Akilesh K

k.akilesh123@gmail.com

Data engineering - Batch 1

Date: 03-02-24

DAY 11 - PYSPARK INSTALLATION

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

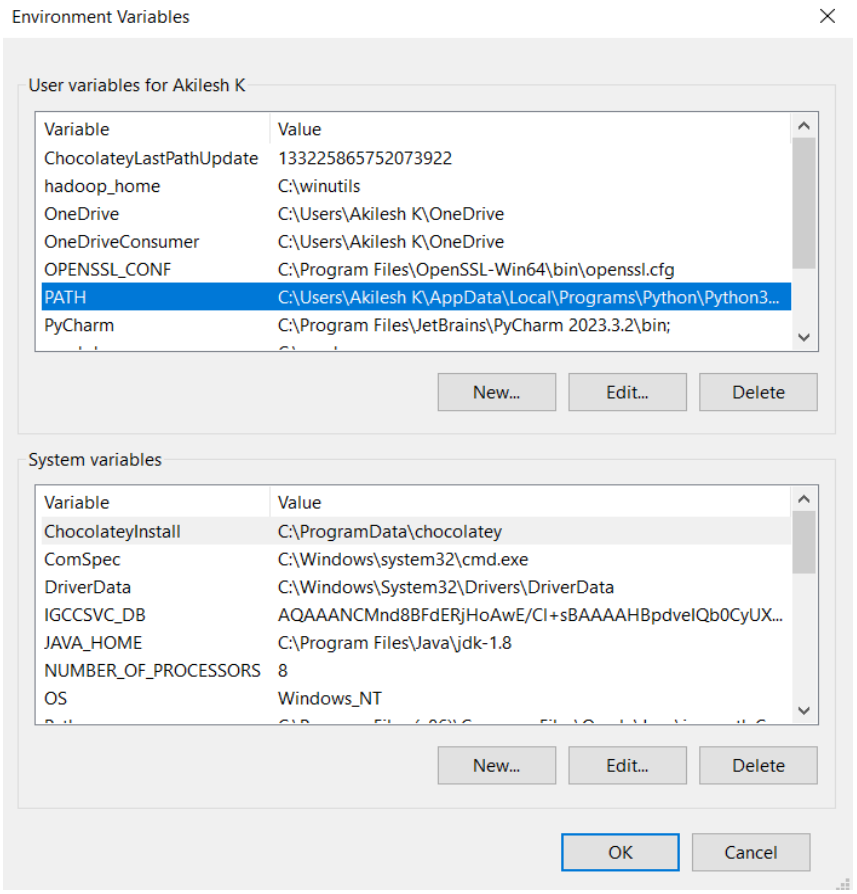
Java needs to be installed

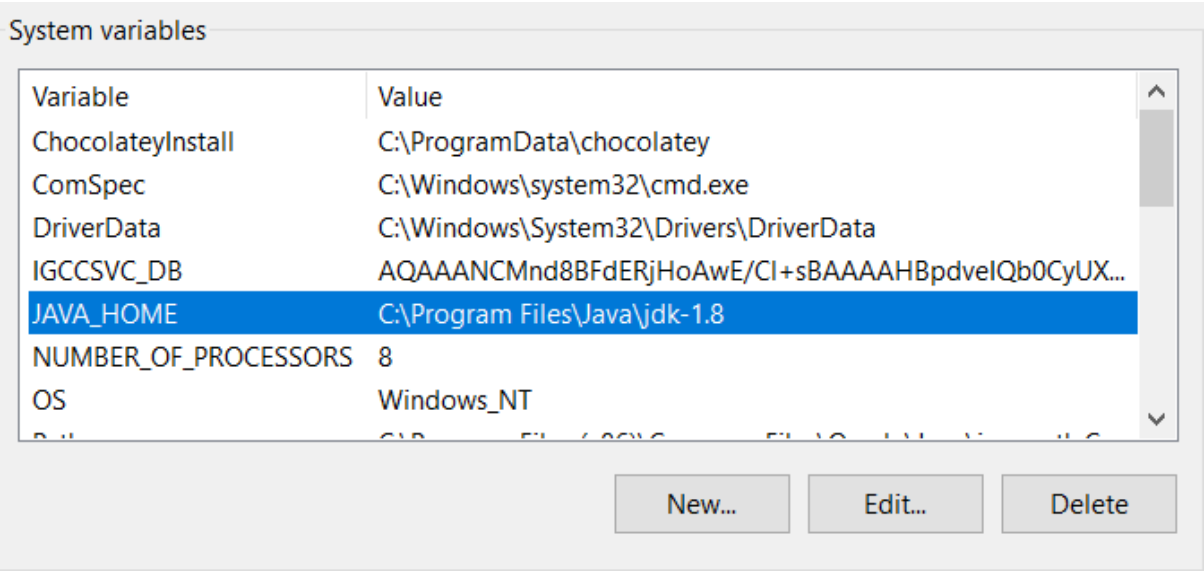
```
C:\Users\Akilesh K>java -version
java version "1.8.0_401"
Java(TM) SE Runtime Environment (build 1.8.0_401-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.401-b10, mixed mode)
```

