PySpark Installation:

Step 1: Install PySpark

Open a terminal or command prompt and run:

```
pip install pyspark
```

Verify Installation

Check if PySpark is installed correctly by running:

```
pyspark –version
```

Step 2: Install Java (if not installed)

PySpark requires **Java 8 or 17**. To check if Java is installed, run:

```
java -version
```

If Java is missing, download and install **OpenJDK 8 or 17**

```
sudo apt install default-jdk
```

Step 3: Install and Configure Findspark

Findspark helps Jupyter locate PySpark.

Install it:

```
        pip install findspark
```

Open Jupyter Notebook:

```
        jupyter notebook
```

In a new Jupyter Notebook cell, run:

```
import findspark
        findspark.init()
```

Step 4: Initialize PySpark in Jupyter Notebook

After running findspark.init(), create a **Spark session**:

```
        from pyspark.sql import SparkSession
        # Create SparkSession
        spark = SparkSession.builder.appName("JupyterPySpark").getOrCreate()
        # Check Spark version
        print(spark.version)
```

If this runs successfully, PySpark is ready in Jupyter Notebook

Step 5: Test PySpark with DataFrames

Try creating a simple DataFrame:

```
data = [("Alice", 25), ("Bob", 30), ("Charlie", 35)]

columns = ["Name", "Age"]

df = spark.createDataFrame(data, columns)

df.show()
```

Experiments :

1. Write PySpark program to perform the following operations
    a. Read data from CSV File
    b. Get basic statistics like count, mean, stddev, min, and max
    c. Count the total number of rows
    d. Find the number of unique values in a column
    e. Update specific value
    f. Write to CSV File
2. Write PySpark program to perform the following operations
    a. Read data from JSON file
    b. Count the total number of records
    c. Insert new record
    d. Update a specific record
    e. Write to JSON file