Assignment

Installation and Setup of Apache Spark:

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.
- It is a data processing framework that can quickly perform processing tasks on very large data sets.
- It can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools.

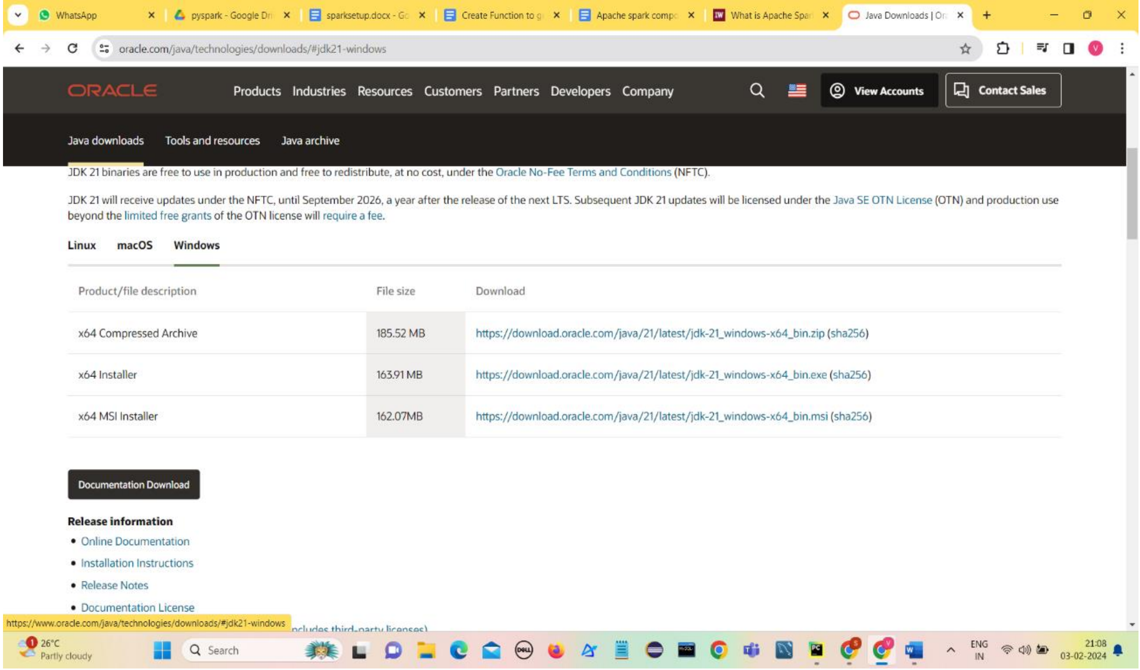
PySpark : It is the Python API for Apache Spark, allowing you to write Spark applications using Python.

Python+Spark=PySpark

Steps to be followed for Installation and Setup

1) Install Java

- Firstly, we have to install java jdk version which is compatible to our system. It is advisable to download and install latest standard version of java**.**
  Download Jdk from the official website oracle.com



WhatsApp
pyspark - Google Dri
sparksetup.docx - Go
Create Function to g
Apache spark compo
What is Apache Spar
Java Downloads | Or
oracle.com/java/technologies/downloads/#jdk21-windows
ORACLE
Products Industries Resources Customers Partners Developers Company
View Accounts
Contact Sales
Java downloads Tools and resources Java archive

JDK 21 binaries are free to use in production and free to redistribute, at no cost, under the Oracle No-Fee Terms and Conditions (NFTC).

JDK 21 will receive updates under the NFTC, until September 2026, a year after the release of the next LTS. Subsequent JDK 21 updates will be licensed under the Java SE OTN License (OTN) and production use beyond the limited free grants of the OTN license will require a fee.

Linux macOS Windows

Product/file description	File size	Download
x64 Compressed Archive	185.52 MB	https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.zip (sha256)
x64 Installer	163.91 MB	https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.exe (sha256)
x64 MSI Installer	162.07MB	https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.msi (sha256)