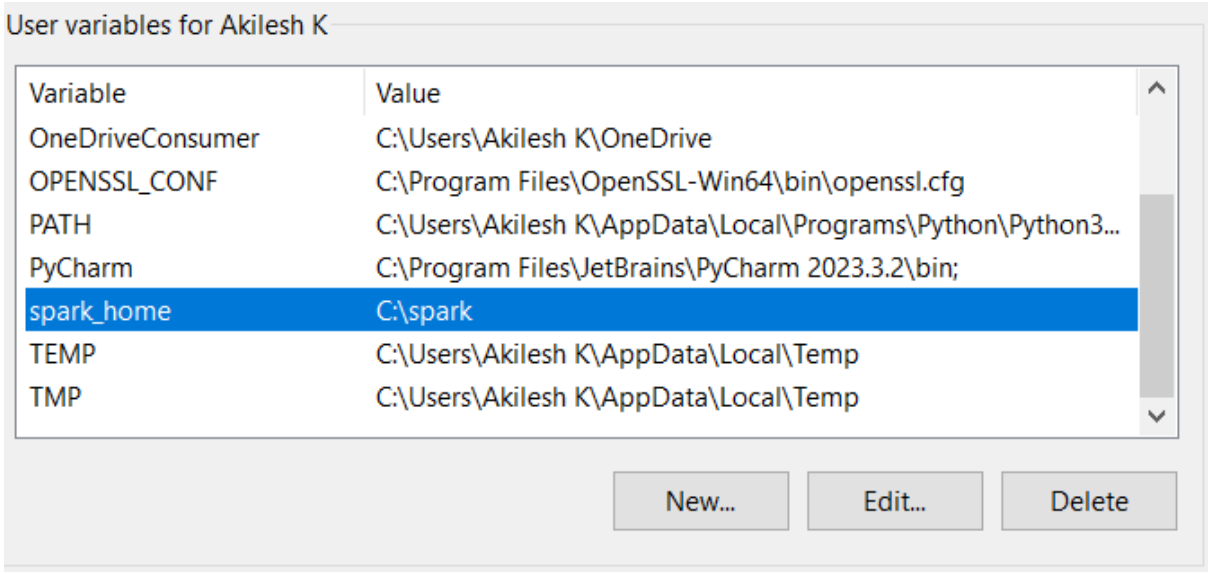
Python is installed

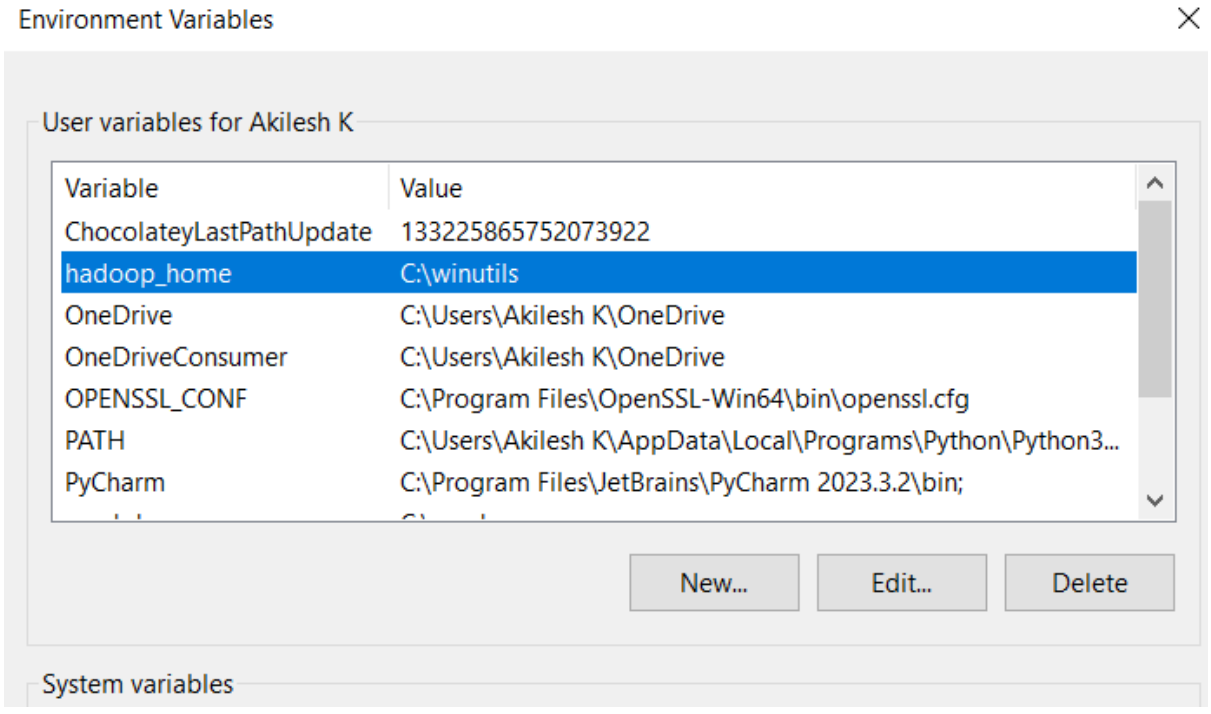
```
C:\Users\Akilesh K>python --version
Python 3.12.1
```

