Pyspark installation in windows

In this video I am explaining how to install Pyspark in windows. How to create environment to practice pyspark. Explaining in this document.

If you want to install Pyspark, first you must download these three softwares from internet

https://dlcdn.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz

https://archive.apache.org/dist/hadoop/core/hadoop-3.2.2/hadoop-3.2.2.tar.gz

https://github.com/aitraining/datasets/blob/main/hadoop-dependencies-3.2.2.zip

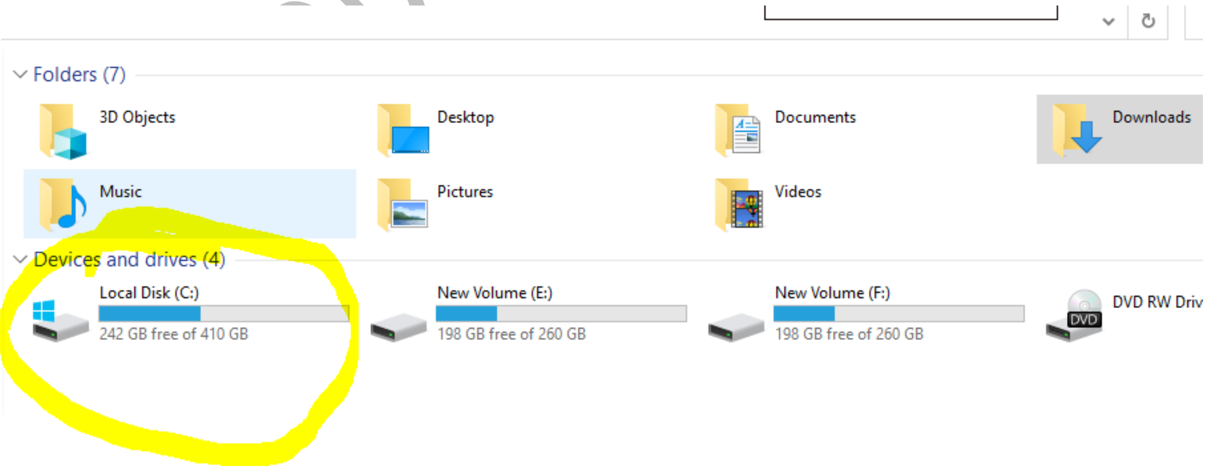
https://repo.anaconda.com/archive/Anaconda3-2021.05-Windows-x86_64.exe

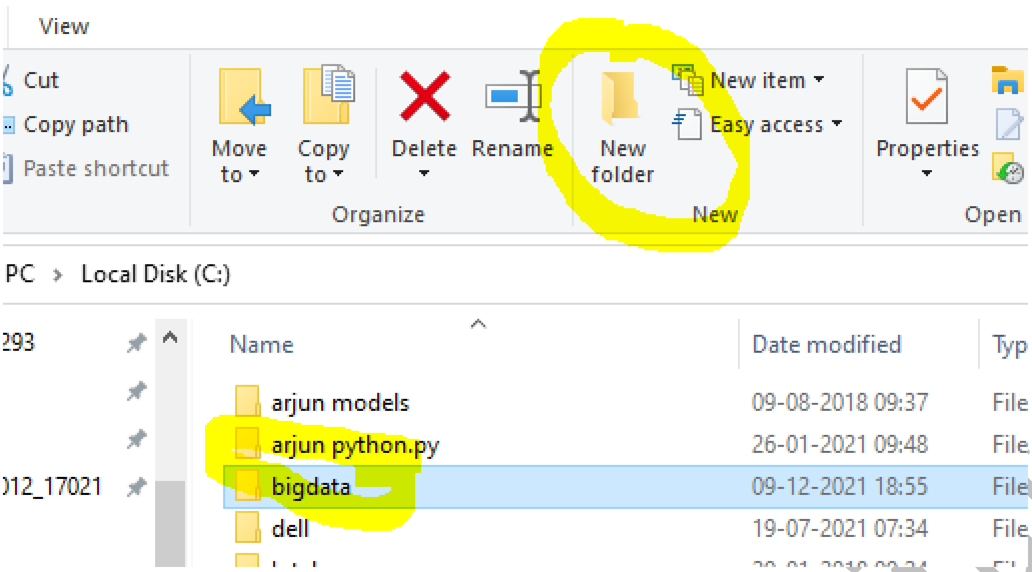
https://www.jetbrains.com/pycharm/download/

https://www.oracle.com/in/java/technologies/javase/jdk11-archive-downloads.html

In your computer you have multiple drives like C, D , E where you have maximum space in that drive create a folder called bigdata.

Important note: Don't create "big data" must create "bigdata" without any space. Otherwise its create new problems.

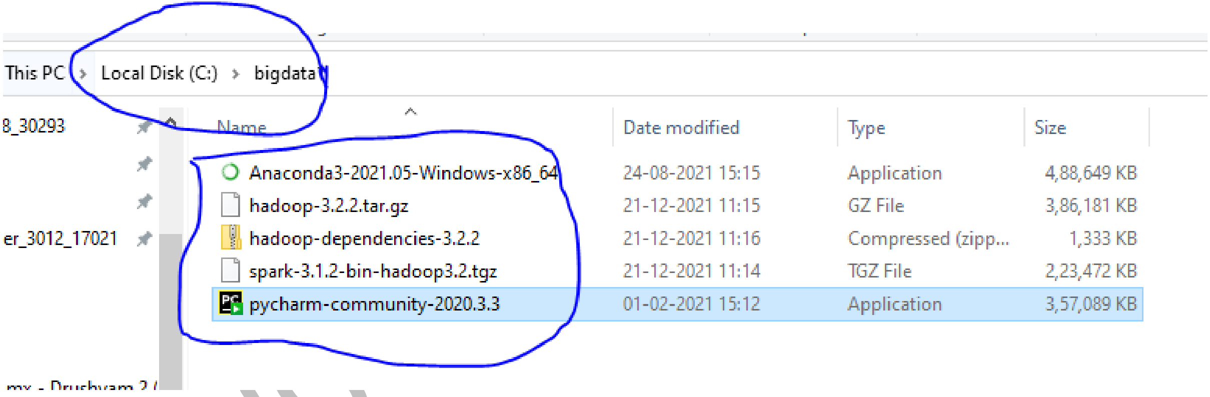




Name	Date modified	Typ
arjun models	09-08-2018 09:37	File
arjun python.py	26-01-2021 09:48	File
bigdata	09-12-2021 18:55	File
dell	19-07-2021 07:34	File

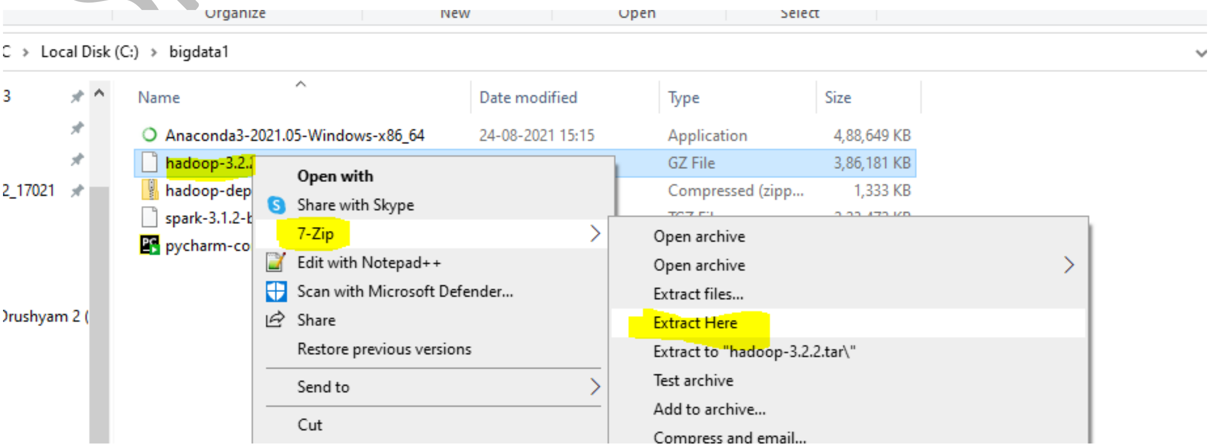
Now.

After download above softwares place these folders in bigdata folders like this

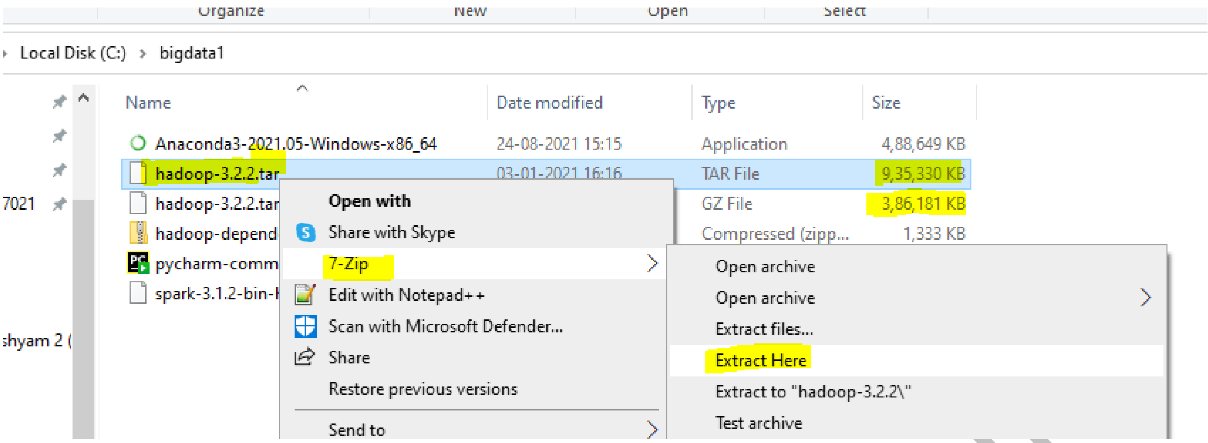


Name	Date modified	Type	Size
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
spark-3.1.2-bin-hadoop3.2.tgz	21-12-2021 11:14	TGZ File	2,23,472 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB

After download please extract all folders. Right click and .. 7zip> extract here.



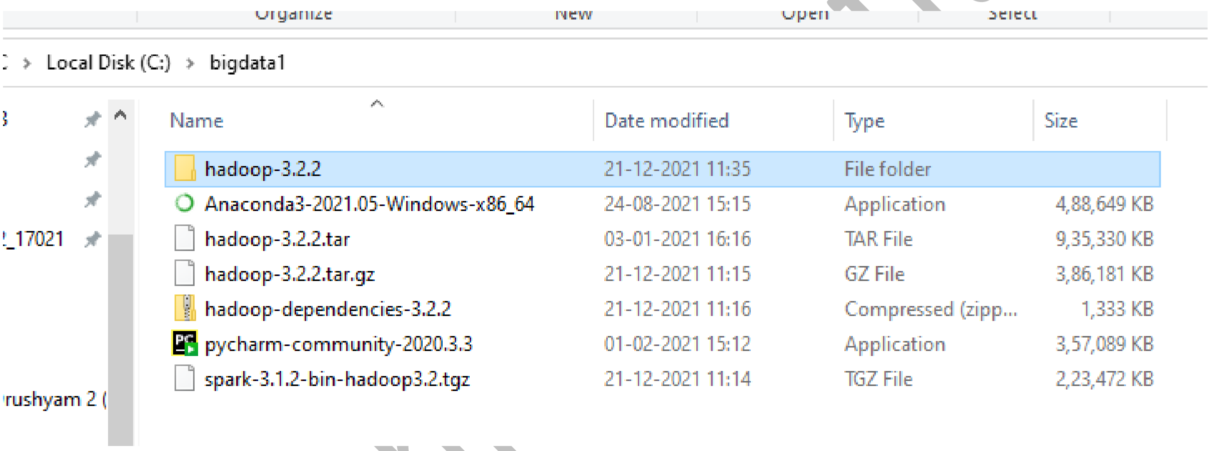
After extract you will get another zip folder. Please extract it again. Please check size its changed



Local Disk (C:) › bigdata1

Name	Date modified	Type	Size
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar		GZ File	3,86,181 KB
hadoop-depend		Compressed (zipp...	1,333 KB
pycharm-comm			
spark-3.1.2-bin-l			

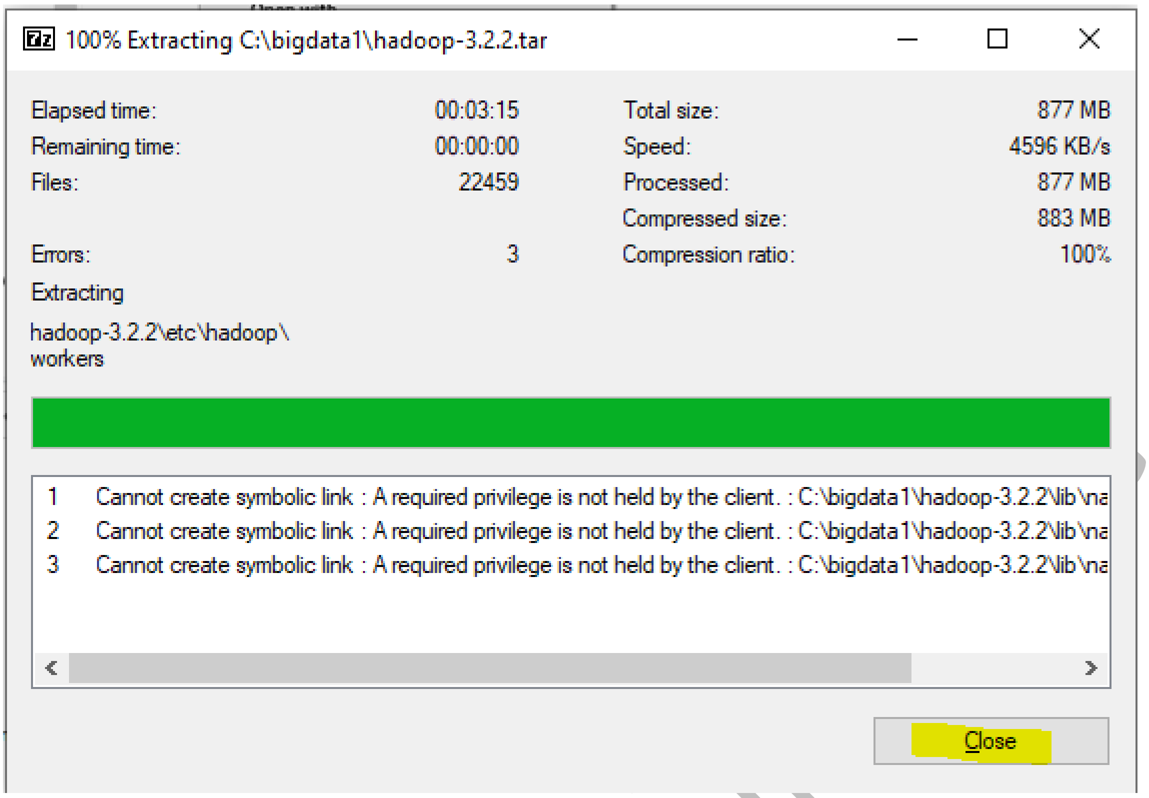
Now you will get folder like this.



Local Disk (C:) › bigdata1

Name	Date modified	Type	Size
hadoop-3.2.2	21-12-2021 11:35	File folder	
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB
spark-3.1.2-bin-hadoop3.2.tgz	21-12-2021 11:14	TGZ File	2,23,472 KB

After 5 minutes you will get a small warning like this. Please ignore, simply click close.

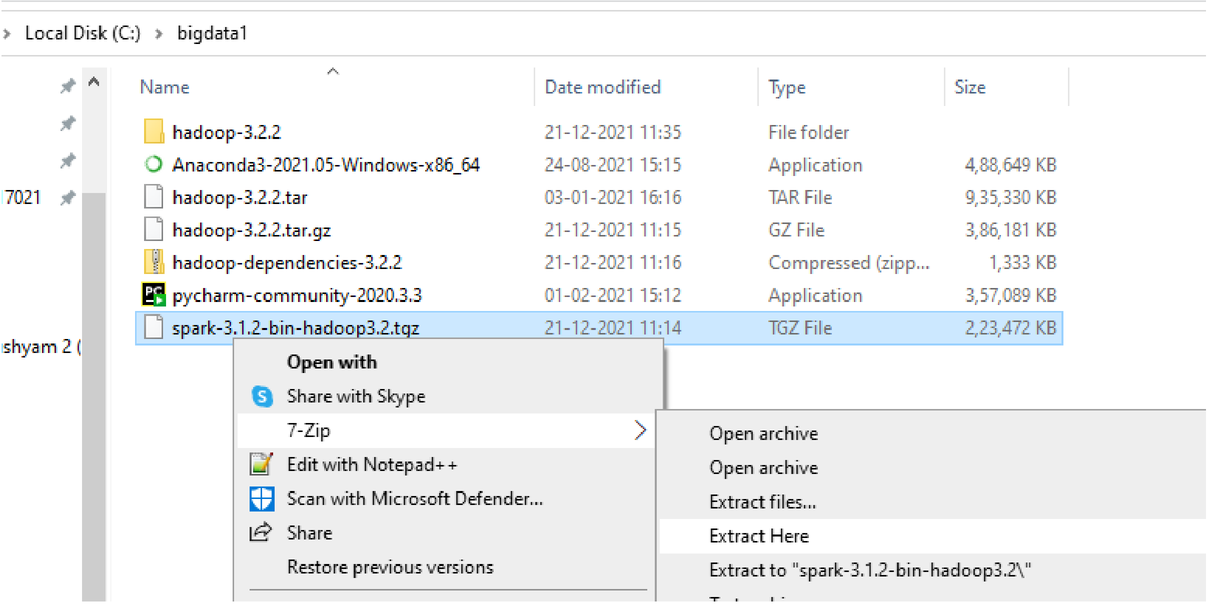


100% Extracting C:\bigdata1\hadoop-3.2.2.tar

```
Elapsed time:        00:03:15    Total size:          877 MB
Remaining time:      00:00:00    Speed:             4596 KB/s
Files:                  22459    Processed:           877 MB
                                 Compressed size:     883 MB
Errors:                     3    Compression ratio:     100%
Extracting
hadoop-3.2.2\etc\hadoop\
workers
```

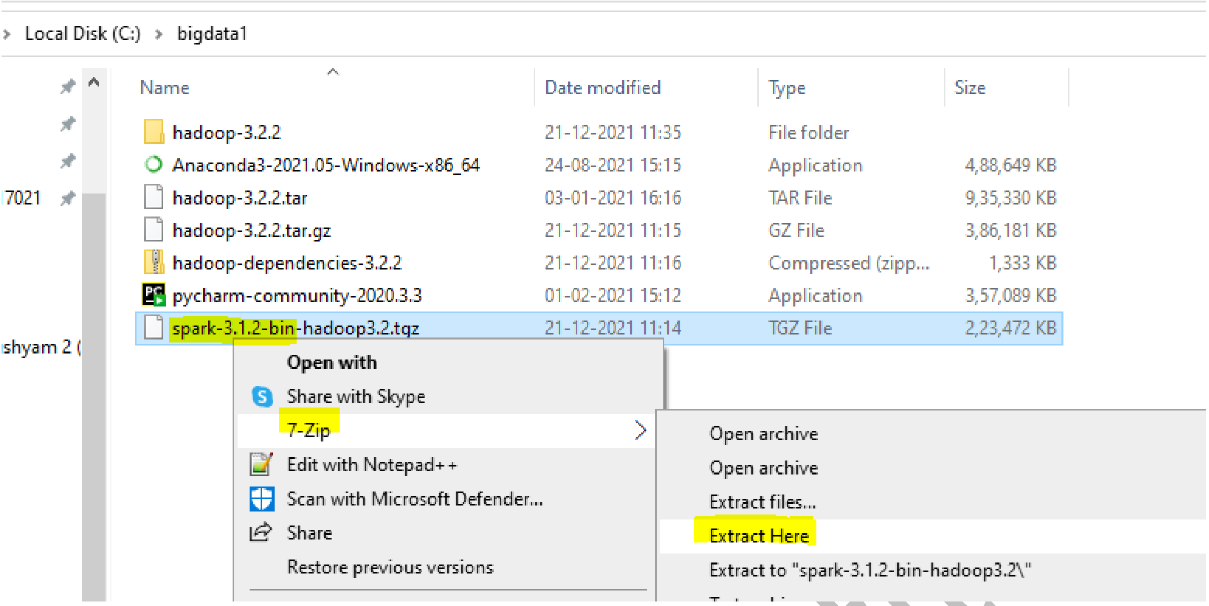
```
1   Cannot create symbolic link : A required privilege is not held by the client. : C:\bigdata1\hadoop-3.2.2\lib\na
2   Cannot create symbolic link : A required privilege is not held by the client. : C:\bigdata1\hadoop-3.2.2\lib\na
3   Cannot create symbolic link : A required privilege is not held by the client. : C:\bigdata1\hadoop-3.2.2\lib\na
```

Same way extract Spark folder as well.



Local Disk (C:) › bigdata1

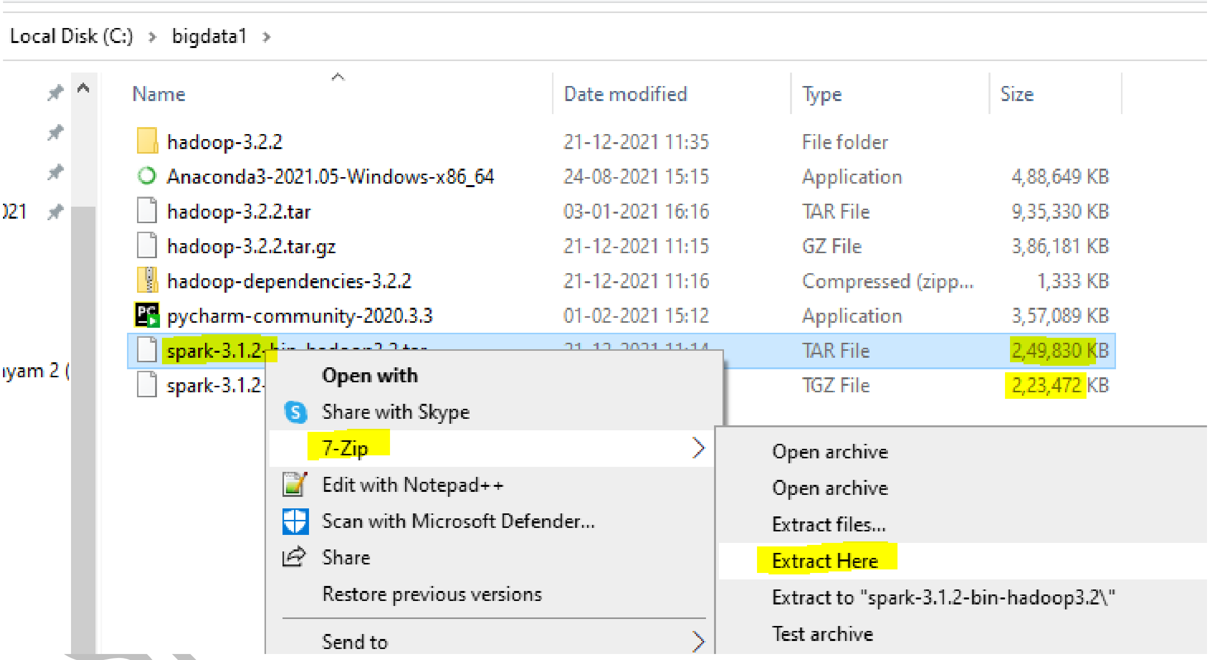
Name	Date modified	Type	Size
hadoop-3.2.2	21-12-2021 11:35	File folder	
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB
spark-3.1.2-bin-hadoop3.2.tgz	21-12-2021 11:14	TGZ File	2,23,472 KB



Local Disk (C:) > bigdata1

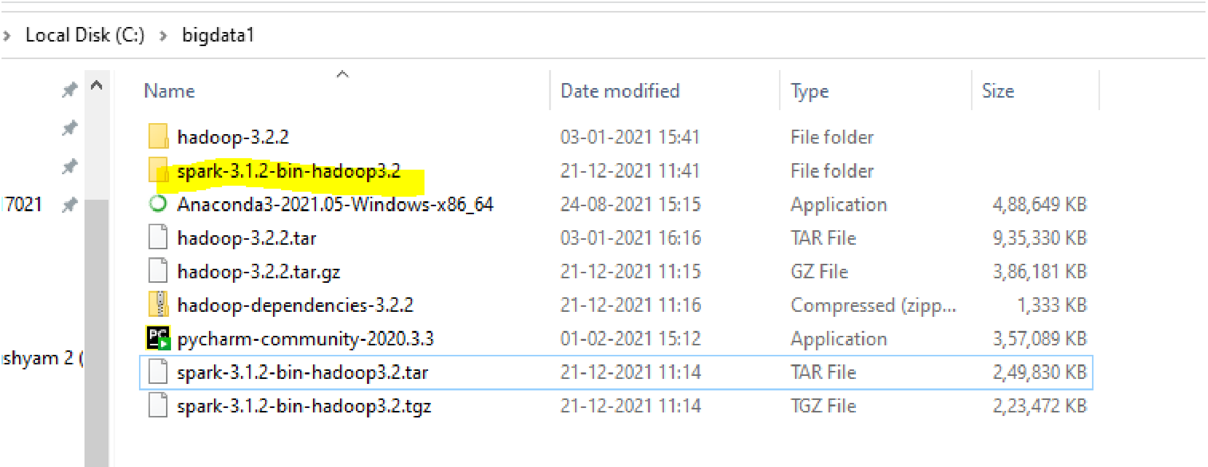
Name	Date modified	Type	Size
hadoop-3.2.2	21-12-2021 11:35	File folder	
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB
spark-3.1.2-bin-hadoop3.2.tgz	21-12-2021 11:14	TGZ File	2,23,472 KB

You will get another file. Please extract again



Local Disk (C:) > bigdata1 >

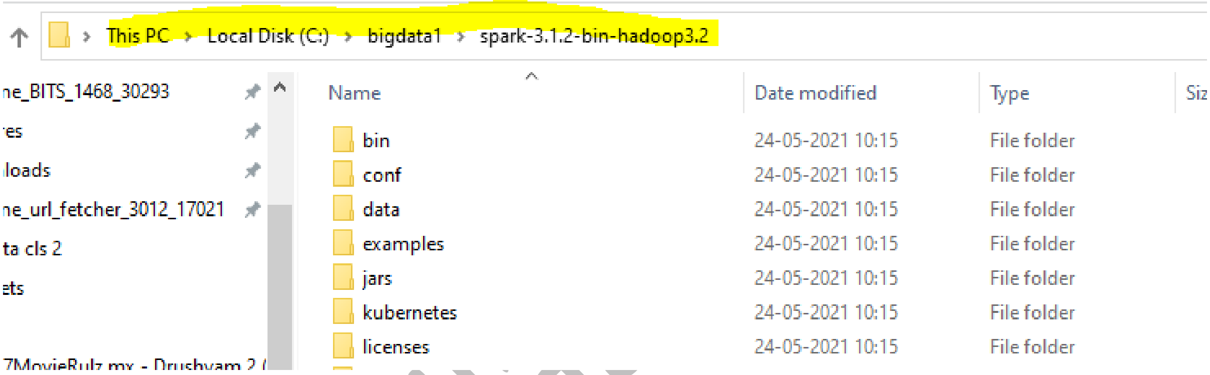
Name	Date modified	Type	Size
hadoop-3.2.2	21-12-2021 11:35	File folder	
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB
spark-3.1.2-bin-hadoop3.2.tar	21-12-2021 11:14	TAR File	2,49,830 KB
spark-3.1.2-		TGZ File	2,23,472 KB



› Local Disk (C:) › bigdata1

Name	Date modified	Type	Size
hadoop-3.2.2	03-01-2021 15:41	File folder	
spark-3.1.2-bin-hadoop3.2	21-12-2021 11:41	File folder	
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB
spark-3.1.2-bin-hadoop3.2.tar	21-12-2021 11:14	TAR File	2,49,830 KB
spark-3.1.2-bin-hadoop3.2.tgz	21-12-2021 11:14	TGZ File	2,23,472 KB

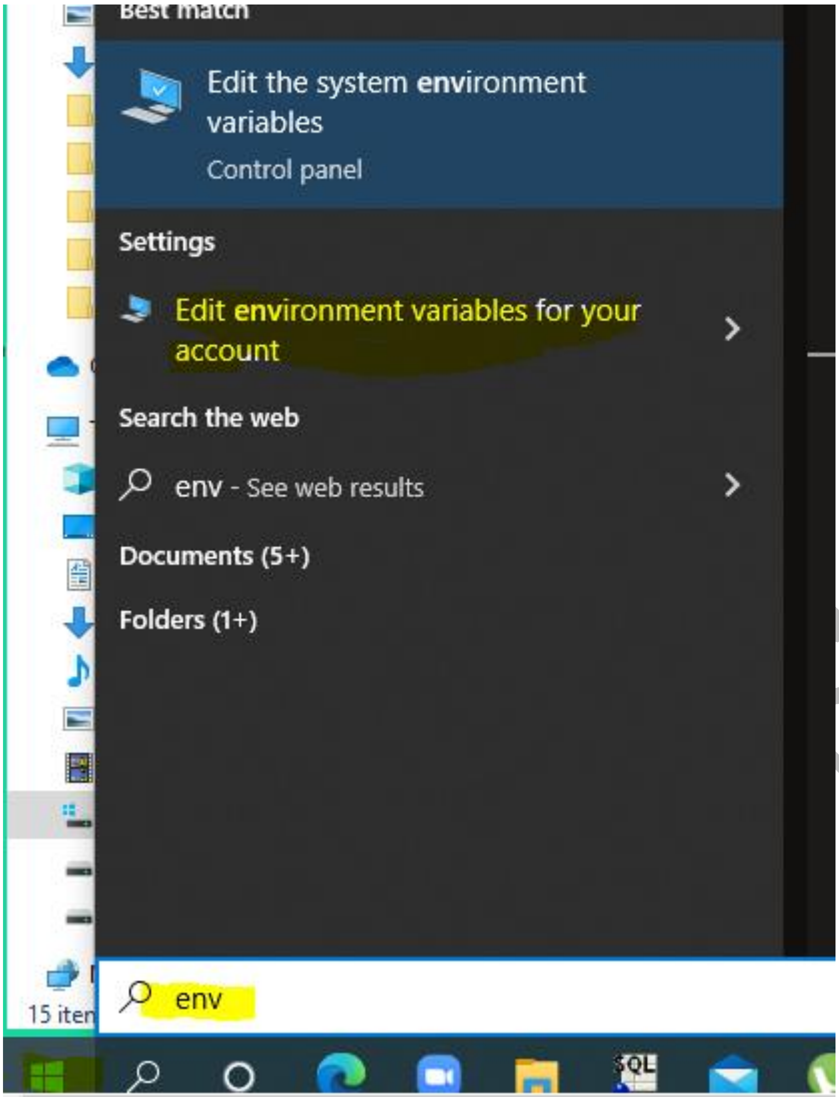
Now configure in environment variable, that's y copy this path



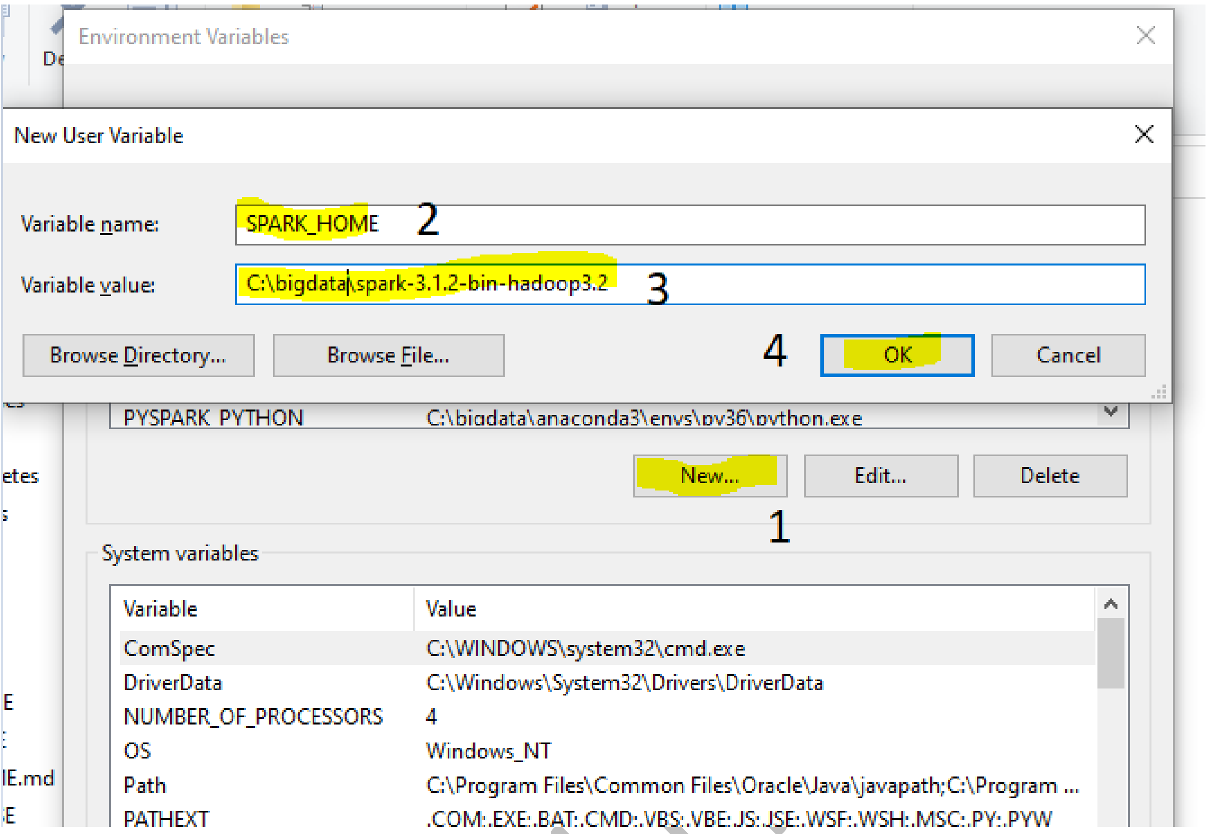
This PC › Local Disk (C:) › bigdata1 › spark-3.1.2-bin-hadoop3.2

