# EXPERIMENT NO. 6

## AIM:

## THEORY:

ETL (Extract, Transform, Load) is a core process in data warehousing and data integration that involves the following stages:

(a) Extract : Collecting raw data from differant sources such as databases, flat files, APIs and real time streaming.

(b) Transform : Applying various transformations to convert the raw data into a useful and meaningful format. Transformations include data cleaning, filtering, aggregation, sorting and generating derived columns.

(c) Load : Storing the transformed data into a target system, such as data warehouse, database or a flat file for analysis and reporting.

The pandas library provides powerful tools for data manipulation and transformation, which are essential for ETL processes.

EXPT. NO. NAME

M T W T F S S

Page No.: 18

Date: 25/2/25

YOUVA

(1) Data Creation :

pd. DataFrame () : Creates a DataFrame object from the given dictionary of lists. Each key in the dictionary represents a column and the corresponding list contains the column values.

(2) Adding a new column :

str .upper() : Converts all string values in the 'Name' column to upper case and adds them as a new column 'Name -Upper'.

(3) Multicasting ( copying DataFrames) :

.copy () : Creates a deep copy of the DataFrame. Changes made to the copy do not affect the original DataFrame.

(4) Conditional split :

(df [condition]) : Filters the data into two subsets based on the condition. The 'Sales' column is used to segregate data into 'High sales' and 'Low sales'.

(5) Aggregation :

.groupby() and sum(): Groups the data by the 'Country' column and computes the sum of 'Sales' for each country. reset_index () converts the result into a DataFrame.

(6) Sorting :

.sort_values (): Sorts the data in descending order of the 'Sales' column. The ascending=False parameter ensures the highest sales appear first.

(7) Derived column :

.apply() with lambda function : Creates a new column 'Sales - Category' based on a conditional rule, classifying sales as 'High' or 'Low'.

CONCLUSION:

The experiment demonstrates ETL transformations using Python & Pandas including data enrichment through new columns, independent data copies for parallel processing, conditional data filtering, aggregation for insights, sorting to prioritize metrics & generating derived columns.