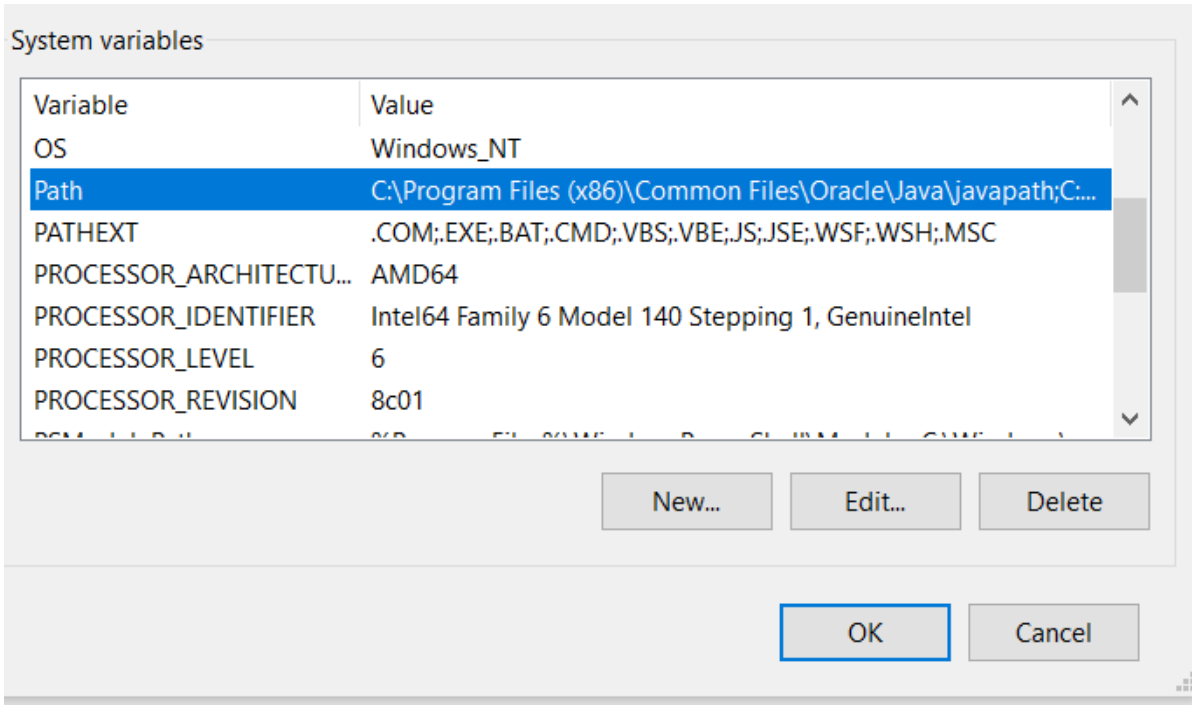
Path variables needs to be mentioned

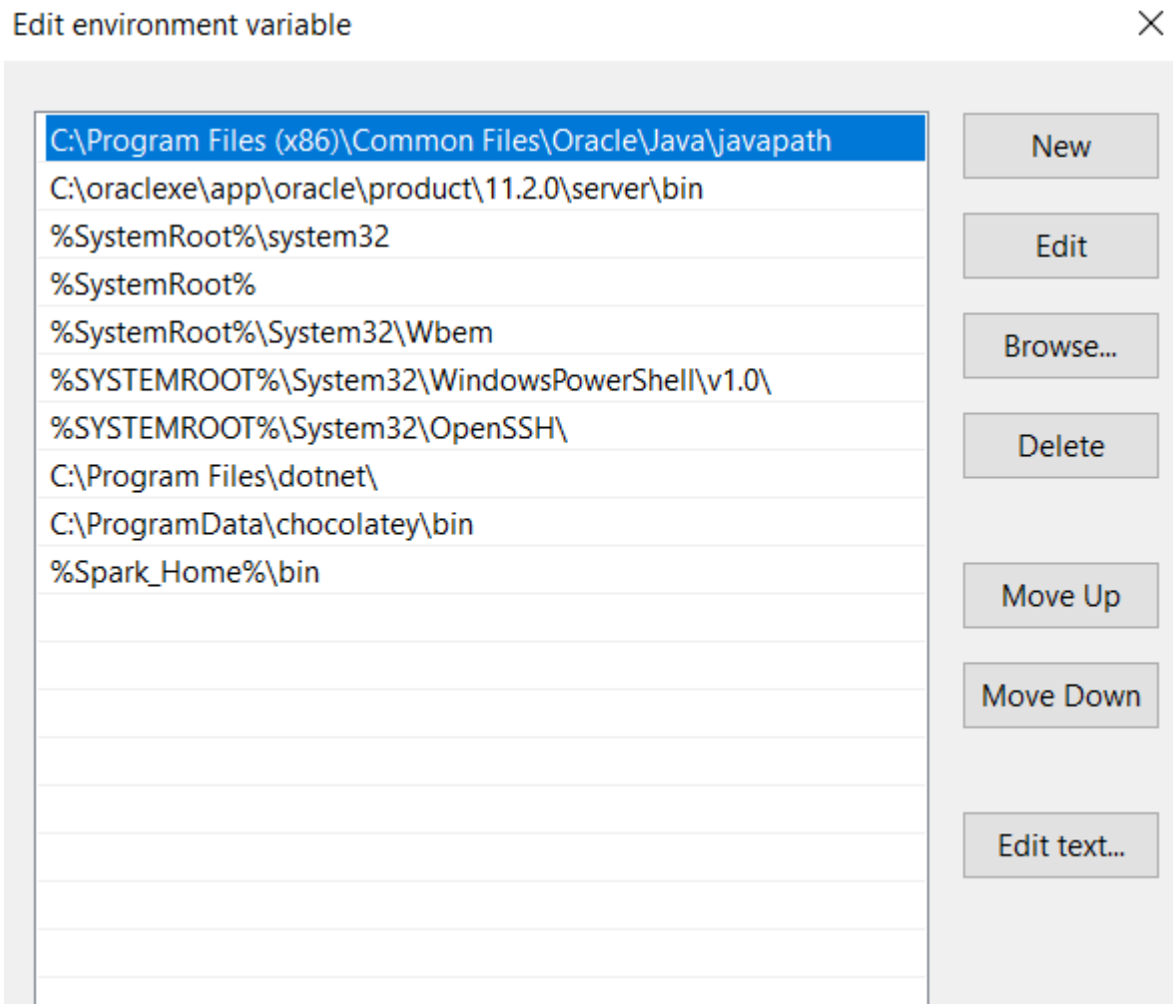





User variables for Akilesh K
Variable
Value
OneDriveConsumer
C:\Users\Akilesh K\OneDrive
OPENSSL_CONF
C:\Program Files\OpenSSL-Win64\bin\openssl.cfg
PATH
C:\Users\Akilesh K\AppData\Local\Programs\Python\Python3...
PyCharm
C:\Program Files\JetBrains\PyCharm 2023.3.2\bin;
spark_home
C:\spark
TEMP
C:\Users\Akilesh K\AppData\Local\Temp
TMP
C:\Users\Akilesh K\AppData\Local\Temp
New...
Edit...
Delete


Environment Variables
User variables for Akilesh K
Variable
Value
ChocolateyLastPathUpdate
133225865752073922
hadoop_home
C:\winutils
OneDrive
C:\Users\Akilesh K\OneDrive
OneDriveConsumer
