PYSPARK INSTALLATION - ON WINDOWS

Compatibility check: Please ensure the s/w compatibility with other softwares, you can check the compatibility on this following link:

https://community.cloudera.com/t5/Community-Articles/Spark-Python-Supportability-Matrix/ta-p/379144

Necessary s/w: For spark installation we need this following s/w:
Spark, Python, Hadoop, Java

Download links:

Java 17 - https://download.oracle.com/java/17/archive/jdk-17.0.12_windows-x64_bin.msi

Python 3.10 - https://www.python.org/ftp/python/3.10.0/python-3.10.0-amd64.exe

Spark 3.4.0 - https://archive.apache.org/dist/spark/spark-3.4.0/spark-3.4.0-bin-hadoop3.tgz

Hadoop - https://codeload.github.com/kontext-tech/winutils/zip/refs/heads/master

Note: We don't have to use the latest version of s/w, choose the ones that are compatible with each other. So choose the above links with these, you can continue with installation.

Installation:

- Install java and python, open cmd(command prompt) type java —version : will return the version of java that is installed in your system

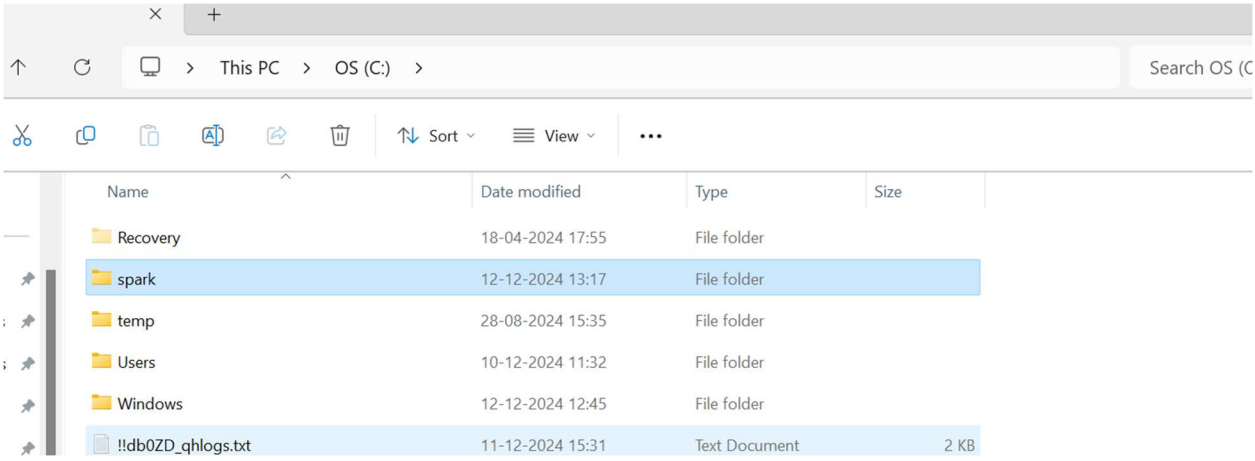
```
C:\Users\gaurav.kumar>java --version
java 17.0.12 2024-07-16 LTS
Java(TM) SE Runtime Environment (build 17.0.12+8-LTS-286)
Java HotSpot(TM) 64-Bit Server VM (build 17.0.12+8-LTS-286, mixed mode, sharing)
```

Do the same with python, type python –version to check the version of python installed.

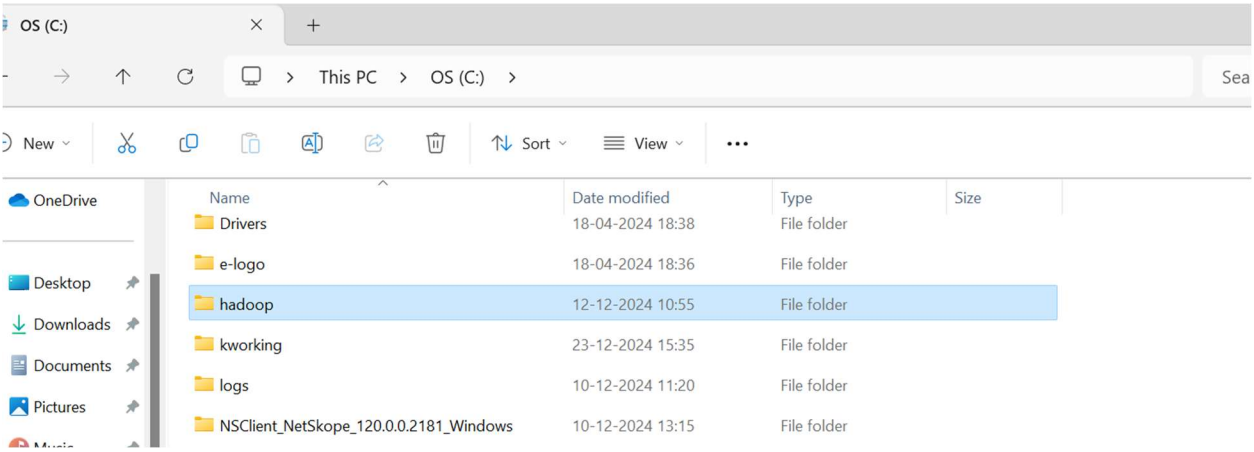
```
C:\Users\gaurav.kumar>python --version
Python 3.10.0
```

- As you will notice that spark and hadoop files are downloaded as compressed files. So we need to extract them.
  Spark file name - spark-3.4.0-bin-hadoop3.tgz
  Hadoop file name - winutils-master.zip
- Now go to the c drive and create folders named spark and hadoop.

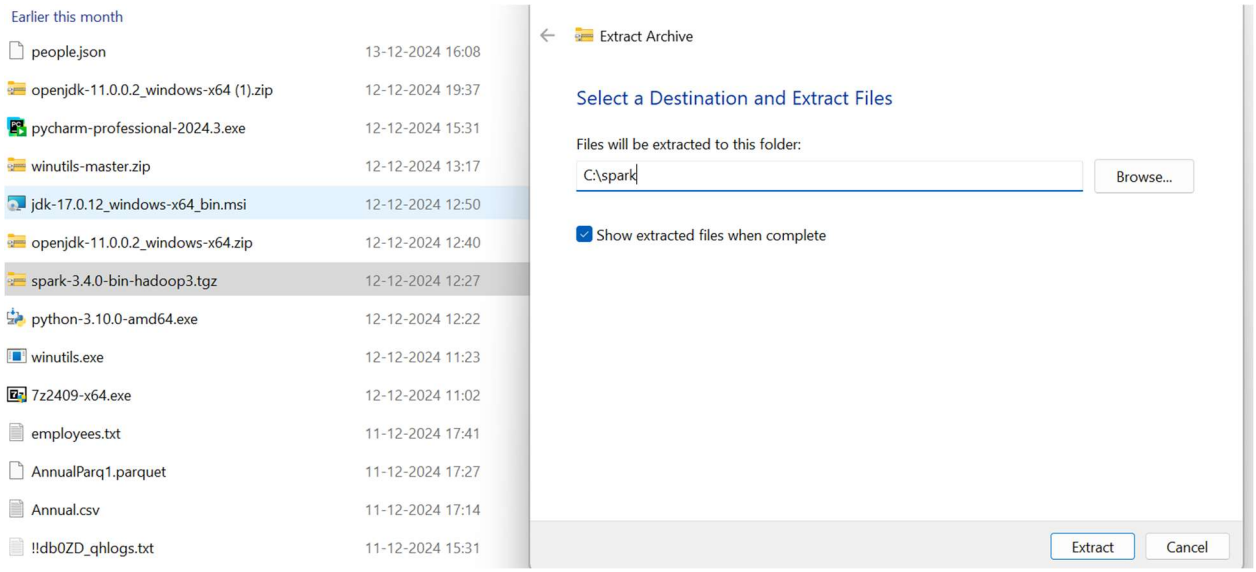
Spark



Hadoop



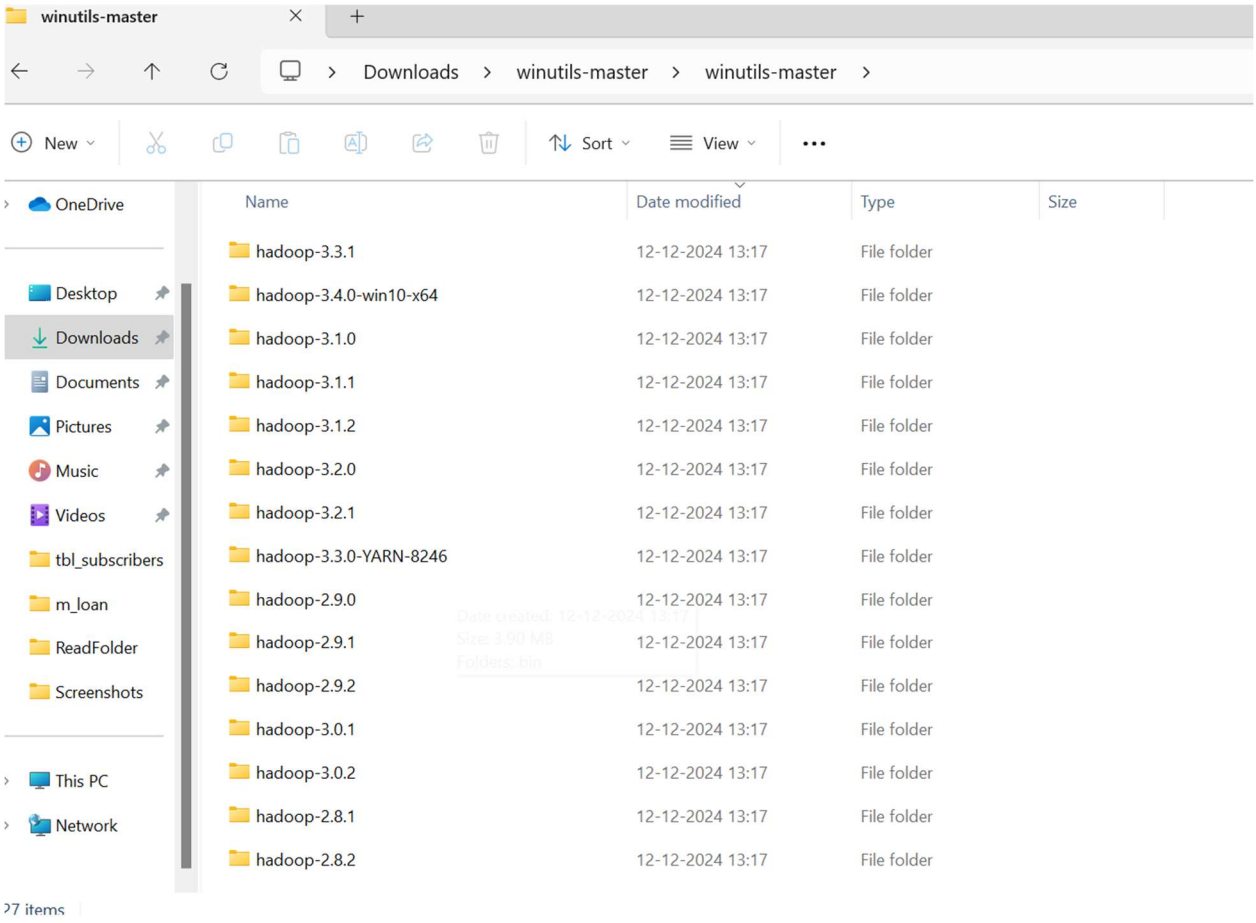
- Now extract the file name spark-3.4.0-bin-hadoop3.tgz in the spark folder.



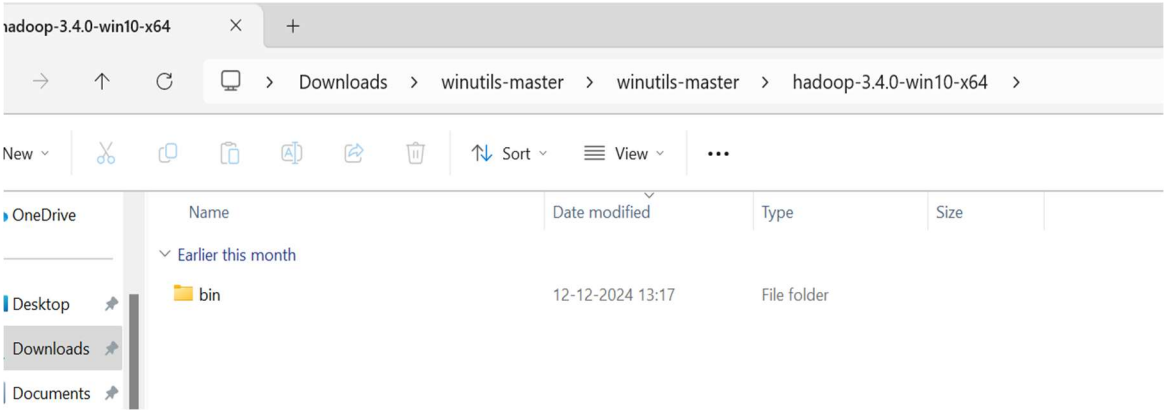
Now extract the winutils-master.zip

- It will create a folder named winutils-master, shown above

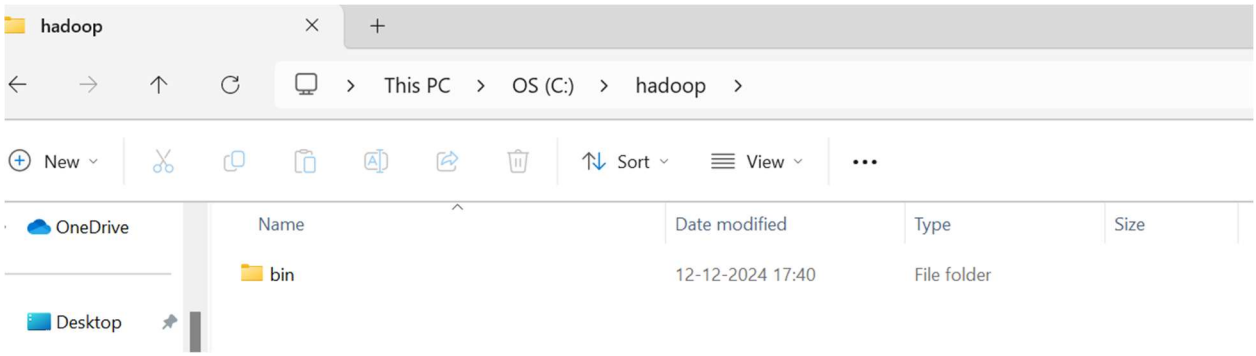
- Now open the winutils-master folder, it will show you a list of directories.



- Now select the folder named hadoop-3.4.0-win10-x64.



- Copy this bin folder and paste it inside the hadoop folder that we created in c drive.



After this we need to edit the system variables.

Edit System environment variables:

- Create variables SPARK_HOME, PYTHON_HOME, HADOOP_HOME inside the system variable.

  Assign Paths:

  SPARK_HOME

  C:\spark\spark-3.4.0-bin-hadoop3

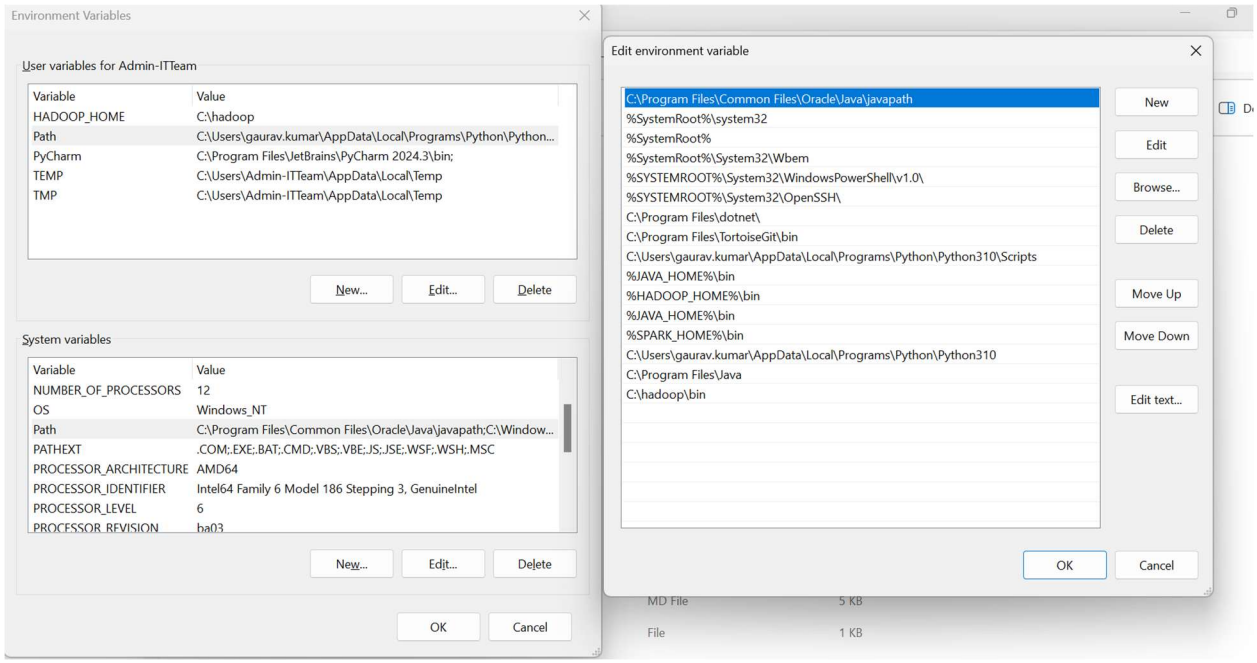
  PYTHON_HOME

  C:\Users\user_name\AppData\Local\Programs\Python\Python310\python.exe

  HADOOP_HOME

  C:\hadoop

- Now change the system path variable.



Please follow the above screenshots and add every mentioned variables in system variable named ‘Path’

- After this, open(restart it if it is open) the cmd(command prompt) and type spark.

```
C:\Users\gaurav.kumar>pyspark
Python 3.10.0 (tags/v3.10.0:b494f59, Oct  4 2021, 19:00:18) [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).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Python version 3.10.0 (tags/v3.10.0:b494f59, Oct  4 2021 19:00:18)
Spark context Web UI available at http://ES-PUN-LT-1414.swtpl.com:4040
Spark context available as 'sc' (master = local[*], app id = local-1735031875007).
SparkSession available as 'spark'.
>>>
```

- If this does not work, please consult your colleagues.

The End