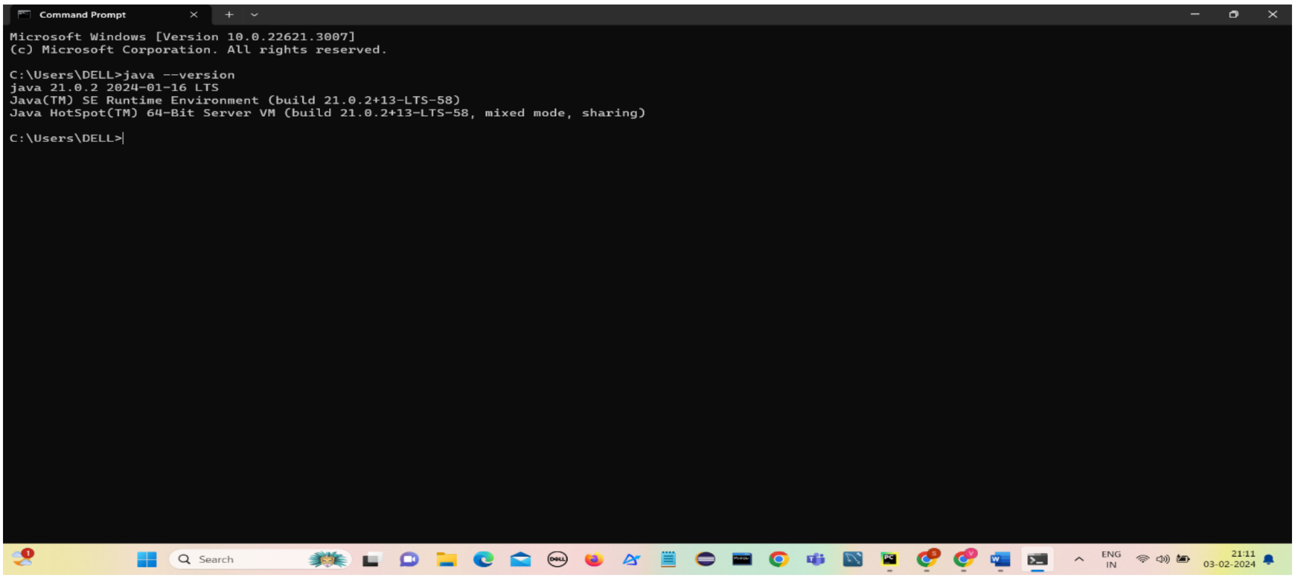
Release information

- Online Documentation
- Installation Instructions
- Release Notes
- Documentation License

https://www.oracle.com/java/technologies/downloads/#jdk21-windows
ncludes third-party licenses)
26°C
Partly cloudy
Search
ENG
IN
21:08
03-02-2024

Check the version of Java installed

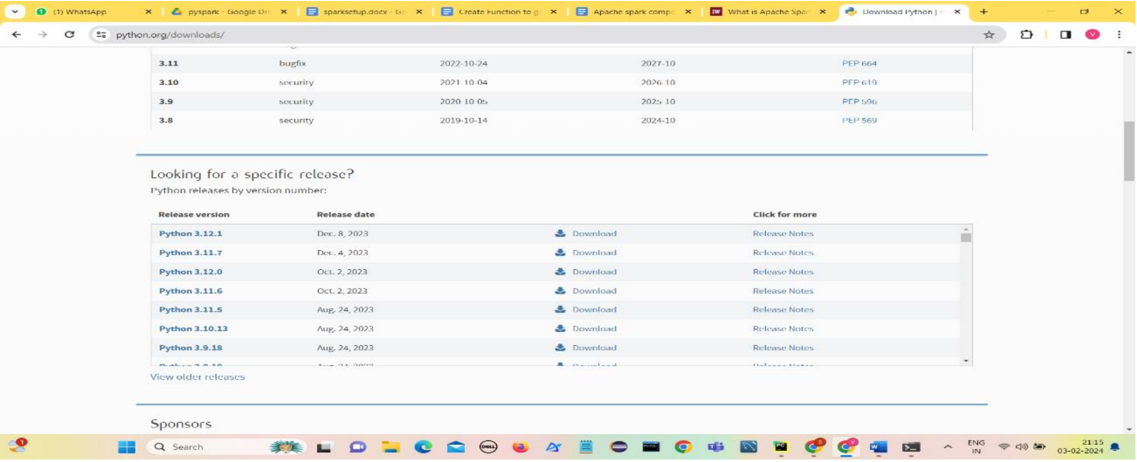
- After that we have to check in command prompt by typing **java –version** which displays the version of java which you downloaded.
- In means that you have successfully downloaded and installed in your system.



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

C:\Users\DELL>java --version
java 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\DELL>
```

2) Install Python

- Along with java, we have to install python environment into our system. Its good to download and install latest and standard version of python.
  Download python from official website in the following way



python.org/downloads/

3.11	bugfix	2022-10-24	2027-10	PEP 664
3.10	security	2021-10-04	2026-10	PEP 619
3.9	security	2020-10-05	2025-10	PEP 596
3.8	security	2019-10-14	2024-10	PEP 569

Looking for a specific release?
Python releases by version number:

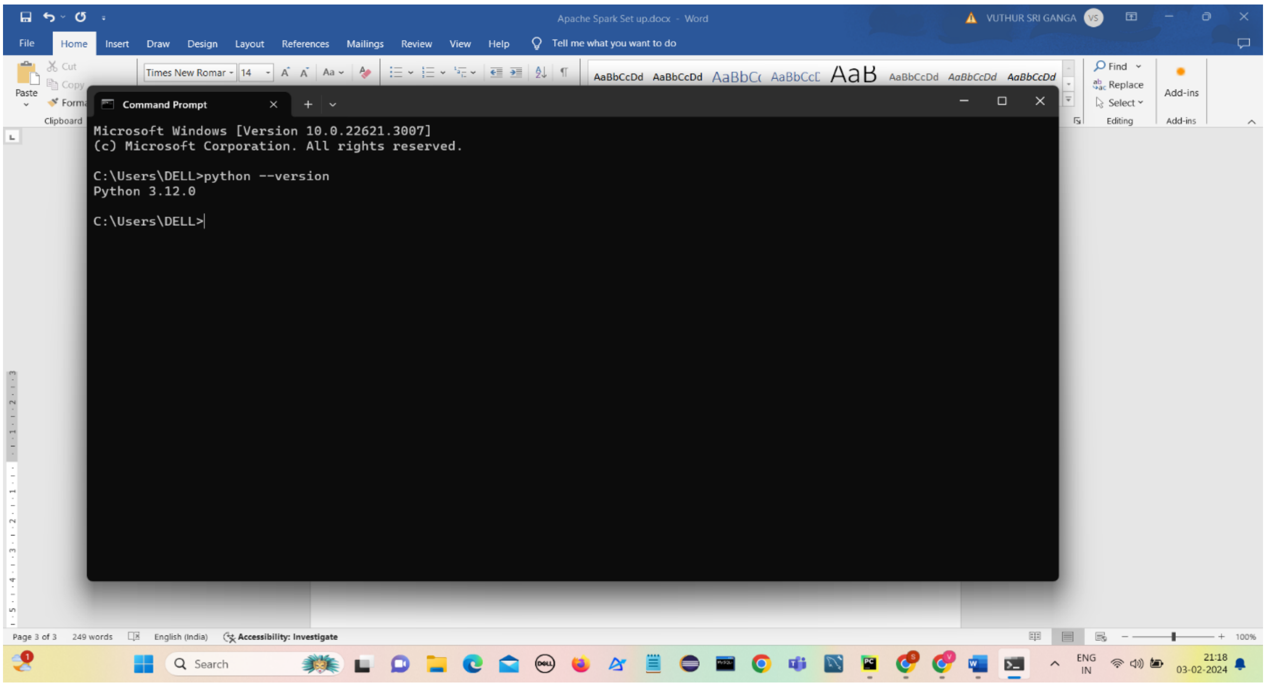
Release version	Release date		Click for more
Python 3.12.1	Dec. 8, 2023	Download	Release Notes
Python 3.11.7	Dec. 4, 2023	Download	Release Notes
Python 3.12.0	Oct. 2, 2023	Download	Release Notes
Python 3.11.6	Oct. 2, 2023	Download	Release Notes
Python 3.11.5	Aug. 24, 2023	Download	Release Notes
Python 3.10.13	Aug. 24, 2023	Download	Release Notes
Python 3.9.18	Aug. 24, 2023	Download	Release Notes

View older releases

Sponsors

- After that we have to check in command prompt by typing **python –version** which displays the version of python which you downloaded.
- In means that you have successfully downloaded and installed in your system.

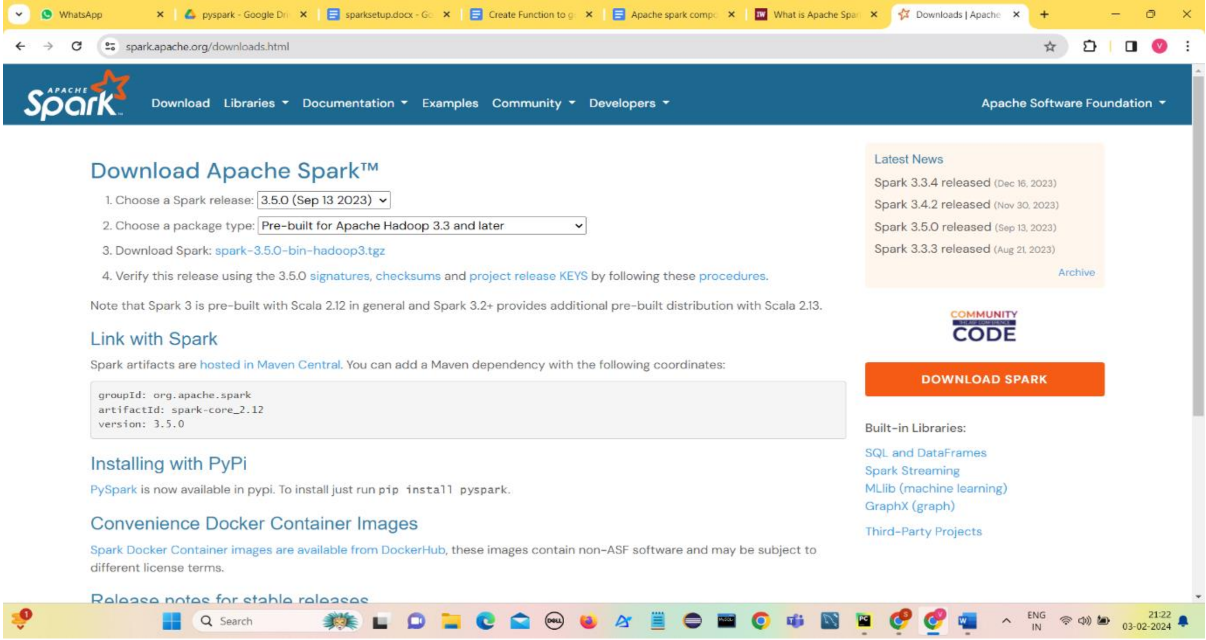
Check the version of Python in Command Prompt



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

C:\Users\DELL>python --version
Python 3.12.0

C:\Users\DELL>
```

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.



spark.apache.org/downloads.html

Download Apache Spark™

1. Choose a Spark release: 3.5.0 (Sep 13 2023)
2. Choose a package type: Pre-built for Apache Hadoop 3.3 and later
3. Download Spark: spark-3.5.0-bin-hadoop3.tgz
4. Verify this release using the 3.5.0 signatures, checksums and project release KEYS by following these procedures.

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

Link with Spark

Spark artifacts are hosted in Maven Central. You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.12
version: 3.5.0
```

Installing with PyPi

PySpark is now available in pypi. To install just run `pip install pyspark`.

Convenience Docker Container Images

Spark Docker Container images are available from DockerHub, these images contain non-ASF software and may be subject to different license terms.

Release notes for stable releases

Latest News

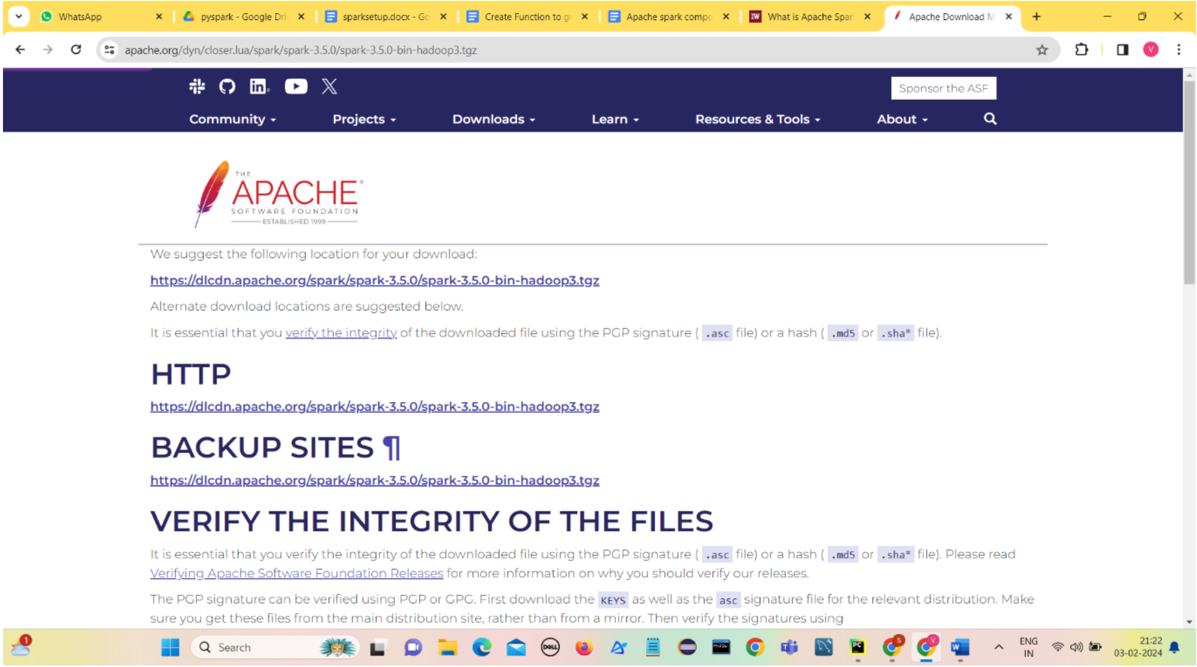
Spark 3.3.4 released (Dec 16, 2023)

Spark 3.4.2 released (Nov 30, 2023)

Spark 3.5.0 released (Sep 13, 2023)

Spark 3.3.3 released (Aug 21, 2023)

Archive



We suggest the following location for your download:

https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

Alternate download locations are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file).

HTTP

https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

BACKUP SITES ¶

https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

VERIFY THE INTEGRITY OF THE FILES

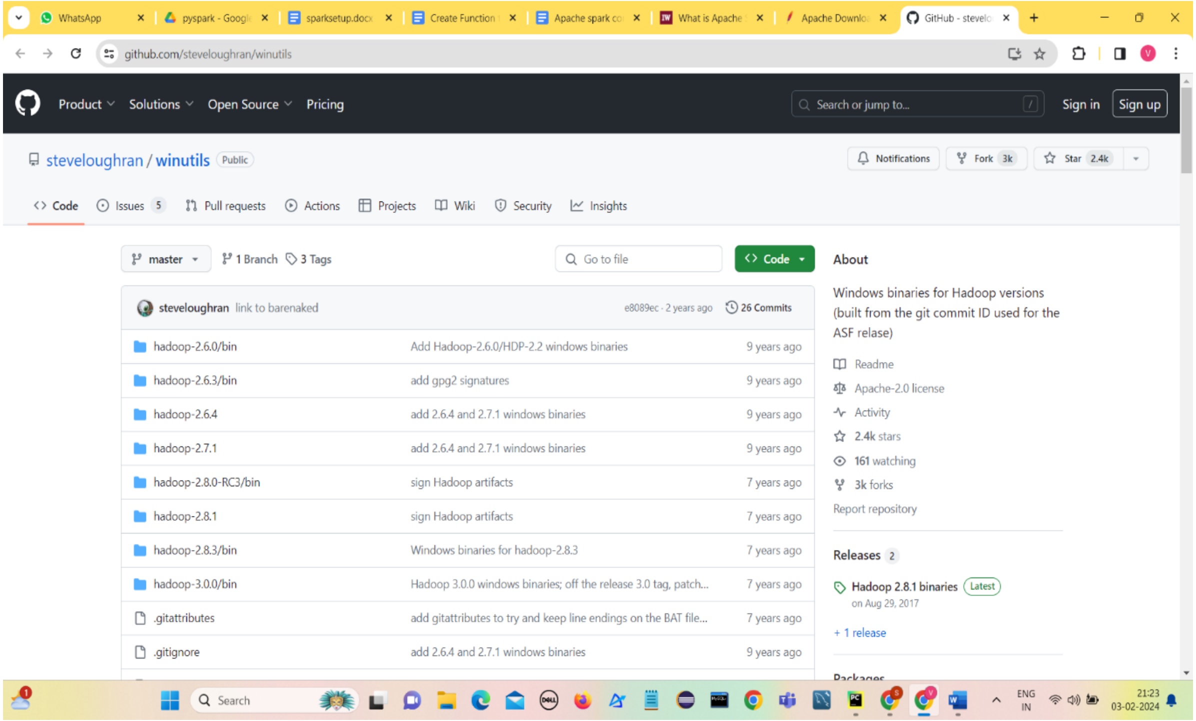
It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file). Please read Verifying Apache Software Foundation Releases for more information on why you should verify our releases.

The PGP signature can be verified using PGP or GPG. First download the KEYS as well as the asc signature file for the relevant distribution. Make sure you get these files from the main distribution site, rather than from a mirror. Then verify the signatures using

Install winutils from the following website

https://github.com/steveloughran/winutils

Ensure that the version of winutils should be same as that of Apache Spark.



steveloughran / winutils Public

steveloughran link to barenaked e8089ec · 2 years ago 26 Commits

hadoop-2.6.0/bin	Add Hadoop-2.6.0/HDP-2.2 windows binaries	9 years ago
hadoop-2.6.3/bin	add gpg2 signatures	9 years ago
hadoop-2.6.4	add 2.6.4 and 2.7.1 windows binaries	9 years ago
hadoop-2.7.1	add 2.6.4 and 2.7.1 windows binaries	9 years ago
hadoop-2.8.0-RC3/bin	sign Hadoop artifacts	7 years ago
hadoop-2.8.1	sign Hadoop artifacts	7 years ago
hadoop-2.8.3/bin	Windows binaries for hadoop-2.8.3	7 years ago
hadoop-3.0.0/bin	Hadoop 3.0.0 windows binaries; off the release 3.0 tag, patch...	7 years ago
.gitattributes	add gitattributes to try and keep line endings on the BAT file...	7 years ago
.gitignore	add 2.6.4 and 2.7.1 windows binaries	9 years ago

About

Windows binaries for Hadoop versions (built from the git commit ID used for the ASF relase)

Readme
Apache-2.0 license
Activity
2.4k stars
161 watching
3k forks
Report repository

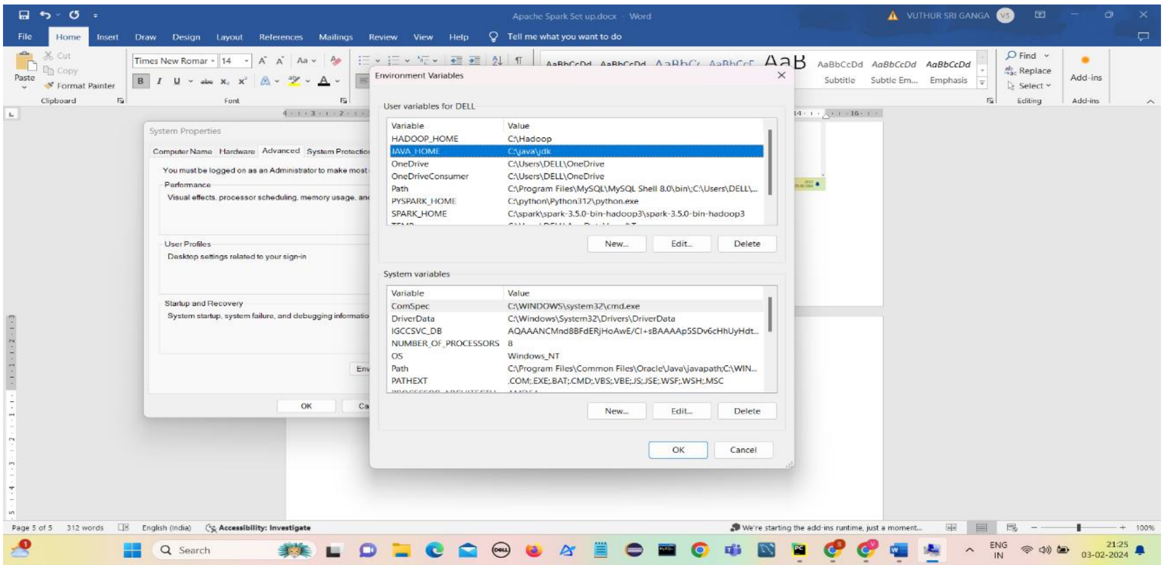
Releases 2

Hadoop 2.8.1 binaries Latest
on Aug 29, 2017

+ 1 release

Packages

Set Environment Variables



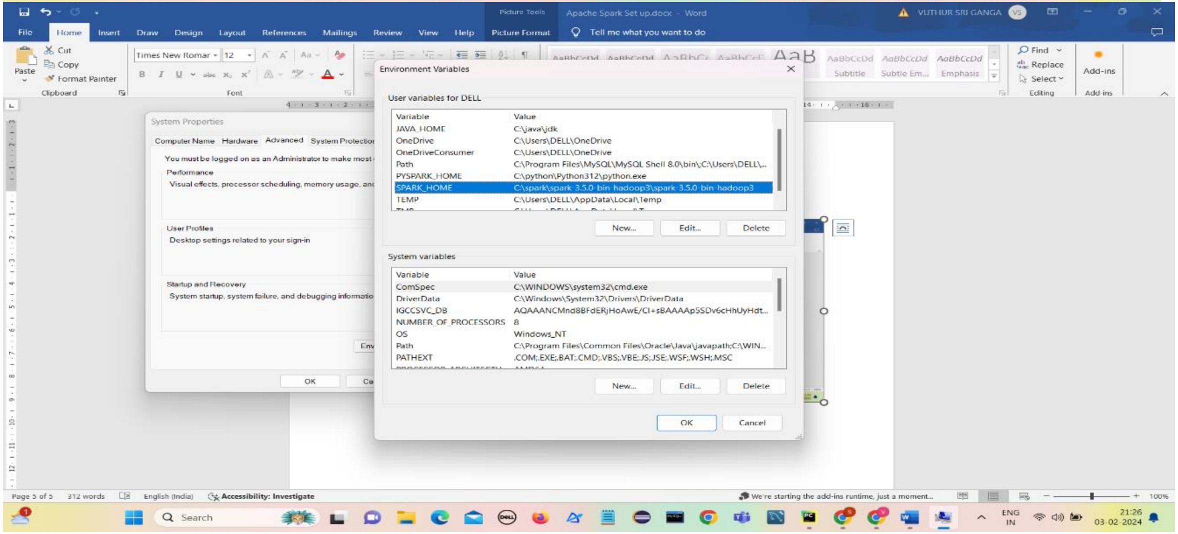
Environment Variables

User variables for DELL

Variable	Value
HADOOP_HOME	C:\Hadoop
JAVA_HOME	C:\java\jdk
OneDrive	C:\Users\DELL\OneDrive
OneDriveConsumer	C:\Users\DELL\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\DELL\...
PYSPARK_HOME	C:\python\Python312\python.exe
SPARK_HOME	C:\spark\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-hadoop3

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
IGCCSVC_DB	AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAAp5SDv6cHhUyHdt...
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC



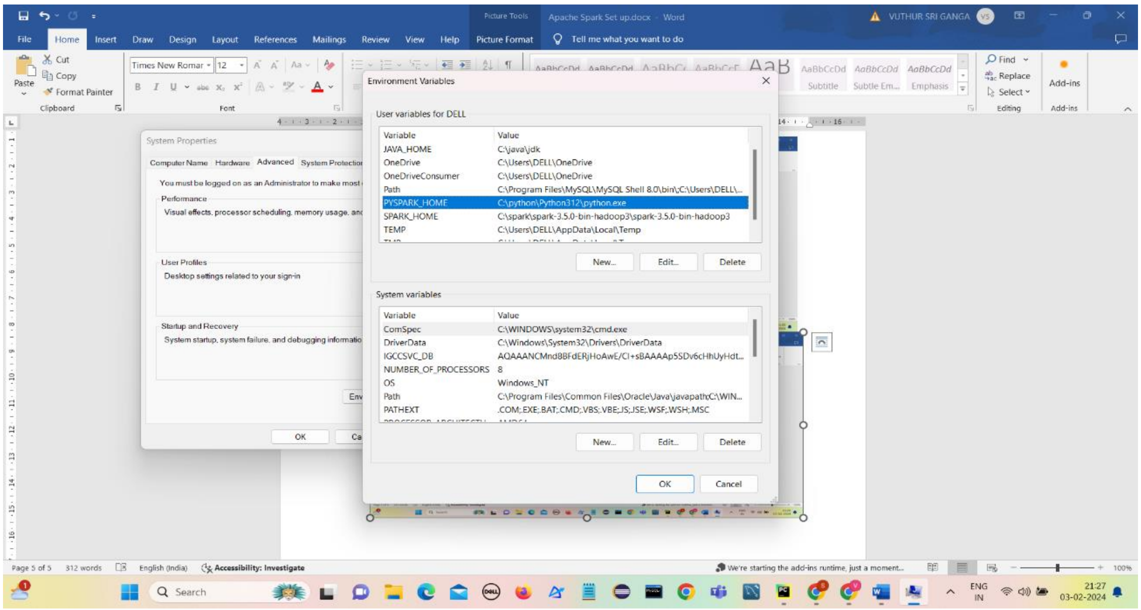
Environment Variables

User variables for DELL

Variable	Value
JAVA_HOME	C:\java\jdk
OneDrive	C:\Users\DELL\OneDrive
OneDriveConsumer	C:\Users\DELL\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\DELL\...
PYSPARK_HOME	C:\python\Python312\python.exe
SPARK_HOME	C:\spark\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-hadoop3
TEMP	C:\Users\DELL\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
IGCCSVC_DB	AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAAp5SDv6cHhUyHdt...
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC



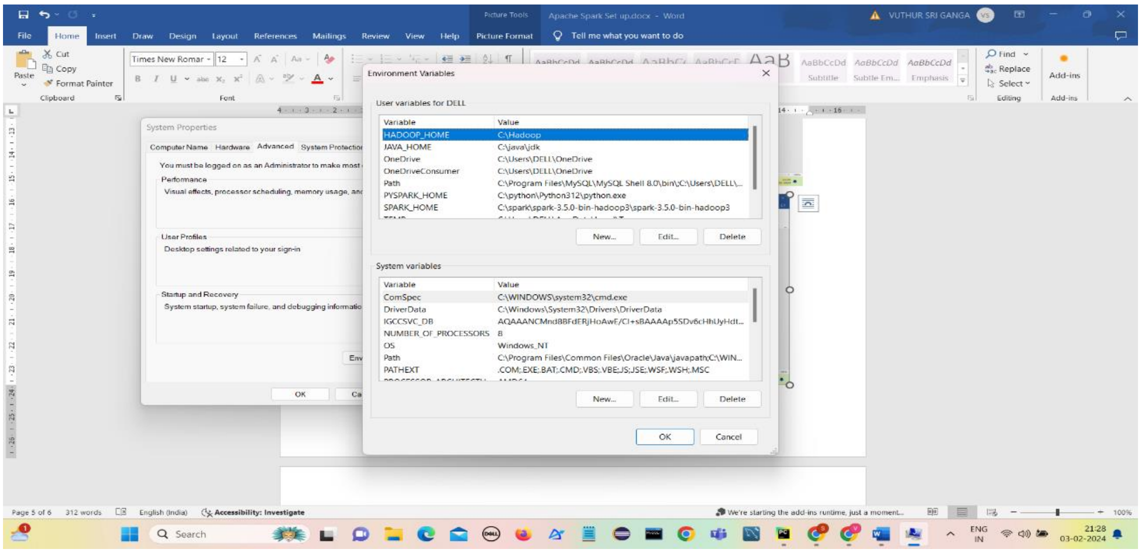
Environment Variables

User variables for DELL

Variable	Value
JAVA_HOME	C:\java\jdk
OneDrive	C:\Users\DELL\OneDrive
OneDriveConsumer	C:\Users\DELL\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\DELL\...
PYSPARK_HOME	C:\python\Python312\python.exe
SPARK_HOME	C:\spark\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-hadoop3
TEMP	C:\Users\DELL\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
IGCCSVC_DB	AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAAp5SDv6cHhUyHdt...
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC

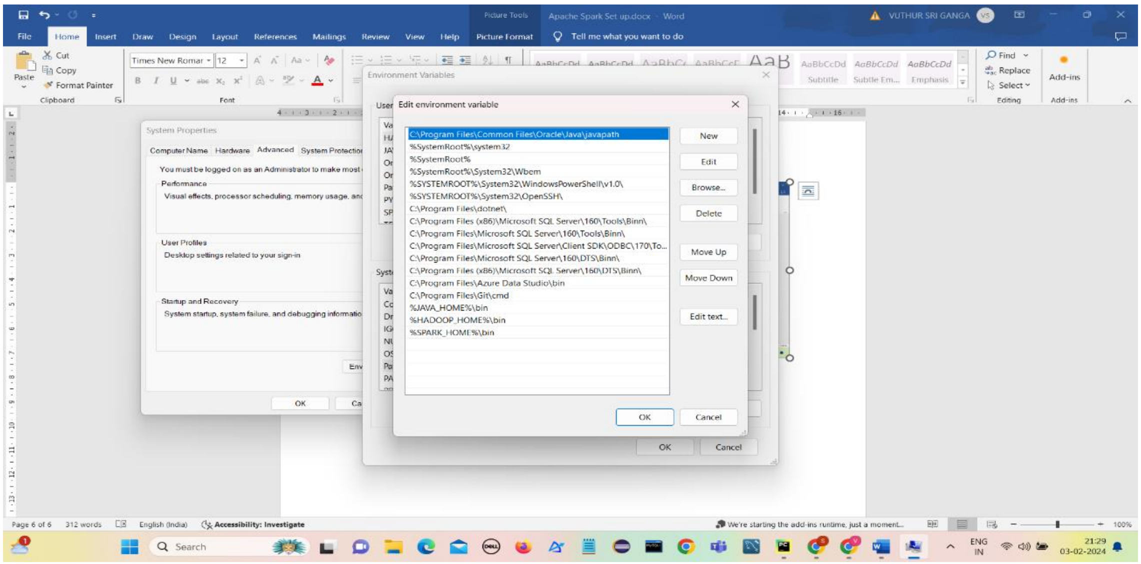


User variables for DELL

Variable	Value
HADOOP_HOME	C:\Hadoop
JAVA_HOME	C:\java\jdk
OneDrive	C:\Users\DELL\OneDrive
OneDriveConsumer	C:\Users\DELL\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\DELL\...
PYSPARK_HOME	C:\python\Python312\python.exe
SPARK_HOME	C:\spark\spark-3.5.0-bin-hadoop3\spark-3.5.0-bin-hadoop3

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
IGCCSVC_DB	AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAAp5SDv6cHhUyHdt...
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC

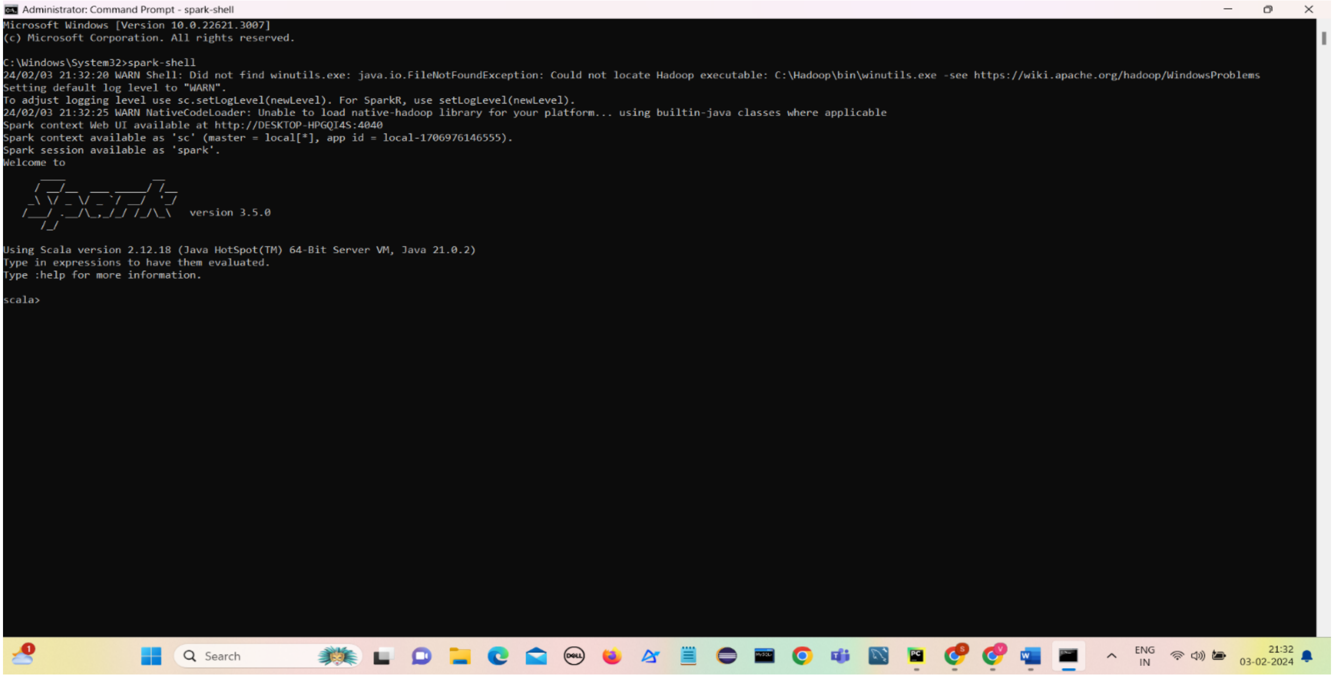


```
C:\Program Files\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files\dotnet\
C:\Program Files (x86)\Microsoft SQL Server\160\Tools\Binn\
C:\Program Files\Microsoft SQL Server\160\Tools\Binn\
C:\Program Files\Microsoft SQL Server\Client SDK\ODBC\170\To...
C:\Program Files\Microsoft SQL Server\160\DTS\Binn\
C:\Program Files (x86)\Microsoft SQL Server\160\DTS\Binn\
C:\Program Files\Azure Data Studio\bin
C:\Program Files\Git\cmd
%JAVA_HOME%\bin
%HADOOP_HOME%\bin
%SPARK_HOME%\bin
```

Click OK and exit.

To check whether spark is installed or not.

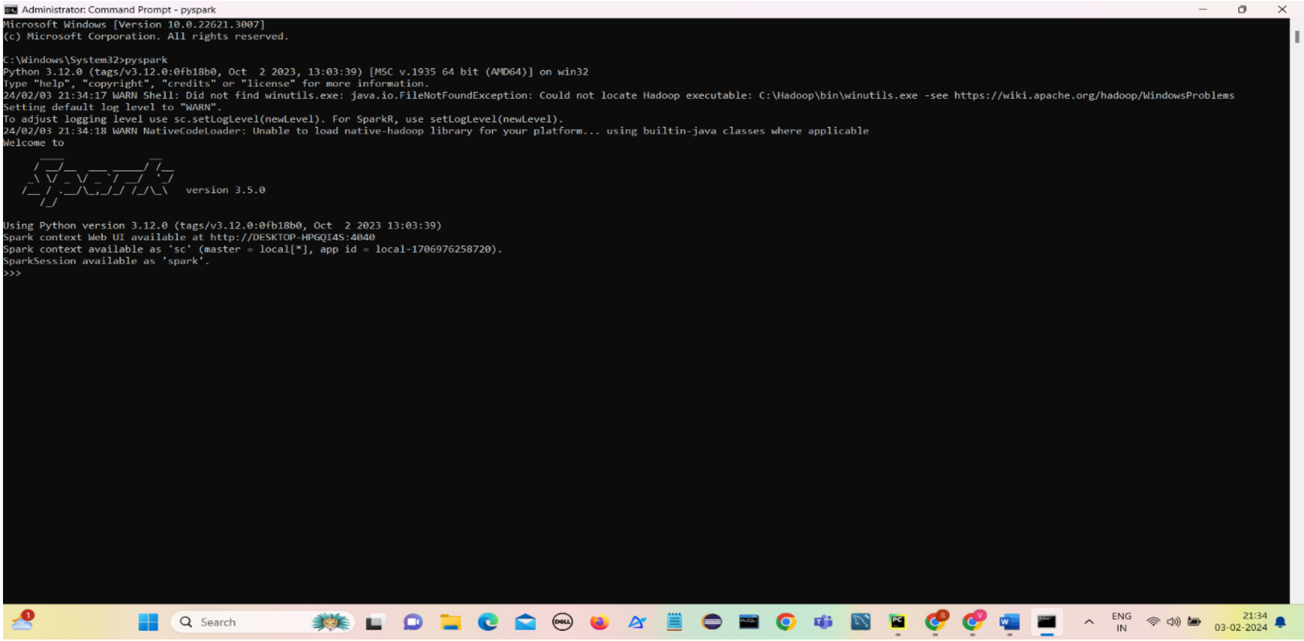
After setting up the environment variables,we need to save all of them and have to go to the command prompt and type **spark-shell** as below.



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

C:\Windows\System32>spark-shell
24/02/03 21:32:20 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Hadoop\bin\winutils.exe -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 21:32:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://DESKTOP-HPGQI4S:4040
Spark context available as 'sc' (master = local[*], app id = local-1706976146555).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Scala version 2.12.18 (Java HotSpot(TM) 64-Bit Server VM, Java 21.0.2)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

To check whether pyspark is installed or not.

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:\Windows\System32>pyspark
Python 3.12.0 (tags/v3.12.0:0fb18b0, Oct  2 2023, 13:03:39) [MSC v.1935 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
24/02/03 21:34:17 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Hadoop\bin\winutils.exe -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 21:34:18 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 3.12.0 (tags/v3.12.0:0fb18b0, Oct  2 2023 13:03:39)
Spark context Web UI available at http://DESKTOP-HPGQI4S:4040
Spark context available as 'sc' (master = local[*], app id = local-1706976258720).
SparkSession available as 'spark'.
>>>
```