Lab3. Apache Spark 환경설정하기

1. Python 환경설정

1)Python Version 확인

-Windows

```
> python --version
```

-macOS

```
$ python3 --version
```

2)Python 3.11 Installation

-각 OS 별 설치 참조.

-Linux의 경우

https://github.com/swacademy/SuwonUniv_DevOps/3주차/Lab3.Installation Jupyter Notebook on Ubuntu.pdf 참조

3)Python 3.11을 Python 기본 버전으로 지정

-macOS

```
$ ls -l /usr/local/bin/python*
$ which python3.11
    /usr/local/bin/python3.11
$ ln -s -f /usr/local/bin/python3.11 /usr/local/bin/python
```

-Windows

--설치시 "Add python.exe to PATH"를 체크했다면 자동으로 운영체제 PATH의 제일 앞에 위치하기 때문에 자동으로 기본 버전으로 셋업됨.

4)Python Version 확인

-Windows

```
> python3 --version
Python 3.11.2
```

-macOS

```
$ python3 --version
or
$ python --version
```

2. DataLake Python 가상환경 구성하기

1)pip3(pip) Version 확인

-Windows

```
> pip --version
pip 22.3.1 from C:\Program Files\Python311\Lib\site-packages\pip (python 3.11)
```

-macOS

```
$ pip3 --version
```

2)virtualenv 설치

```
$ pip3 install virtualenv      <---Windows도 가능
Collecting virtualenv
    Downloading virtualenv-20.21.0-py3-none-any.whl (8.7 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.7/8.7 MB 31.0 MB/s eta 0:00:00
Collecting distlib<1,>=0.3.6
    Downloading distlib-0.3.6-py2.py3-none-any.whl (468 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 468.5/468.5 kB ? eta 0:00:00
Collecting filelock<4,>=3.4.1
    Downloading filelock-3.10.0-py3-none-any.whl (9.9 kB)
Collecting platformdirs<4,>=2.4
    Downloading platformdirs-3.1.1-py3-none-any.whl (14 kB)
Installing collected packages: distlib, platformdirs, filelock, virtualenv
Successfully installed distlib-0.3.6 filelock-3.10.0 platformdirs-3.1.1 virtualenv-20.21.0

[notice] A new release of pip available: 22.3.1 -> 23.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

3)virtualenv PATH에 추가하기(Windows에서는 안함)

```
$ echo $PATH
$ export PATH=$PATH:/Library/Frameworks/Python.framework/Versions/3.11/bin
```

4)DataLake 가상환경 구성 및 해당 가상환경으로 이동하기

-Windows

--C:/로 이동

```
> cd C:/
```

--가상환경 설치

```
C:\>virtualenv DataLake
created virtual environment CPython3.11.2.final.0-64 in 4322ms
    creator CPython3Windows(dest=D:\DataLake, clear=False, no_vcs_ignore=False, global=False)
seeder FromAppData(download=False, pip=bundle, setuptools=bundle, wheel=bundle, via=copy,
app_data_dir=C:\Users\MZC-USER\AppData\Local\pypa\virtualenv)
        added seed packages: pip==23.0.1, setuptools==67.4.0, wheel==0.38.4
```

```
            activators BashActivator,BatchActivator,FishActivator,NushellActivator,PowerShellActivator,PythonActivator

        --가상환경으로 이동
            C:\>cd DataLake/Scripts
            C:\DataLake\Scripts>activate.bat
            (DataLake) C:\DataLake\Scripts>
```

-macOS

```
    $ virtualenv DataLake
    $ source DataLake/bin/activate
```

3. DataLake 가상환경에 pyspark 설치하기

1)pip3 list 확인

```
    (DataLake) ... $ pip3 list
    Package    Version
    ---------- -------
    pip        23.0.1
    setuptools 67.4.0
    wheel      0.38.4
```

2)pyspark 설치

```
    (DataLake) ... $ pip3 install pyspark
    (DataLake) C:\DataLake\Scripts>pip3 install pyspark
    Collecting pyspark
      Downloading pyspark-3.3.2.tar.gz (281.4 MB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 281.4/281.4 MB 1.8 MB/s eta 0:00:00
      Preparing metadata (setup.py) ... done
    Collecting py4j==0.10.9.5
      Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
         ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 199.7/199.7 kB 6.1 MB/s eta 0:00:00
    Building wheels for collected packages: pyspark
      Building wheel for pyspark (setup.py) ... done
      Created wheel for pyspark: filename=pyspark-3.3.2-py2.py3-none-any.whl size=281824045
      sha256=d6b9e54867e6f72b516e89d6e9a740e6da69326a80f56e6fc6fd63ae8a39a0e9
      Stored in directory:
      c:\users\mzc-user\appdata\local\pip\cache\wheels\47\69\84\c7c7776e2287a654536f5cba7dc54c904c03aa2c3e29206f0f
    Successfully built pyspark
    Installing collected packages: py4j, pyspark
    Successfully installed py4j-0.10.9.5 pyspark-3.3.2
```

4. Visual Studio Code에서 테스트

1)Visual Studio Code 런치 후

2)[Extensions : MARKETPLACE]에서 Python Extension Pack 설치

3)pyspark-test.py 새 파일 생성

4)다음과 같이 코드 입력 후 pyspark 테스트

```
    import pyspark
    from pyspark.sql import SparkSession

    conf = pyspark.SparkConf()
    conf.set('spark.driver.host', '127.0.0.1')

    spark = SparkSession.builder \
        .config(conf=conf) \
        .appName('Pyspark Test Program') \
        .getOrCreate()

    print(f'Hadoop Version = {spark._jvm.org.apache.hadoop.util.VersionInfo.getVersion()}')
```

5. Windows에서 위의 코드를 실행하면 다음과 같은 오류가 발생할 수 있다.

```
PS C:\DataLake> & C:/DataLake/Scripts/Activate.ps1 & : 이 시스템에서 스크립트를 실행할 수 없으므로 C:\DataLake\Scripts\Activate.ps1
파일을 로드할 수 없습니다. 자세한 내용은 about_Execution_Policies(https://go.microsoft.com/fwlink/?LinkID=135170)를 참조하십시오.
위치 줄:1 문자:3
+ & d:/DataLake/Scripts/Activate.ps1
+   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
    + CategoryInfo          : 보안 오류: (:) [], PS
Sec    urityException
    + FullyQualifiedErrorId : UnauthorizedAccess

PS C:\DataLake> & C:/DataLake/Scripts/python.exe C:/DataLake/pyspartk-test.py
Java not found and JAVA_HOME environment variable is not set.
Install Java and set JAVA_HOME to point to the Java installation directory.
Traceback (most recent call last):
  File "d:\DataLake\pyspartk-test.py", line 10, in <module>
    .getOrCreate()
     ^^^^^^^^^^^^^
```

```
  File "C:\DataLake\Lib\site-packages\pyspark\sql\session.py", line 269, in getOrCreate
    sc = SparkContext.getOrCreate(sparkConf)
         ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "C:\DataLake\Lib\site-packages\pyspark\context.py", line 483, in getOrCreate
    SparkContext(conf=conf or SparkConf())
  File "C:\DataLake\Lib\site-packages\pyspark\context.py", line 195, in __init__
    SparkContext._ensure_initialized(self, gateway=gateway, conf=conf)
  File "C:\DataLake\Lib\site-packages\pyspark\context.py", line 417, in _ensure_initialized
    SparkContext._gateway = gateway or launch_gateway(conf)
  File "C:\DataLake\Lib\site-packages\pyspark\java_gateway.py", line 106, in launch_gateway    raise RuntimeError("Java gateway process exited before sending its port number")RuntimeError: Java gateway process exited before sending its port number
```

1)Powershell 문제 해결

- -Powershell을 관리자 권한으로 실행
    - --시작 > Windows Powershell > 관리자 권한으로 실행

```
PS C:\....> Set-ExecutionPolicy Unrestricted
```

[실행 규칙 변경]창

"실행 정책은 신뢰하지 않는 스크립트로부터 사용자를 보호합니다. 실행 정책을 변경하면 about_Execution_Policies 도움말 항목(https://go.microsoft.com/fwlink/?LinkID=135170)에 설명된 보안 위험에 노출될 수 있습니다. 실행 정책을 변경하시겠습니까?"

--> [예] 클릭

2)How to Setup and Use Pyspark in Python (Windows 10)

-원인

Error: Exception: Java gateway process exited before sending the driver its port number

-설치 순서

a.pip install pyspark   <---이미 설치했음

b.pip install sparksql-magic

```
(DataLake) C:\DataLake>pip install sparksql-magic
Collecting sparksql-magic
  Downloading sparksql_magic-0.0.3-py36-none-any.whl (4.3 kB)
Requirement already satisfied: pyspark>=2.3.0 in d:\datalake\lib\site-packages (from sparksql-magic) (3.3.2)
Collecting ipython>=7.4.0
  Downloading ipython-8.11.0-py3-none-any.whl (793 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 793.3/793.3 kB 10.0 MB/s eta 0:00:00
Collecting backcall
  Downloading backcall-0.2.0-py2.py3-none-any.whl (11 kB)
Collecting decorator
  Downloading decorator-5.1.1-py3-none-any.whl (9.1 kB)
Collecting jedi>=0.16
  Downloading jedi-0.18.2-py2.py3-none-any.whl (1.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 50.3 MB/s eta 0:00:00
Collecting matplotlib-inline
  Downloading matplotlib_inline-0.1.6-py3-none-any.whl (9.4 kB)
Collecting pickleshare
  Downloading pickleshare-0.7.5-py2.py3-none-any.whl (6.9 kB)
Collecting prompt-toolkit!=3.0.37,<3.1.0,>=3.0.30
  Downloading prompt_toolkit-3.0.38-py3-none-any.whl (385 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 385.8/385.8 kB ? eta 0:00:00
Collecting pygments>=2.4.0
  Downloading Pygments-2.14.0-py3-none-any.whl (1.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.1/1.1 MB 35.9 MB/s eta 0:00:00
Collecting stack-data
  Downloading stack_data-0.6.2-py3-none-any.whl (24 kB)
Collecting traitlets>=5
  Downloading traitlets-5.9.0-py3-none-any.whl (117 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 117.4/117.4 kB ? eta 0:00:00
Collecting colorama
  Downloading colorama-0.4.6-py2.py3-none-any.whl (25 kB)
Requirement already satisfied: py4j==0.10.9.5 in d:\datalake\lib\site-packages (from pyspark>=2.3.0->sparksql-magic) (0.10.9.5)
Collecting parso<0.9.0,>=0.8.0
  Downloading parso-0.8.3-py2.py3-none-any.whl (100 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100.8/100.8 kB ? eta 0:00:00
Collecting wcwidth
  Downloading wcwidth-0.2.6-py2.py3-none-any.whl (29 kB)
Collecting executing>=1.2.0
  Downloading executing-1.2.0-py2.py3-none-any.whl (24 kB)
Collecting asttokens>=2.1.0
  Downloading asttokens-2.2.1-py2.py3-none-any.whl (26 kB)
Collecting pure-eval
  Downloading pure_eval-0.2.2-py3-none-any.whl (11 kB)
Collecting six
  Downloading six-1.16.0-py2.py3-none-any.whl (11 kB)
Installing collected packages: wcwidth, pure-eval, pickleshare, executing, backcall, traitlets, six, pygments, prompt-toolkit, parso, decorator, colorama, matplotlib-inline, jedi, asttokens, stack-data, ipython, sparksql-magic
Successfully installed asttokens-2.2.1 backcall-0.2.0 colorama-0.4.6 decorator-5.1.1 executing-1.2.0 ipython-8.11.0 jedi-0.18.2 matplotlib-inline-0.1.6 parso-0.8.3 pickleshare-0.7.5 prompt-toolkit-3.0.38 pure-eval-0.2.2 pygments-2.14.0 six-1.16.0 sparksql-magic-0.0.3 stack-data-0.6.2 traitlets-5.9.0 wcwidth-0.2.6
```

c.install Java
※JAVA_HOME문제
Spark 홈페이지에 다음과 같이 Note가 있다.
Note that PySpark requires Java 8 or later with JAVA_HOME properly set. If using JDK 11, set -Dio.netty.tryReflectionSetAccessible=true for Arrow related features and refer to Downloading.
-Googling "download java for windows"
--https://www.oracle.com/kr/java/technologies/downloads/ 방문
--[Java SE Development Kit 17.0.6 downloads]의 Windows 버전 Download
--x64 MSI Installer(https://download.oracle.com/java/17/latest/jdk-17_windows-x64_bin.msi (sha256)))
--Download받은 jdk-17_windows-x64_bin.msi 설치
-JAVA_HOME 설정하기
--Windows 환경변수로 이동하여
---Windows Key + R <--- 실행창을 띄운 후,
---SystemPropertiesAdvanced > 확인버튼 클릭 <---시스템속성 고급탭으로 이동
---[시작 및 복구] > [환경 변수]
--[시스템 변수] > [새로 만들기]
---변수 이름 : JAVA_HOME
---변수 값 : C:\Program Files\Java\jdk-17
---[확인] 버튼 클릭
--[시스템 변수] 목록에서 "Path"를 찾아서 선택한 후, [편집] 버튼 클릭
---[환경 변수 편집] 팝업창에서 [새로 만들기] 클릭
---C:\Program Files\Java\jdk-17\bin 넣고 엔터
---[확인] 버튼 클릭
---[확인] 버튼을 클릭하여 [환경 변수] 팝업창 닫기
--[확인] 버튼 클릭하여 [시스템 속성]창 닫기
--Command 창을 관리자 권한으로 오픈하여 다음 명령어 실행

```
C:\Users\{User}>set JAVA_HOME
C:\Users\{User}>set PATH
```

d. Install latest Apache Spark
-http://spark.apache.org/downloads.html 이동
-[Download Apache Spark] 페이지에서
1. Choose a Spark release:3.3.2(Feb 17 2023)
2. Choose a packge type: Pre-built for Apache Hadoop 3.3 and later
3. Download Spark: spark-3.3.2-bin-hadoop3.tgz

-다운로드 받은 spark-3.3.2-bin-hadoop3.tgz의 압축을 풀고
-spark-3.3.2-bin-hadoop3 폴더를 C:\Program Files\에 붙여넣는다.
-Windows 환경변수로 이동하여
--Windows Key + R <--- 실행창을 띄운 후,
--SystemPropertiesAdvanced > 확인버튼 클릭 <---시스템속성 고급탭으로 이동
--[시작 및 복구] > [환경 변수]
-[시스템 변수] > [새로 만들기]
--변수 이름 : SPARK_HOME
--변수 값 : C:\Program Files\spark-3.3.2-bin-hadoop3
--[확인] 버튼 클릭
-[시스템 변수] 목록에서 "Path"를 찾아서 선택한 후, [편집] 버튼 클릭
--[환경 변수 편집] 팝업창에서 [새로 만들기] 클릭
--C:\Program Files\spark-3.3.2-bin-hadoop3\bin 넣고 엔터
--[확인] 버튼 클릭
--[확인] 버튼을 클릭하여 [환경 변수] 팝업창 닫기
-[확인] 버튼 클릭하여 [시스템 속성]창 닫기
-Command 창을 관리자 권한으로 오픈하여 다음 명령어 실행

```
C:\Users\{User}>set SPARK_HOME
C:\Users\{User}>set PATH
```

3)Problems running Hadoop on Windows
-Refer to https://cwiki.apache.org/confluence/display/HADOOP2/WindowsProblems
-원인
Hadoop requires native libraries on Windows to work properly -that includes to access the file:// filesystem, where Hadoop uses some Windows APIs to implement posix-like file access permissions.
This is implemented in HADOOP.DLL and WINUTILS.EXE.
In particular, %HADOOP_HOME%\BIN\WINUTILS.EXE must be locatable.
If it is not, Hadoop or an application built on top of Hadoop will fail.

-How to fix a missing WINUTILS.EXE
You can fix this problem in two ways
a. Install a full native windows Hadoop version. The ASF does not currently (September 2015) release such a version; releases are available externally.
b. Or: get the WINUTILS.EXE binary from a Hadoop redistribution. There is a repository of this for some Hadoop versions on github(https://github.com/steveloughran/winutils).
-다음과 같은 사이트에서 winutils.exe 다운로드할 것
--https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe
-C:\Program Files\에 winutils 폴더를 생성하고 그 하위에 bin 폴더를 생성한다.
-다운로드 받은 winutils.exe를 방금 생성한 bin 폴더에 붙여넣는다.
--즉 경로는 다음과 같다.
C:\Program Files\winutils\bin\winutils.exe

-Then
--Set the environment variable %HADOOP_HOME% to point to the directory above the BIN dir containing WINUTILS.EXE.
--Or: run the Java process with the system property hadoop.home.dir set to the home directory.
--Windows 환경변수로 이동하여

---Windows Key + R   <--- 실행창을 띄운 후,
----SystemPropertiesAdvanced > 확인버튼 클릭 <---시스템속성 고급탭으로 이동
----[시작 및 복구] > [환경 변수]
--[시스템 변수] > [새로 만들기]
---변수 이름 : HADOOP_HOME
---변수 값 : C:\Program Files\winutils
---[확인] 버튼 클릭
--[시스템 변수] 목록에서 "Path"를 찾아서 선택한 후, [편집] 버튼 클릭
---[환경 변수 편집] 팝업창에서 [새로 만들기] 클릭
---C:\Program Files\winutils\bin 넣고 엔터
---[확인] 버튼 클릭
---[확인] 버튼을 클릭하여 [환경 변수] 팝업창 닫기
--[확인] 버튼 클릭하여 [시스템 속성]창 닫기
---Command 창을 관리자 권한으로 오픈하여 다음 명령어 실행

```
C:\Users\{User}>set HADOOP_HOME
C:\Users\{User}>set PATH
```

4)지금 현재의 가상환경에서 deactivate로 나온 후, 커맨드 창을 닫는다.
5)새로 커맨드 창을 오픈하고 다시 위에서 생성한 가상환경으로 들어가서 Visual Studio Code를 오픈하여 코드를 다시 실행한다.
6)만일 [Windows 보안 경고] 창이 나오면 [액세스 허용]을 클릭한다.
7)위의 pyspark-test.py를 Windows에서 실행하면 다음과 같은 결과를 받는다.

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/03/21 13:11:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Hadoop Version = 3.3.2
(DataLake) PS C:\DataLake> 성공: PID 7268인 프로세스(PID 4740인 자식 프로세스)가 종료되었습니다.
성공: PID 4740인 프로세스(PID 7448인 자식 프로세스)가 종료되었습니다.
성공: PID 7448인 프로세스(PID 9344인 자식 프로세스)가 종료되었습니다.
```

8)다음과 같은 오류를 만날 수 있다.
-Error Message

```
(DataLake) PS C:\DataLake> 23/03/21 13:37:07 WARN SparkEnv: Exception while deleting Spark temp dir: C:\Users\MZC-USER\AppData\Local\Temp\spark-a844ef07-148d-4c27-bbd3-d02d5ee3e05a\userFiles-bc2d42a3-d1c5-46ad-ba27-d64b085f8cee
java.io.IOException: Failed to delete: C:\Users\MZC-USER\AppData\Local\Temp\spark-a844ef07-148d-4c27-bbd3-d02d5ee3e05a\userFiles-bc2d42a3-d1c5-46ad-ba27-d64b085f8cee\org.wildfly.openssl_wildfly-openssl-1.0.7.Final.jar
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:144)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:128)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:91)
        at org.apache.spark.util.Utils$.deleteRecursively(Utils.scala:1206)
        at org.apache.spark.SparkEnv.stop(SparkEnv.scala:108)
        at org.apache.spark.SparkContext.$anonfun$stop$23(SparkContext.scala:2150)
        at org.apache.spark.util.Utils$.tryLogNonFatalError(Utils.scala:1484)
        at org.apache.spark.SparkContext.stop(SparkContext.scala:2150)
        at org.apache.spark.SparkContext.$anonfun$new$35(SparkContext.scala:670)
        at org.apache.spark.util.SparkShutdownHook.run(ShutdownHookManager.scala:214)
        at org.apache.spark.util.SparkShutdownHookManager.$anonfun$runAll$2(ShutdownHookManager.scala:188)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at org.apache.spark.util.Utils$.logUncaughtExceptions(Utils.scala:2066)
        at org.apache.spark.util.SparkShutdownHookManager.$anonfun$runAll$1(ShutdownHookManager.scala:188)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at scala.util.Try$.apply(Try.scala:213)
        at org.apache.spark.util.SparkShutdownHookManager.runAll(ShutdownHookManager.scala:188)
        at org.apache.spark.util.SparkShutdownHookManager$$anon$2.run(ShutdownHookManager.scala:178)
        at java.base/java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:539)
        at java.base/java.util.concurrent.FutureTask.run(FutureTask.java:264)
        at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1136)
        at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:635)
        at java.base/java.lang.Thread.run(Thread.java:833)
23/03/21 13:37:07 ERROR ShutdownHookManager: Exception while deleting Spark temp dir: C:\Users\MZC-USER\AppData\Local\Temp\spark-a844ef07-148d-4c27-bbd3-d02d5ee3e05a\userFiles-bc2d42a3-d1c5-46ad-ba27-d64b085f8cee
java.io.IOException: Failed to delete: C:\Users\MZC-USER\AppData\Local\Temp\spark-a844ef07-148d-4c27-bbd3-d02d5ee3e05a\userFiles-bc2d42a3-d1c5-46ad-ba27-d64b085f8cee\org.wildfly.openssl_wildfly-openssl-1.0.7.Final.jar
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:144)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:128)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:91)
        at org.apache.spark.util.Utils$.deleteRecursively(Utils.scala:1206)
        at org.apache.spark.util.ShutdownHookManager$.$anonfun$new$4(ShutdownHookManager.scala:65)
        at org.apache.spark.util.ShutdownHookManager$.$anonfun$new$4$adapted(ShutdownHookManager.scala:62)
        at scala.collection.IndexedSeqOptimized.foreach(IndexedSeqOptimized.scala:36)
        at scala.collection.IndexedSeqOptimized.foreach$(IndexedSeqOptimized.scala:33)
        at scala.collection.mutable.ArrayOps$ofRef.foreach(ArrayOps.scala:198)
        at org.apache.spark.util.ShutdownHookManager$.$anonfun$new$2(ShutdownHookManager.scala:62)
        at org.apache.spark.util.SparkShutdownHook.run(ShutdownHookManager.scala:214)
        at org.apache.spark.util.SparkShutdownHookManager.$anonfun$runAll$2(ShutdownHookManager.scala:188)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at org.apache.spark.util.Utils$.logUncaughtExceptions(Utils.scala:2066)
        at org.apache.spark.util.SparkShutdownHookManager.$anonfun$runAll$1(ShutdownHookManager.scala:188)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at scala.util.Try$.apply(Try.scala:213)
        at org.apache.spark.util.SparkShutdownHookManager.runAll(ShutdownHookManager.scala:188)
        at org.apache.spark.util.SparkShutdownHookManager$$anon$2.run(ShutdownHookManager.scala:178)
        at java.base/java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:539)
        at java.base/java.util.concurrent.FutureTask.run(FutureTask.java:264)
        at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1136)
        at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:635)
        at java.base/java.lang.Thread.run(Thread.java:833)
23/03/21 13:37:07 ERROR ShutdownHookManager: Exception while deleting Spark temp dir: C:\Users\MZC-USER\AppData\Local\Temp\spark-a844ef07-148d-4c27-bbd3-d02d5ee3e05a
java.io.IOException: Failed to delete: C:\Users\MZC-USER\AppData\Local\Temp\spark-a844ef07-148d-4c27-bbd3-d02d5ee3e05a\userFiles-bc2d42a3-d1c5-46ad-ba27-d64b085f8cee\org.wildfly.openssl_wildfly-openssl-1.0.7.Final.jar
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:144)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:128)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursivelyUsingJavaIO(JavaUtils.java:128)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:118)
        at org.apache.spark.network.util.JavaUtils.deleteRecursively(JavaUtils.java:91)
        at org.apache.spark.util.Utils$.deleteRecursively(Utils.scala:1206)
        at org.apache.spark.util.ShutdownHookManager$.$anonfun$new$4(ShutdownHookManager.scala:65)
        at org.apache.spark.util.ShutdownHookManager$.$anonfun$new$4$adapted(ShutdownHookManager.scala:62)
        at scala.collection.IndexedSeqOptimized.foreach(IndexedSeqOptimized.scala:36)
        at scala.collection.IndexedSeqOptimized.foreach$(IndexedSeqOptimized.scala:33)
        at scala.collection.mutable.ArrayOps$ofRef.foreach(ArrayOps.scala:198)
        at org.apache.spark.util.ShutdownHookManager$.$anonfun$new$2(ShutdownHookManager.scala:62)
        at org.apache.spark.util.SparkShutdownHook.run(ShutdownHookManager.scala:214)
        at org.apache.spark.util.SparkShutdownHookManager.$anonfun$runAll$2(ShutdownHookManager.scala:188)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at org.apache.spark.util.Utils$.logUncaughtExceptions(Utils.scala:2066)
        at org.apache.spark.util.SparkShutdownHookManager.$anonfun$runAll$1(ShutdownHookManager.scala:188)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
        at scala.util.Try$.apply(Try.scala:213)
        at org.apache.spark.util.SparkShutdownHookManager.runAll(ShutdownHookManager.scala:188)
        at org.apache.spark.util.SparkShutdownHookManager$$anon$2.run(ShutdownHookManager.scala:178)
        at java.base/java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:539)
        at java.base/java.util.concurrent.FutureTask.run(FutureTask.java:264)
        at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1136)
        at java.base/java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:635)
        at java.base/java.lang.Thread.run(Thread.java:833)
```

-원인

살펴보니 Spark를 실행하는 동안 C:\Users\{User}\AppData\Local\Temp 폴더에 폴더를 생성하는데, 그것을 지울 수 없다는 내용이다.

-해결방법

--Windows 환경에서 Visual Studio Code로 실행할 시에는 반드시 [Run and Debug] 또는 [Run without Debug]로 실행하면 된다.

--물론, 수동으로 OS의 %TEMP% 또는 %TMP% 폴더의 Spark 생성파일들은 삭제해야 한다.