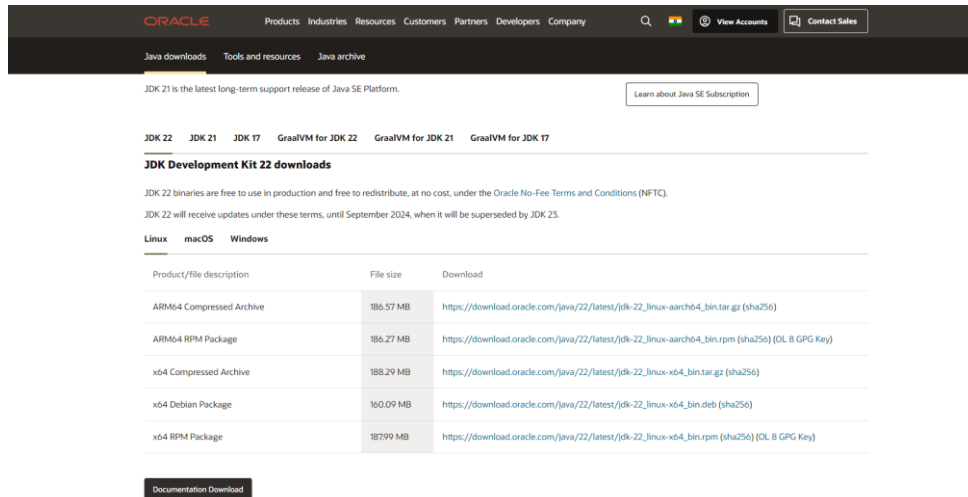


Spark Installation Guide (Windows)

1. Installation of Java and Python necessary:

- Java Download:
- <https://www.oracle.com/java/technologies/java-se-glance.html>



2. Check Python and Java Versions using Windows CMD:

- java -version
- python --version

```
C:\WINDOWS\system32\cmd. x + v
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Users\sachi>java -version
java version "17.0.6" 2023-01-17 LTS
Java(TM) SE Runtime Environment (build 17.0.6+9-LTS-190)
Java HotSpot(TM) 64-Bit Server VM (build 17.0.6+9-LTS-190, mixed mode, sharing)

C:\Users\sachi>python --version
Python 3.11.8

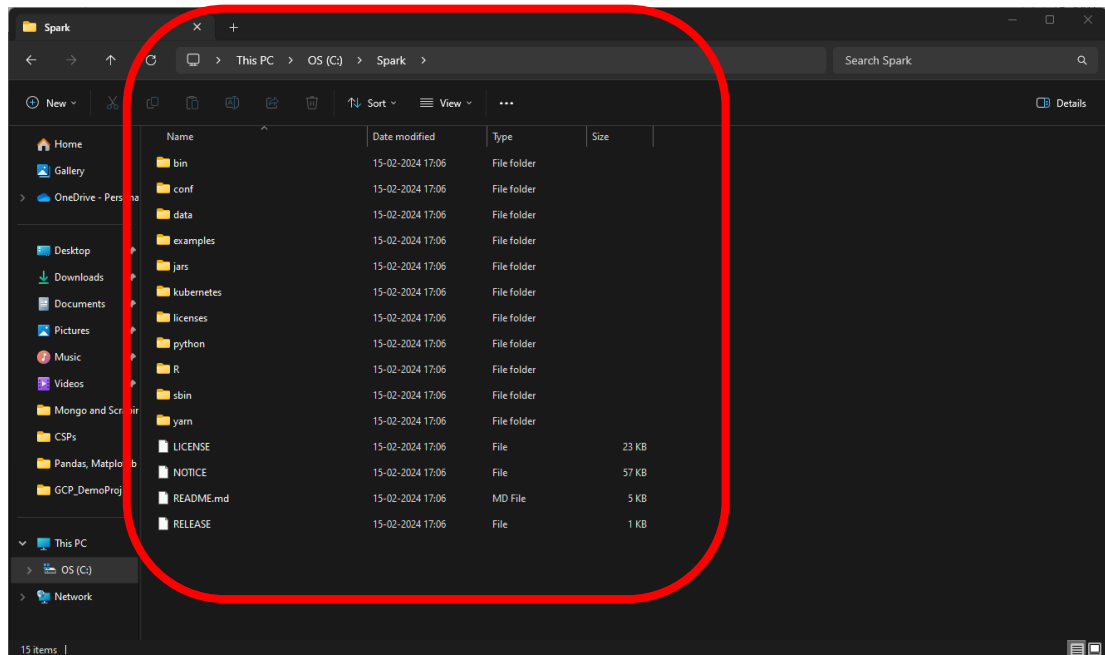
C:\Users\sachi>
```

3. Download Spark Files and copy it in C:/SPARK folder:

- <https://spark.apache.org/downloads.html>

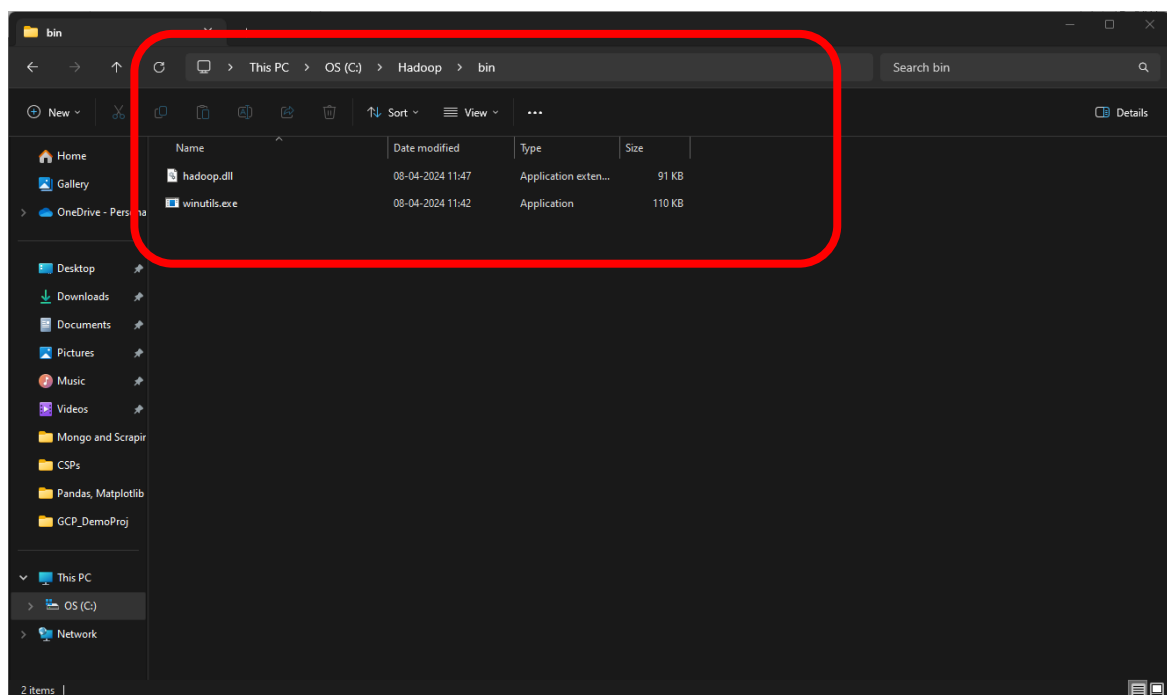
Alternatively, we can use direct download link:

- <https://d1cdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz>



4. Download WinUtils and copy it in C:/HADOOP/BIN folder:

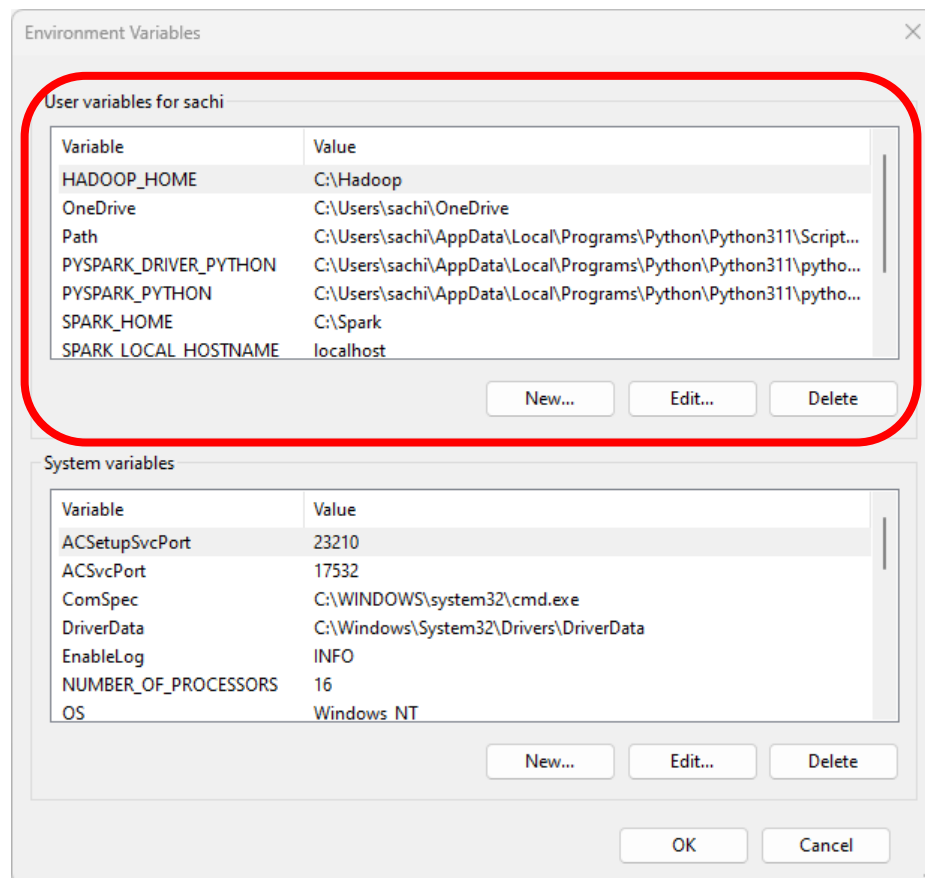
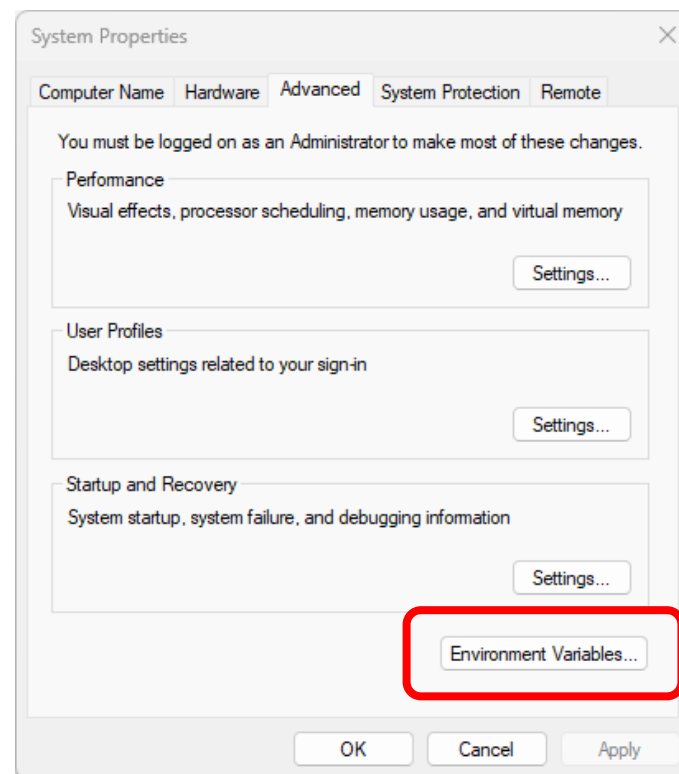
- <https://github.com/steveloughran/winutils/tree/master/hadoop-3.0.0/bin>
- <https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe>
- <https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/hadoop.dll>



5. Set the ENVIRONMENT PATH:

- SPARK_HOME = /**PATH**/
- HADOOP_HOME = /**PATH**/

- SPARK_LOCAL_HOSTNAME = localhost
- PYSARK_PYTHON = /**PYTHON EXECUTABLE**/
- PYSARK_DRIVER_PYTHON = /**PYTHON EXECUTABLE**/



Edit User Variable

Variable name: HADOOP_HOME

Variable value: C:\Hadoop

Browse Directory... Browse File... OK Cancel

Edit User Variable

Variable name: PYSPARK_DRIVER_PYTHON

Variable value: C:\Users\sachi\AppData\Local\Programs\Python\Python311\python.exe

Browse Directory... Browse File... OK Cancel

Edit User Variable

Variable name: PYSPARK_PYTHON

Variable value: C:\Users\sachi\AppData\Local\Programs\Python\Python311\python.exe

Browse Directory... Browse File... OK Cancel

Edit User Variable

Variable name: SPARK_HOME

Variable value: C:\Spark

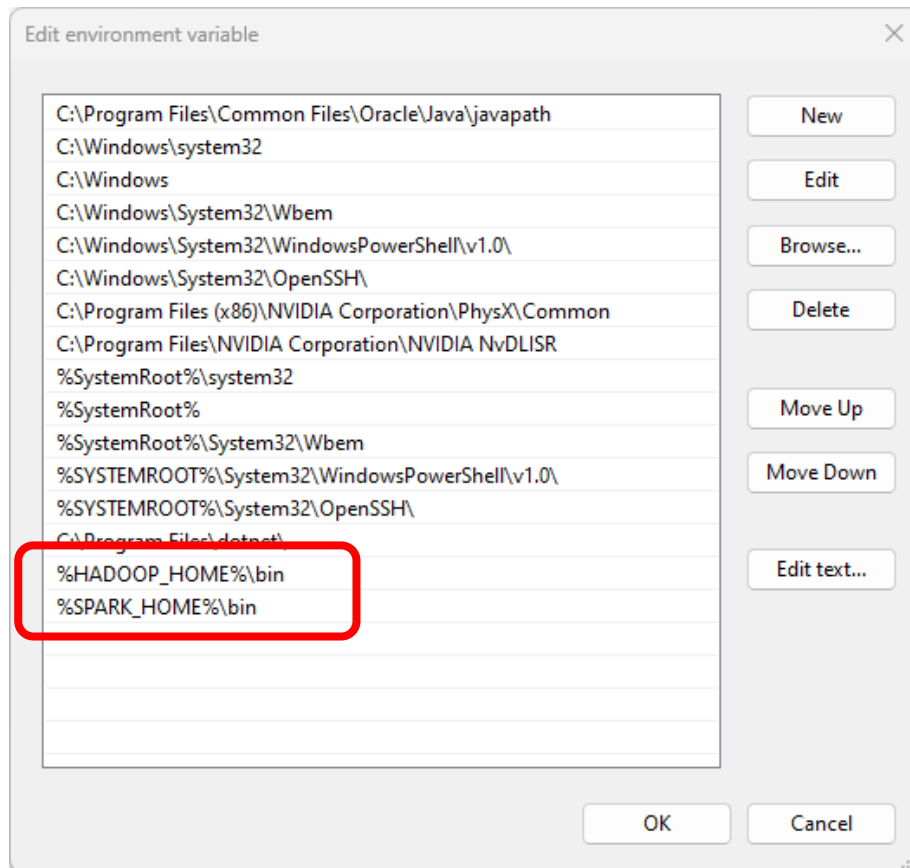
Browse Directory... Browse File... OK Cancel

Edit User Variable

Variable name: SPARK_LOCAL_HOSTNAME

Variable value: localhost

Browse Directory... Browse File... OK Cancel



6. Run Command pyspark in cmd terminal to check if Spark is loaded

```
C:\WINDOWS\system32\cmd. X + -
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Users\sachi>pyspark
Python 3.11.8 (tags/v3.11.8:db85d51, Feb 6 2024, 22:03:32) [MSC v.1937 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).
24/04/08 15:57:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
24/04/08 15:57:46 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Welcome to

  ____      _
 / ___|  __| | | |
 \___ \  | | | | | |
  ___) | | | | | | |
 |____|_|_|_|_|_|_|_|

version 3.5.1

Using Python version 3.11.8 (tags/v3.11.8:db85d51, Feb 6 2024 22:03:32)
Spark context Web UI available at http://localhost:4041
Spark context available as 'sc' (master = local[*], app id = local-1712572066704).
SparkSession available as 'spark'.
>>>
```