PANDAS THEORY - DAY 1

1. What is Pandas?

Pandas is an open-source Python library used for data manipulation, data analysis,

data cleaning, and data transformation. It is widely used in Data Science, Machine Learning,

Data Analytics, and ETL processes. Pandas is built on top of NumPy.

2. Why Pandas is Needed?

Raw data is often messy and unstructured. Pandas helps to clean, structure, and

prepare data for Machine Learning models and analysis.

3. Core Data Structures in Pandas:

A) Series:

A Series is a one-dimensional labeled array that can store data such as numbers and strings.

It has an index and values.

B) DataFrame:

A DataFrame is a two-dimensional labeled data structure (rows and columns).

It is like an Excel table. Each column in a DataFrame is a Series.

4. Features of Pandas:

- Fast and efficient data handling
- Easy handling of missing values
- Powerful grouping using groupby
- Merge and join support
- File handling (CSV, Excel, SQL)
- Data filtering and sorting

- Statistical operations

5. Difference Between Series and DataFrame:

Series: 1D, single column structure

DataFrame: 2D, table structure with rows and columns

6. Applications of Pandas:

- Data cleaning
- Exploratory Data Analysis (EDA)
- Financial analysis
- Sales analysis
- Machine Learning preprocessing
- Dashboard backend data preparation