting Data from PDF File**

The task is to extract Data (Image, text) from PDF in Python. we would have to install the PyMuPDF library using Pillow.

*pip install PyMuPDF Pillow*

*# import module*

*import fitz*

*# Reading our pdf file docu=fitz.open('file.pdf')*

*# Initializing an empty list where we will put all text text\_list=[]*

*# Looping through all pages of the pdf file for i in range(docu.pageCount):*

*# Loading each page pg=docu.loadPage(i)*

*# Extracting text from each page*

*pg\_txt=pg.getText('text')*

*# Appending text to the empty list text\_list.append(pg\_txt)*

*# Cleaning the text by removing useless*

*# empty strings and unicode character '\u200b' text\_list=[i.replace(u'\u200b','') for i in text\_list[0].split('\n') if len(i.strip())! = 0]*

*print(text\_list)*