Name	Date modified	Type
bin	24-05-2021 10:15	File folder
conf	24-05-2021 10:15	File folder
data	24-05-2021 10:15	File folder
examples	24-05-2021 10:15	File folder
jars	24-05-2021 10:15	File folder
kubernetes	24-05-2021 10:15	File folder
licenses	24-05-2021 10:15	File folder

Now Start button > env > Environment variable >



Best match
Edit the system environment variables
Control panel
Settings
Edit environment variables for your account
Search the web
env - See web results
Documents (5+)
Folders (1+)
env
15 iten



Environment Variables

New User Variable

Variable name: SPARK_HOME 2

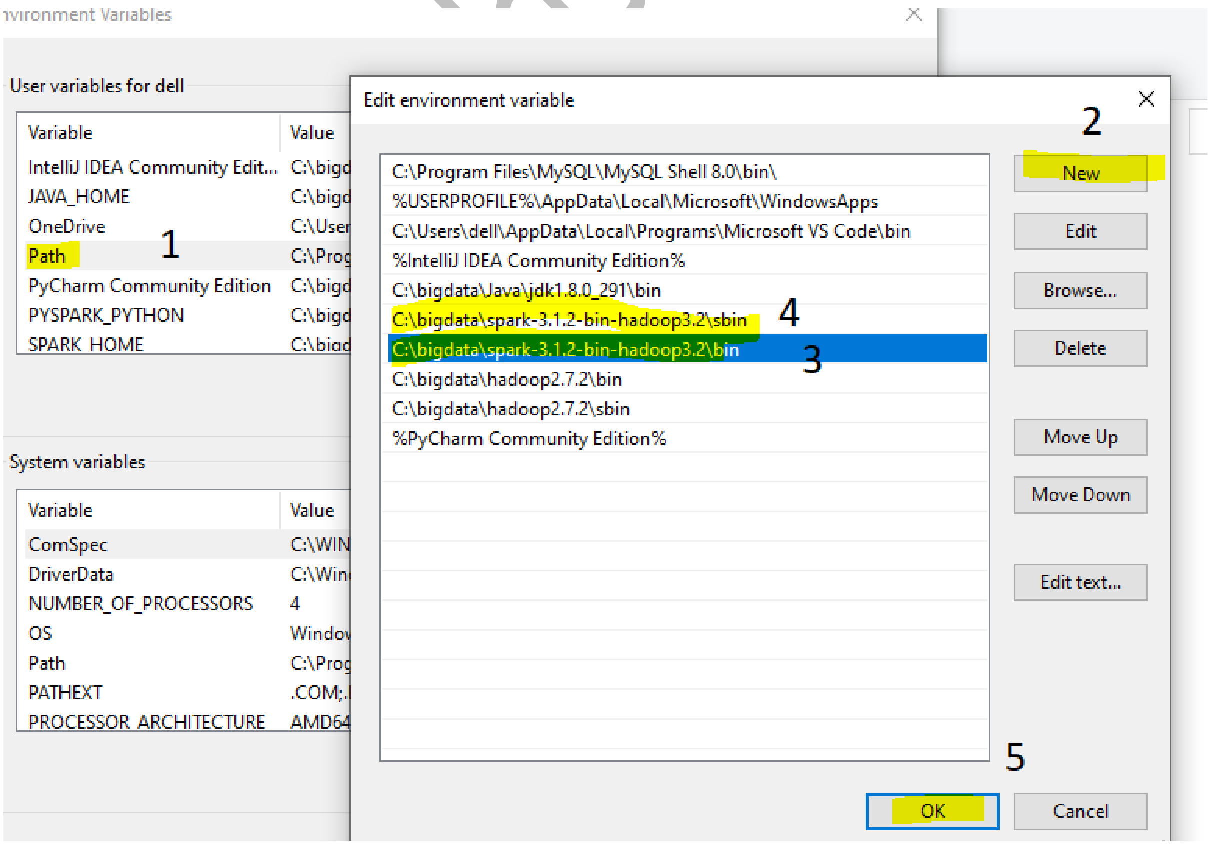
Variable value: C:\bigdata\spark-3.1.2-bin-hadoop3.2 3

PYSPARK_PYTHON C:\bigdata\anaconda3\envs\py36\python.exe

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Program ...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW

After this step



nvironment Variables

User variables for dell

Variable	Value
IntelliJ IDEA Community Edit...	C:\bigd
JAVA_HOME	C:\bigd
OneDrive	C:\User
Path 1	C:\Prog
PyCharm Community Edition	C:\bigd
PYSPARK_PYTHON	C:\bigd
SPARK_HOME	C:\bigd

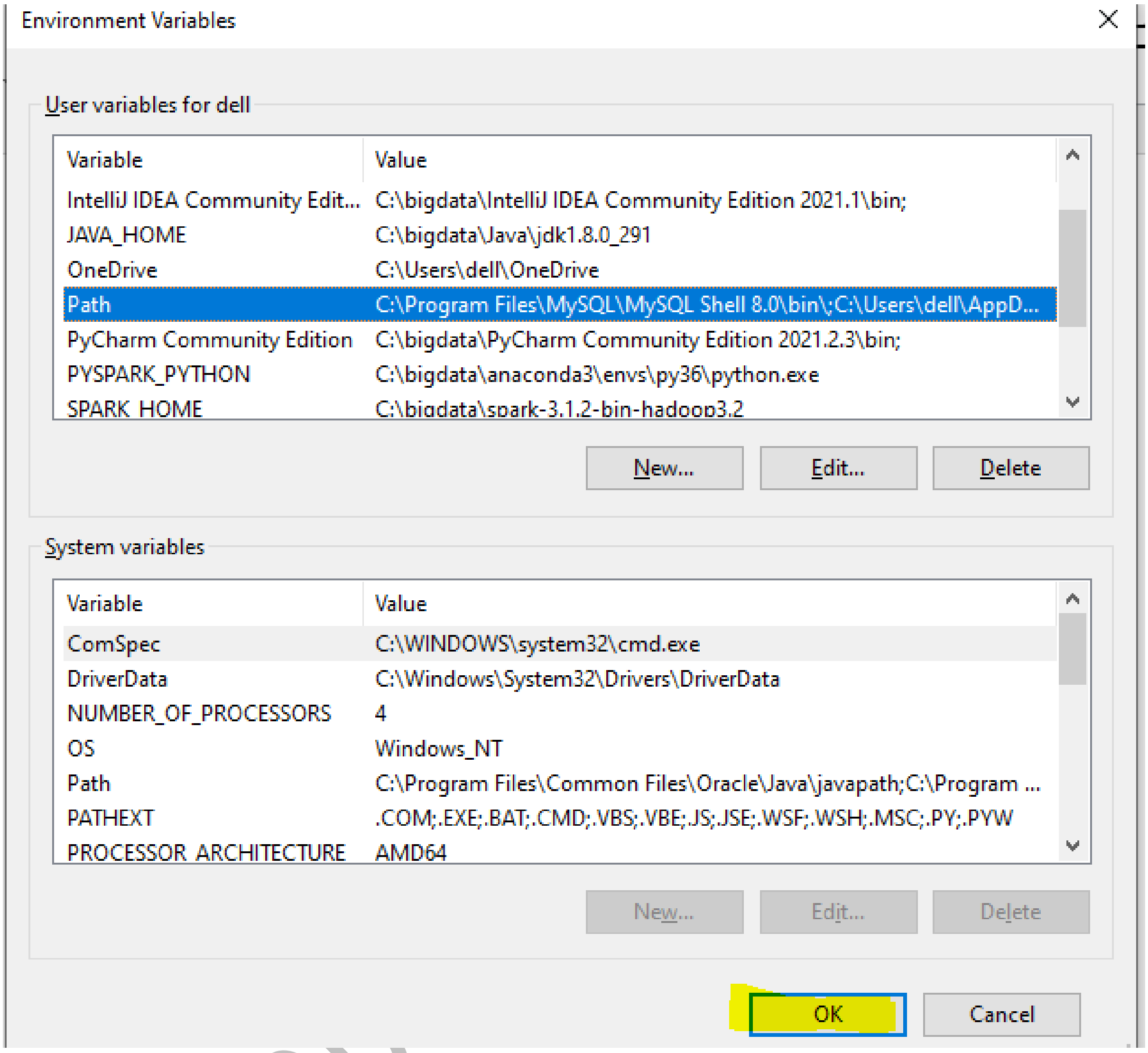
System variables

Variable	Value
ComSpec	C:\WIN
DriverData	C:\Win
NUMBER_OF_PROCESSORS	4
OS	Windov
Path	C:\Prog
PATHEXT	.COM;.
PROCESSOR_ARCHITECTURE	AMD64

Edit environment variable

```
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Users\dell\AppData\Local\Programs\Microsoft VS Code\bin
%IntelliJ IDEA Community Edition%
C:\bigdata\Java\jdk1.8.0_291\bin
C:\bigdata\spark-3.1.2-bin-hadoop3.2\sbin   4
C:\bigdata\spark-3.1.2-bin-hadoop3.2\bin    3
C:\bigdata\hadoop2.7.2\bin
C:\bigdata\hadoop2.7.2\sbin
%PyCharm Community Edition%
```

Finally click ok



Environment Variables

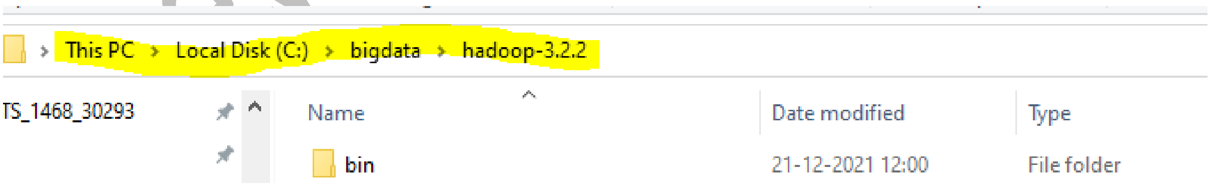
User variables for dell

Variable	Value
IntelliJ IDEA Community Edit...	C:\bigdata\IntelliJ IDEA Community Edition 2021.1\bin;
JAVA_HOME	C:\bigdata\Java\jdk1.8.0_291
OneDrive	C:\Users\dell\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\dell\AppD...
PyCharm Community Edition	C:\bigdata\PyCharm Community Edition 2021.2.3\bin;
PYSPARK_PYTHON	C:\bigdata\anaconda3\envs\py36\python.exe
SPARK_HOME	C:\bigdata\spark-3.1.2-bin-hadoop3.2

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Program ...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
PROCESSOR_ARCHITECTURE	AMD64

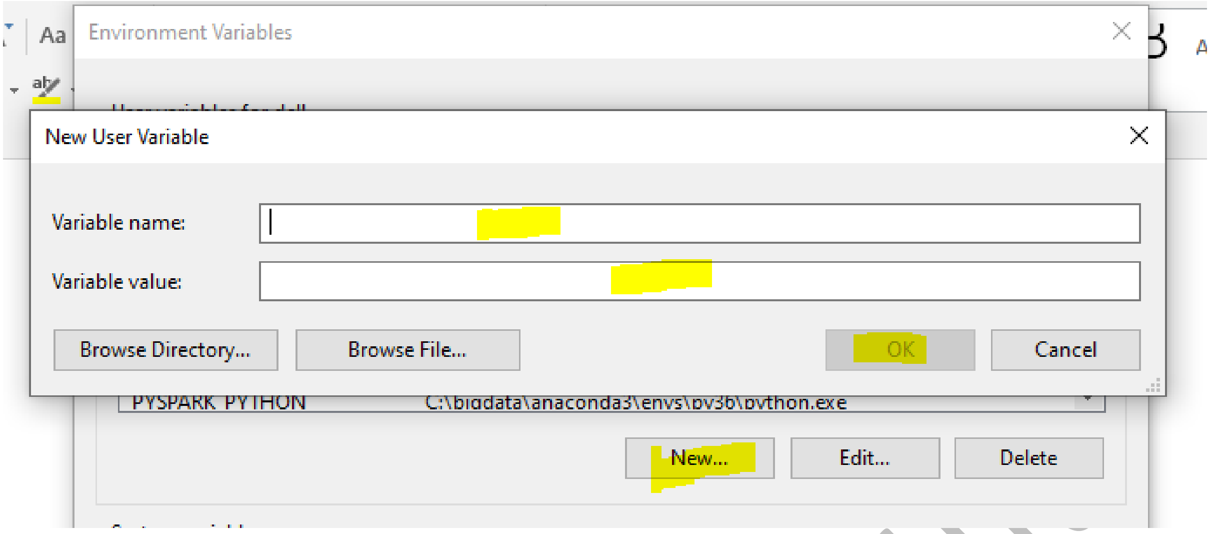
Similarly Hadoop too

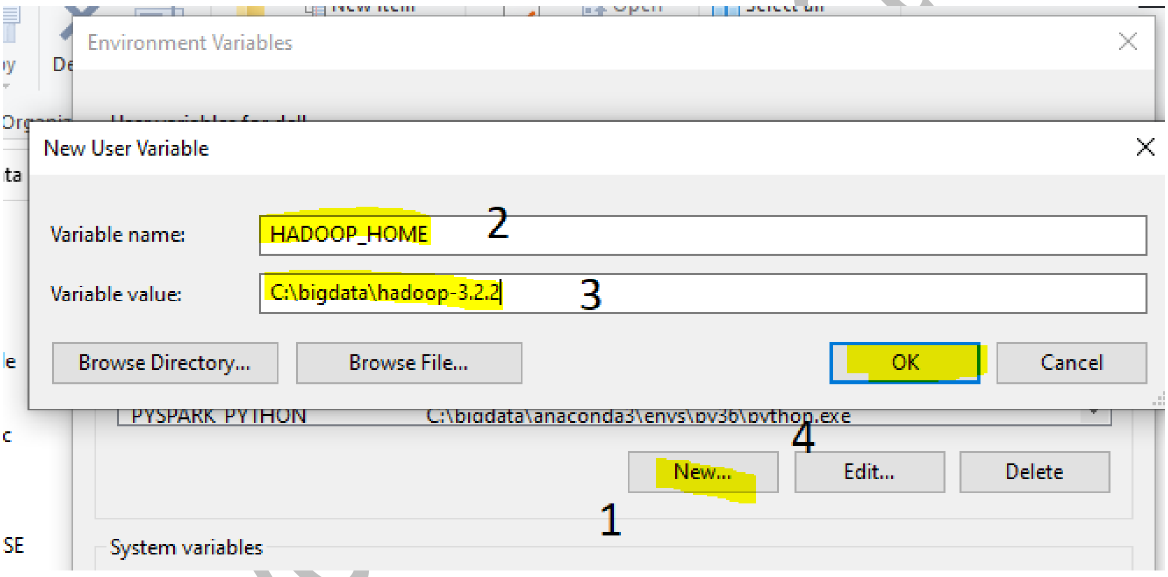


This PC > Local Disk (C:) > bigdata > hadoop-3.2.2

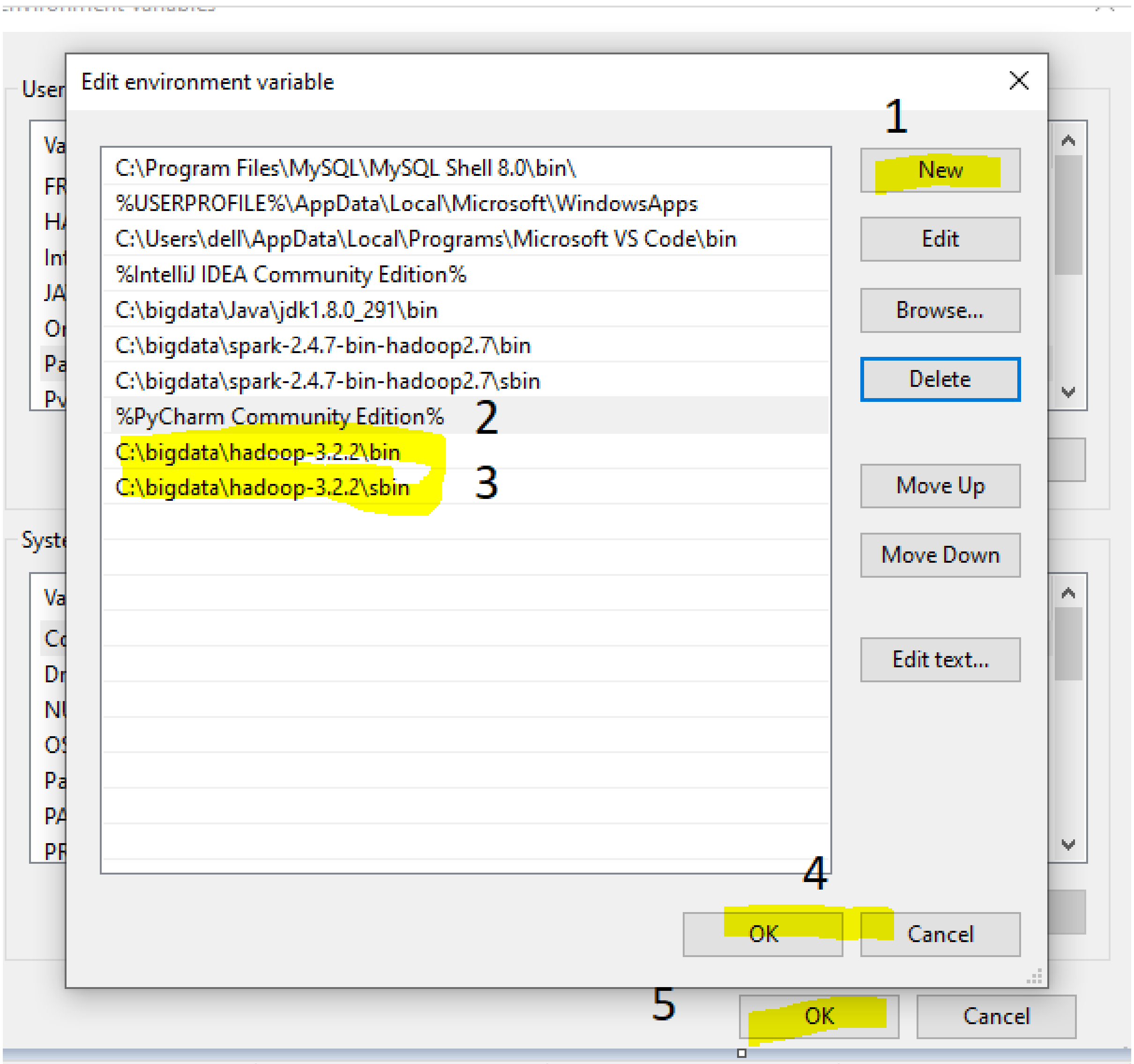
Name	Date modified	Type
bin	21-12-2021 12:00	File folder

Copy this path and paste in environment variable





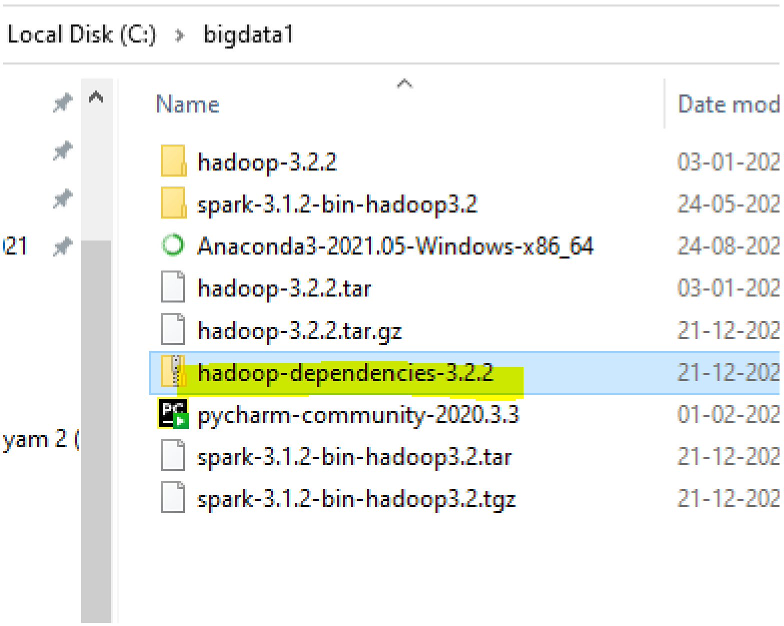
Next Same way path also just paste in path



Edit environment variable

C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Users\dell\AppData\Local\Programs\Microsoft VS Code\bin
%IntelliJ IDEA Community Edition%
C:\bigdata\Java\jdk1.8.0_291\bin
C:\bigdata\spark-2.4.7-bin-hadoop2.7\bin
C:\bigdata\spark-2.4.7-bin-hadoop2.7\sbin
%PyCharm Community Edition%
C:\bigdata\hadoop-3.2.2\bin
C:\bigdata\hadoop-3.2.2\sbin

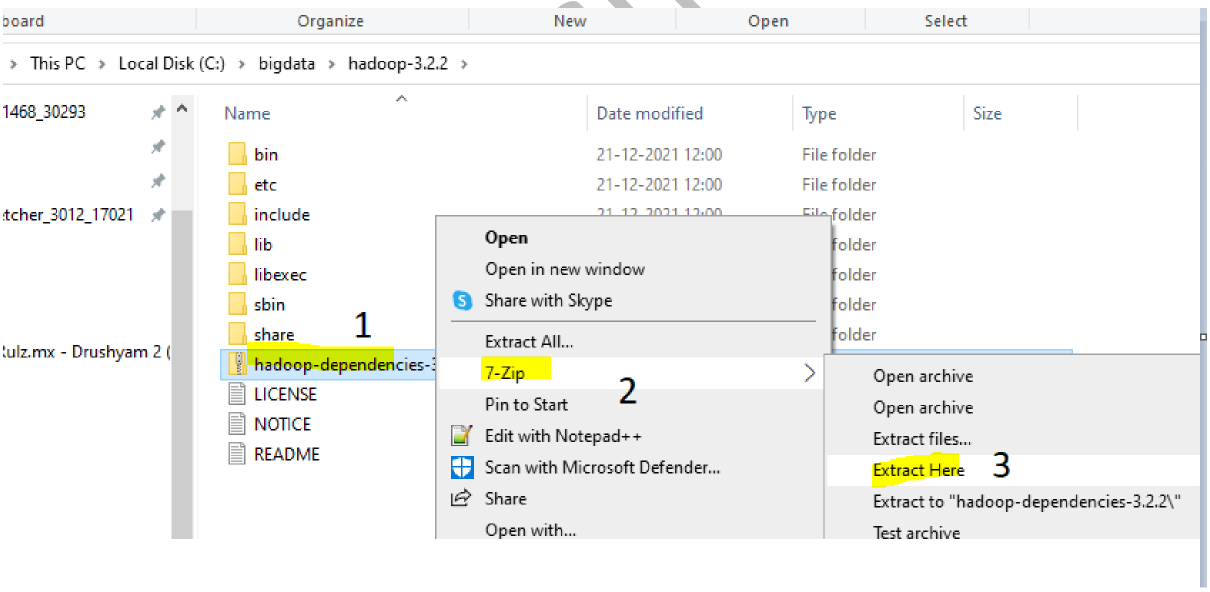
New


Local Disk (C:) › bigdata1

Name	Date mod
hadoop-3.2.2	03-01-202
spark-3.1.2-bin-hadoop3.2	24-05-202
Anaconda3-2021.05-Windows-x86_64	24-08-202
hadoop-3.2.2.tar	03-01-202
hadoop-3.2.2.tar.gz	21-12-202
hadoop-dependencies-3.2.2	21-12-202
pycharm-community-2020.3.3	01-02-202
spark-3.1.2-bin-hadoop3.2.tar	21-12-202
spark-3.1.2-bin-hadoop3.2.tgz	21-12-202

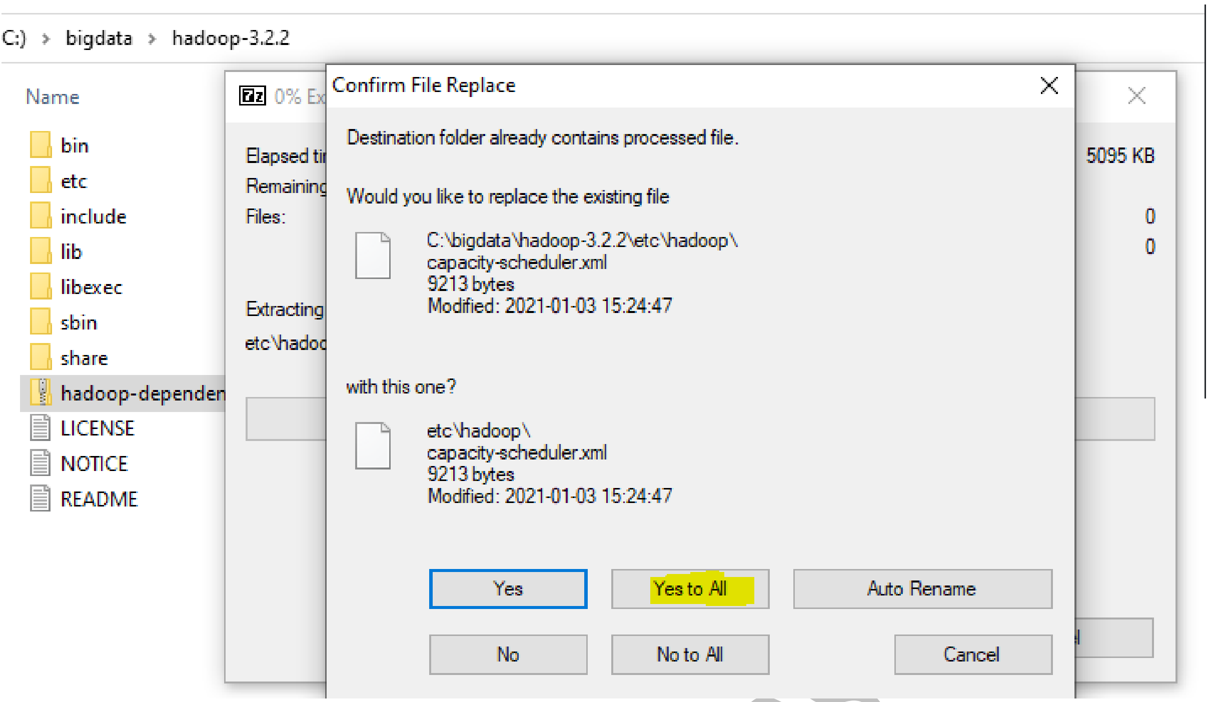
C:\bigdata1\hadoop-3.2.2 …

Paste this Hadoop-dependencies-322 zip file in Hadoop 3.2.2 folder … now right click and extract it.


› This PC › Local Disk (C:) › bigdata › hadoop-3.2.2 ›

Name	Date modified	Type	Size
bin	21-12-2021 12:00	File folder	
etc	21-12-2021 12:00	File folder	
include	21-12-2021 12:00	File folder	
lib		folder	
libexec		folder	
sbin		folder	
share		folder	
hadoop-dependencies-3			
LICENSE			
NOTICE			
README			

You will get replace all , Plese click replace all



Confirm File Replace

Destination folder already contains processed file.

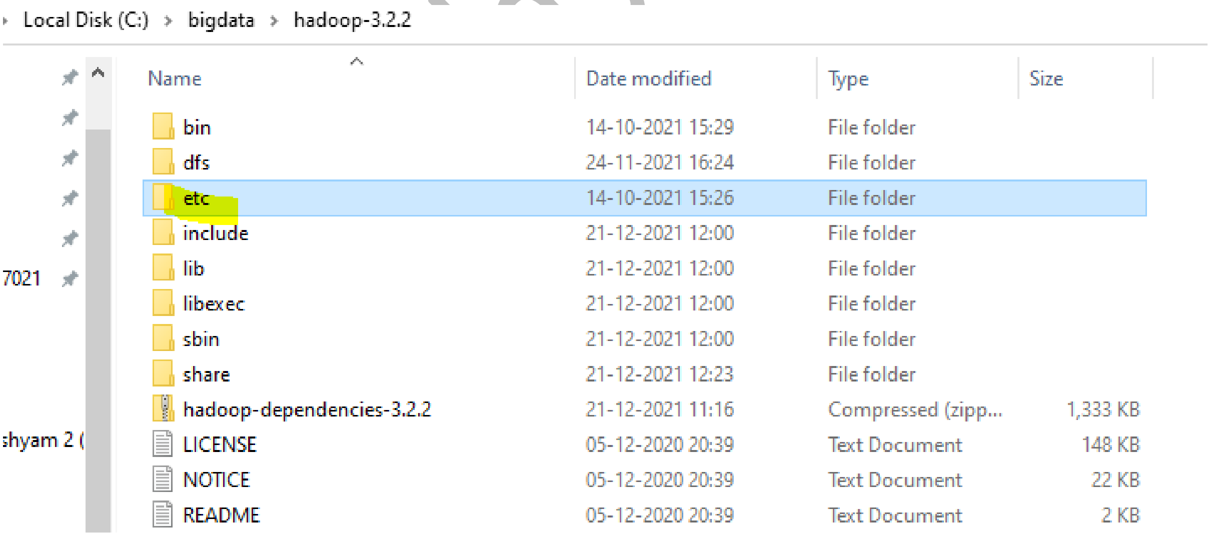
Would you like to replace the existing file

C:\bigdata\hadoop-3.2.2\etc\hadoop\
capacity-scheduler.xml
9213 bytes
Modified: 2021-01-03 15:24:47

with this one?

etc\hadoop\
capacity-scheduler.xml
9213 bytes
Modified: 2021-01-03 15:24:47

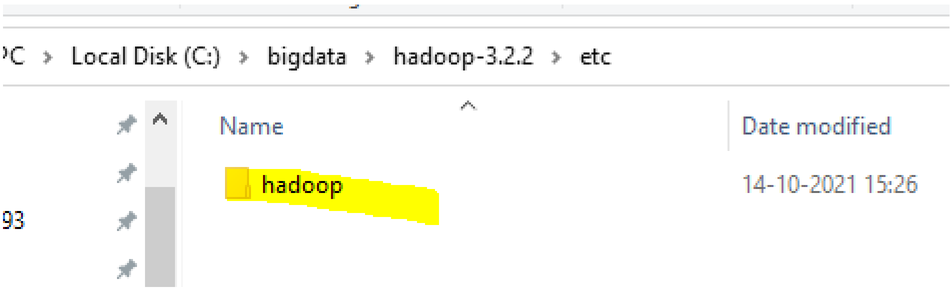
Now somethime if you want to change namenode and datanode place please modify these



Local Disk (C:) › bigdata › hadoop-3.2.2

Name	Date modified	Type	Size
bin	14-10-2021 15:29	File folder	
dfs	24-11-2021 16:24	File folder	
etc	14-10-2021 15:26	File folder	
include	21-12-2021 12:00	File folder	
lib	21-12-2021 12:00	File folder	
libexec	21-12-2021 12:00	File folder	
sbin	21-12-2021 12:00	File folder	
share	21-12-2021 12:23	File folder	
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
LICENSE	05-12-2020 20:39	Text Document	148 KB
NOTICE	05-12-2020 20:39	Text Document	22 KB
README	05-12-2020 20:39	Text Document	2 KB

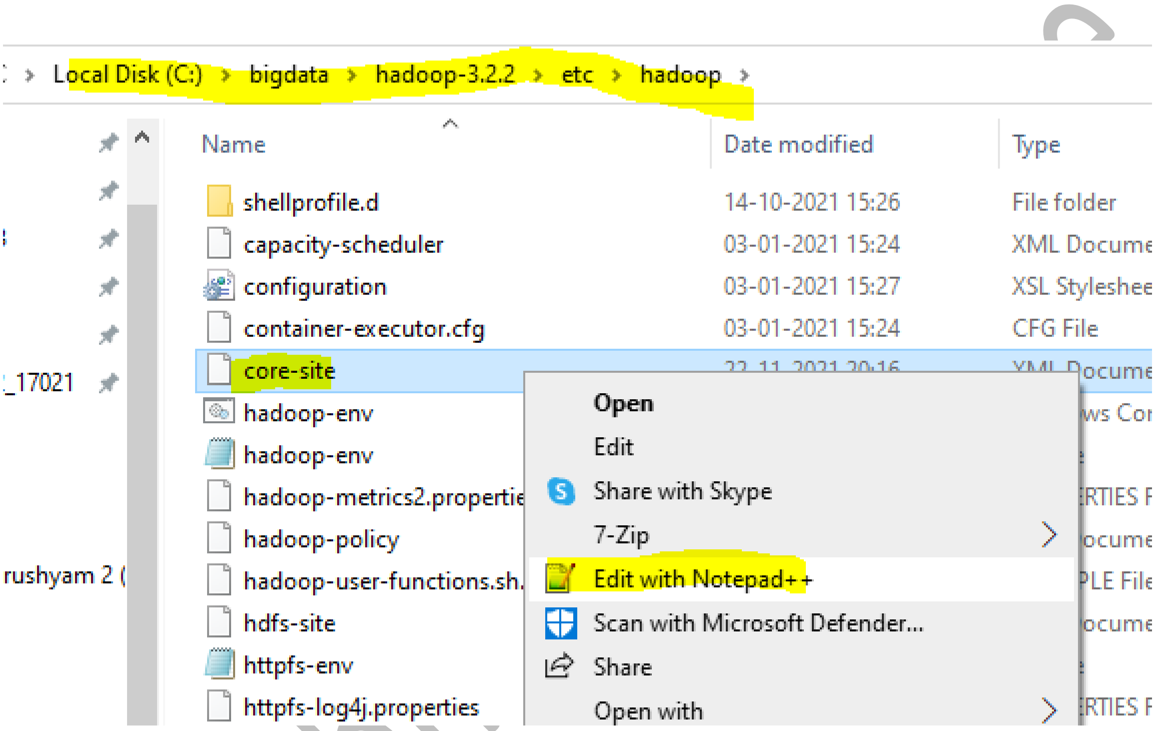
Click Hadoop



PC > Local Disk (C:) > bigdata > hadoop-3.2.2 > etc

Name	Date modified
hadoop	14-10-2021 15:26

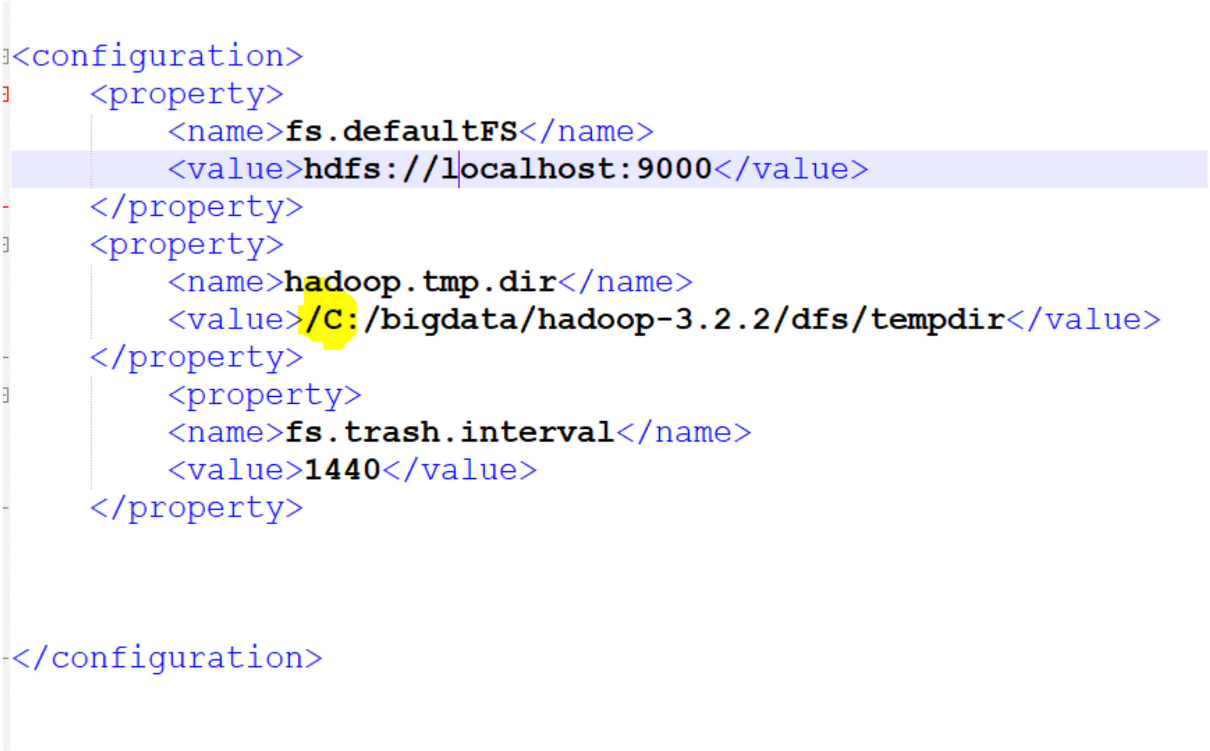
Now right click core.site.xml edit with notepad++



: > Local Disk (C:) > bigdata > hadoop-3.2.2 > etc > hadoop >

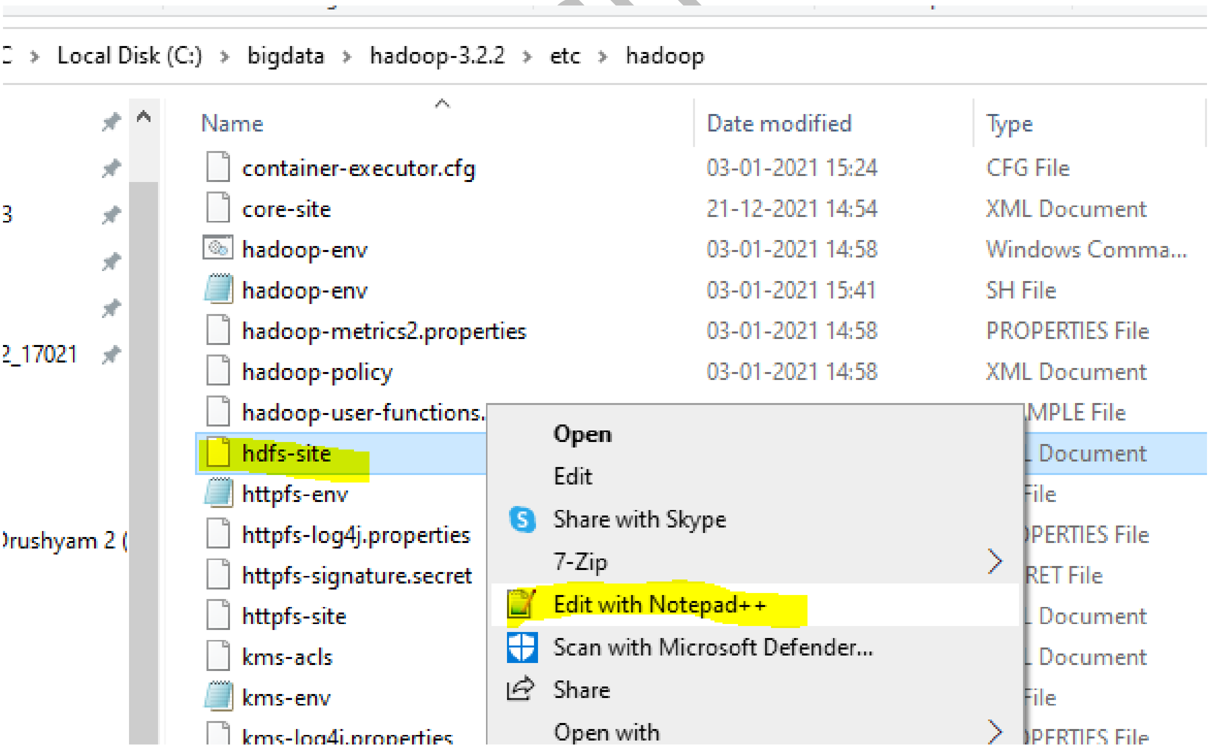
Name	Date modified	Type
shellprofile.d	14-10-2021 15:26	File folder
capacity-scheduler	03-01-2021 15:24	XML Docume
configuration	03-01-2021 15:27	XSL Styleshee
container-executor.cfg	03-01-2021 15:24	CFG File
core-site	22-11-2021 20:16	XML Docume
hadoop-env		ws Cor
hadoop-env		
hadoop-metrics2.propertie		RTIES F
hadoop-policy		ocume
hadoop-user-functions.sh.		PLE File
hdfs-site		ocume
httpfs-env		
httpfs-log4j.properties		RTIES F

If you want to change namenodde or datanode other information pls change these



```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/C:/bigdata/hadoop-3.2.2/dfs/tempdir</value>
    </property>
        <property>
        <name>fs.trash.interval</name>
        <value>1440</value>
    </property>


</configuration>
```

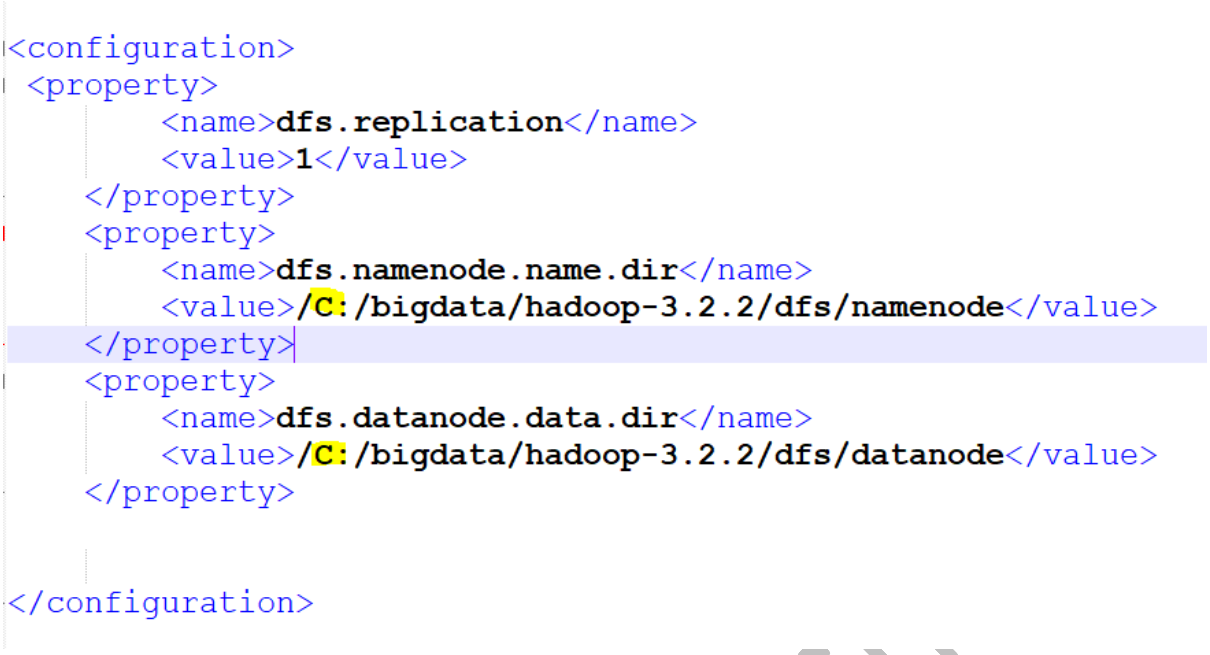
In hdfs-site.xml also pls change



Local Disk (C:) › bigdata › hadoop-3.2.2 › etc › hadoop

Name	Date modified	Type
container-executor.cfg	03-01-2021 15:24	CFG File
core-site	21-12-2021 14:54	XML Document
hadoop-env	03-01-2021 14:58	Windows Comma...
hadoop-env	03-01-2021 15:41	SH File
hadoop-metrics2.properties	03-01-2021 14:58	PROPERTIES File
hadoop-policy	03-01-2021 14:58	XML Document
hadoop-user-functions.		MPLE File
hdfs-site		Document
httpfs-env		File
httpfs-log4j.properties		PERTIES File
httpfs-signature.secret		RET File
httpfs-site		Document
kms-acls		Document
kms-env		File
kms-log4j.properties		PERTIES File

If you want to change pls change



```
<configuration>
 <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/C:/bigdata/hadoop-3.2.2/dfs/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/C:/bigdata/hadoop-3.2.2/dfs/datanode</value>
    </property>


</configuration>
```

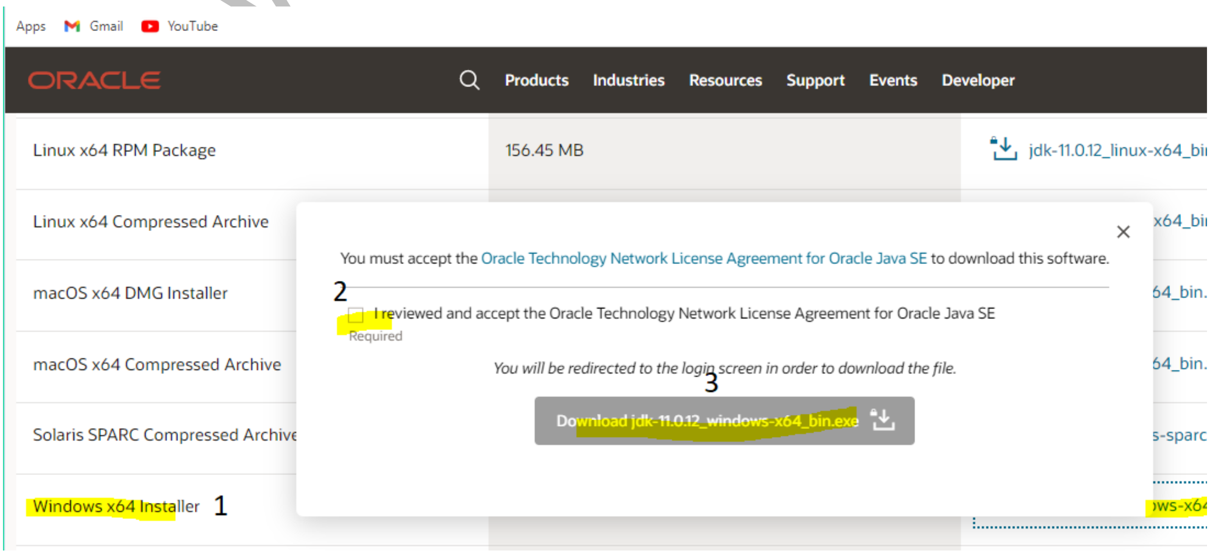
Now save core-site.xml and hdfs-site.xml

That’s it 99.9% Hadoop installation completed. Verification do it later

Please note if you want to do anything Java 11 is mandatory so install **Java 11 and Anaconda**

Download java from this link

https://www.oracle.com/in/java/technologies/javase/jdk11-archive-downloads.html



Linux x64 RPM Package	156.45 MB	jdk-11.0.12_linux-x64_bi
Linux x64 Compressed Archive		x64_bi
macOS x64 DMG Installer		64_bin.
macOS x64 Compressed Archive		64_bin.
Solaris SPARC Compressed Archive		s-sparc
Windows x64 Installer 1		ws-x64

You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

2 □ I reviewed and accept the Oracle Technology Network License Agreement for Oracle Java SE

Required

You will be redirected to the login screen in order to download the file.

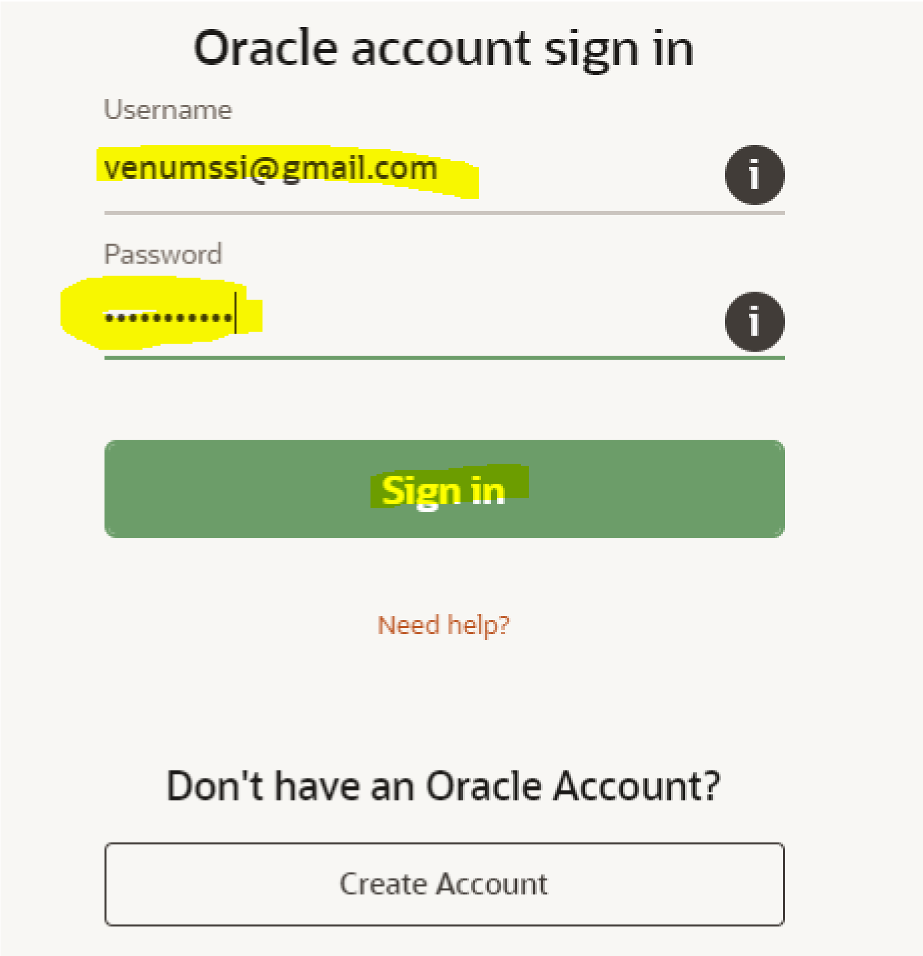
3

Download jdk-11.0.12_windows-x64_bin.exe

Now you will get popup like this enter these credentials

venumssi@gmail.om

Opassword.1



Oracle account sign in

Username

venumssi@gmail.com

Password

••••••••••••

Sign in

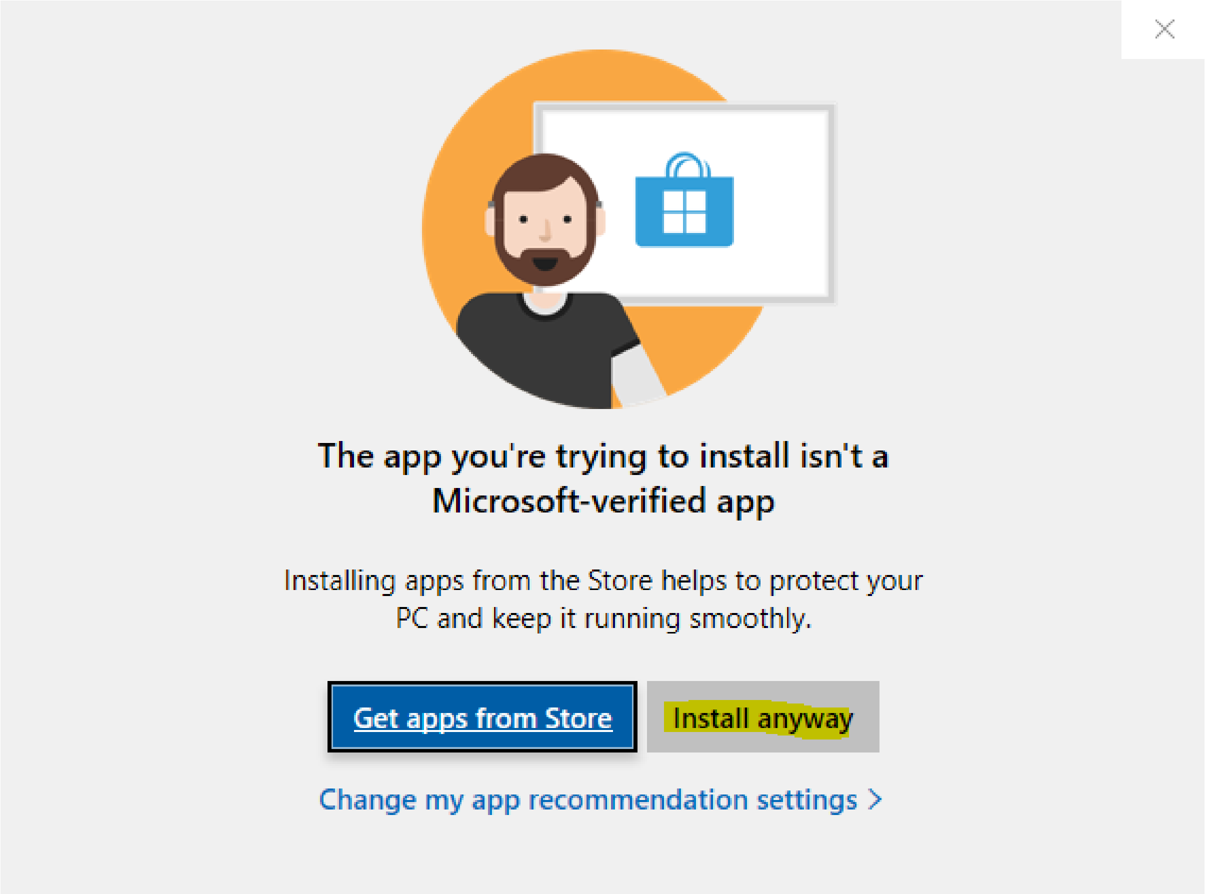
Need help?

Don't have an Oracle Account?

Create Account

Now after download click to install java 11

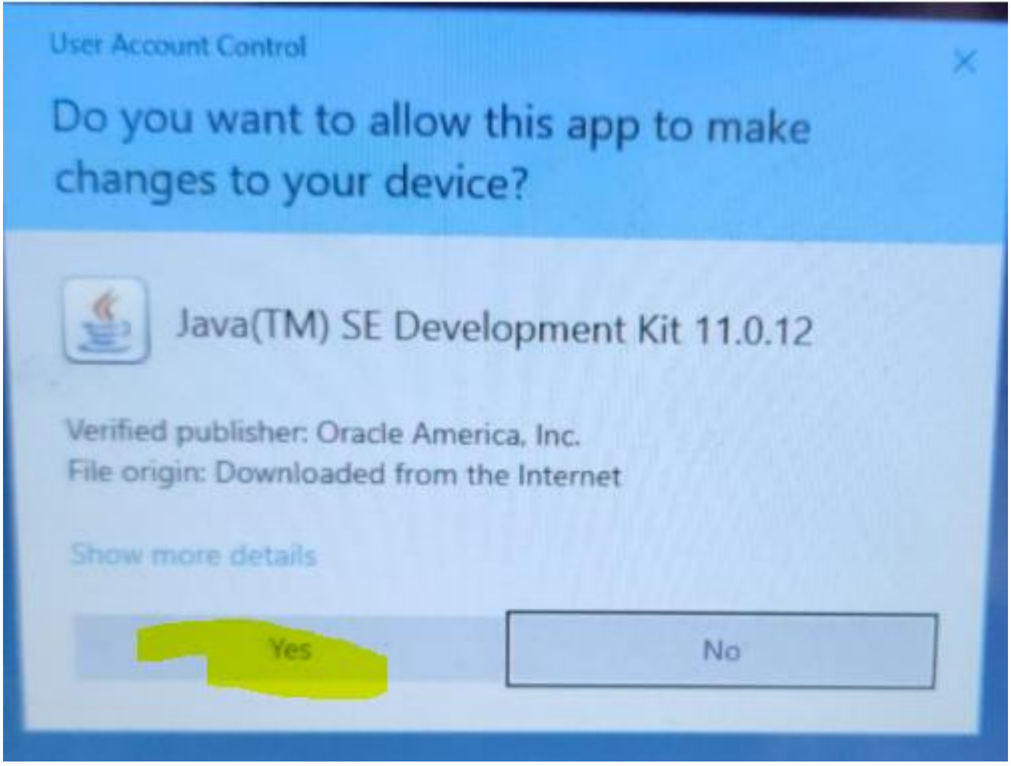
You will get like this



The app you're trying to install isn't a Microsoft-verified app

Installing apps from the Store helps to protect your PC and keep it running smoothly.

Install anyway next



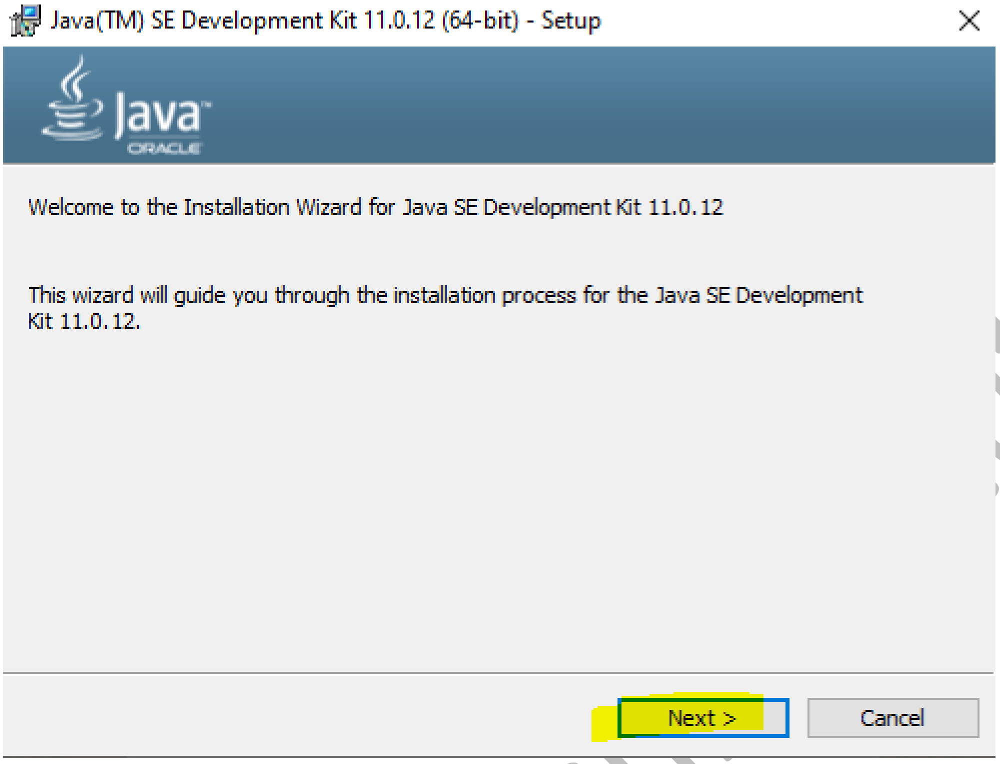
User Account Control

Do you want to allow this app to make changes to your device?

Java(TM) SE Development Kit 11.0.12

Verified publisher: Oracle America, Inc.
File origin: Downloaded from the Internet

Click yes

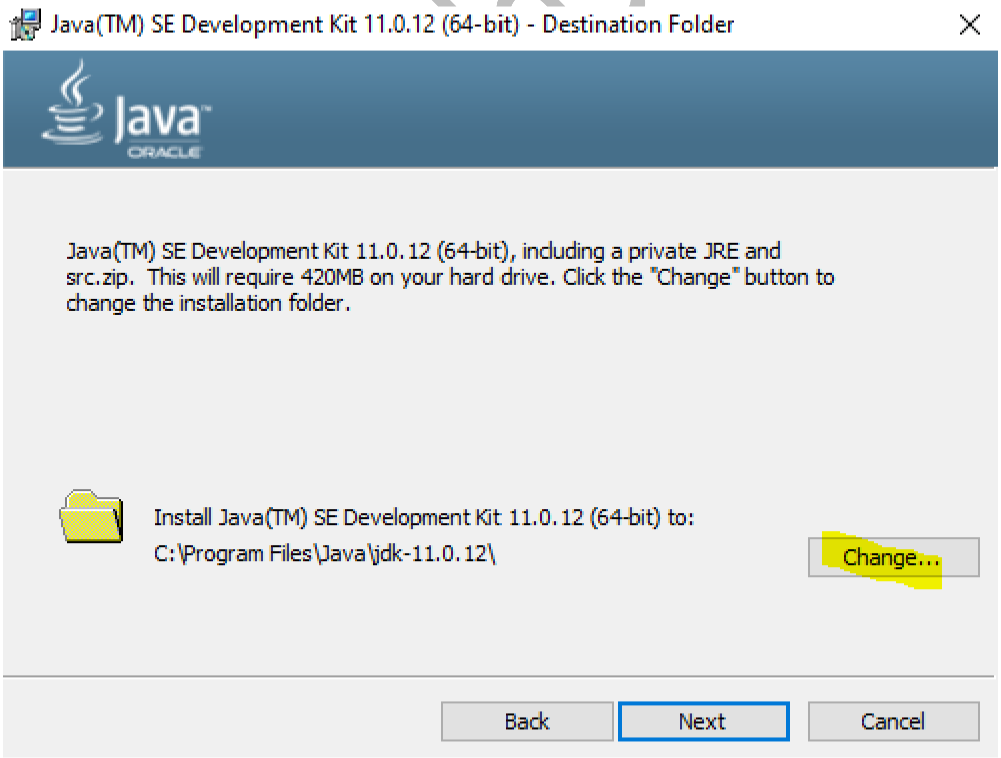


Java(TM) SE Development Kit 11.0.12 (64-bit) - Setup

Welcome to the Installation Wizard for Java SE Development Kit 11.0.12

This wizard will guide you through the installation process for the Java SE Development Kit 11.0.12.

Next



Java(TM) SE Development Kit 11.0.12 (64-bit) - Destination Folder

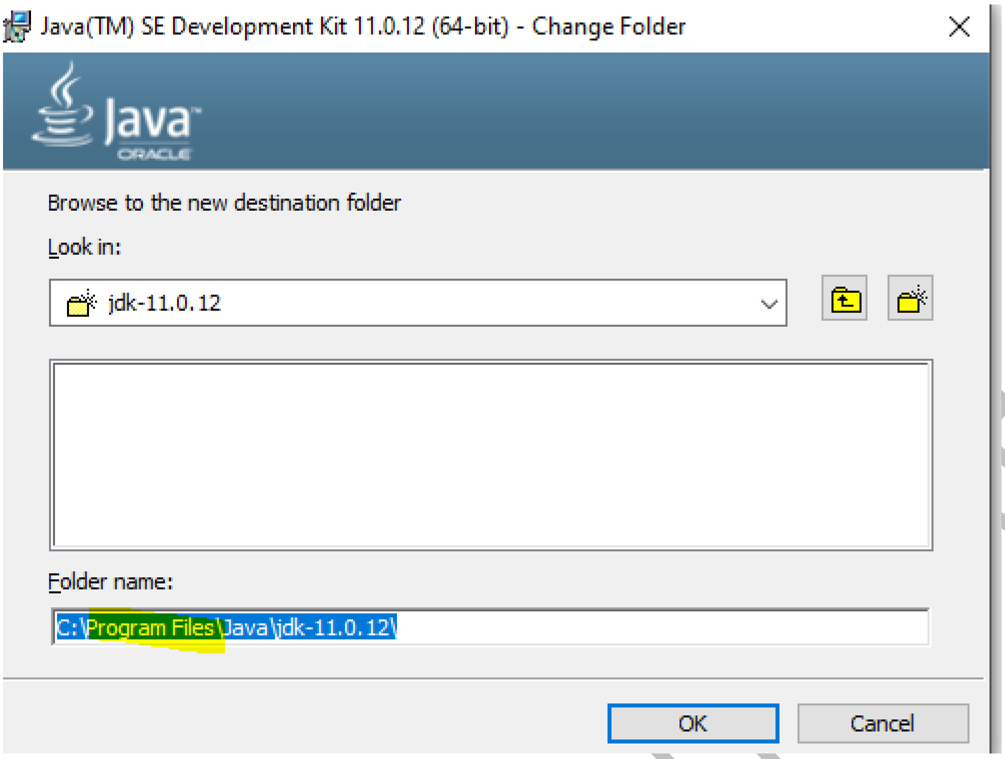
Java(TM) SE Development Kit 11.0.12 (64-bit), including a private JRE and src.zip. This will require 420MB on your hard drive. Click the "Change" button to change the installation folder.

