

NAME : AKULA SHARATH CHANDRA

BATCH: DATA ENGINEERING

TOPIC: HOW TO INSTALL AND SETUP APACHE SPARK

DATE: 03-02-2024

Installation of Apache Spark and Setup...

Apache PySpark..

Apache Spark is an open-source distributed computing system that provides a fast and

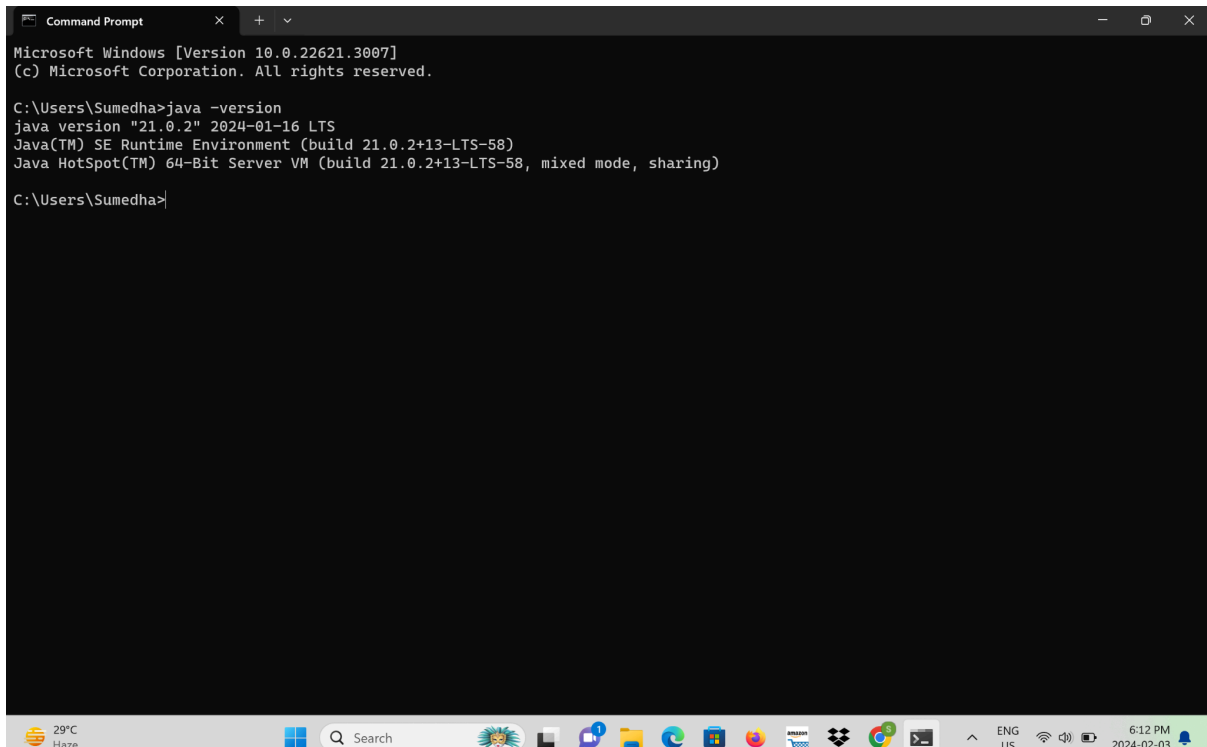
general-purpose cluster-computing framework for big data processing. PySpark is the Python

API for Apache Spark, allowing you to write Spark applications using Python.

Python+Spark=PySpark..

Install Java

- Firstly, we have to install a java jdk version which is compatible with your system. It's good to download and install latest standard version of java.
- After that we have to check in the command prompt by typing **java -version** which displays the version of java which you downloaded.
- In means that you have successfully downloaded and installed it in your system.



