IT609: Big Data Processing
Assignment 04: Configuring PySpark in Windows

ID - 202218061
NAME - JATAN SAHU

1. Download and Install JAVA

→As Spark uses Java Virtual Machine internally, it has a dependency on JAVA

—>Link - https://www.java.com/en/download/

->Already downloaded

2. Download and Install Python

→ Already installed python

→

```
C:\Users\M.K. COMPUTERS>python
Python 3.10.5 (tags/v3.10.5:f377153, Jun  6 2022, 16:14:13) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

3. Download PySpark .tgz file from

https://spark.apache.org/downloads.html . Download the appropriate version. Here, I have downloaded Release 3.3.2 Prebuilt for Hadoop 2.7. Extract the file. The extracted file is stored in 'D:\BDP\spark-3.3.2-bin-hadoop2'

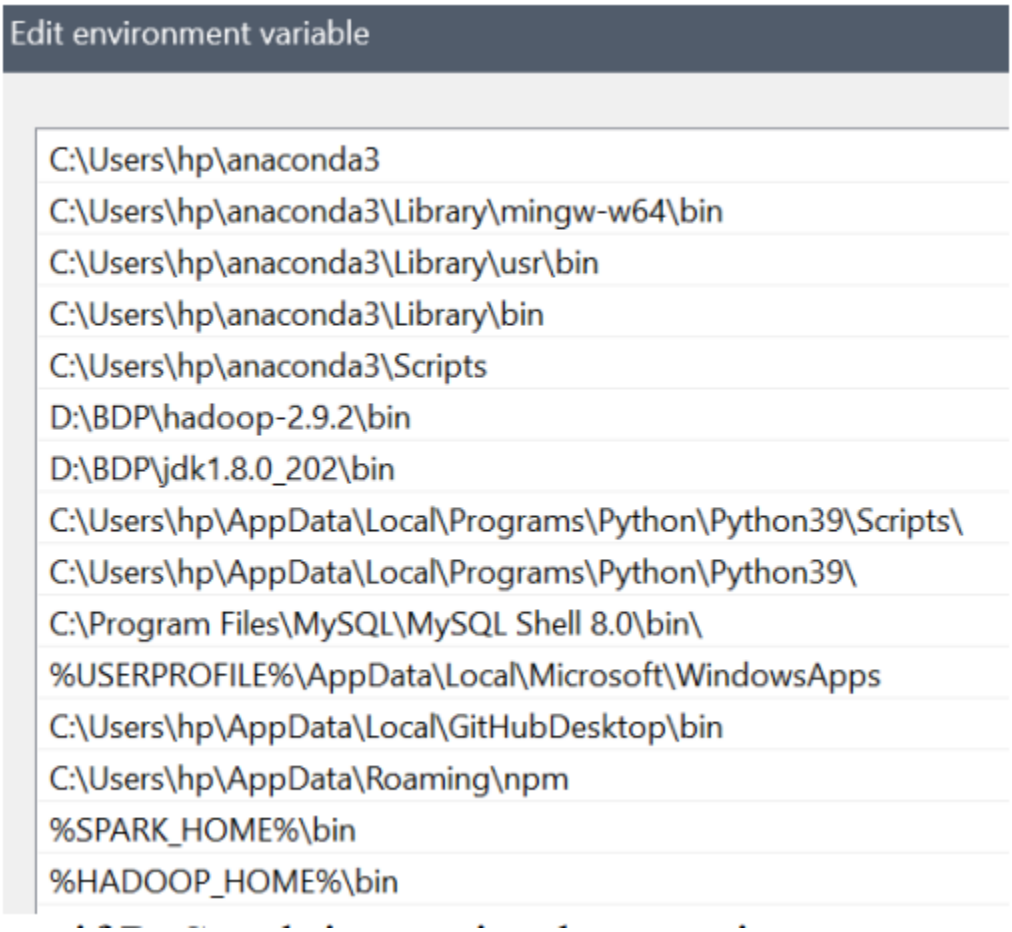
4. Download winutils.exe from

Download winutils.exe from https://github.com/cdarlint/winutils using the correct Hadoop version (here, 2.9.2). Create a new folder structure hadoop/bin in D:\BDP\spark-3.3.2-bin-hadoop2 and save the winutils file here.

5.Setting Environment Variables: Go to Search and open ‘Advanced System Settings.’ Click on Environment Variables. Add the following variables:

Variable Name	Value
SPARK_HOME	D:\BDP\spark-3.3.2-bin-hadoop2
HADOOP_HOME	%SPARK_HOME%/hadoop
PYSPARK_PYTHON	C:\Users\hp\anaconda3

Click on the “Path” variable add the following two values:
%SPARK_HOME%\bin %HADOOP_HOME%\bin



6. Test if PySpark is running by opening command prompt and typing ‘pyspark’.