

# Steps to install PySpark on windows:

## 1. Install Java Development Kit (JDK):

- Download and install the latest version of JDK from the Oracle website: Java SE Downloads.
- After installation, set the JAVA\_HOME environment variable to the JDK installation directory.

## 2. Install Apache Spark:

- Download the latest version of Apache Spark from the official website: Apache Spark Downloads.
- Extract the downloaded Spark archive to a directory on your system, for example, C:\spark.

## 3. Install FindSpark (if using Anaconda):

- FindSpark is a Python library that helps to locate Spark installed on the system.
- Install FindSpark using pip:
- Copy code
- pip install findspark

## 4. Set Environment Variables:

- Add Spark and Hadoop binary directories to the system's PATH environment variable.
- Create a new system environment variable named SPARK\_HOME and set its value to the Spark installation directory (e.g., C:\spark).
- If you installed Hadoop, create a new system environment variable named HADOOP\_HOME and set its value to the Hadoop installation directory (e.g., C:\hadoop).

## 5. Install PySpark:

- If you are using Anaconda, PySpark is included by default. You can directly import and use it in your Python scripts.
- If you are not using Anaconda, you can install PySpark using pip:
- Copy code
- pip install pyspark

## 6. Verify Installation:

- Open a command prompt and run the following command to launch PySpark:
- Copy code
- `pyspark`
- If PySpark starts without any errors and you see the Spark shell, it means the installation was successful.
-