Install Java(TM) SE Development Kit 11.0.12 (64-bit) to:
C:\Program Files\Java\jdk-11.0.12\

Click Change

Next

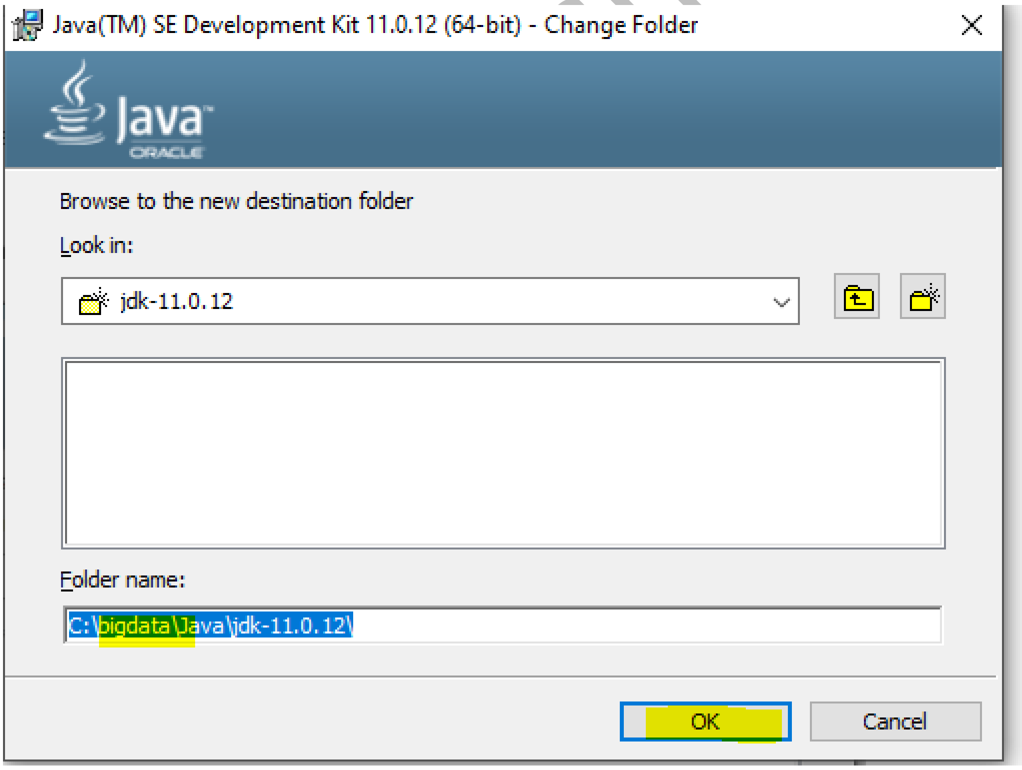
Now rename this as bigdata instead of Program Files


Java(TM) SE Development Kit 11.0.12 (64-bit) - Change Folder

Browse to the new destination folder

Look in: jdk-11.0.12

Folder name: C:\Program Files\Java\jdk-11.0.12\

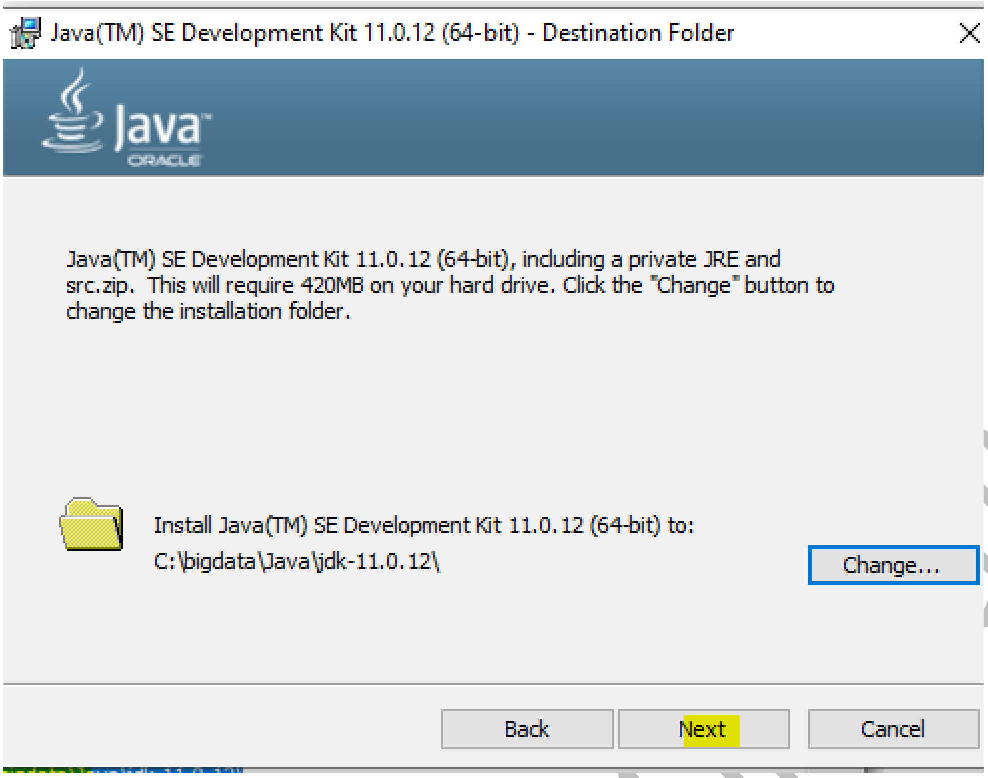

Java(TM) SE Development Kit 11.0.12 (64-bit) - Change Folder

Browse to the new destination folder

Look in: jdk-11.0.12

Folder name: C:\bigdata\Java\jdk-11.0.12\

Now click Ok



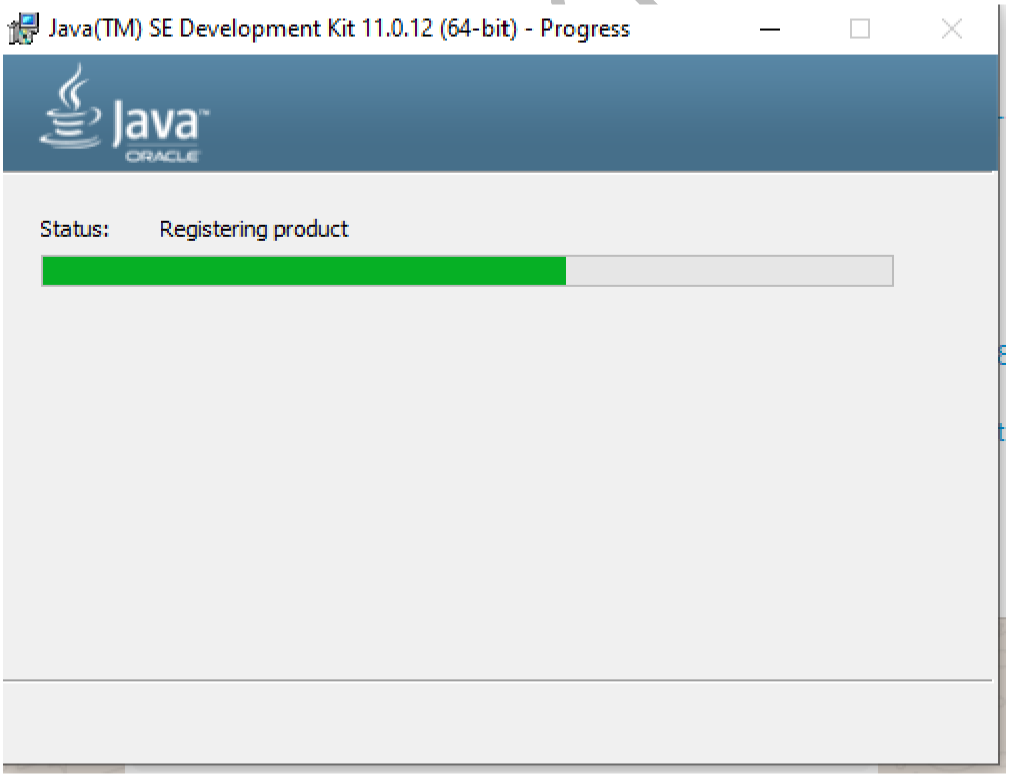
Java(TM) SE Development Kit 11.0.12 (64-bit) - Destination Folder

Java(TM) SE Development Kit 11.0.12 (64-bit), including a private JRE and src.zip. This will require 420MB on your hard drive. Click the "Change" button to change the installation folder.

Install Java(TM) SE Development Kit 11.0.12 (64-bit) to:
C:\bigdata\Java\jdk-11.0.12\

Now click Next

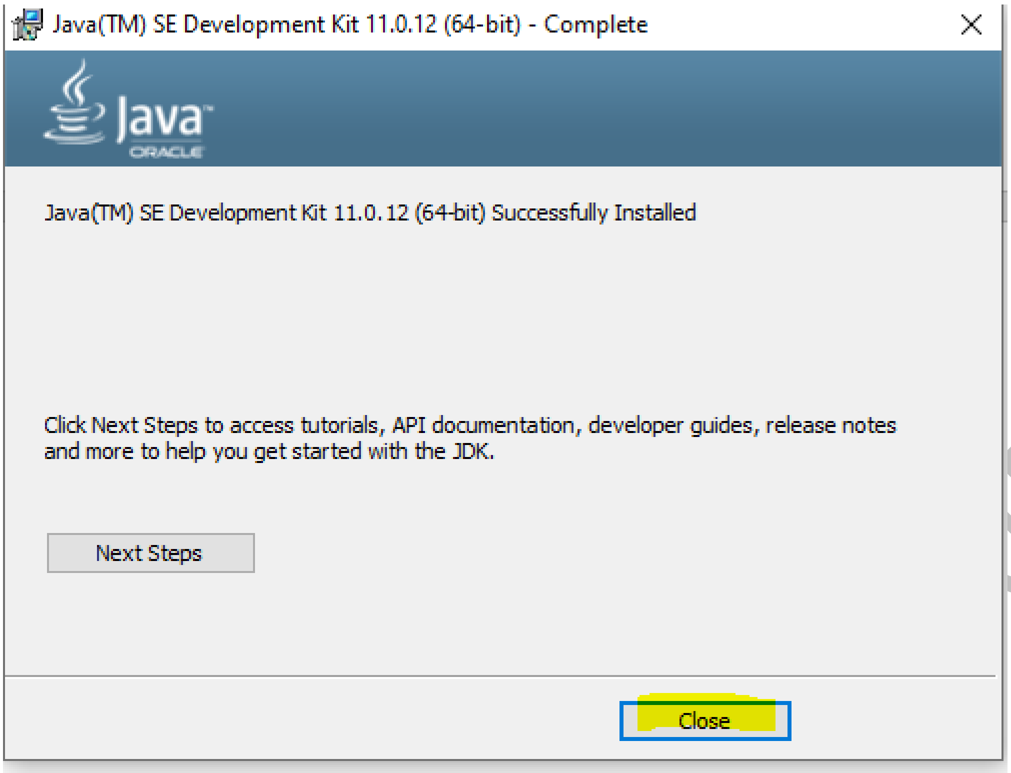
Now java 11 installing in ur system wait sometime



Java(TM) SE Development Kit 11.0.12 (64-bit) - Progress

Status: Registering product

Now click Close



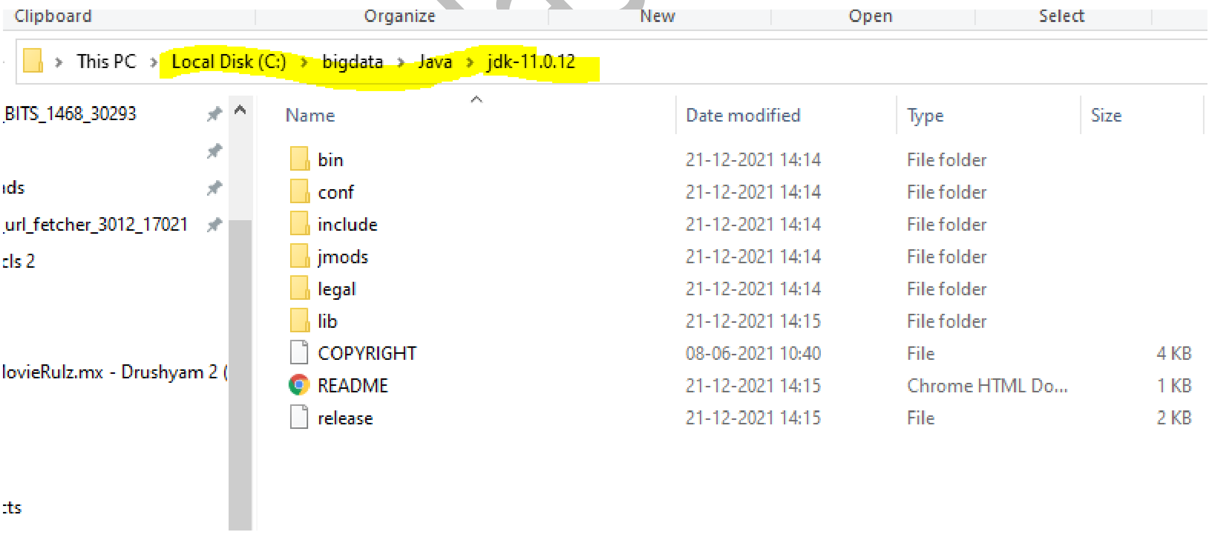
Java(TM) SE Development Kit 11.0.12 (64-bit) - Complete

Java(TM) SE Development Kit 11.0.12 (64-bit) Successfully Installed

Click Next Steps to access tutorials, API documentation, developer guides, release notes and more to help you get started with the JDK.

Now java installation completed, I recommend configure in environment variable

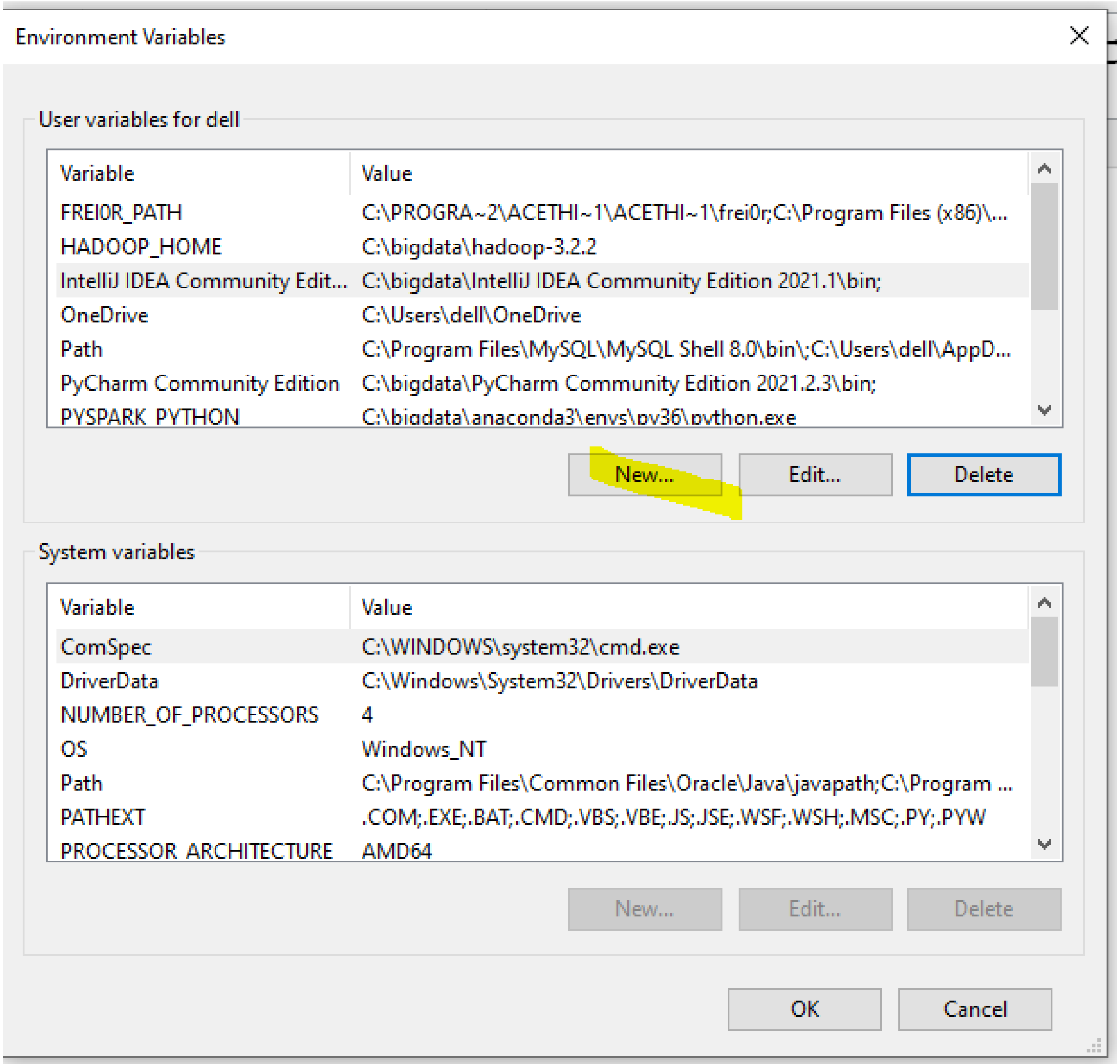
Means



This PC › Local Disk (C:) › bigdata › Java › jdk-11.0.12

Name	Date modified	Type	Size
bin	21-12-2021 14:14	File folder	
conf	21-12-2021 14:14	File folder	
include	21-12-2021 14:14	File folder	
jmods	21-12-2021 14:14	File folder	
legal	21-12-2021 14:14	File folder	
lib	21-12-2021 14:15	File folder	
COPYRIGHT	08-06-2021 10:40	File	4 KB
README	21-12-2021 14:15	Chrome HTML Do...	1 KB
release	21-12-2021 14:15	File	2 KB

Copy this path and paste inenvironment variable



Environment Variables

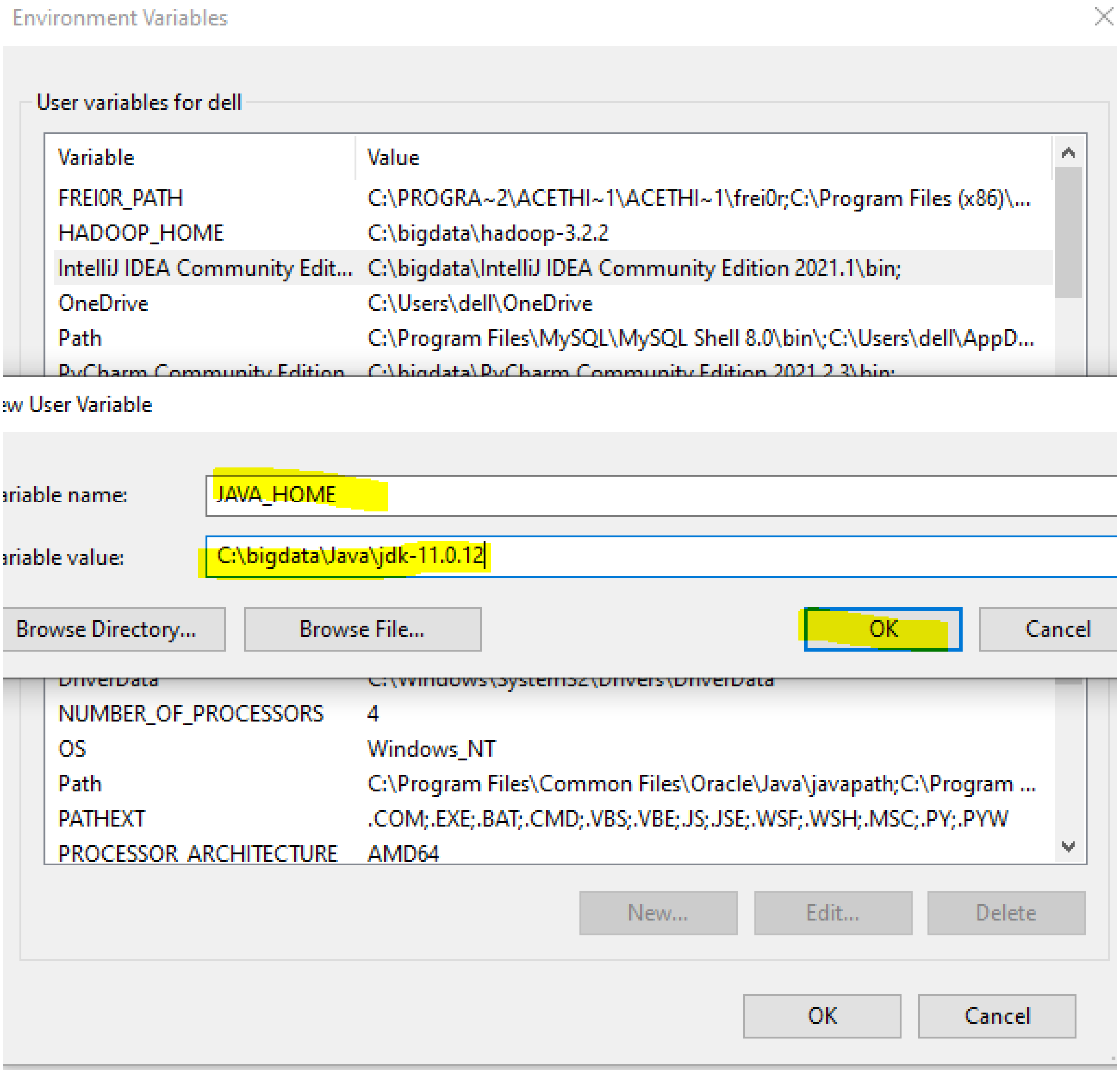
User variables for dell

Variable	Value
FREI0R_PATH	C:\PROGRA~2\ACETHI~1\ACETHI~1\frei0r;C:\Program Files (x86)\...
HADOOP_HOME	C:\bigdata\hadoop-3.2.2
IntelliJ IDEA Community Edit...	C:\bigdata\IntelliJ IDEA Community Edition 2021.1\bin;
OneDrive	C:\Users\dell\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\dell\AppD...
PyCharm Community Edition	C:\bigdata\PyCharm Community Edition 2021.2.3\bin;
PYSPARK_PYTHON	C:\bigdata\anaconda3\envs\py36\python.exe

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Program ...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
PROCESSOR_ARCHITECTURE	AMD64

JAVA_HOME must be capital (case sensitive)



Environment Variables

User variables for dell

Variable	Value
FREI0R_PATH	C:\PROGRA~2\ACETHI~1\ACETHI~1\frei0r;C:\Program Files (x86)\...
HADOOP_HOME	C:\bigdata\hadoop-3.2.2
IntelliJ IDEA Community Edit...	C:\bigdata\IntelliJ IDEA Community Edition 2021.1\bin;
OneDrive	C:\Users\dell\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\dell\AppD...
PyCharm Community Edition	C:\bigdata\PyCharm Community Edition 2021.2.3\bin;

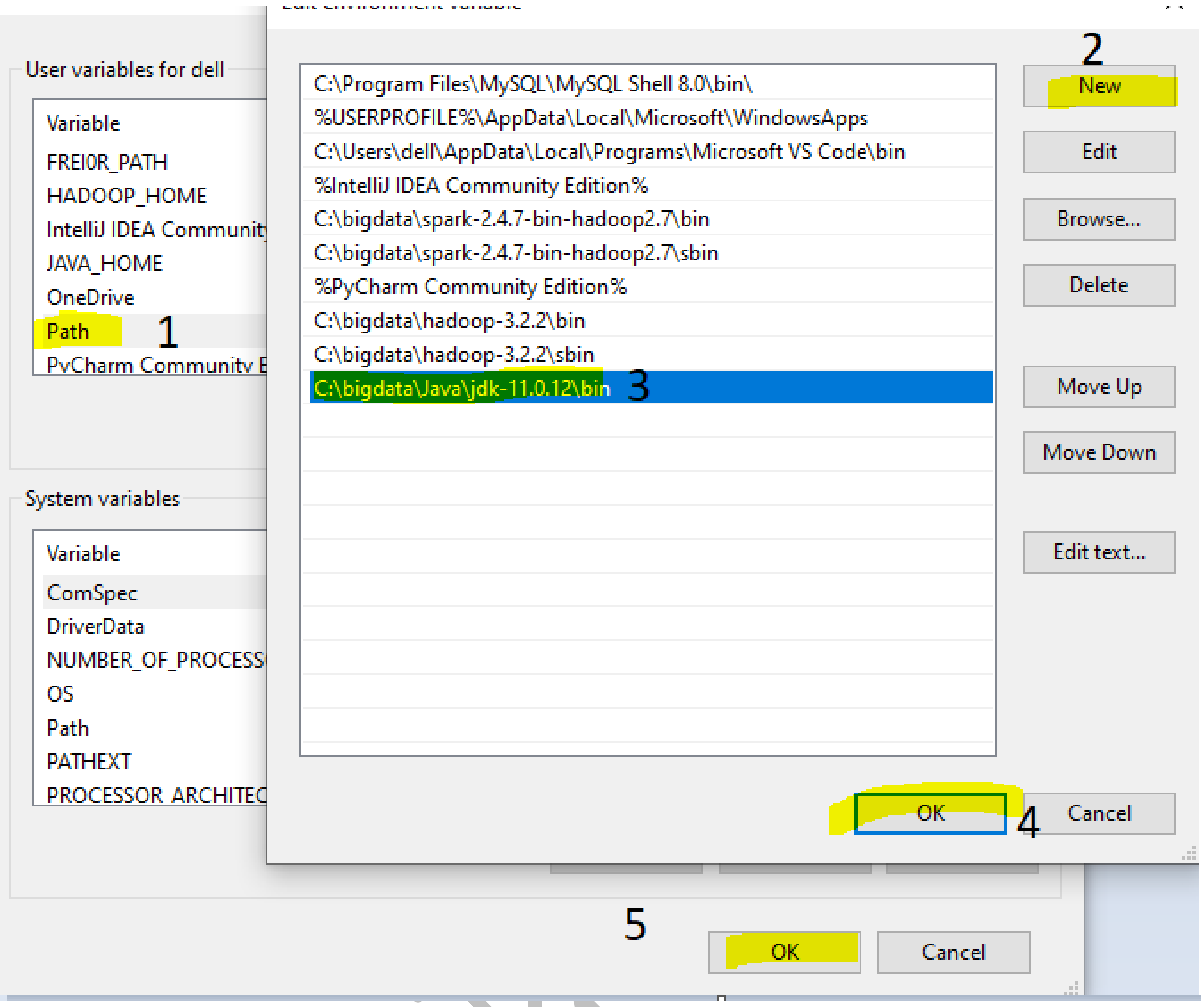
ew User Variable

ariable name: JAVA_HOME

ariable value: C:\bigdata\Java\jdk-11.0.12

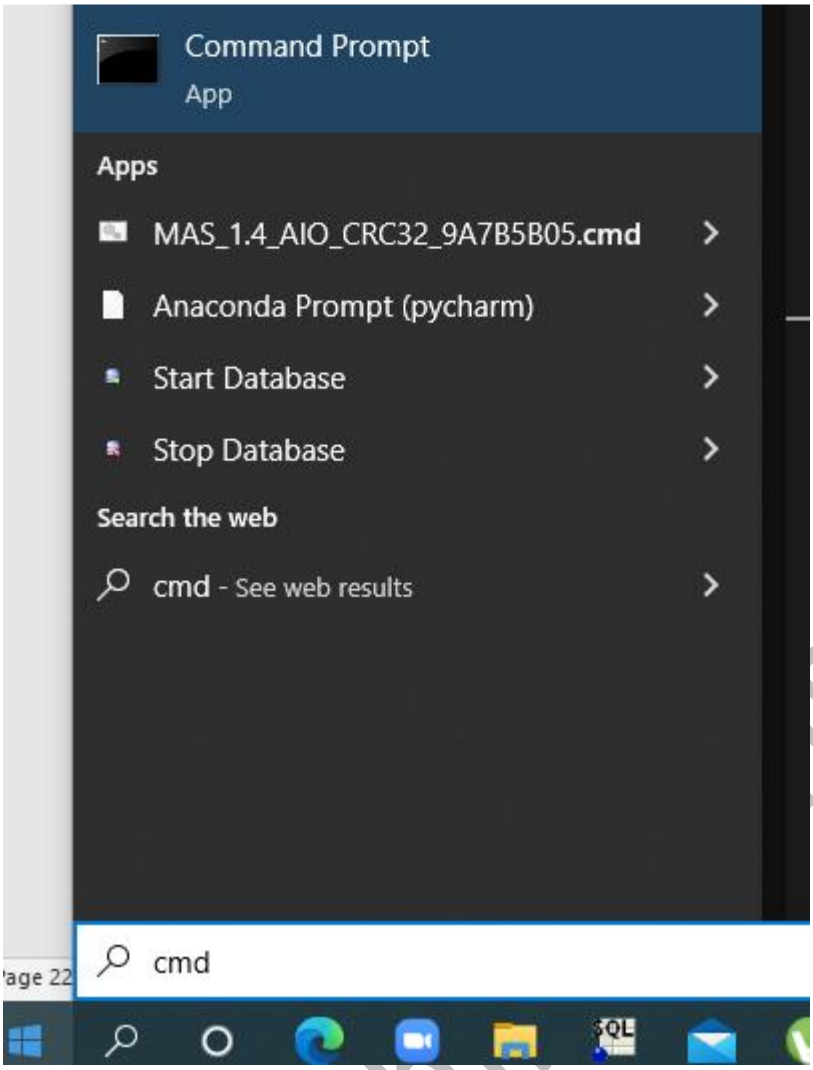
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Program ...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
PROCESSOR_ARCHITECTURE	AMD64

Configure in path as well

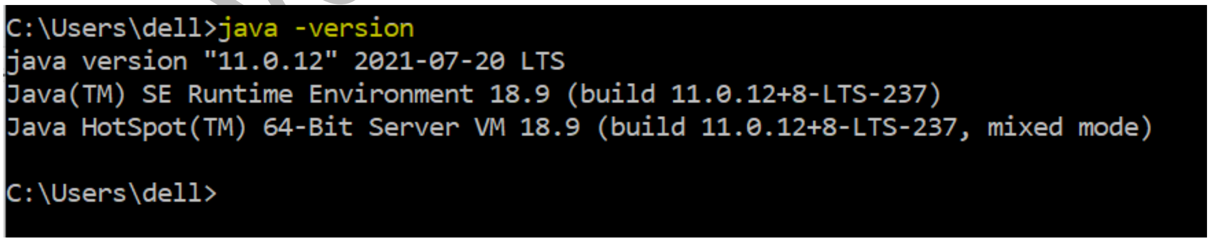


```
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Users\dell\AppData\Local\Programs\Microsoft VS Code\bin
%IntelliJ IDEA Community Edition%
C:\bigdata\spark-2.4.7-bin-hadoop2.7\bin
C:\bigdata\spark-2.4.7-bin-hadoop2.7\sbin
%PyCharm Community Edition%
C:\bigdata\hadoop-3.2.2\bin
C:\bigdata\hadoop-3.2.2\sbin
C:\bigdata\Java\jdk-11.0.12\bin
```

Testing purpose java installed or not click on start … cmd …

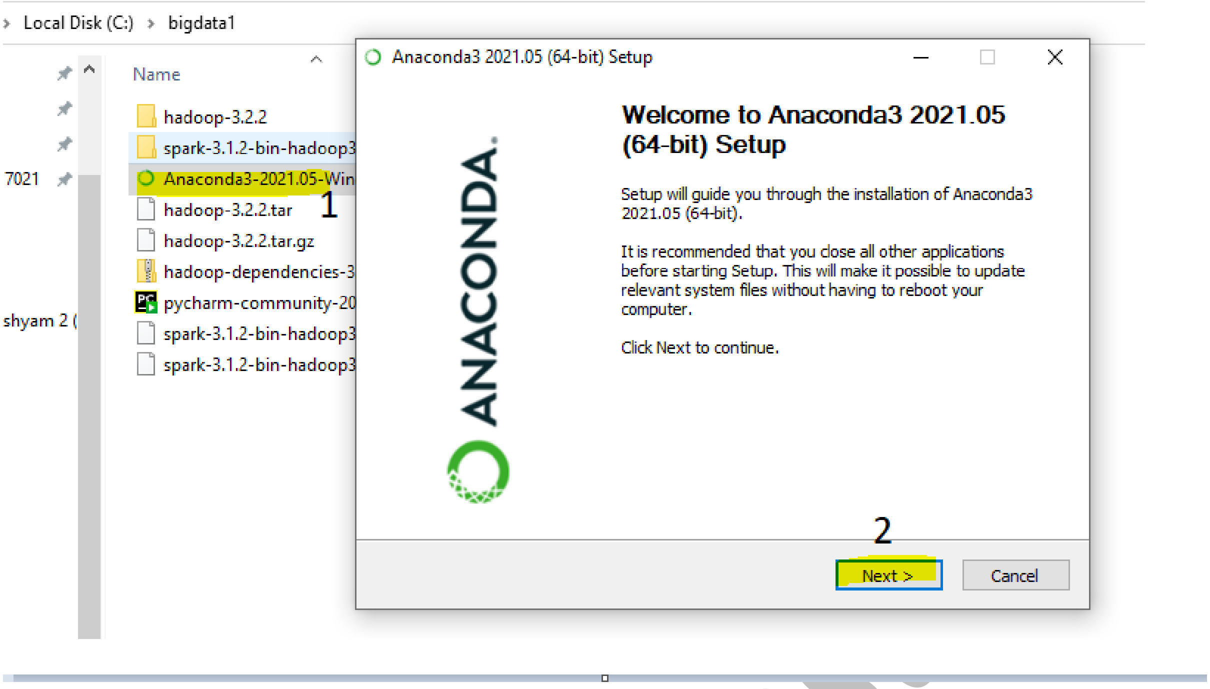


Java -version



```
C:\Users\dell>java -version
java version "11.0.12" 2021-07-20 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.12+8-LTS-237)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.12+8-LTS-237, mixed mode)

C:\Users\dell>
```

Please install anaconda as well

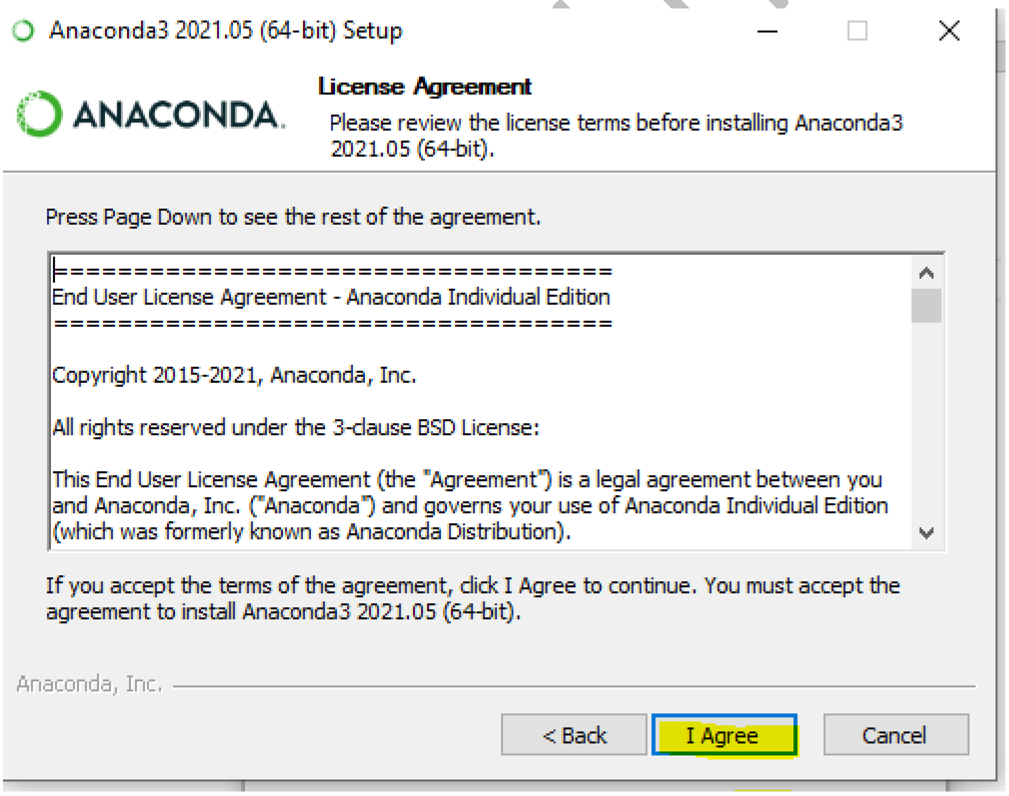


Welcome to Anaconda3 2021.05 (64-bit) Setup

Setup will guide you through the installation of Anaconda3 2021.05 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.



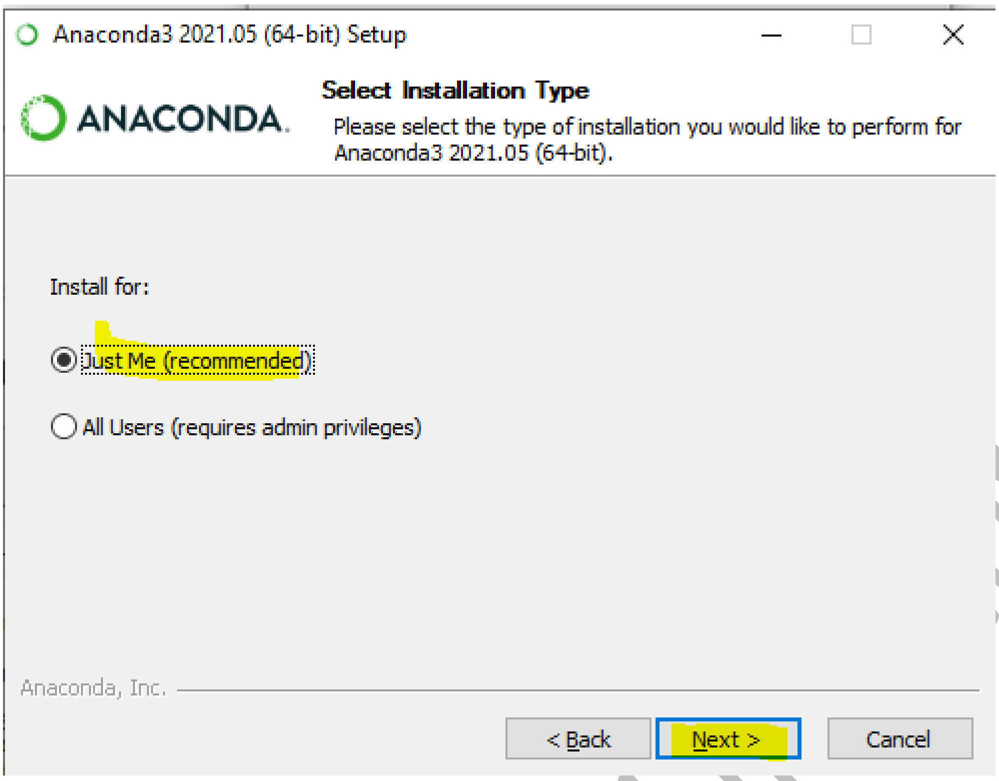
License Agreement

Please review the license terms before installing Anaconda3 2021.05 (64-bit).

Press Page Down to see the rest of the agreement.

```
==================================
End User License Agreement - Anaconda Individual Edition
==================================

Copyright 2015-2021, Anaconda, Inc.

All rights reserved under the 3-clause BSD License:

This End User License Agreement (the "Agreement") is a legal agreement between you
and Anaconda, Inc. ("Anaconda") and governs your use of Anaconda Individual Edition
(which was formerly known as Anaconda Distribution).
```

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Anaconda3 2021.05 (64-bit).



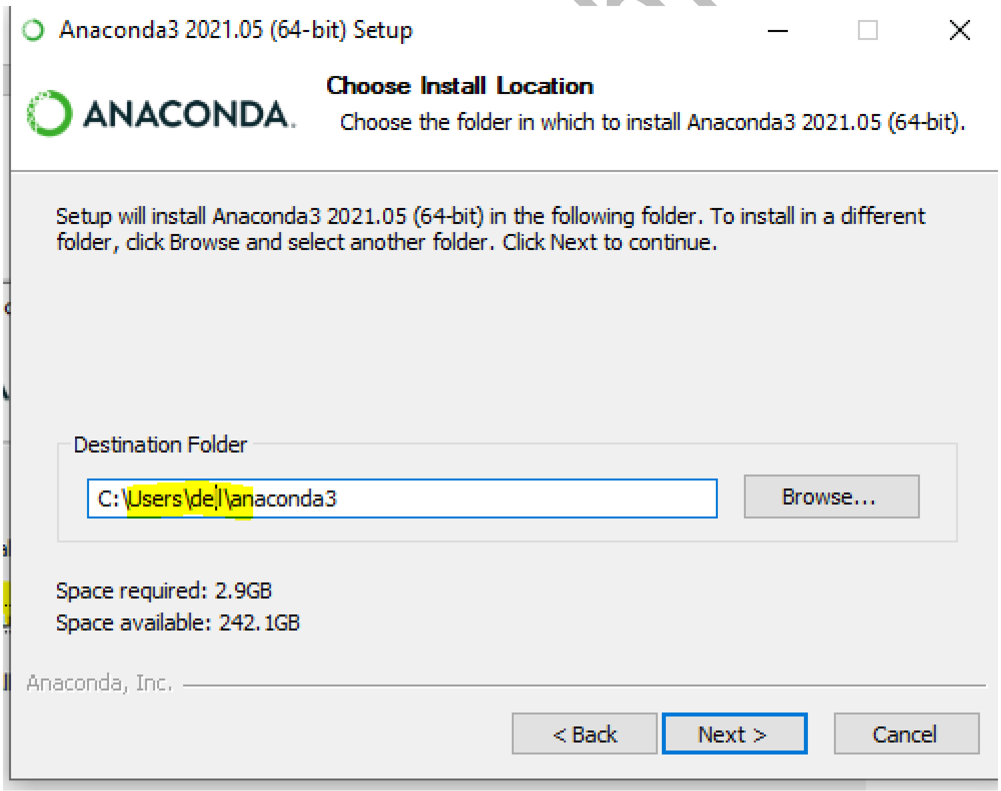
Select Installation Type

Please select the type of installation you would like to perform for Anaconda3 2021.05 (64-bit).

Install for:

● Just Me (recommended)

○ All Users (requires admin privileges)



Choose Install Location

Choose the folder in which to install Anaconda3 2021.05 (64-bit).

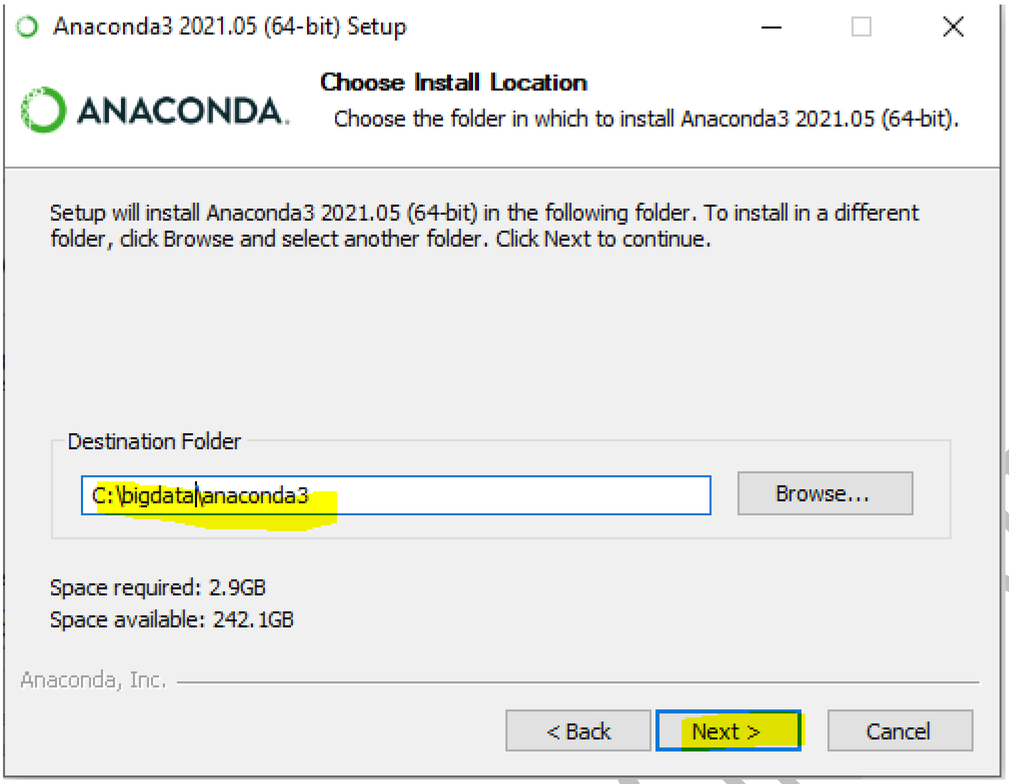
Setup will install Anaconda3 2021.05 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

C:\Users\dell\anaconda3

Space required: 2.9GB

Space available: 242.1GB



Anaconda3 2021.05 (64-bit) Setup

Choose Install Location

Choose the folder in which to install Anaconda3 2021.05 (64-bit).

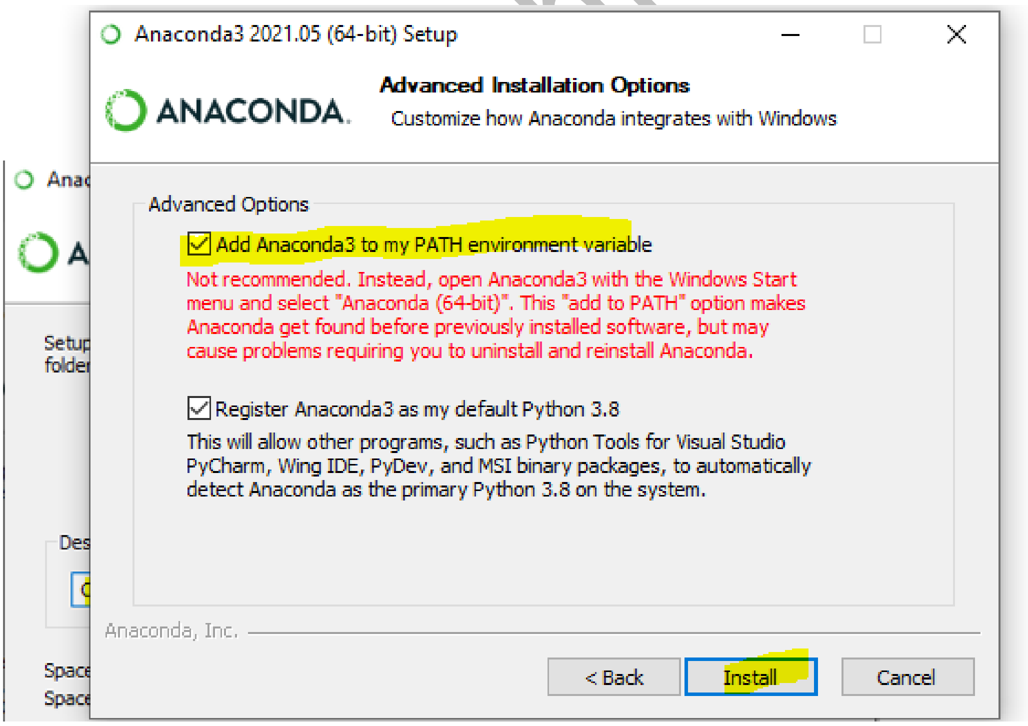
Setup will install Anaconda3 2021.05 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

C:\bigdata\anaconda3

Space required: 2.9GB
Space available: 242.1GB

Anaconda, Inc.



Anaconda3 2021.05 (64-bit) Setup

Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

☑ Add Anaconda3 to my PATH environment variable

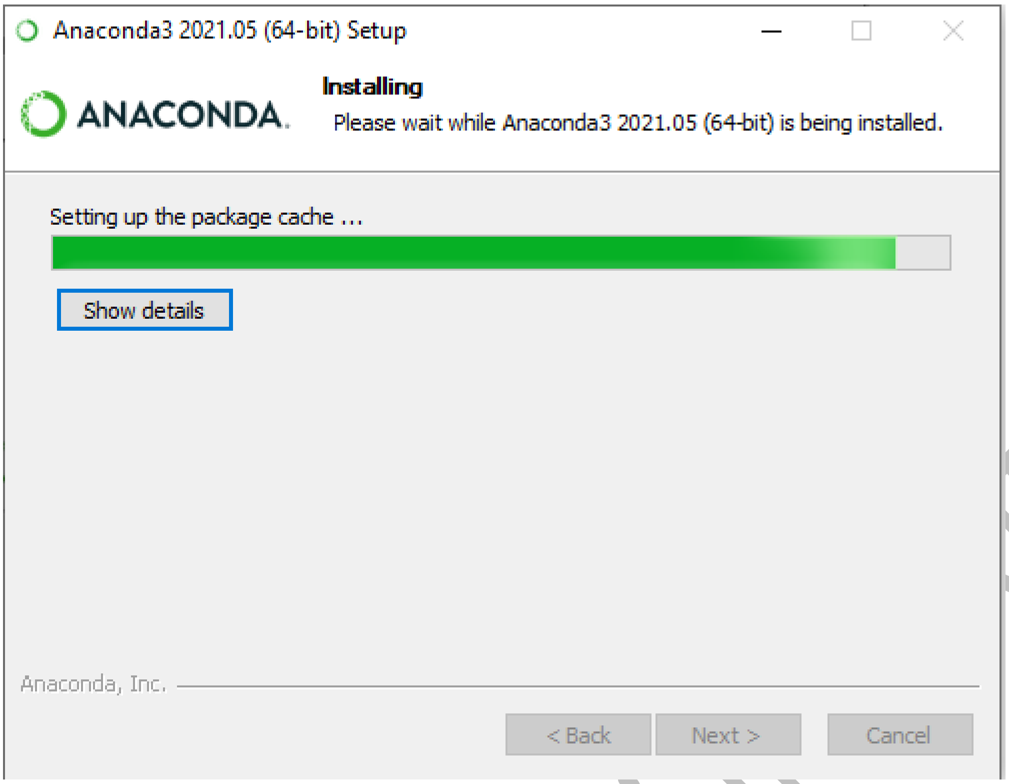
Not recommended. Instead, open Anaconda3 with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

☑ Register Anaconda3 as my default Python 3.8

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.8 on the system.

Anaconda, Inc.

Its take a lot of time



Anaconda3 2021.05 (64-bit) Setup

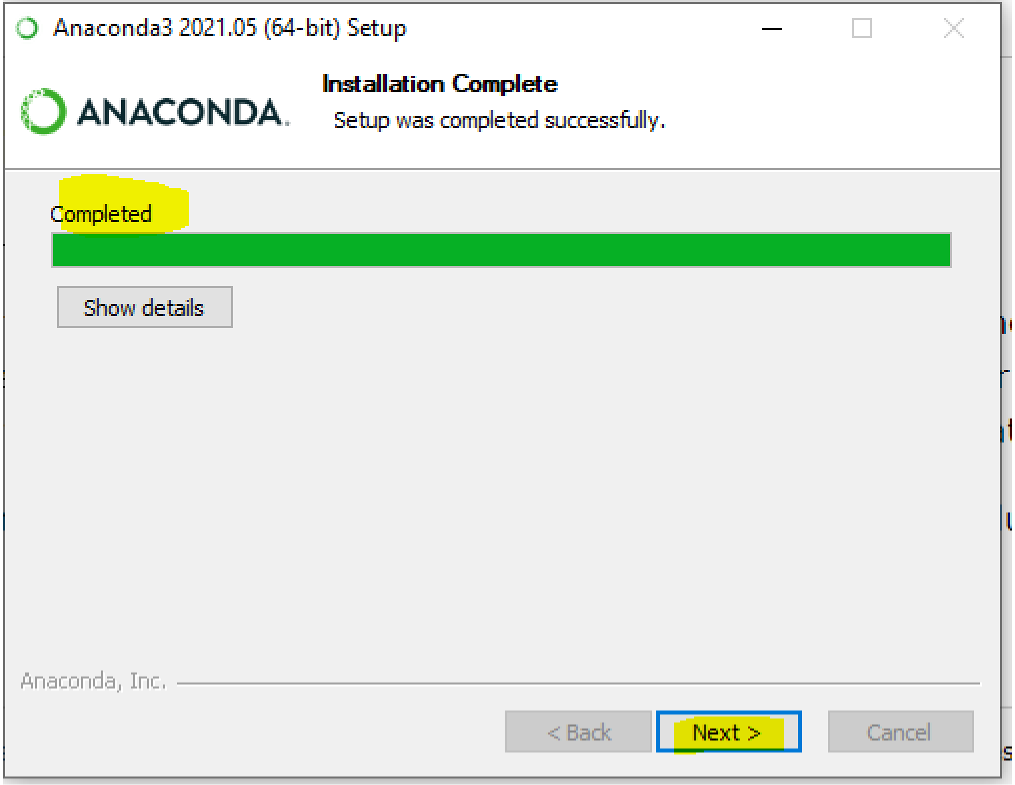
Installing

Please wait while Anaconda3 2021.05 (64-bit) is being installed.

Setting up the package cache ...

After 30 minutes anaconda installed

Finally you will get like this



Anaconda3 2021.05 (64-bit) Setup

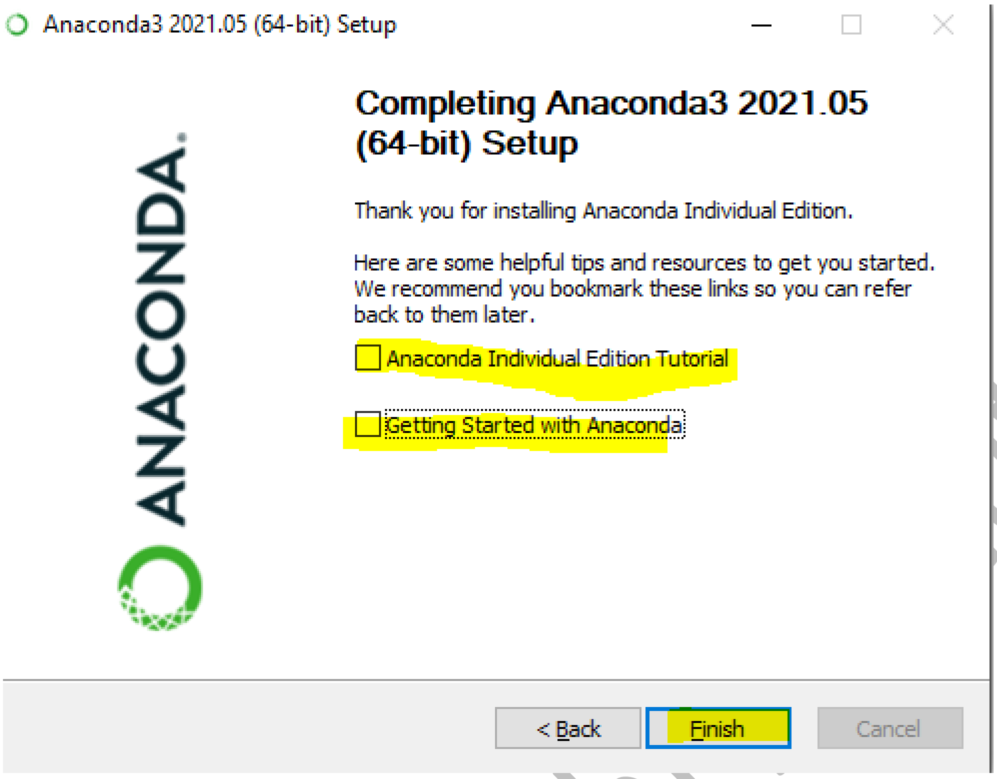
Installation Complete

Setup was completed successfully.

Completed

Click Next

Finally uncheck and click finish



Anaconda3 2021.05 (64-bit) Setup

Completing Anaconda3 2021.05 (64-bit) Setup

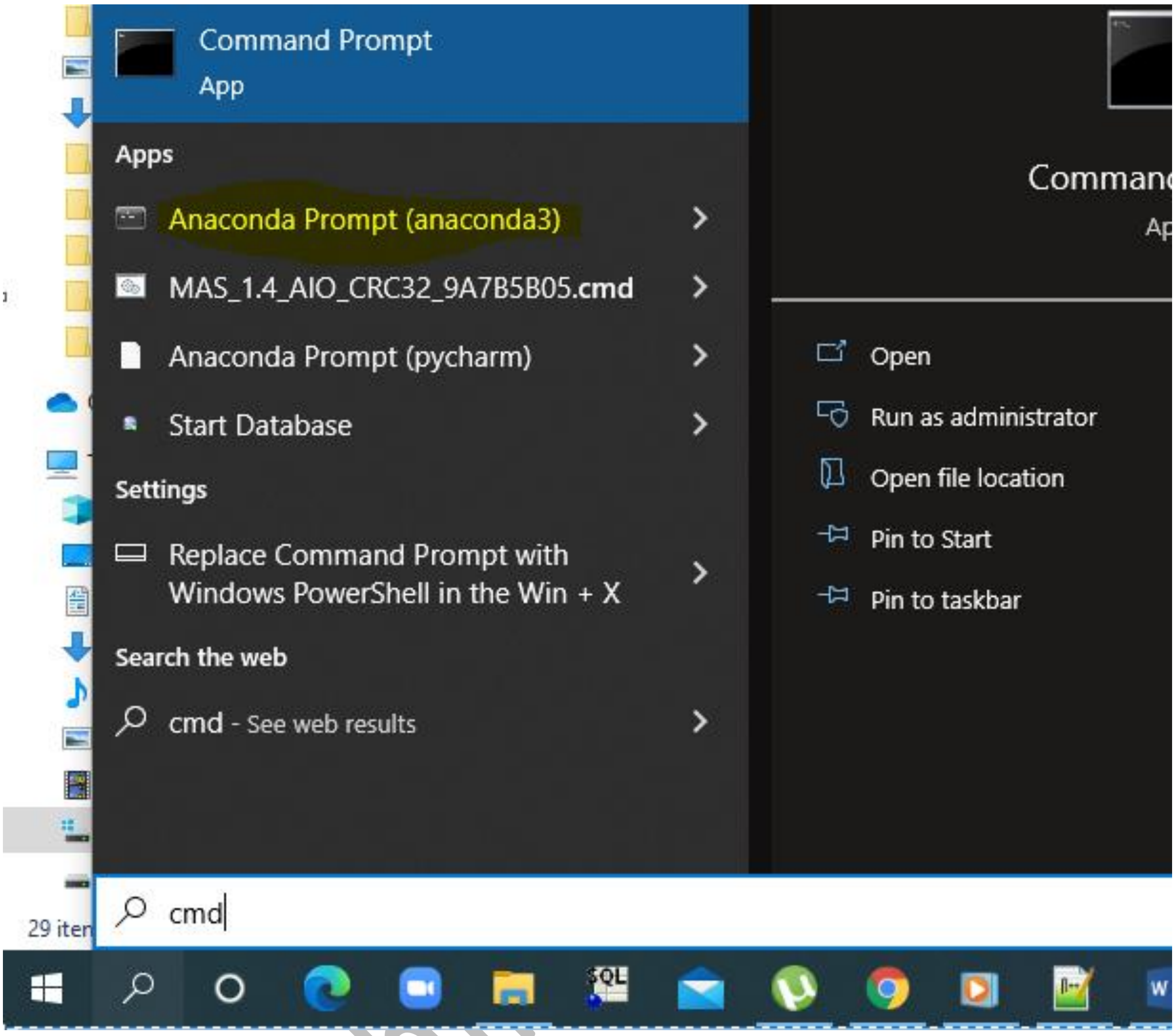
Thank you for installing Anaconda Individual Edition.

Here are some helpful tips and resources to get you started. We recommend you bookmark these links so you can refer back to them later.

□ Anaconda Individual Edition Tutorial

□ Getting Started with Anaconda

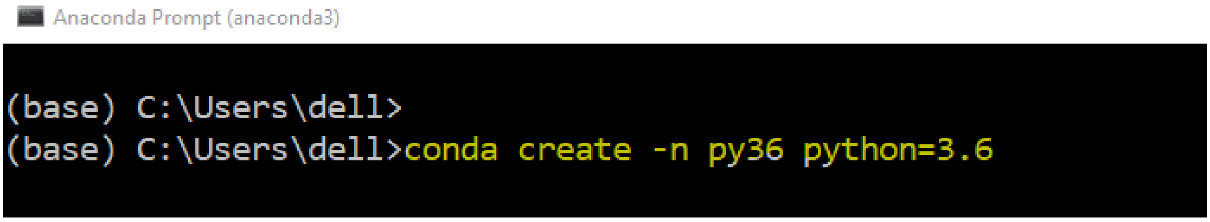
Testing purpose anaconda installed or not verify like this



If you get like this Anaconda Prompt (anaconda3) .. congrats anaconda installed.

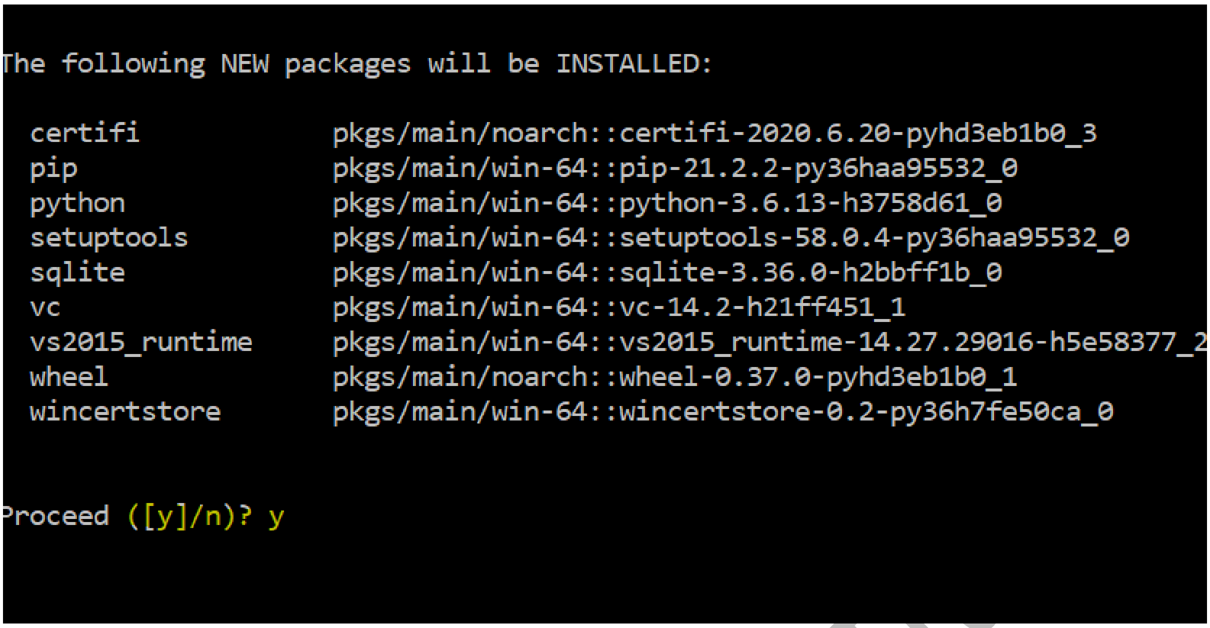
Now python 3.6 recommended that's y execute this command.

conda create -n py36 python=3.6



```
(base) C:\Users\dell>
(base) C:\Users\dell>conda create -n py36 python=3.6
```

Sometime you will get like this enter Y

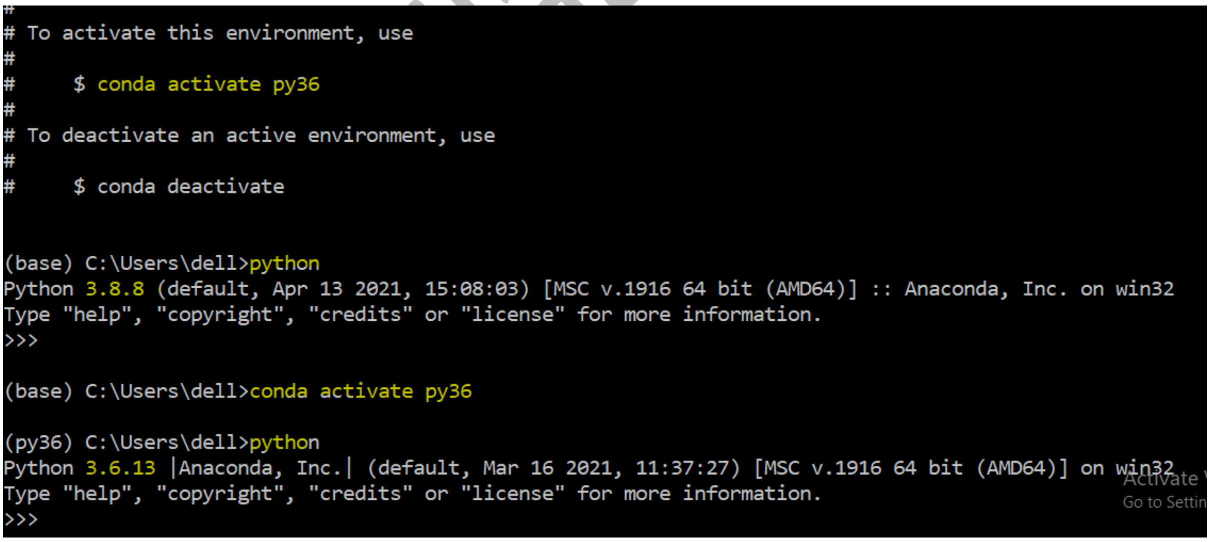


```
The following NEW packages will be INSTALLED:

  certifi            pkgs/main/noarch::certifi-2020.6.20-pyhd3eb1b0_3
  pip                pkgs/main/win-64::pip-21.2.2-py36haa95532_0
  python             pkgs/main/win-64::python-3.6.13-h3758d61_0
  setuptools         pkgs/main/win-64::setuptools-58.0.4-py36haa95532_0
  sqlite             pkgs/main/win-64::sqlite-3.36.0-h2bbff1b_0
  vc                 pkgs/main/win-64::vc-14.2-h21ff451_1
  vs2015_runtime     pkgs/main/win-64::vs2015_runtime-14.27.29016-h5e58377_2
  wheel              pkgs/main/noarch::wheel-0.37.0-pyhd3eb1b0_1
  wincertstore       pkgs/main/win-64::wincertstore-0.2-py36h7fe50ca_0


Proceed ([y]/n)? y
```

After 10 minutes py36 installed you can practice python 3.6 version in py36 environment.

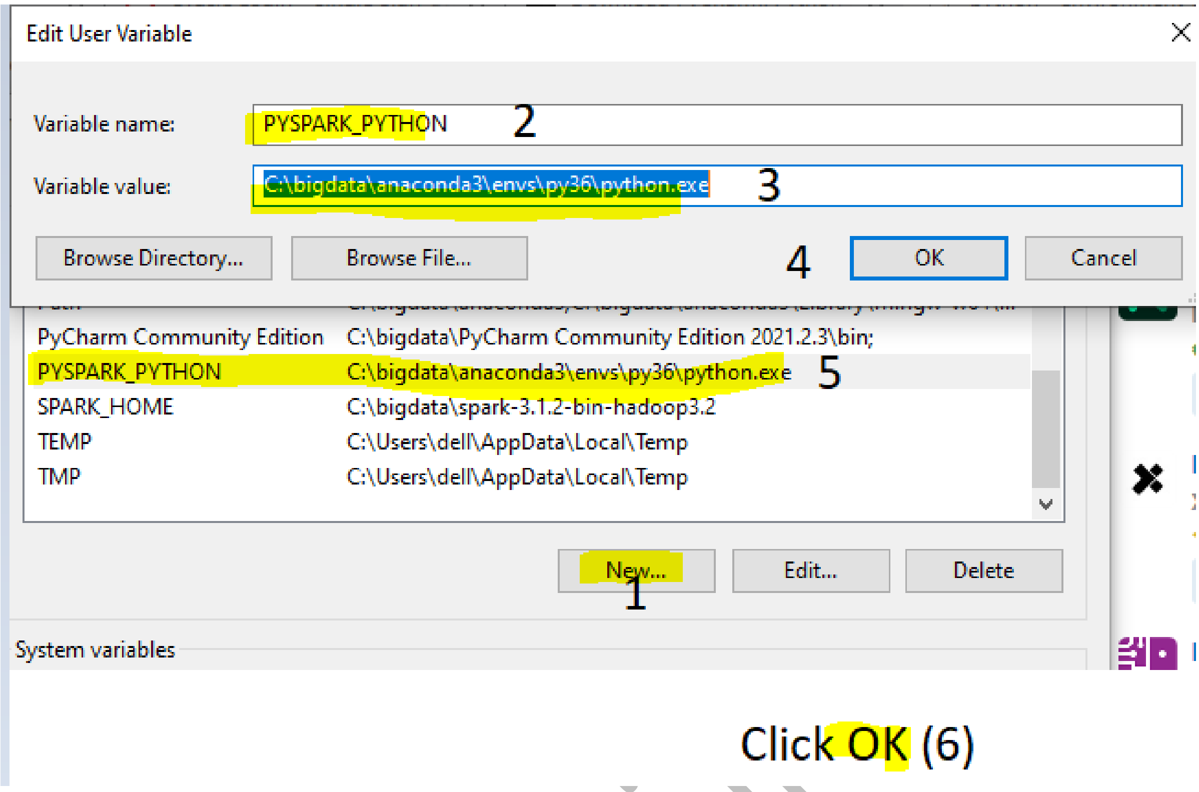
By default you will get like this

By default python (base) you will get 3.8 or 3.9 but I want to practice python 3.6 (pyspark support by default 3.6 that's y) activate py36 using this command.



```
#
# To activate this environment, use
#
#     $ conda activate py36
#
# To deactivate an active environment, use
#
#     $ conda deactivate


(base) C:\Users\dell>python
Python 3.8.8 (default, Apr 13 2021, 15:08:03) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>

(base) C:\Users\dell>conda activate py36

(py36) C:\Users\dell>python
Python 3.6.13 |Anaconda, Inc.| (default, Mar 16 2021, 11:37:27) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Now Its recommended not mandatory



Variable name: PYSPARK_PYTHON

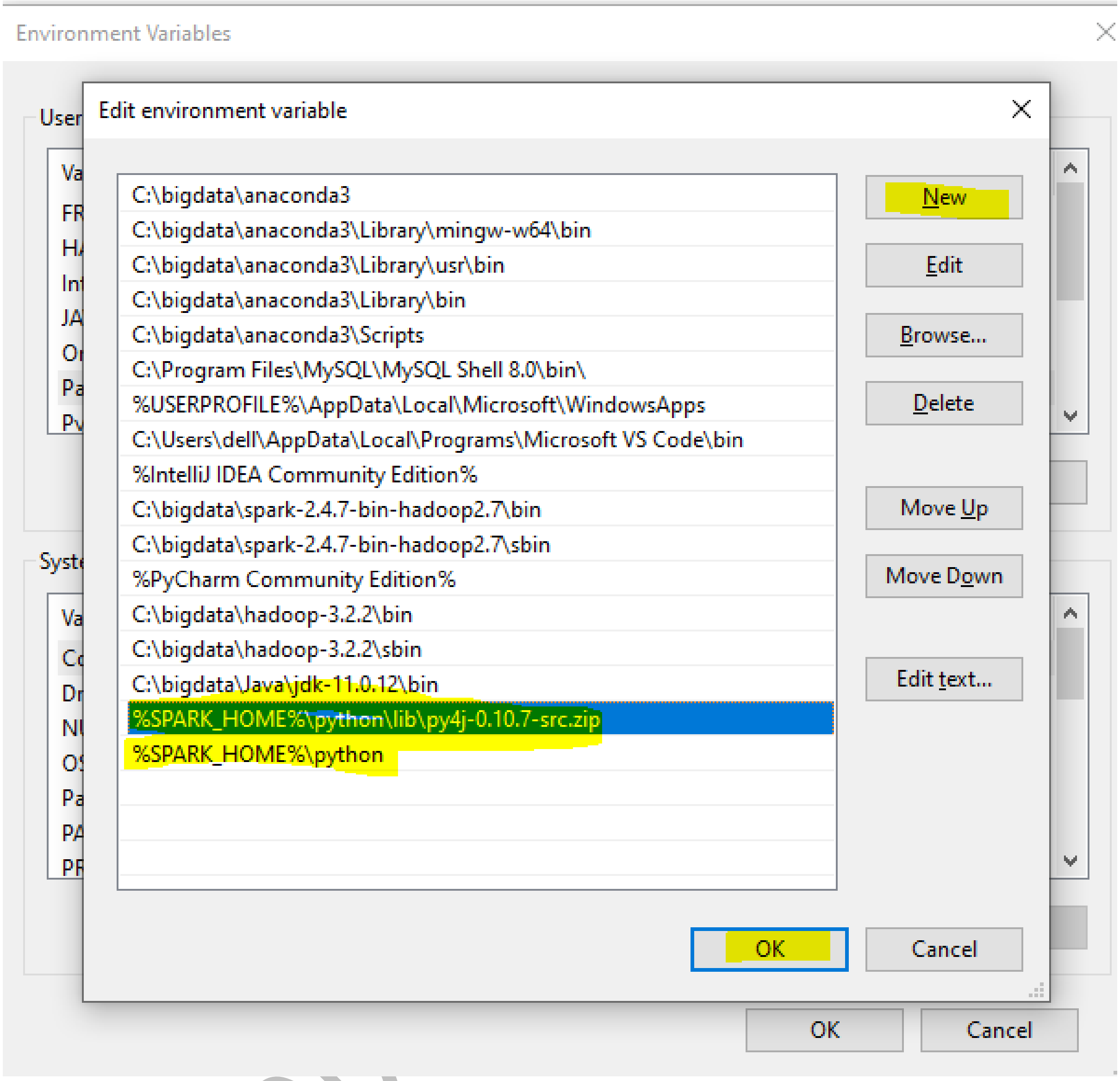
Variable value: C:\bigdata\anaconda3\envs\py36\python.exe

PyCharm Community Edition	C:\bigdata\PyCharm Community Edition 2021.2.3\bin;
PYSPARK_PYTHON	C:\bigdata\anaconda3\envs\py36\python.exe
SPARK_HOME	C:\bigdata\spark-3.1.2-bin-hadoop3.2
TEMP	C:\Users\dell\AppData\Local\Temp
TMP	C:\Users\dell\AppData\Local\Temp

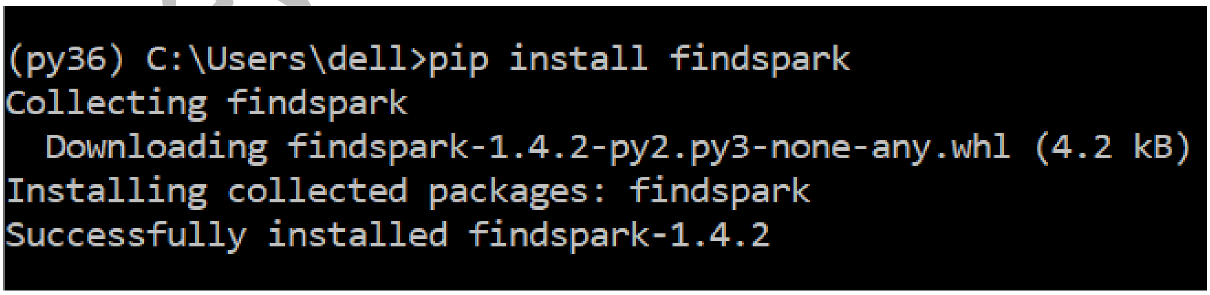
Click OK (6)

%SPARK_HOME%\python

%SPARK_HOME%\python\lib\py4j-0.10.7-src.zip



```
C:\bigdata\anaconda3
C:\bigdata\anaconda3\Library\mingw-w64\bin
C:\bigdata\anaconda3\Library\usr\bin
C:\bigdata\anaconda3\Library\bin
C:\bigdata\anaconda3\Scripts
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Users\dell\AppData\Local\Programs\Microsoft VS Code\bin
%IntelliJ IDEA Community Edition%
C:\bigdata\spark-2.4.7-bin-hadoop2.7\bin
C:\bigdata\spark-2.4.7-bin-hadoop2.7\sbin
%PyCharm Community Edition%
C:\bigdata\hadoop-3.2.2\bin
C:\bigdata\hadoop-3.2.2\sbin
C:\bigdata\Java\jdk-11.0.12\bin
%SPARK_HOME%\python\lib\py4j-0.10.7-src.zip
%SPARK_HOME%\python
```

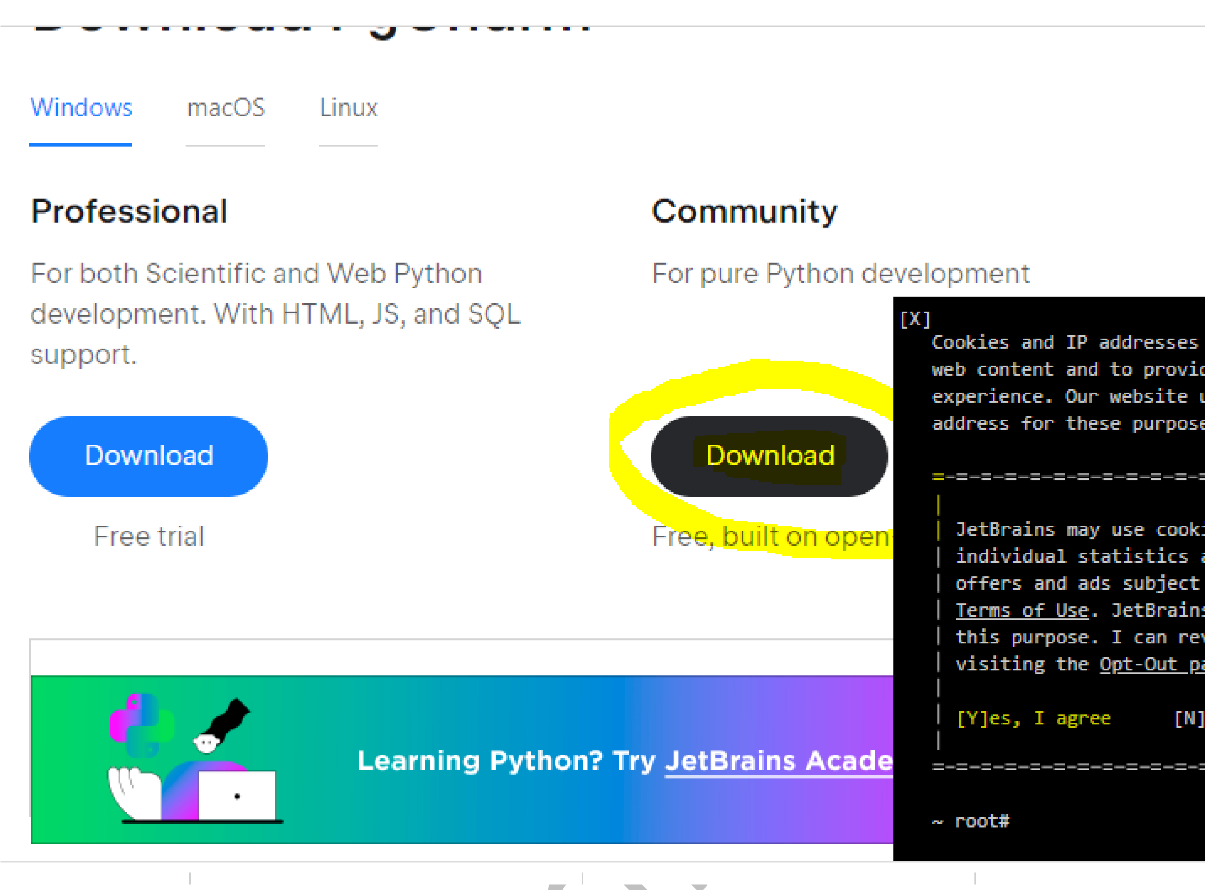


```
(py36) C:\Users\dell>pip install findspark
Collecting findspark
  Downloading findspark-1.4.2-py2.py3-none-any.whl (4.2 kB)
Installing collected packages: findspark
Successfully installed findspark-1.4.2
```

Within this time install pycharm

https://www.jetbrains.com/pycharm/download/

in this link download community version only



Windows macOS Linux

Professional

For both Scientific and Web Python development. With HTML, JS, and SQL support.

Download

Free trial

Community

For pure Python development

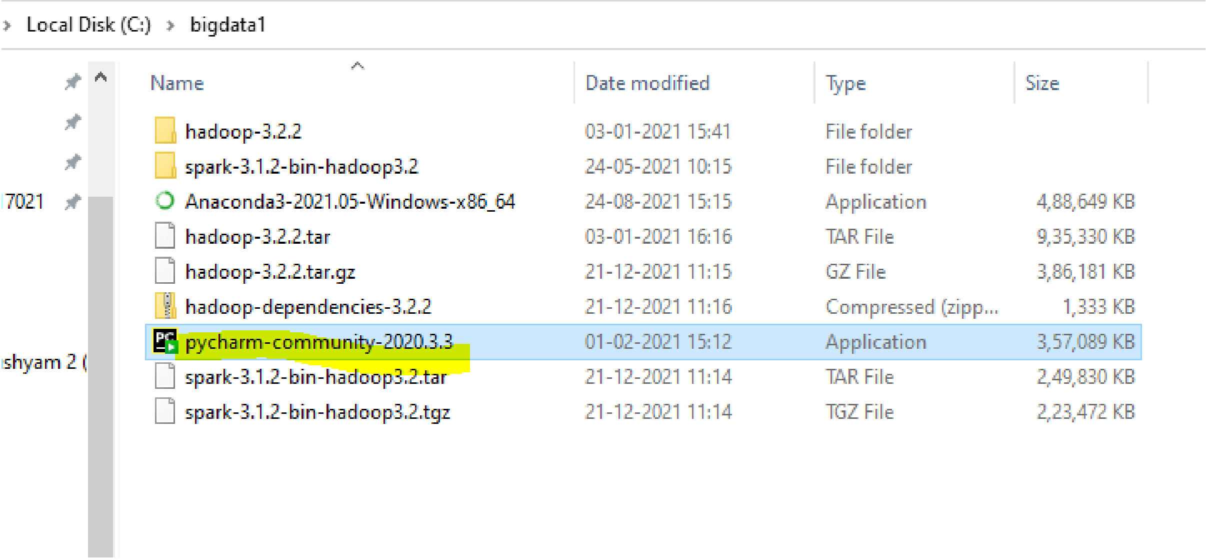
Download

Free, built on open

Learning Python? Try JetBrains Acade

```
[X]
  Cookies and IP addresses
  web content and to provic
  experience. Our website u
  address for these purpose

  =-=-=-=-=-=-=-=-=-=-=-=-=
  |
  | JetBrains may use cook
  | individual statistics a
  | offers and ads subject
  | Terms of Use. JetBrains
  | this purpose. I can rev
  | visiting the Opt-Out pa
  |
  | [Y]es, I agree      [N]
  |
  =-=-=-=-=-=-=-=-=-=-=-=-=