C:\Users\Akilesh K\OneDrive
OPENSSL_CONF
C:\Program Files\OpenSSL-Win64\bin\openssl.cfg
PATH
C:\Users\Akilesh K\AppData\Local\Programs\Python\Python3...
PyCharm
C:\Program Files\JetBrains\PyCharm 2023.3.2\bin;
New...
Edit...
Delete
System variables

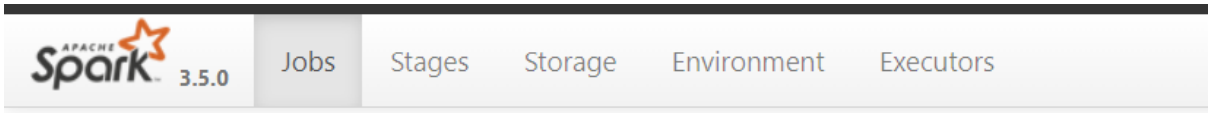

System variables
Variable
Value
OS
Windows_NT
Path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTU...
AMD64
PROCESSOR_IDENTIFIER
Intel64 Family 6 Model 140 Stepping 1, GenuineIntel
PROCESSOR_LEVEL
6
PROCESSOR_REVISION
8c01
New...
Edit...
Delete
OK
Cancel



Spark running successfully

```
C:\Users\Akilesh K>pyspark
Python 3.12.1 (tags/v3.12.1:2305ca5, Dec  7 2023, 22:03:25) [MSC v.1937 64 bit (AMD64)] on
Type "help", "copyright", "credits" or "license" for more information.
24/02/04 19:38:27 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: Hac
ist: C:\winutils\bin -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)
24/02/04 19:38:28 WARN NativeCodeLoader: Unable to load native-hadoop library for your plat
asses where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Python version 3.12.1 (tags/v3.12.1:2305ca5, Dec  7 2023 22:03:25)
Spark context Web UI available at http://DESKTOP-A98NUHP:4040
Spark context available as 'sc' (master = local[*], app id = local-1707055710044).
SparkSession available as 'spark'.
```



Spark Jobs (?)

User: Akilesh K
Total Uptime: 35 s
Scheduling Mode: FIFO

▸ Event Timeline