## Setup of Python and Pyspark on Windows machines local

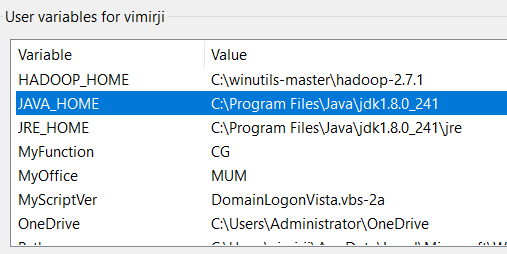
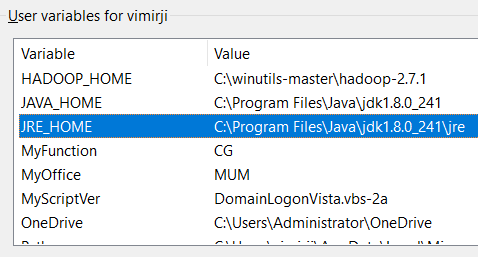
### Install 64-bit JDK 1.8 for Windows operating system from Oracle.

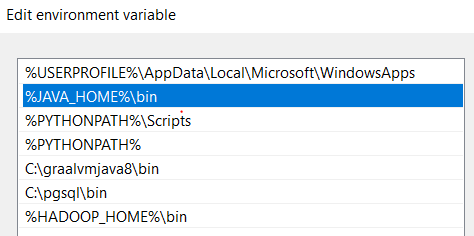
Download URL – <https://www.oracle.com/java/technologies/javase-jdk8-downloads.html>

#### Set below environment variables pointing to home directory of java

(Please remember the path mentioned below may be different for you)

JAVA\_HOME=C:\Program Files\Java\jdk1.8.0\_241

  
  
JRE\_HOME=%JAVA\_HOME%\jre  


PATH=%JAVA\_HOME%\bin  


### Verify Java compiler and JRE

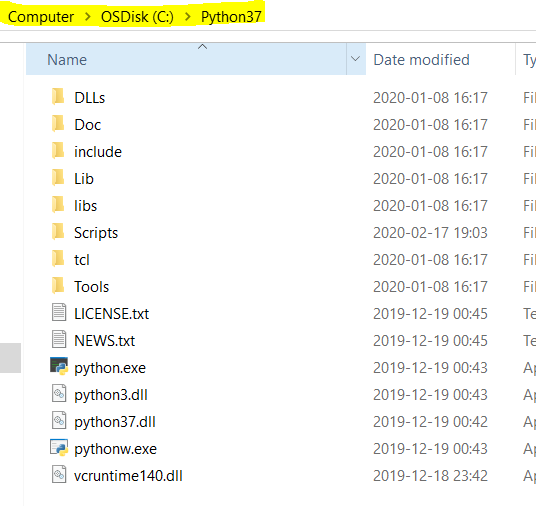
#### javac -version

#### java -version

### Install Python 3.x

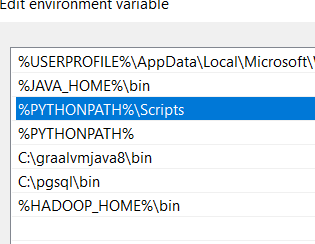
#### Download Windows binaries from URL – <https://www.python.org/downloads/>

#### Installation location



#### Declare environment variables

PYTHONPATH=C:\Program Files\Python37  
PATH=%PYTHONPATH%;%PYTHONPATH%\Scripts



#### Verify Python on Windows

### Install Hadoop windows binaries

#### Download URL

<https://github.com/steveloughran/winutils>

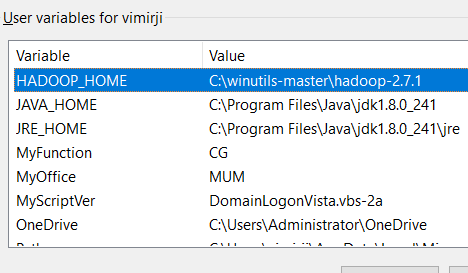
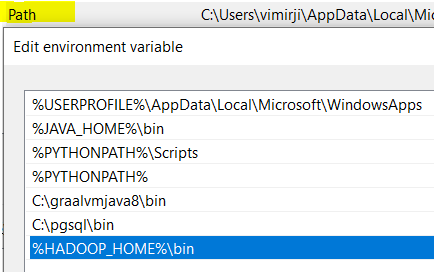
Download the zip file and unzip to location as given below –

C:\winutils-master

#### Setup environment variables –

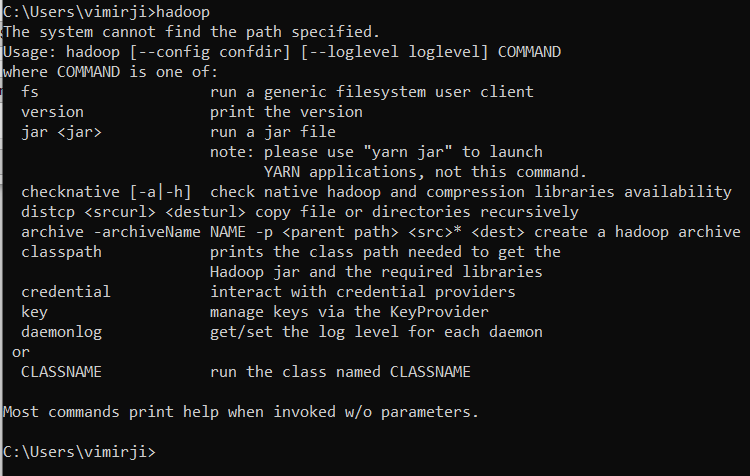
HADOOP\_HOME=C:\winutils-master\hadoop-2.7.1

PATH=%PATH%;%HADOOP\_HOME%\bin

#### Verify Hadoop on Windows –

Type **hadoop** command on prompt to see help message



### Install Python PyPI package to create virtual environments using pip

#### Install package virtualenvwrapper-win

pip install virtualenvwrapper-win

### Create virtual environment aws\_env

#### Create a folder “clinical\_trial\_ctgov” in Documents

#### Create virtual environment

***Syntax to create virtual environment –***

mkvirtualenv -a <full-path-to-project-folder> -p <full-path-to-python.exe> <environment-name>

***Example –***

mkvirtualenv -a “C:\Users\vimirji\Documents\clinical\_trial\_ctgov” -p “C:\Python37\python.exe” aws\_env

#### Switch to virtual environment

workon aws\_env

#### Switch to project folder

Cdproject

### Install pyspark & boto3 packages

#### Switch to virtual environment

workon aws\_env

#### Switch to project folder

cdproject

#### Install pyspark version 2.4.3

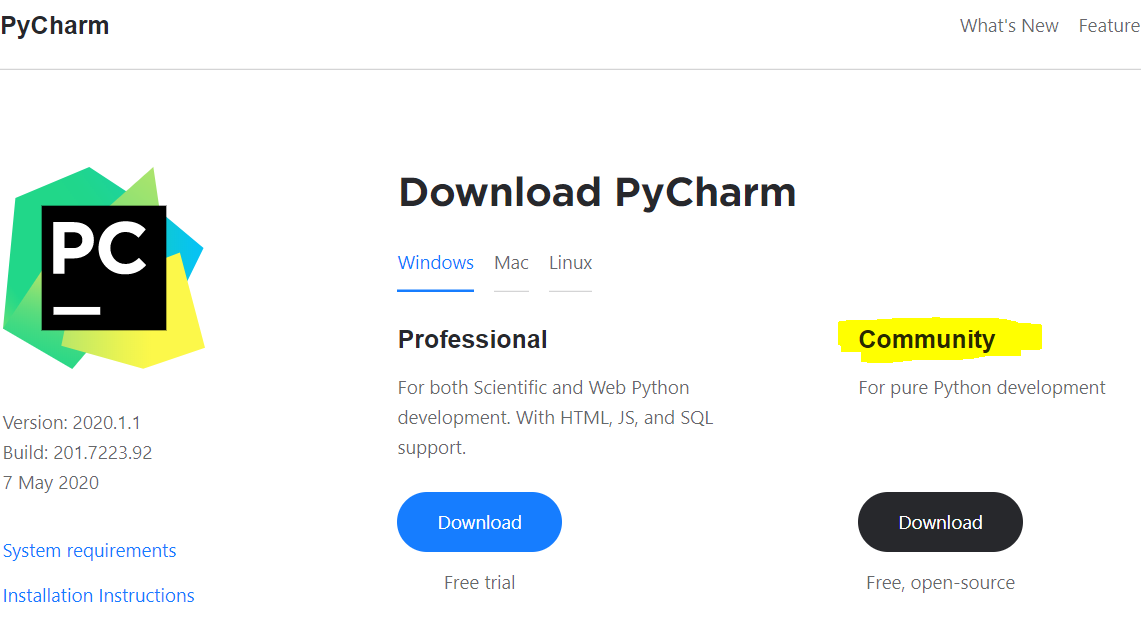
pip install pyspark==2.4.3

#### Install boto3

pip install boto3

### Install Pycharm

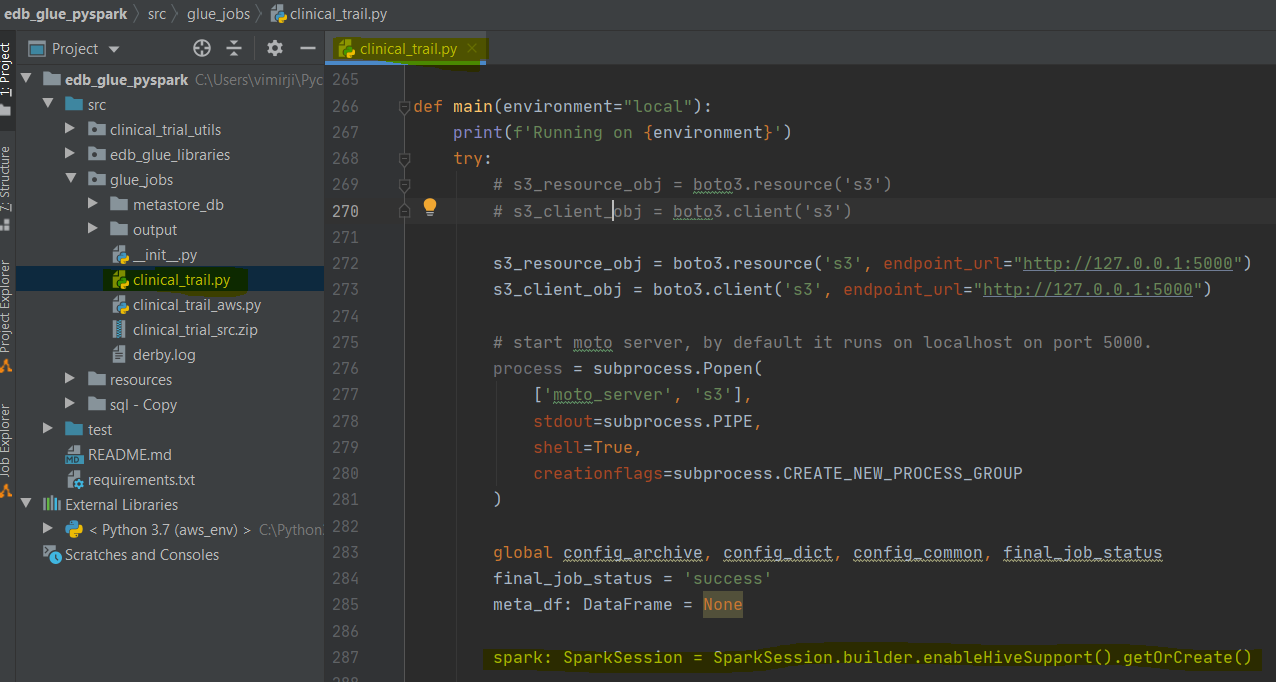
* Download URL – <https://www.jetbrains.com/pycharm/download/>

Please note: Download community Edition, Professional edition is paid tool  


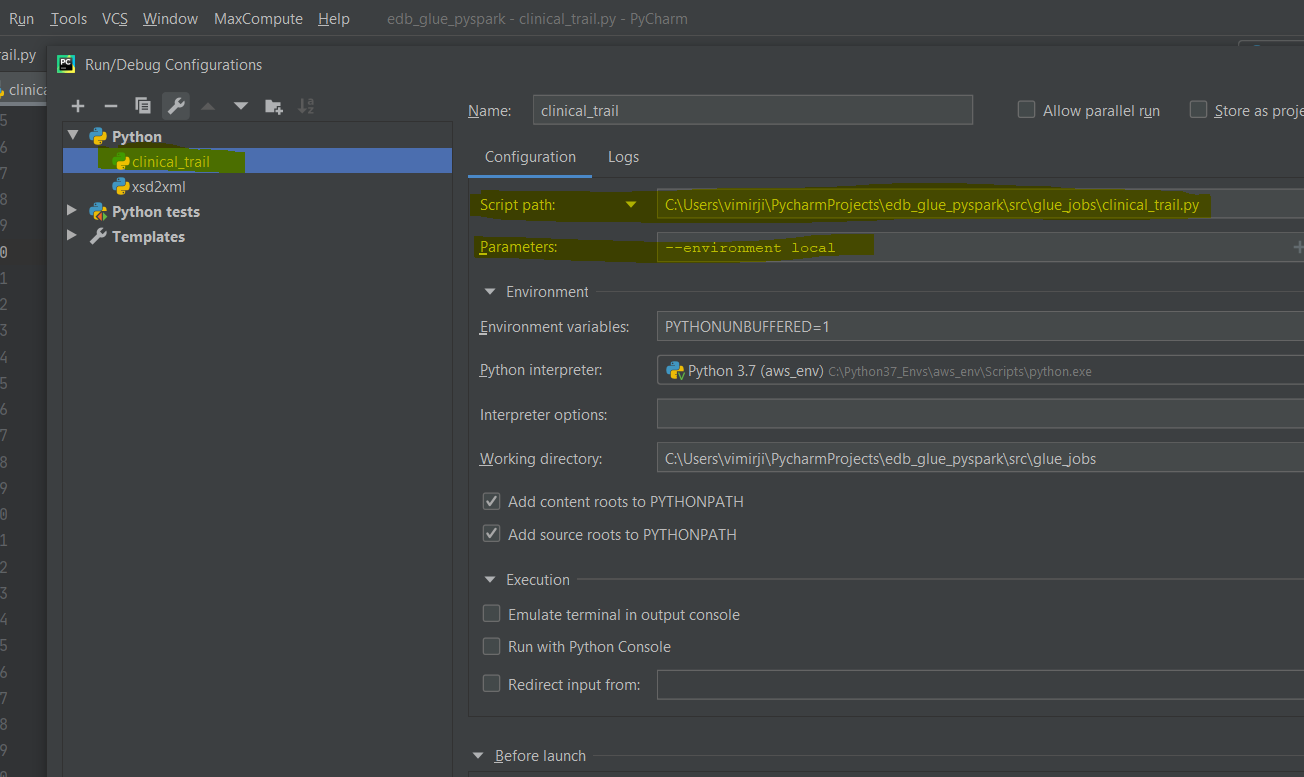
#### Open the project in PyCharm Example – (Please note your project name & folder may be different)

#### Create a python file

For example, here it is clinical\_trial.py



#### Execution on PyCharm

* From menu Run -> Edit Configurations...  
  Click + button under Python and add Run configuration as shown below  
  
* Click play button to Run

