Project Title: Heart Disease Dataset Analysis Project

❖ Project Description

This project performs an Exploratory Data Analysis (EDA) on a heart disease dataset using Python. The code identifies trends, relationships, and patterns in the dataset using statistical methods and data visualization tools. Additionally, the code outlines how to implement distributed data processing in Big Data environments using master and worker nodes.

❖ Features and Functionality

The project analyzes the dataset along the dimensions of the **3Vs of Big Data (Volume, Velocity, Variety)** and performs the following tasks:

1. Volume:

- Loads and analyzes structured data of various sizes.

- Extracts information about dataset dimensions and memory usage.

2. Variety:

- Processes and visualizes a mix of categorical and numerical features (e.g., age, gender, chest pain type).

3. Velocity:

- Performs efficient statistical operations and generates insights quickly.

4. Veracity

- Identifies missing values, ensures data quality, and visualizes correlations for trustworthy results.

❖ Dataset Description

The dataset (`heart_disease.csv`) contains the following key features:

- age: Age of the patient.

- sex: Gender (1: Male, 0: Female).

- cp: Chest pain type (0-3).

- target: Target variable (1: Presence of heart disease, 0: No disease).

- Other features include exercise-induced angina, oldpeak, and more.

❖ Installation and Prerequisites

1. Software Requirements

- Python 3.x
- Libraries:
  - `pandas`
  - `numpy`
  - `matplotlib`
  - `seaborn`
  -`PySpark`
  - `Jupyter Notebook`

2. Hardware Requirements

A cluster setup with one master node and one or more worker nodes.

Worker nodes should have PySpark installed and accessible.

Setting Up PySpark with Jupyter Notebook

Step 1: Install PySpark and Jupyter

Install the required packages on all nodes.

Step 2: Start the Master Node

Set the master node to run Spark in standalone mode.

Step 3: Start the Worker Nodes

Connect worker nodes to the master.

Verify that the worker nodes are connected by visiting the Spark Web UI.

Step 4: Start Jupyter Notebook

On the master node, launch Jupyter Notebook:

❖ **Running the Code in Jupyter with PySpark**

1. Initialize the SparkSession

Start by importing PySpark and initializing the SparkSession:

from pyspark.sql import SparkSession bash

2. Load the Dataset

Load the heart_disease.csv dataset into a Spark DataFrame:

3. Perform Analysis

Perform all the necessary requirements.

4. Stop the Spark Session

Once the analysis is complete, stop the Spark session:

❖ **Outputs**

Local Outputs

- Statistical summaries (mean, median, standard deviation).
- Visualizations:
    - **Target Class Distribution**: Count of patients with and without heart disease.
    - **Age Distribution**: Histogram of patient ages.
    - **Gender Distribution**: Male vs. female patients.
    - **Correlation Heatmap**: Relationships between numerical features.
    - **Chest Pain Analysis**: Impact of chest pain types on heart disease.