  ~ root#
```

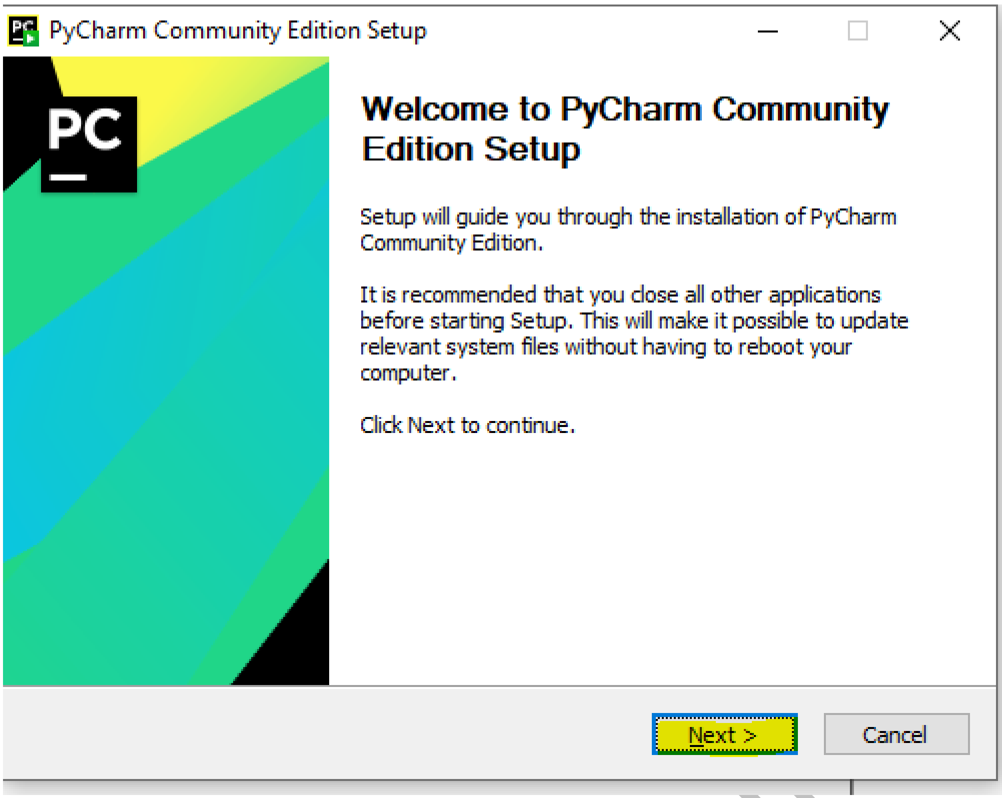
Paste that Pycharm in bigdata folder



› Local Disk (C:) › bigdata1

Name	Date modified	Type	Size
hadoop-3.2.2	03-01-2021 15:41	File folder	
spark-3.1.2-bin-hadoop3.2	24-05-2021 10:15	File folder	
Anaconda3-2021.05-Windows-x86_64	24-08-2021 15:15	Application	4,88,649 KB
hadoop-3.2.2.tar	03-01-2021 16:16	TAR File	9,35,330 KB
hadoop-3.2.2.tar.gz	21-12-2021 11:15	GZ File	3,86,181 KB
hadoop-dependencies-3.2.2	21-12-2021 11:16	Compressed (zipp...	1,333 KB
pycharm-community-2020.3.3	01-02-2021 15:12	Application	3,57,089 KB
spark-3.1.2-bin-hadoop3.2.tar	21-12-2021 11:14	TAR File	2,49,830 KB
spark-3.1.2-bin-hadoop3.2.tgz	21-12-2021 11:14	TGZ File	2,23,472 KB

Double click



PyCharm Community Edition Setup

Welcome to PyCharm Community Edition Setup

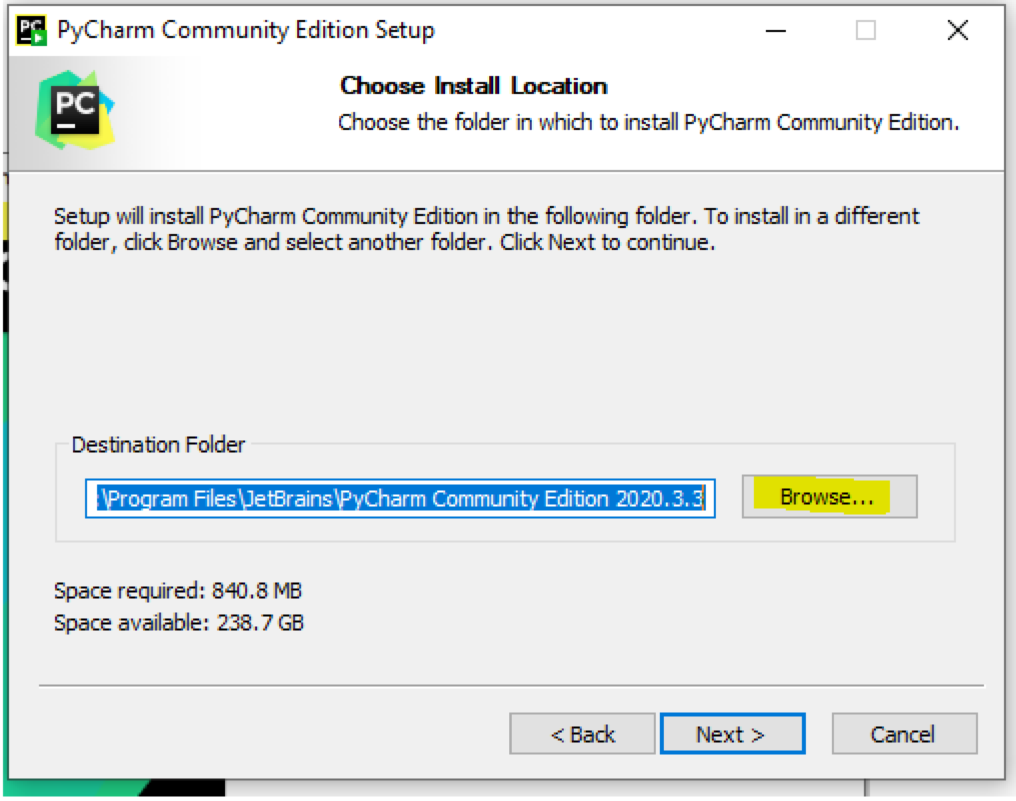
Setup will guide you through the installation of PyCharm Community Edition.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Click Next

My recommendation this is also install in bigdata folder



PyCharm Community Edition Setup

Choose Install Location

Choose the folder in which to install PyCharm Community Edition.

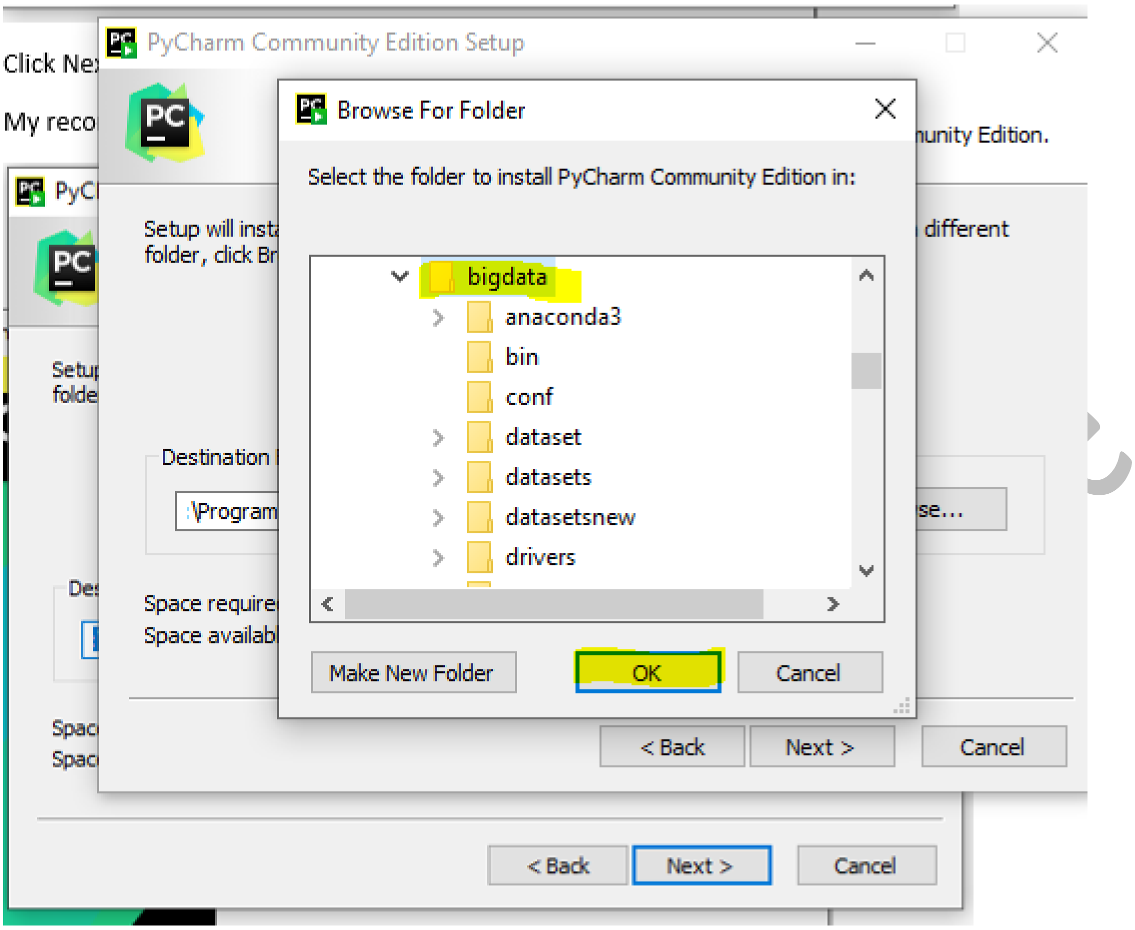
Setup will install PyCharm Community Edition in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

\Program Files\JetBrains\PyCharm Community Edition 2020.3.3

Space required: 840.8 MB
Space available: 238.7 GB

Finally go to C drive bigdata folder click OK

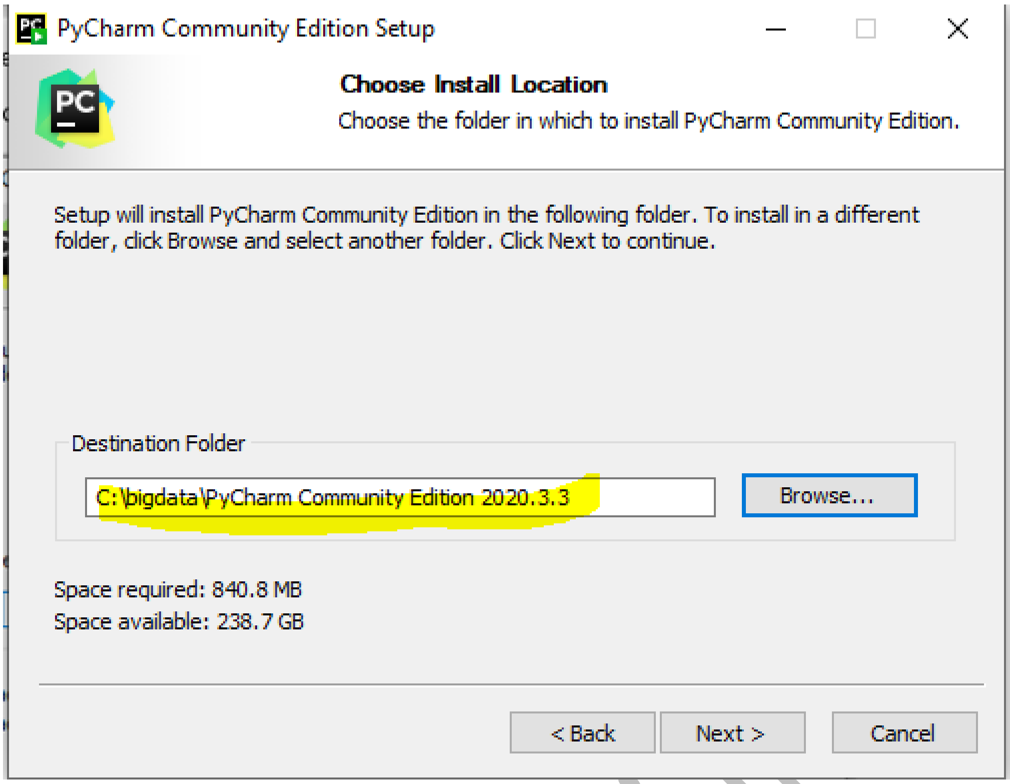


Browse For Folder

Select the folder to install PyCharm Community Edition in:

Now you will get like this

It means pycharm u r installing in bigdata folder


PyCharm Community Edition Setup

Choose Install Location

Choose the folder in which to install PyCharm Community Edition.

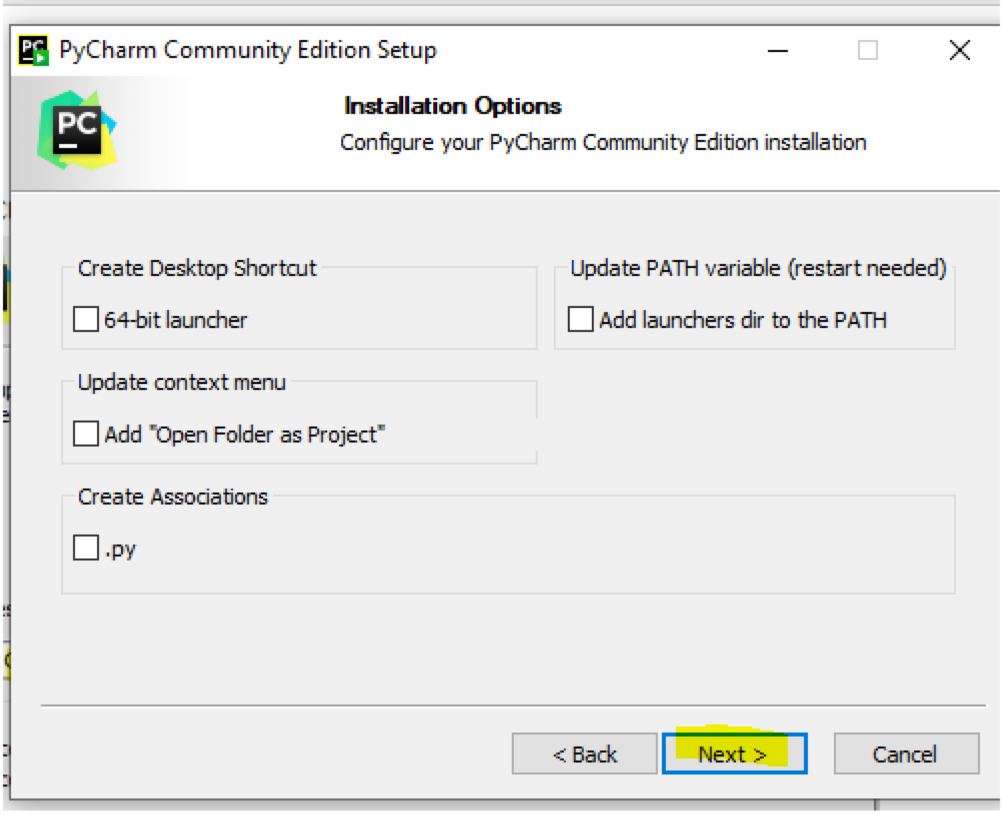
Setup will install PyCharm Community Edition in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

C:\bigdata\PyCharm Community Edition 2020.3.3

Space required: 840.8 MB
Space available: 238.7 GB

All these are optional pls ignore


PyCharm Community Edition Setup

Installation Options

Configure your PyCharm Community Edition installation

Create Desktop Shortcut

□ 64-bit launcher

Update PATH variable (restart needed)

□ Add launchers dir to the PATH

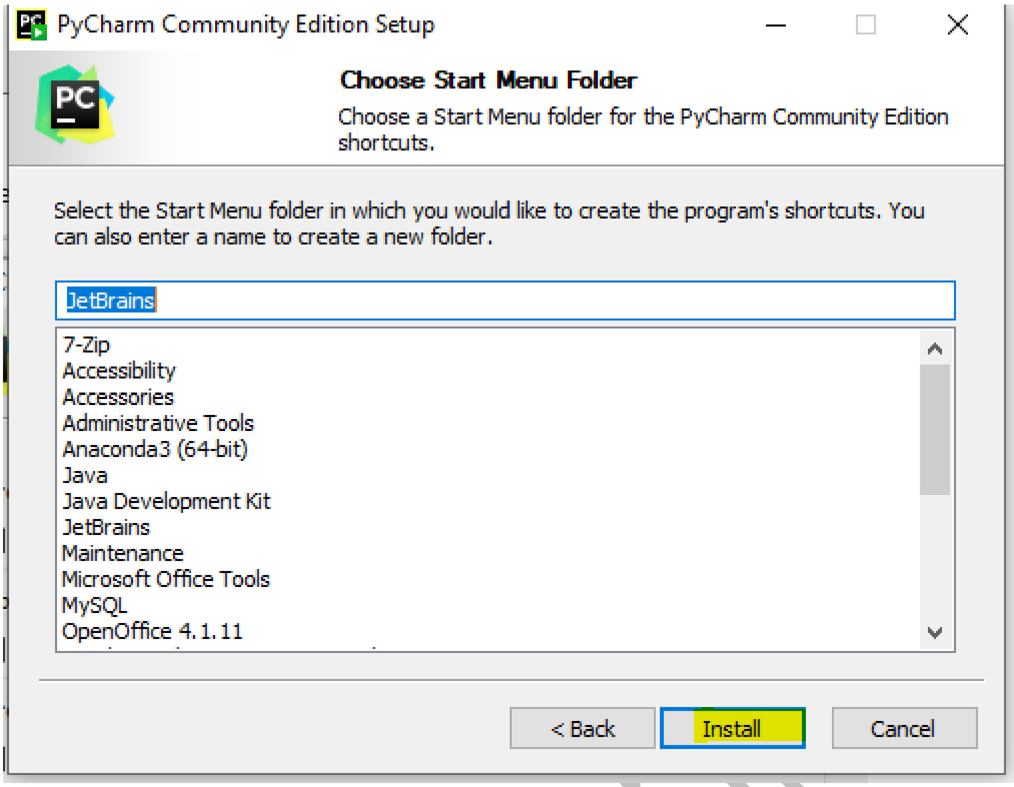
Update context menu

□ Add "Open Folder as Project"

Create Associations

□ .py

Click install


PyCharm Community Edition Setup

Choose Start Menu Folder

Choose a Start Menu folder for the PyCharm Community Edition shortcuts.

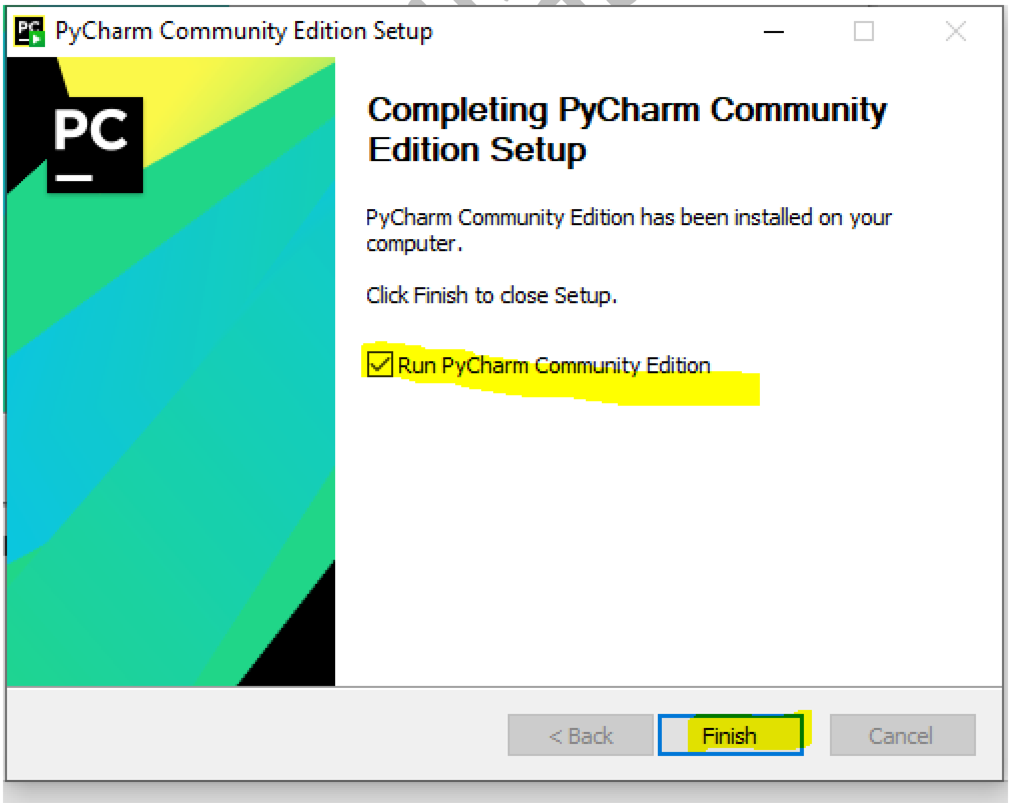
Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.

JetBrains

7-Zip
Accessibility
Accessories
Administrative Tools
Anaconda3 (64-bit)
Java
Java Development Kit
JetBrains
Maintenance
Microsoft Office Tools
MySQL
OpenOffice 4.1.11

It will take 5 minutes maximum

Finally you will get like this


PyCharm Community Edition Setup

Completing PyCharm Community Edition Setup

PyCharm Community Edition has been installed on your computer.

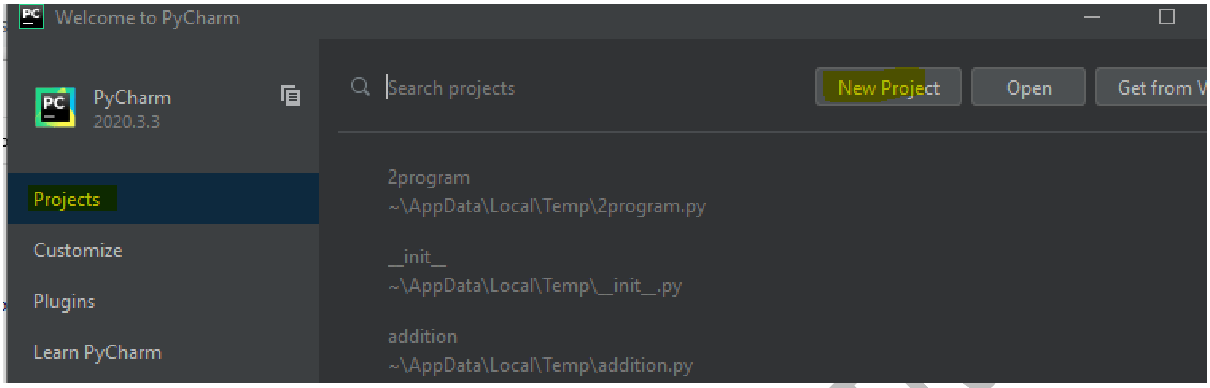
Click Finish to close Setup.

☑ Run PyCharm Community Edition

Finally check Run Pycharm Community edition and Click Finish

Now 99% pycham also installation completed.

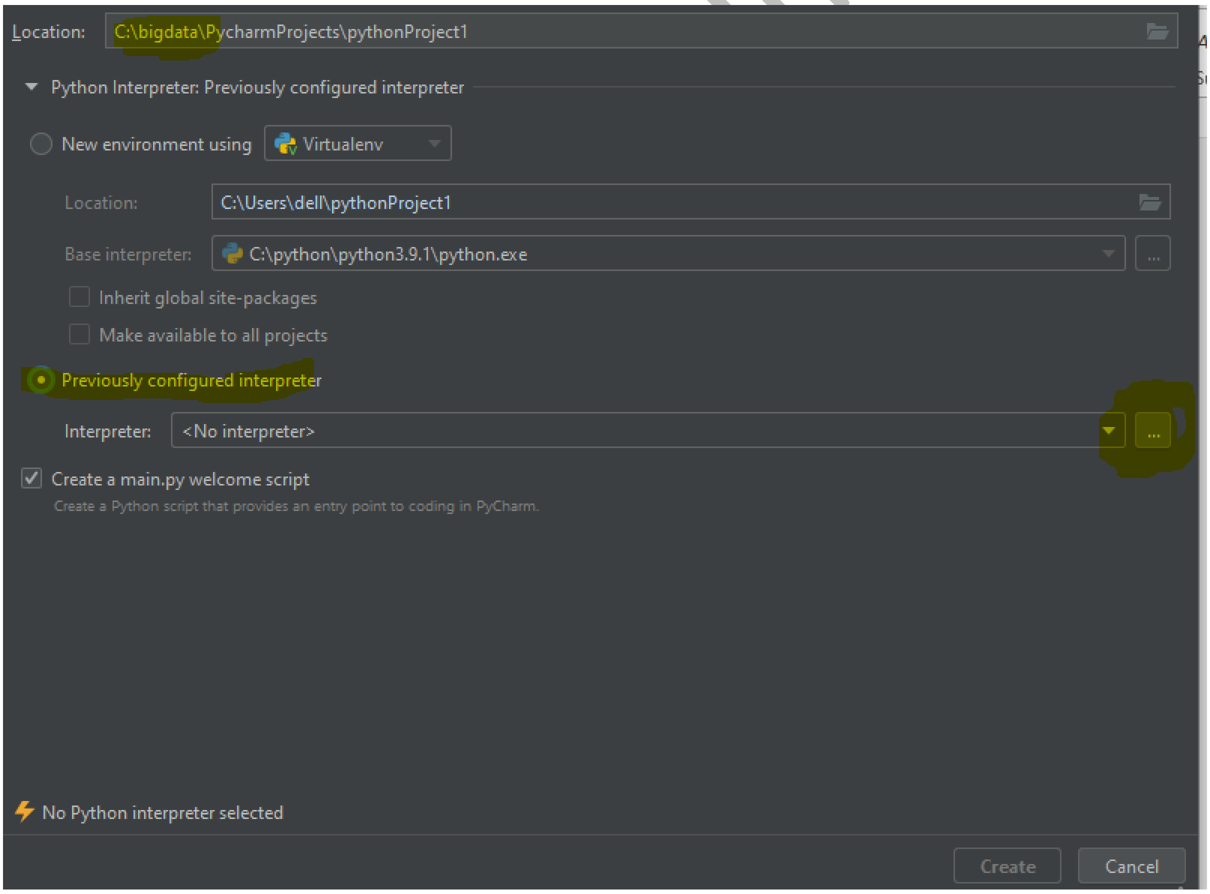
How to practice pyspark ill guide now.



2program
~\AppData\Local\Temp\2program.py

__init__
~\AppData\Local\Temp\__init__.py

addition
~\AppData\Local\Temp\addition.py



Location: C:\bigdata\PycharmProjects\pythonProject1

Python Interpreter: Previously configured interpreter

○ New environment using Virtualenv

Location: C:\Users\dell\pythonProject1

Base interpreter: C:\python\python3.9.1\python.exe

□ Inherit global site-packages

□ Make available to all projects

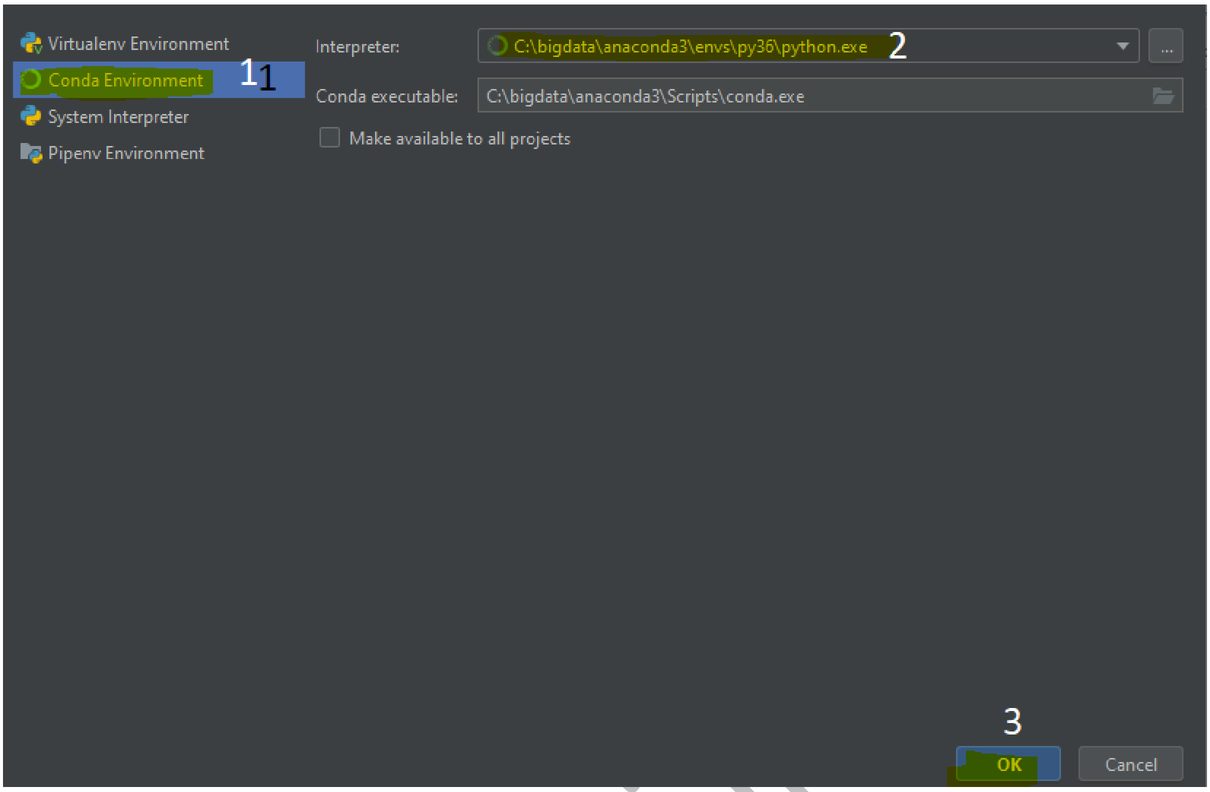
● Previously configured interpreter

Interpreter: <No interpreter>

☑ Create a main.py welcome script

Create a Python script that provides an entry point to coding in PyCharm.

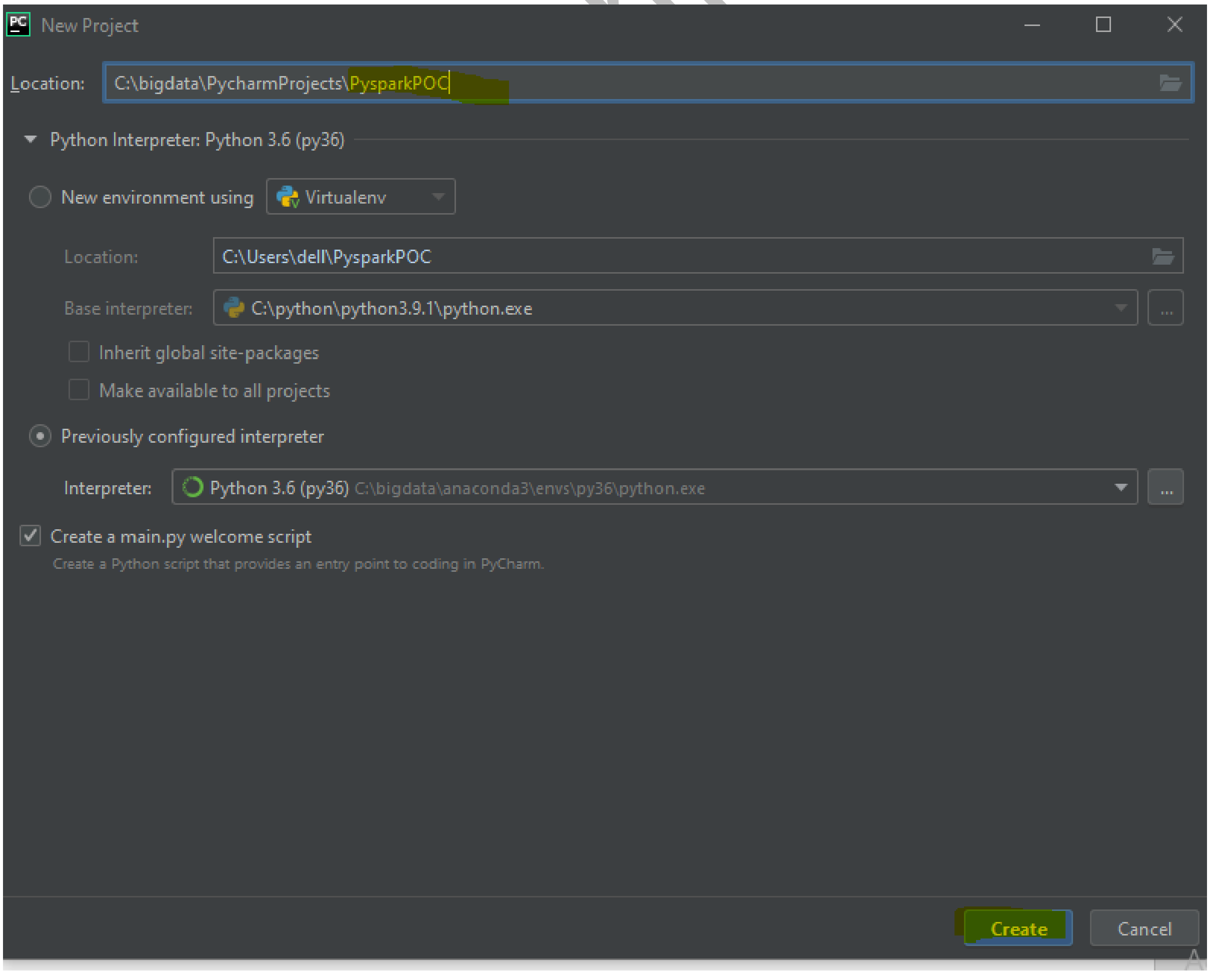
No Python interpreter selected



Interpreter: C:\bigdata\anaconda3\envs\py36\python.exe

Conda executable: C:\bigdata\anaconda3\Scripts\conda.exe

□ Make available to all projects



Location: C:\bigdata\PycharmProjects\PysparkPOC

Python Interpreter: Python 3.6 (py36)

○ New environment using Virtualenv

Location: C:\Users\dell\PysparkPOC

Base interpreter: C:\python\python3.9.1\python.exe

□ Inherit global site-packages

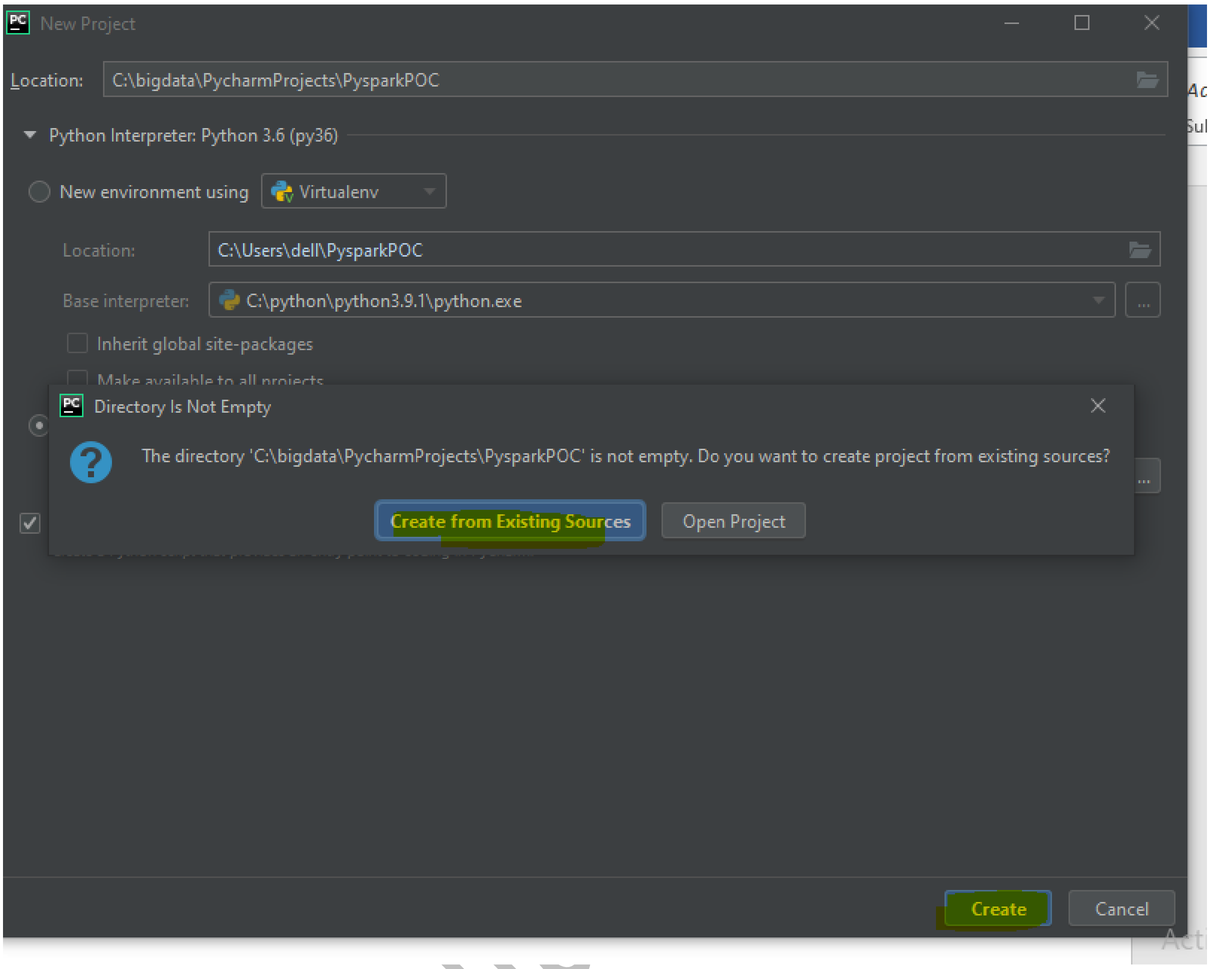
□ Make available to all projects

● Previously configured interpreter

Interpreter: Python 3.6 (py36) C:\bigdata\anaconda3\envs\py36\python.exe

☑ Create a main.py welcome script

Create a Python script that provides an entry point to coding in PyCharm.



Location: C:\bigdata\PycharmProjects\PysparkPOC

Python Interpreter: Python 3.6 (py36)

○ New environment using Virtualenv

Location: C:\Users\dell\PysparkPOC

Base interpreter: C:\python\python3.9.1\python.exe

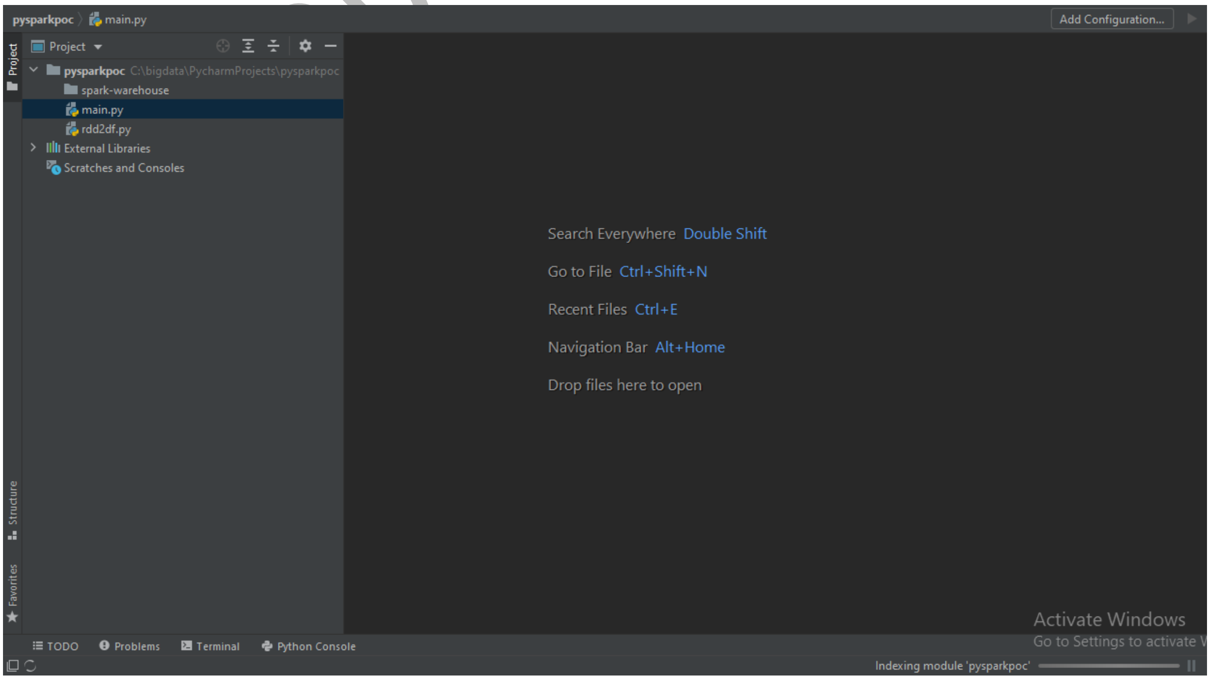
□ Inherit global site-packages

□ Make available to all projects

Directory Is Not Empty

The directory 'C:\bigdata\PycharmProjects\PysparkPOC' is not empty. Do you want to create project from existing sources?

Now you will get home screen like this



Search Everywhere Double Shift

Go to File Ctrl+Shift+N

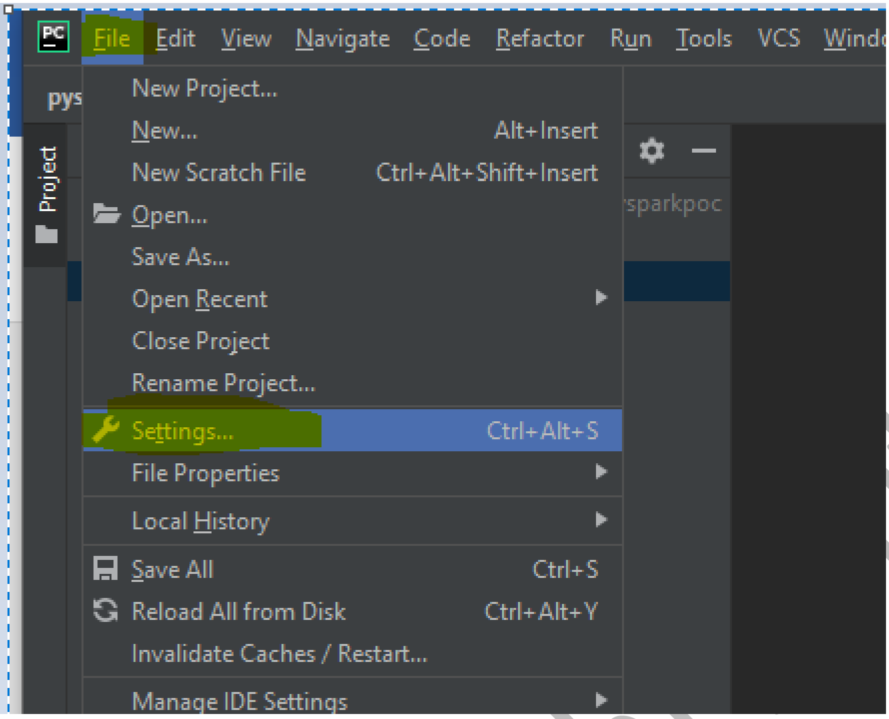
Recent Files Ctrl+E

Navigation Bar Alt+Home

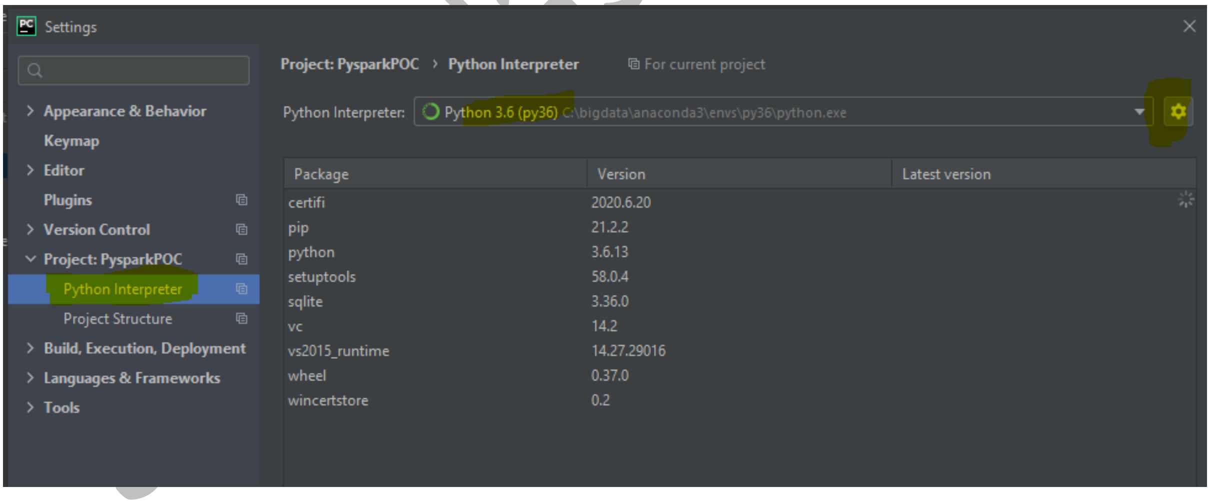
Drop files here to open

Indexing module 'pysparkpoc'

File > setting



If python version wrong here change

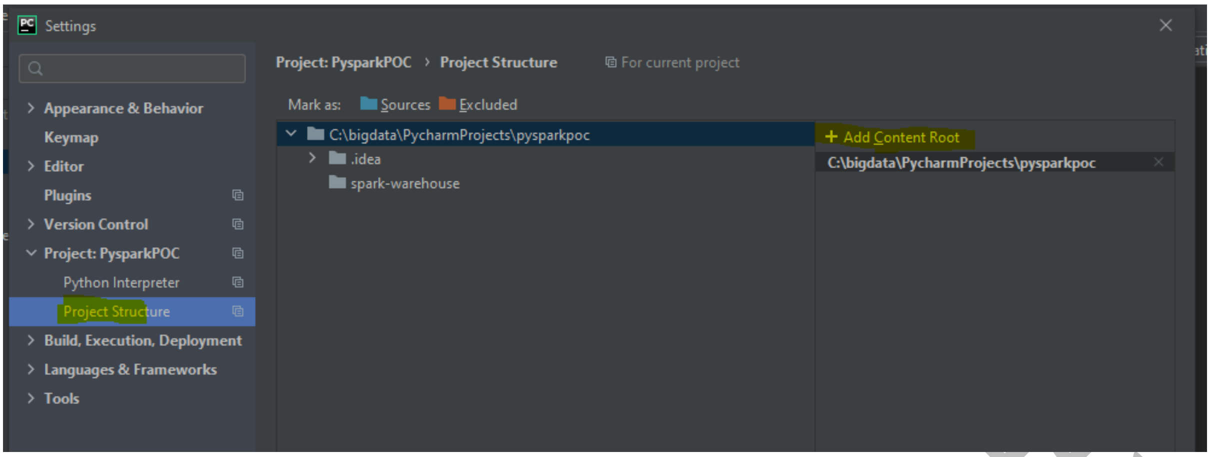


Project: PysparkPOC › Python Interpreter   For current project

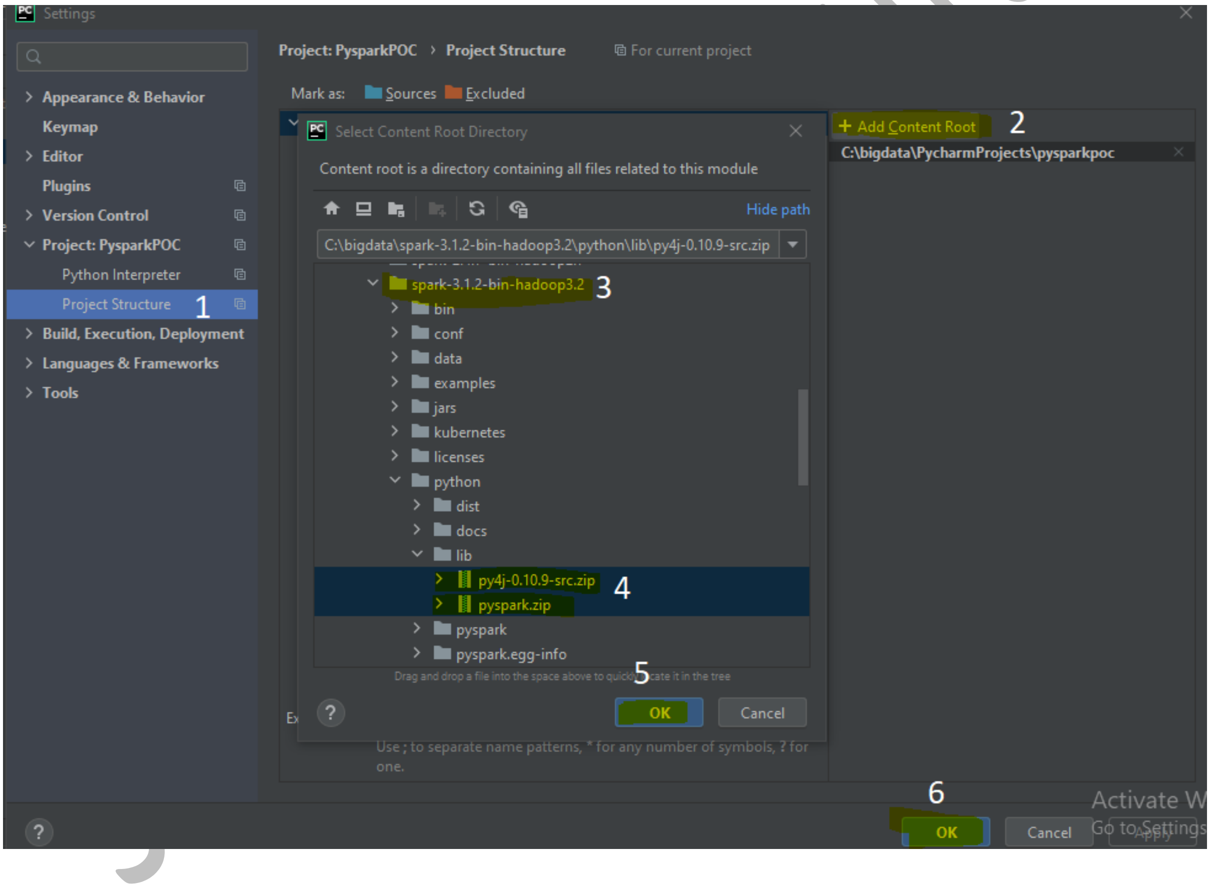
Python Interpreter: Python 3.6 (py36) C:\bigdata\anaconda3\envs\py36\python.exe

Package	Version	Latest version
certifi	2020.6.20	
pip	21.2.2	
python	3.6.13	
setuptools	58.0.4	
sqlite	3.36.0	
vc	14.2	
vs2015_runtime	14.27.29016	
wheel	0.37.0	
wincertstore	0.2	

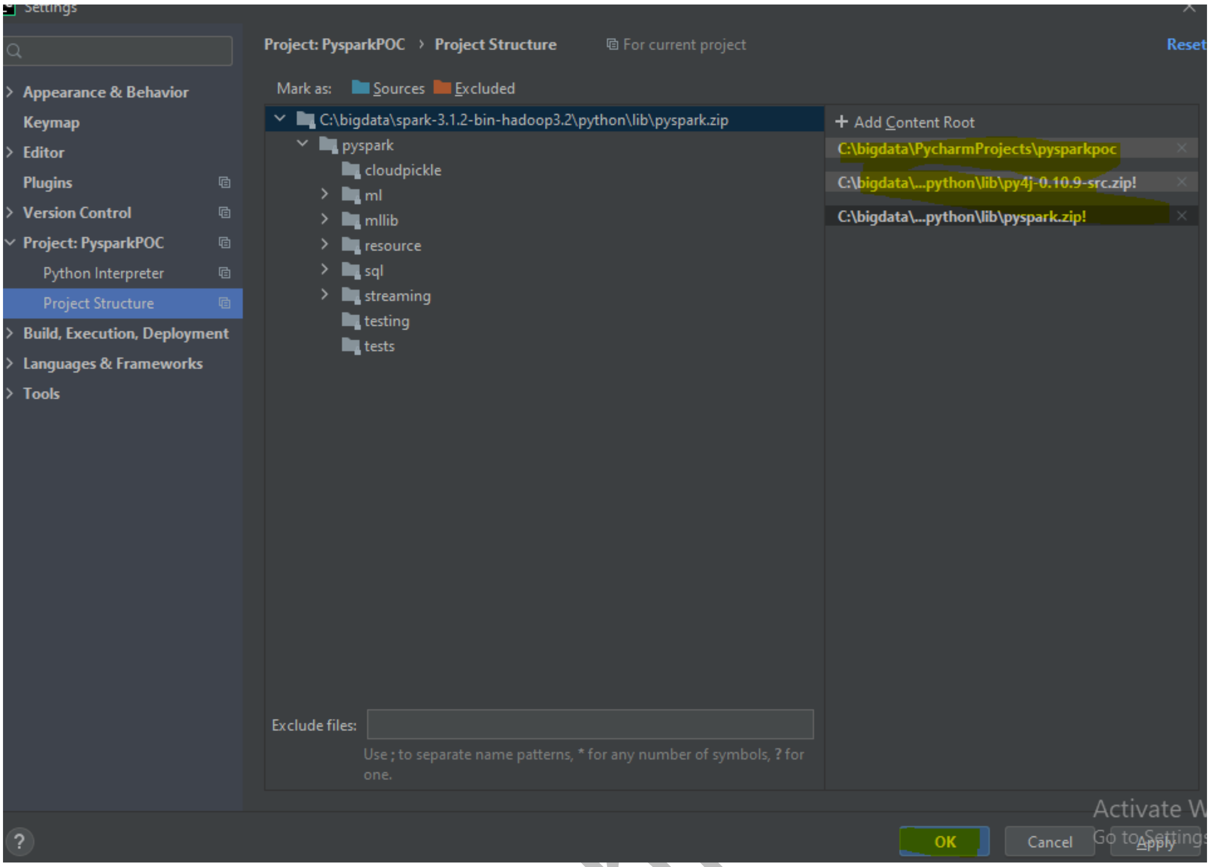
Next



Its important step. Click py4j and pyspark.zip file

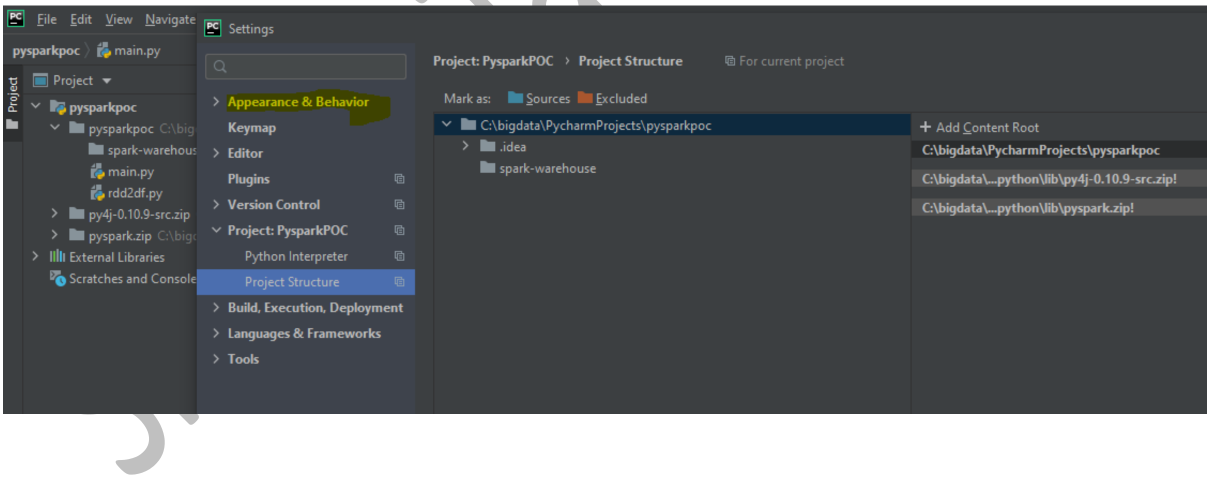

Settings
Appearance & Behavior
Keymap
Editor
Plugins
Version Control
Project: PysparkPOC
Python Interpreter
Project Structure
1
Build, Execution, Deployment
Languages & Frameworks
Tools
Project: PysparkPOC › Project Structure
For current project
Mark as: Sources Excluded
Select Content Root Directory
Content root is a directory containing all files related to this module
Hide path
C:\bigdata\spark-3.1.2-bin-hadoop3.2\python\lib\py4j-0.10.9-src.zip
spark-3.1.2-bin-hadoop3.2
3
bin
conf
data
examples
jars
kubernetes
licenses
python
dist
docs
lib
py4j-0.10.9-src.zip
4
pyspark.zip
pyspark
pyspark.egg-info
Drag and drop a file into the space above to quickly locate it in the tree
5
OK
Cancel
+ Add Content Root
2
C:\bigdata\PycharmProjects\pysparkpoc
Use ; to separate name patterns, * for any number of symbols, ? for one.
6
OK
Cancel
Apply

Now you will get like this click ok



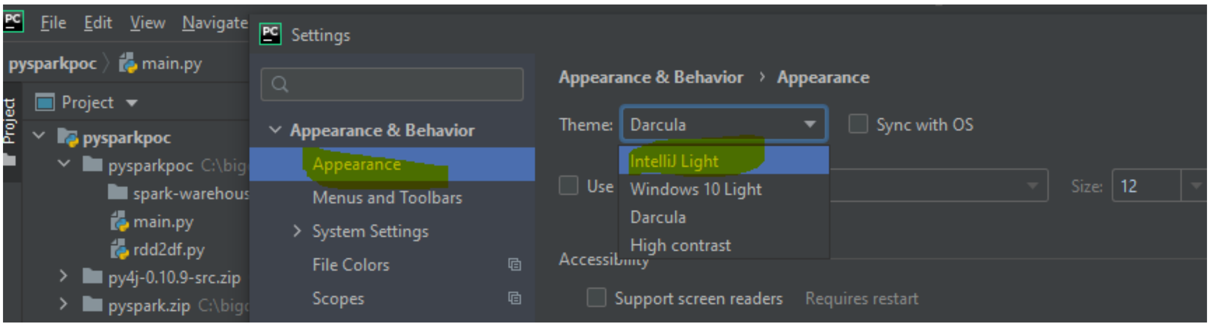
Settings
Appearance & Behavior
Keymap
Editor
Plugins
Version Control
Project: PysparkPOC
Python Interpreter
Project Structure
Build, Execution, Deployment
Languages & Frameworks
Tools
Project: PysparkPOC › Project Structure
For current project
Reset
Mark as: Sources Excluded
C:\bigdata\spark-3.1.2-bin-hadoop3.2\python\lib\pyspark.zip
pyspark
cloudpickle
ml
mllib
resource
sql
streaming
testing
tests
+ Add Content Root
C:\bigdata\PycharmProjects\pysparkpoc
C:\bigdata\...python\lib\py4j-0.10.9-src.zip!
C:\bigdata\...python\lib\pyspark.zip!
Exclude files:
Use ; to separate name patterns, * for any number of symbols, ? for one.
OK
Cancel
Apply

If you don't prefer black colour screen(theme) change like this (its optional)



Click Appearance & behaviour

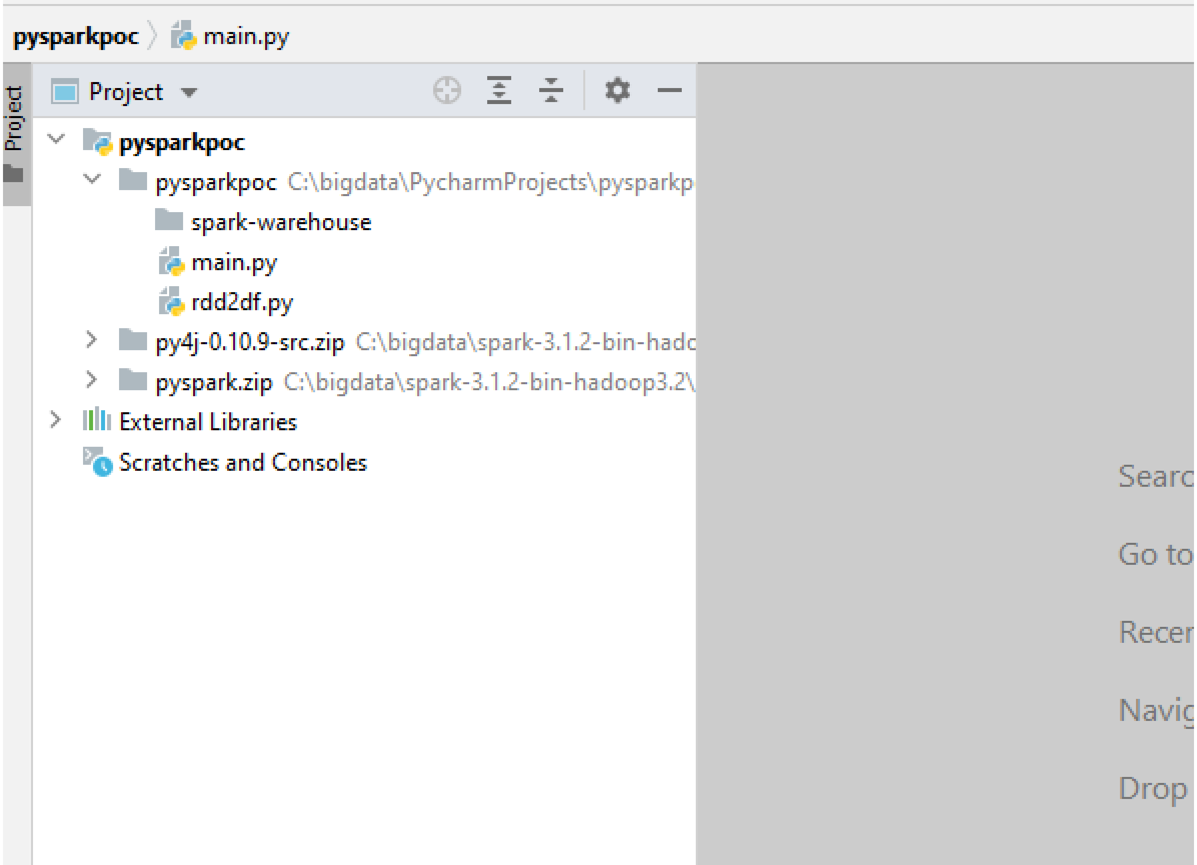
Now click Appearance ..
Under Theme … instead of Darcula choose IntelliJ Light … Click Ok



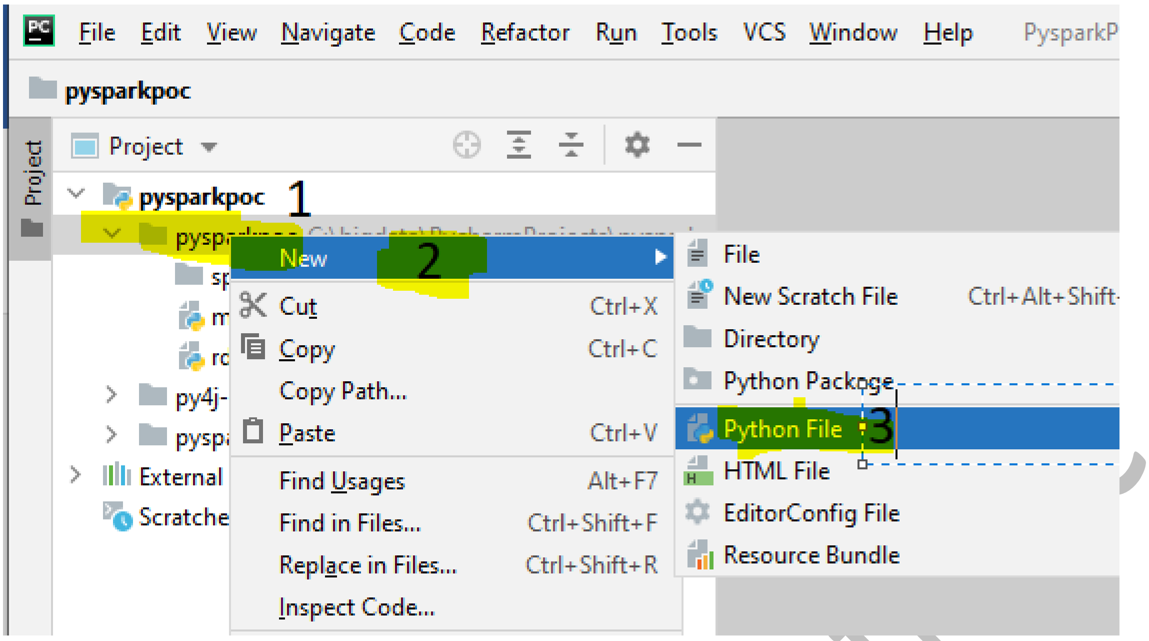
Theme: Darcula

- IntelliJ Light
- Windows 10 Light
- Darcula
- High contrast

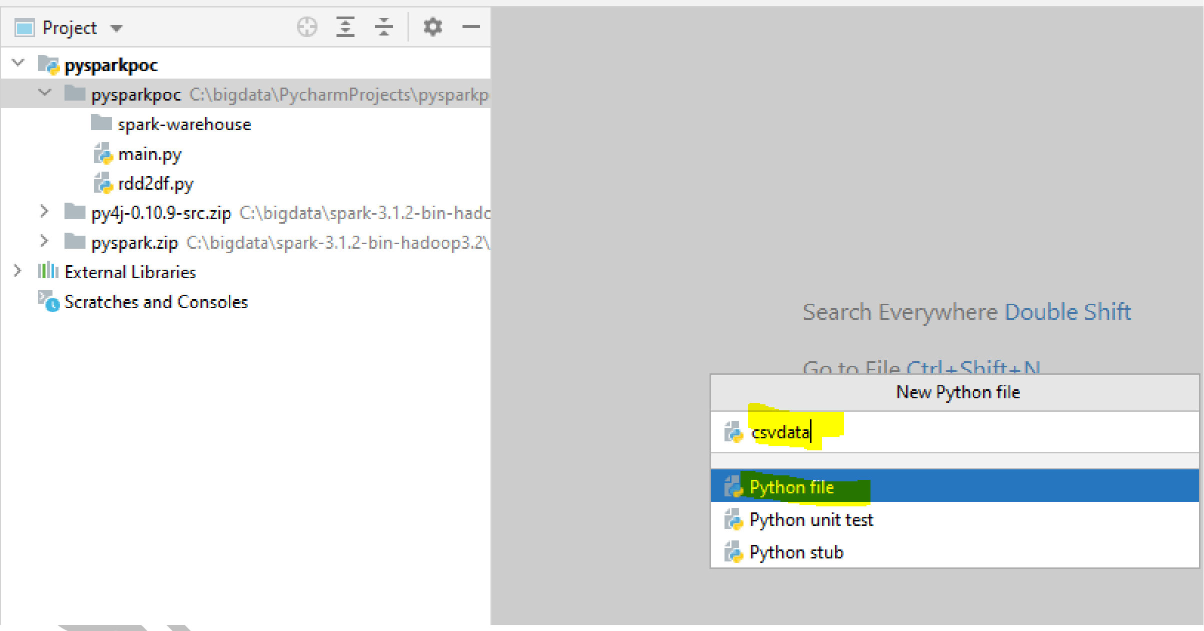
You will get like this like light colour


pysparkpoc 〉 main.py
Project
Project
pysparkpoc
pysparkpoc C:\bigdata\PycharmProjects\pysparkp
spark-warehouse
main.py
rdd2df.py
py4j-0.10.9-src.zip C:\bigdata\spark-3.1.2-bin-hadc
pyspark.zip C:\bigdata\spark-3.1.2-bin-hadoop3.2\
External Libraries
Scratches and Consoles
Searc
Go to
Recer
Navig
Drop

Right click on Pysparkpoc … Click New …Choose Python File



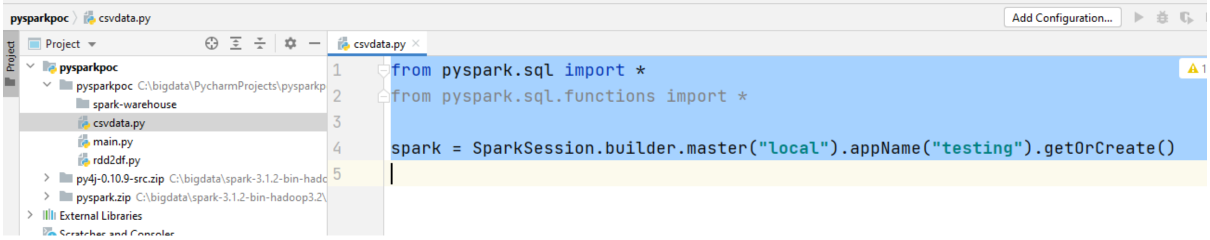
Write something python file name like csvdata


Project
pysparkpoc
pysparkpoc C:\bigdata\PycharmProjects\pysparkp
spark-warehouse
main.py
rdd2df.py
py4j-0.10.9-src.zip C:\bigdata\spark-3.1.2-bin-hadc
pyspark.zip C:\bigdata\spark-3.1.2-bin-hadoop3.2\
External Libraries
Scratches and Consoles
Search Everywhere Double Shift
Go to File Ctrl+Shift+N
New Python file
Python file
Python unit test
Python stub

csvdata

Now write something program like this

