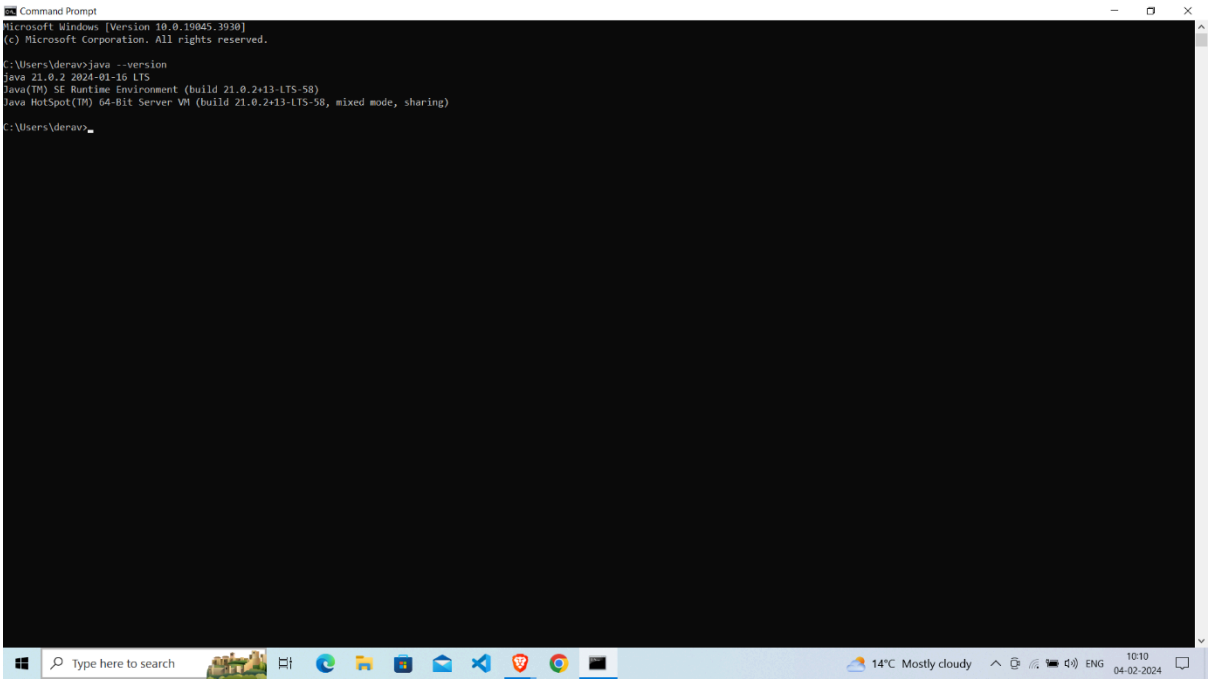
Installation of Apache Spark on my system

Step 1: Install Java 8

Start > type cmd> click Command Prompt.

java -version



If you don't have Java installed:

1. Open a browser window, and navigate to https://java.com/en/download/.

2. Click the Java Download button and save the file to a location of your choice.

3. Once the download finishes double-click the file to install Java.

Step 2: Install Python

1. To install the Python package manager, navigate to https://www.python.org/ in your web browser.

2. Mouse over the Download menu option and click Python 3.11.

3. Once the download finishes, run the file.

4. Near the bottom of the first setup dialog box, check off Add Python 3.11 to PATH. Leave the other box checked.

5. Next, click Customize Installation.

6. You can leave all boxes checked at this step, or you can uncheck the options you do not want.

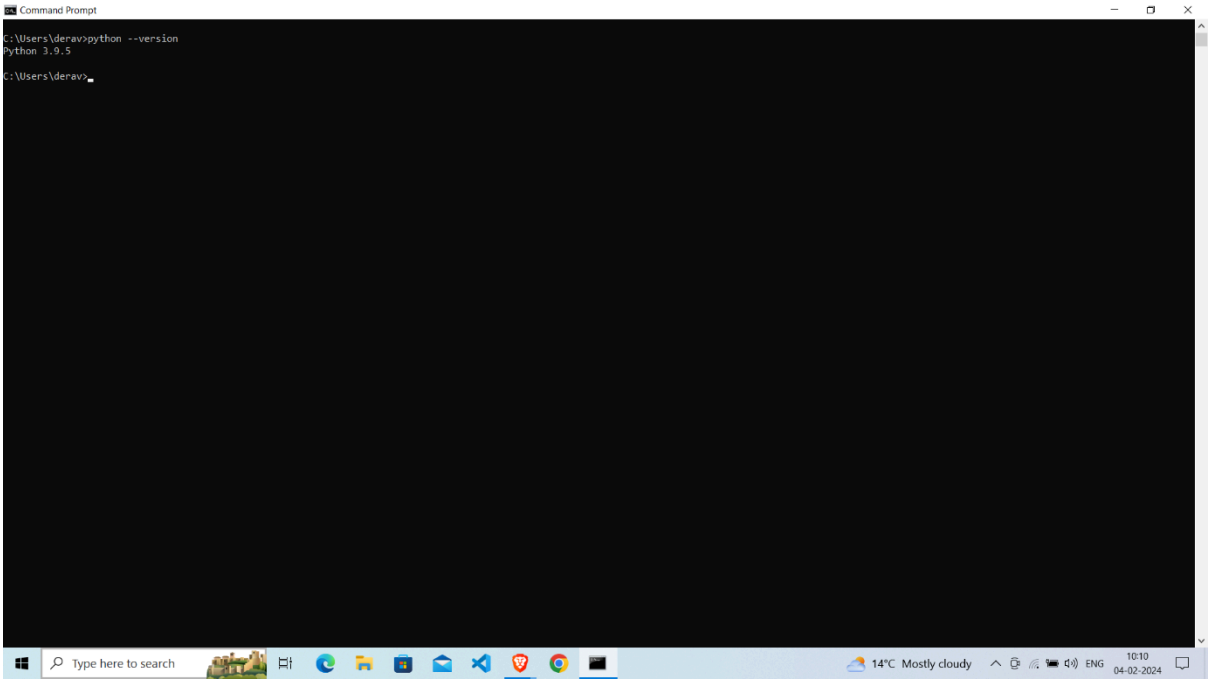
7. Select the box Install for all users and leave other boxes as they are.

8. Under Customize install location, click Browse and navigate to the C drive. Add a new folder and name it Python.

9. Select that folder and click OK.

10. When the installation completes, click the Disable path length limit option at the bottom and then click Close.

11. If you have a command prompt open, restart it. Verify the installation by checking the version of Python:



Step 3: Download Apache Spark

1. Open a browser and navigate to https://spark.apache.org/downloads.html.

2. Under the Download Apache Spark heading, there are two drop-down menus. Use the current non-preview version.

Choose a Spark release -> 3.5.0

Choose a package type -> Pre-built for Apache Hadoop 3.

3. Click the spark-3.5.0-bin-hadoop3.tgz link

4. A page with a list of mirrors loads where you can see different servers to download from. Pick any from the list and save the file to your Downloads folder.

Step 4: Add winutils.exe File

Download the winutils.exe file for the underlying Hadoop version for the Spark installation you downloaded.

1. Navigate to this URL https://github.com/cdarlint/winutils and inside the bin folder, locate winutils.exe, and click it.

2. Find the Download button on the right side to download the file.

3. Now, create new folders Hadoop and bin on C: using Windows Explorer or the Command Prompt.

4. Copy the winutils.exe file from the Downloads folder to C:\Hadoop\bin

Step 5: Configure Environment Variables

Configuring environment variables in Windows adds the Spark and Hadoop locations to your system PATH. It allows you to run the Spark shell directly from a command prompt window.

1. Click Start and type environment.

2. Select the result labelled Edit the system environment variables.

3. A System Properties dialog box appears. In the lower-right corner, click Environment Variables and then click New in the next window.

4. For Variable Name type SPARK_HOME.

5. For Variable Value type C:\Spark\spark-3.5.0-bin-hadoop3 and click OK. If you changed the folder path, use that one instead.

6. In the top box, click the Path entry, then click Edit. Be careful with editing the system path. Avoid deleting any entries already on the list.

7. You should see a box with entries on the left. On the right, click New.

8. The system highlights a new line. Enter the path to the Spark folder C:\Spark\spark-3.5.0-bin-hadoop3\bin. We recommend using %SPARK_HOME%\bin to avoid possible issues with the path.

9. Repeat this process for Hadoop and Java.

For Hadoop, the variable name is HADOOP_HOME and for the value use the path of the folder you created earlier: C:\Hadoop. Add C:\Hadoop\bin to the Path variable field, but we recommend using %HADOOP_HOME%\bin.

For Java, the variable name is JAVA_HOME and for the value use the path to your Java JDK directory (example, C:\Program Files\Java\<jdk_version>).

10. Click OK to close all open windows.

Step 6: Launch Spark

1. Open a new command prompt Window using the right-click and Run as administrator:

2. To start Spark, enter: