

```
1 Lab. Apache Spark 환경설정하기
2
3 1. Python 환경설정
4 1)Python Version 확인
5 $ python3 --version
6
7 2)Python 3.11 Install
8 3)Python 3.11을 Python 기본 버전으로 지정
9 $ ls -l /usr/local/bin/python*
10 $ which python3.11
11 /usr/local/bin/python3.11
12 $ ln -s -f /usr/local/bin/python3.11 /usr/local/bin/python
13
14 4)Python Version 확인
15 $ python --version
16
17
18 2. DataLake Python 가상환경 구성하기
19 1)pip3 Version 확인
20 $ pip3 --version
21
22 2)virtualenv 설치
23 $ pip3 install virtualenv
24
25 3)virtualenv PATH에 추가하기
26 $ echo $PATH
27 $ export PATH=$PATH:/Library/Frameworks/Python.framework/Versions/3.11/bin
28
29 4)DataLake 가상환경 구성하기
30 $ virtualenv DataLake
31 $ source DataLake/bin/activate
32
33
34 3. DataLake 가상환경에 pyspark 설치하기
35 1)pip3 list 확인
36 (DataLake) ... $ pip3 list
37
38 2)pyspark 설치
39 (DataLake) ... $ pip3 install pyspark
40
41
42 4. Visual Studio Code에서 테스트
43 1)Visual Studio Code 런치 후
44 2)DataLake 디렉토리를 Workspace Folder로 지정
45 3)pyspark-test.py 새 파일 생성
46 4)다음과 같이 코드 입력 후 pyspark 테스트
47 import pyspark
48 from pyspark.sql import SparkSession
49
50 conf = pyspark.SparkConf()
51 conf.set('spark.driver.host', '127.0.0.1')
52
53 spark = SparkSession.builder \
54     .config(conf=conf) \
55     .appName('Pyspark Test Program') \
56     .getOrCreate()
57
58 print(f'Hadoop Version = {spark._jvm.org.apache.hadoop.util.VersionInfo.getVersion()}')
59
```