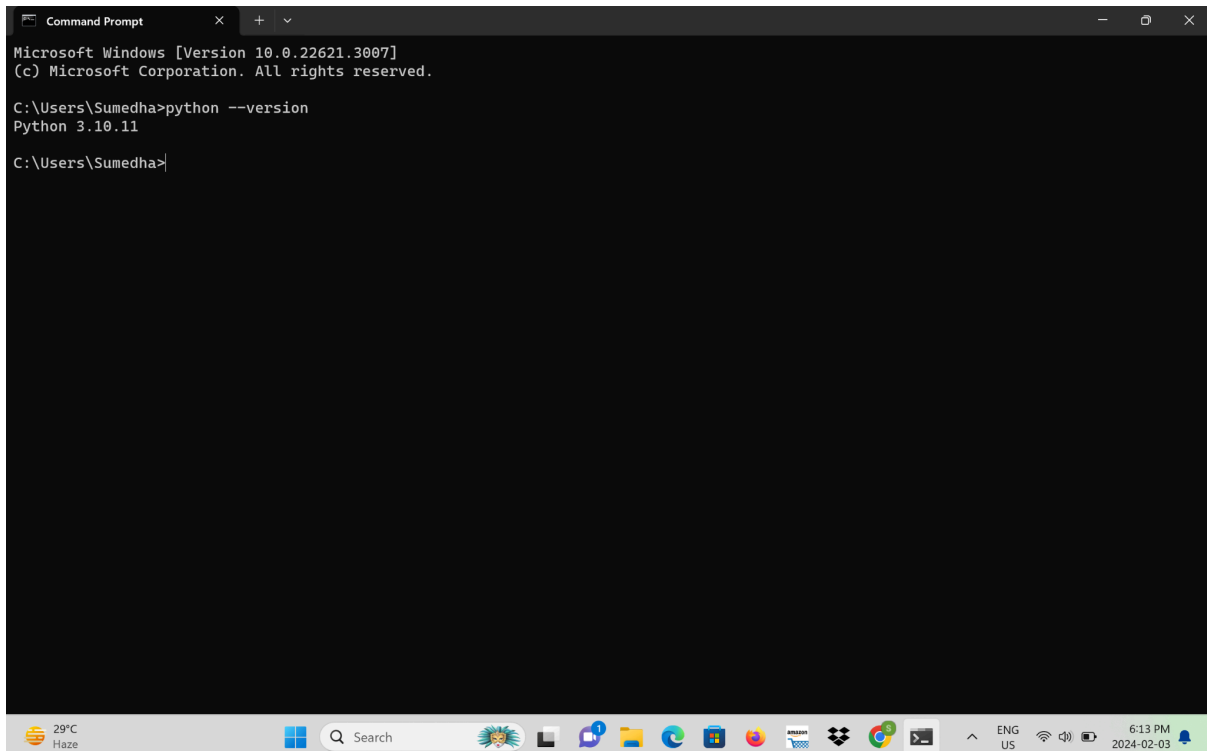
```
Command Prompt
Microsoft Windows [Version 10.0.22621.3007]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Sumedha>java -version
java version "21.0.2" 2024-01-16 LTS
Java(TM) SE Runtime Environment (build 21.0.2+13-LTS-58)
Java HotSpot(TM) 64-Bit Server VM (build 21.0.2+13-LTS-58, mixed mode, sharing)

C:\Users\Sumedha>
```

Install Python

- Along with java, we have to install a python environment into our system. It's good to download and install the latest and standard version of python.
- After that we have to check the command prompt by typing **python-version** which displays the version of java which you downloaded.
- It means that you have successfully downloaded and installed in your system.



```
Command Prompt
Microsoft Windows [Version 10.0.22621.3007]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Sumedha>python --version
Python 3.10.11

C:\Users\Sumedha>
```

The screenshot shows a Windows Command Prompt window. The title bar reads 'Command Prompt'. The window content displays the following text: 'Microsoft Windows [Version 10.0.22621.3007]', '(c) Microsoft Corporation. All rights reserved.', 'C:\Users\Sumedha>python --version', 'Python 3.10.11', and 'C:\Users\Sumedha>'. The Windows taskbar is visible at the bottom, showing the search bar, task view button, and several application icons. The system tray on the right indicates the temperature is 29°C, the weather is 'Haze', and the time is 6:13 PM on 2024-02-03.

Install Apache Spark:

- Visit the Apache Spark download page.
- Choose the latest version of Spark and download the pre-built package for Hadoop. It will be a tarball (.tgz) file.
- Extract the downloaded tarball to a location on your machine.

Set Environment Variables:

Environment Variables

User variables for Sumedha

Variable	Value
HADOOP_HOME	C:\Program Files\HADOOP
JAVA_HOME	C:\Program Files\Java\jdk-21
OneDrive	C:\Users\Sumedha\OneDrive
OneDriveConsumer	C:\Users\Sumedha\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\Sumed...
PYSPARK_HOME	C:\Users\Sumedha\python\python.exe
SPARK_HOME	C:\Program Files\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-had...
TEMP	C:\Users\Sumedha\AppData\Local\Temp

New...

Edit...

Delete

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIND...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 140 Stepping 1, GenuineIntel

New...

Edit...

Delete

OK

Cancel

Environment Variables



User variables for Sumedha

Variable	Value
HADOOP_HOME	C:\Program Files\HADOOP
JAVA_HOME	C:\Program Files\Java\jdk-21
OneDrive	C:\Users\Sumedha\OneDrive
OneDriveConsumer	C:\Users\Sumedha\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\Sumed...
PYSPARK_HOME	C:\Users\Sumedha\python\python.exe
SPARK_HOME	C:\Program Files\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-had...
TEMP	C:\Users\Sumedha\AppData\Local\Temp

New...

Edit...

Delete

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIND...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 140 Stepping 1, GenuineIntel

New...

Edit...

Delete

OK

Cancel

Environment Variables



User variables for Sumedha

Variable	Value
HADOOP_HOME	C:\Program Files\HADOOP
JAVA_HOME	C:\Program Files\Java\jdk-21
OneDrive	C:\Users\Sumedha\OneDrive
OneDriveConsumer	C:\Users\Sumedha\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\Sumed...
PYSPARK_HOME	C:\Users\Sumedha\python\python.exe
SPARK_HOME	C:\Program Files\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-had...
TEMP	C:\Users\Sumedha\AppData\Local\Temp

New...

