Step 1: Install Java

PySpark requires Java (JDK 17 recommended).

1. Download **Java JDK** (or OpenJDK).
2. Install it, e.g., in `C:\Java\jdk-17`
3. Set environment variables:
    - **JAVA_HOME** → `C:\Java\jdk-17`
    - Add `%JAVA_HOME%\bin` to your **PATH**

Check installation in cmd:

java -version

Step 2: Install **Apache Spark**

Download Spark from the official site.

- Choose version **3.5.0** (or latest) with **Hadoop 3**.
- Extract it, e.g., `C:\spark\spark-3.5.0-bin-hadoop3`.

Set environment variables:

- **SPARK_HOME** → `C:\spark\spark-3.5.0-bin-hadoop3`
- Add `%SPARK_HOME%\bin` to your **PATH**

Check installation:

```
spark-shell
```

Step 3: Install Winutils (needed for Hadoop on Windows)

1. Download the `winutils.exe` for Hadoop 3.
   👉 Unofficial binaries link.
2. Paste to "C:\hadoop\bin\winutils.exe"

Set environment variable:

- **HADOOP_HOME** → `C:\hadoop`
- Add `%HADOOP_HOME%\bin` to **PATH**

Execute in cmd

C:\hadoop\bin\winutils.exe

Step 4: Install **PySpark** via pip

Open **CMD** or **PowerShell**:

```
pip install pyspark findspark
```

Step 5: Test PySpark

Open Python and try:

```
import findspark

findspark.init()


from pyspark.sql import SparkSession


spark = SparkSession.builder.appName("LocalTest").getOrCreate()


df = spark.createDataFrame([(1, "Alice"), (2, "Bob")], ["id", "name"])

df.show()
```