

PySpark Installation steps:

- 1) First Step is to download the spark package from sparks official website the file comes with the extension .tgz
- 2) After that we need to install the java and python in our laptops if they are not present in our system
- 3) After that we need to download the winutils.exe file for the latest version of Hadoop
- 4) Then we need to set some environmental variables for java, spark and Hadoop in our local systems

%SPARK_HOME%\bin
%HADDOP_HOME%\bin
%SPARK_HOME%\python
%PYTHONPATH%
C:\spark\spark-3.5.0-bin-hadoop3\python\lib\py4j-0.10.9.7-src.zip

- 5) Also we need to set the system variables as well as follows

SPARK_HOME	C:\spark\spark-3.5.0-bin-hadoop3
HADDOP_HOME	C:\hadoop
JAVA_HOME	C:\Program Files\Java\jdk-21

6) Now to check PySpark is installed we need to type `pyspark` command in cmd then we will get the output as follows

```

Microsoft Windows [Version 10.0.22631.3085]
(c) Microsoft Corporation. All rights reserved.

C:\Users\delldell>pyspark
Python 3.9.6 (tags/v3.9.6:db3fff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
24/02/03 20:04:43 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/02/03 20:04:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to

  ____
 /  __ \
/   /  \
/_____/

 version 3.5.0

Using Python version 3.9.6 (tags/v3.9.6:db3fff76, Jun 28 2021 15:26:21)
Spark context Web UI available at http://PRATIJK:4040
Spark context available as 'sc' (master = local[*], app id = local-1706970888767).
SparkSession available as 'spark'.
>>> 24/02/03 20:04:59 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC),
users should configure it them) to spark.eventlog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetric
s.oldGenerationGarbageCollectors
24/02/03 20:25:46 WARN Executor: Issue communicating with driver in heartbeater
org.apache.spark.rpc.RpcTimeoutException: Futures timed out after [10000 milliseconds]. This timeout is controlled by sp
ark.executor.heartbeatInterval
at org.apache.spark.rpc.RpcTimeout.org$apache$spark$rpc$RpcTimeout$$createRpcTimeoutException(RpcTimeout.scala:4

```

7) If we go on <http://localhost:4040/jobs/> we will get the dashboard as follows