Edit...

Delete

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIND...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 140 Stepping 1, GenuineIntel

New...

Edit...

Delete

OK

Cancel

Environment Variables



User variables for Sumedha

Variable	Value
HADOOP_HOME	C:\Program Files\HADOOP
JAVA_HOME	C:\Program Files\Java\jdk-21
OneDrive	C:\Users\Sumedha\OneDrive
OneDriveConsumer	C:\Users\Sumedha\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\Sumed...
PYSPARK_HOME	C:\Users\Sumedha\python\python.exe
SPARK_HOME	C:\Program Files\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-had...
TEMP	C:\Users\Sumedha\AppData\Local\Temp

New...

Edit...

Delete

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIND...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 140 Stepping 1, GenuineIntel

New...

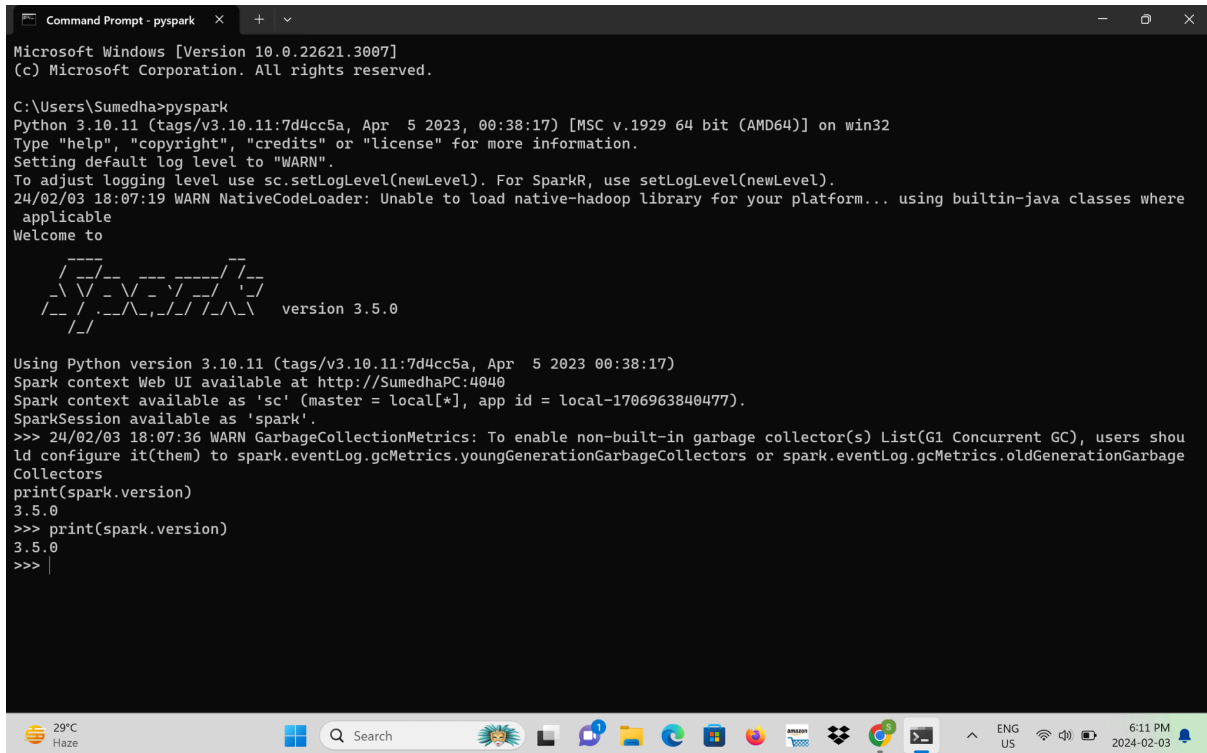
Edit...

Delete

OK

Cancel

After setting up the environment variables, we need to save all of them and have to go to the command prompt and type **pyspark** as below.



```
Microsoft Windows [Version 10.0.22621.3007]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Sumedha>pyspark
Python 3.10.11 (tags/v3.10.11:7d4cc5a, Apr  5 2023, 00:38:17) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/02/03 18:07:19 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 3.10.11 (tags/v3.10.11:7d4cc5a, Apr  5 2023 00:38:17)
Spark context Web UI available at http://SumedhaPC:4040
Spark context available as 'sc' (master = local[*], app id = local-1706963840477).
SparkSession available as 'spark'.
>>> 24/02/03 18:07:36 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.gcMetrics.oldGenerationGarbageCollectors
print(spark.version)
3.5.0
>>> print(spark.version)
3.5.0
>>> |
```

- If it shows like these, you have successfully set up the Apache Spark setup.
- If **http://SumedhaPC:4040** you will get the web page as follows which indicates your pyspark is installed as follows

29°C Haze Search ENG US 6:10 PM 2024-02-03