Pyspark Installation

1) install JAVA 1.7 or 1.8

https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html#license-lightbox

2) Install Anaconda

https://repo.anaconda.com/archive/Anaconda3-2020.02-Windows-x86_64.exe

3) Install Apache Spark

Download spark and extract with in one of the folder.
create folder with any name and i have created pyspark
then extract spark zip file.

C:\pyspark\spark-2.4.5-bin-hadoop2.7\

4) Download winutils.exe file

Download winutils.exe and place into C:\pyspark\spark-2.4.5-bin-hadoop2.7\bin\ folder.

https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe

Pyspark Installation …

5) set environment variables.

set JAVA_HOME=C:\Java\jdk1.7.0_80
set SPARK_HOME=C:\pyspark\spark-2.4.5-bin-hadoop2.7
set HADOOP_HOME=C:\pyspark\spark-2.4.5-bin-hadoop2.7
set PATH=C:\pyspark\spark-2.4.5-bin-hadoop2.7\bin;
set PATH=C:\Java\jdk1.7.0_80\bin;
set PATH=C:\Windows\System32;

6) Install jupyter

Click on Windows and search "Anacoda Prompt".
Open Anaconda prompt and type "python -m pip install findspark".
This package is necessary to run spark from Jupyter notebook.

Now, from the same Anaconda Prompt, type "jupyter notebook" and hit enter.
This would open a jupyter notebook from your browser. From Jupyter notebookàNewàSelect Python3, as shown below.

Pyspark Installation

7) Open CMD with Admin and run below command for granting access to C:\tmp\hive

winutils.exe chmod -R 777 C:\tmp\hive
winutils.exe ls -F C:\tmp\hive

open anaconda prompt

Install Pyspark and findspark

Pip install Pyspark
Pip install findspark

and Type pyspark enter..... To re-verify..