```
from pyspark.sql import *
from pyspark.sql.functions import *

spark =
SparkSession.builder.master("local").appName("testing")
.getOrCreate()
```



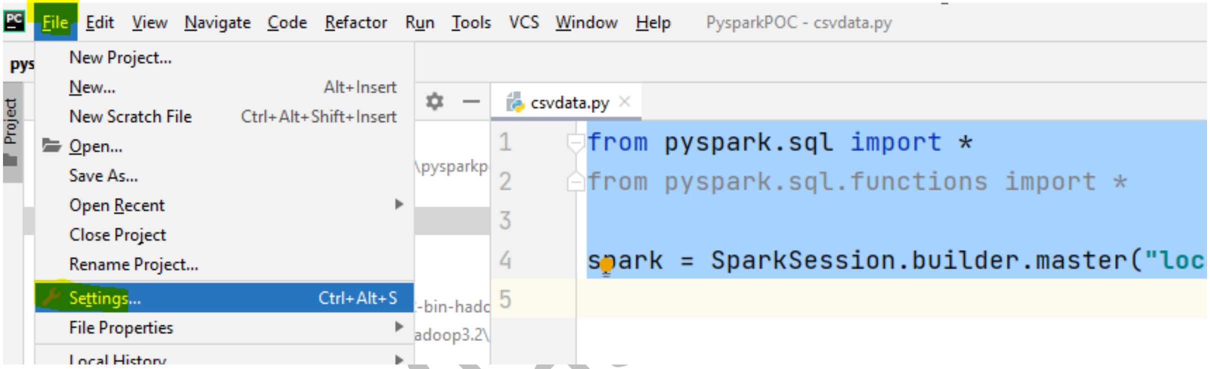
pysparkpoc
csvdata.py
Add Configuration...
Project
pysparkpoc
pysparkpoc C:\bigdata\PycharmProjects\pysparkp
spark-warehouse
csvdata.py
main.py
rdd2df.py
py4j-0.10.9-src.zip C:\bigdata\spark-3.1.2-bin-hadc
pyspark.zip C:\bigdata\spark-3.1.2-bin-hadoop3.2\
External Libraries
Scratches and Consoles

```python
from pyspark.sql import *
from pyspark.sql.functions import *

spark = SparkSession.builder.master("local").appName("testing").getOrCreate()
```

These three lines ur using most frequently in any program so its recommended like this

(its optional)

Goto file .. setting



```python
from pyspark.sql import *
from pyspark.sql.functions import *

spark = SparkSession.builder.master("loc
```

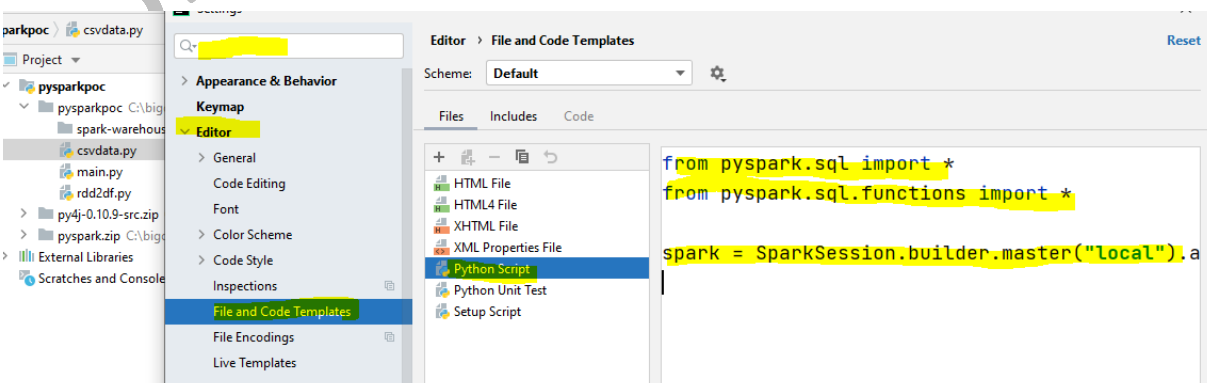
Now click on Editor .. Click on File and Code Templates … Click PythonScript…

I recommend paste this code here

from pyspark.sql import *

from pyspark.sql.functions import *

spark = SparkSession.builder.master("local").appName("testing").getOrCreate()



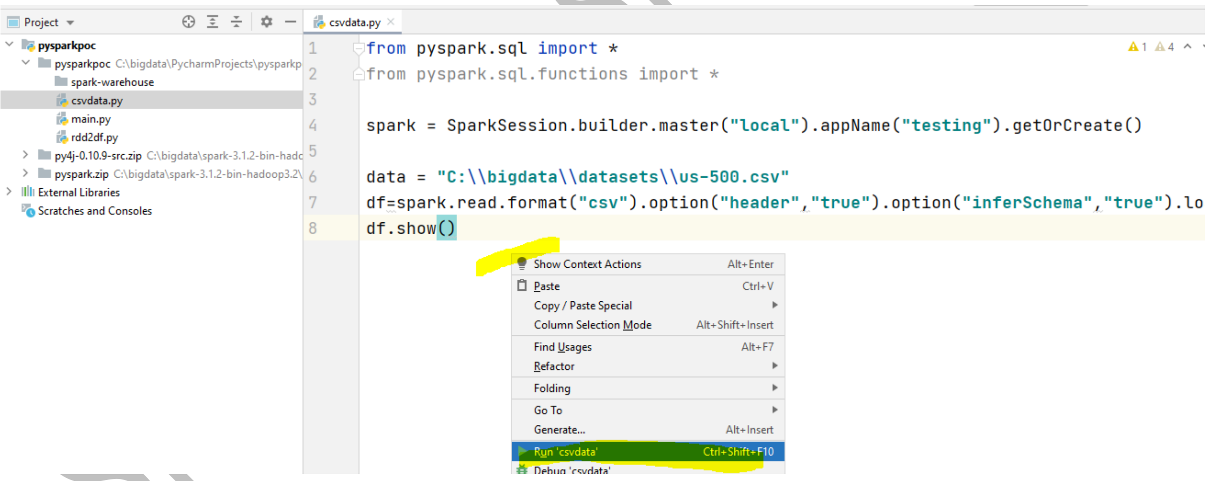
```python
from pyspark.sql import *
from pyspark.sql.functions import *

spark = SparkSession.builder.master("local").a
```

If you paste here what happen, if you create any python object automatically you will get this pyspark code

Testing purpose paste this code and run this program

```
from pyspark.sql import *
from pyspark.sql.functions import *

spark = SparkSession.builder.master("local").appName("testing").getOrCreate()

data = "C:\\bigdata\\datasets\\us-500.csv"
df=spark.read.format("csv").option("header","true").option("inferSchema","true").load(data)
df.show()
```

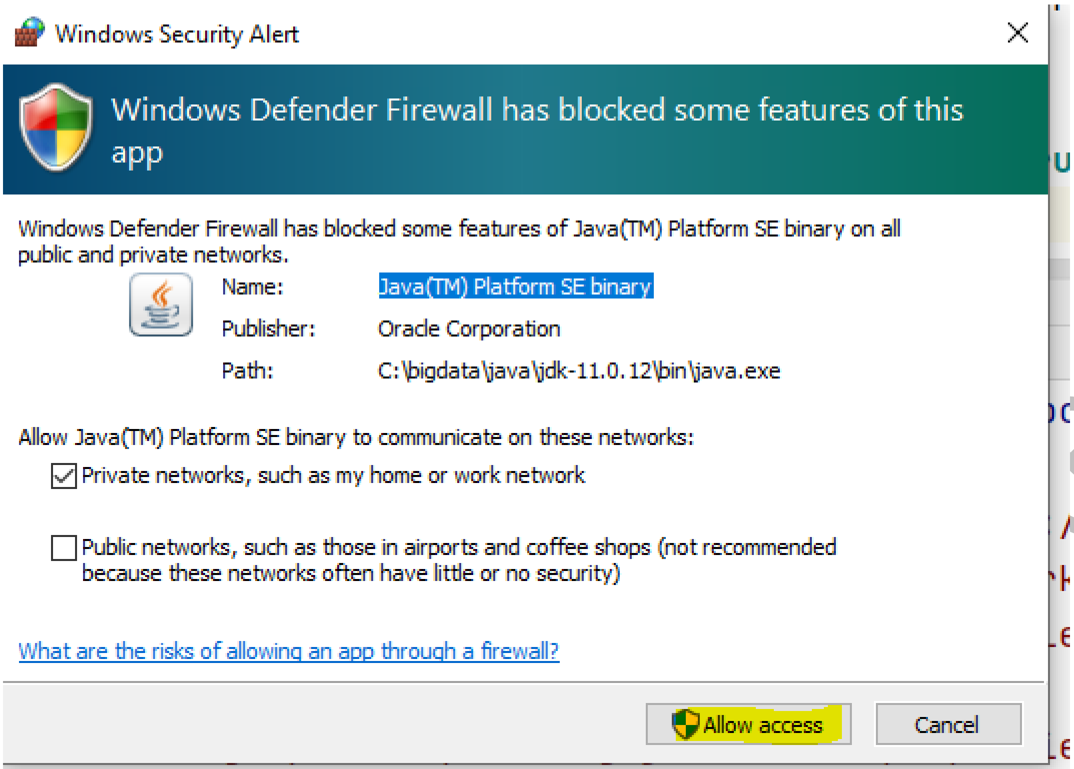
Here you must replace data location data = "C:\\bigdata\\datasets\\us-500.csv"
Similarly must use \\ otherwise python consider as escape character … now right click run this program



```
from pyspark.sql import *
from pyspark.sql.functions import *

spark = SparkSession.builder.master("local").appName("testing").getOrCreate()

data = "C:\\bigdata\\datasets\\us-500.csv"
df=spark.read.format("csv").option("header","true").option("inferSchema","true").lo
df.show()
```

Project
pysparkpoc
pysparkpoc C:\bigdata\PycharmProjects\pysparkp
spark-warehouse
csvdata.py
main.py
rdd2df.py
py4j-0.10.9-src.zip C:\bigdata\spark-3.1.2-bin-hadc
pyspark.zip C:\bigdata\spark-3.1.2-bin-hadoop3.2\
External Libraries
Scratches and Consoles
csvdata.py
Show Context Actions Alt+Enter
Paste Ctrl+V
Copy / Paste Special
Column Selection Mode Alt+Shift+Insert
Find Usages Alt+F7
Refactor
Folding
Go To
Generate... Alt+Insert
Run 'csvdata' Ctrl+Shift+F10
Debug 'csvdata'

If you running first time you will get a popup like this



Windows Security Alert

Windows Defender Firewall has blocked some features of this app

Windows Defender Firewall has blocked some features of Java(TM) Platform SE binary on all public and private networks.

Name: Java(TM) Platform SE binary

Publisher: Oracle Corporation

Path: C:\bigdata\java\jdk-11.0.12\bin\java.exe

Allow Java(TM) Platform SE binary to communicate on these networks:

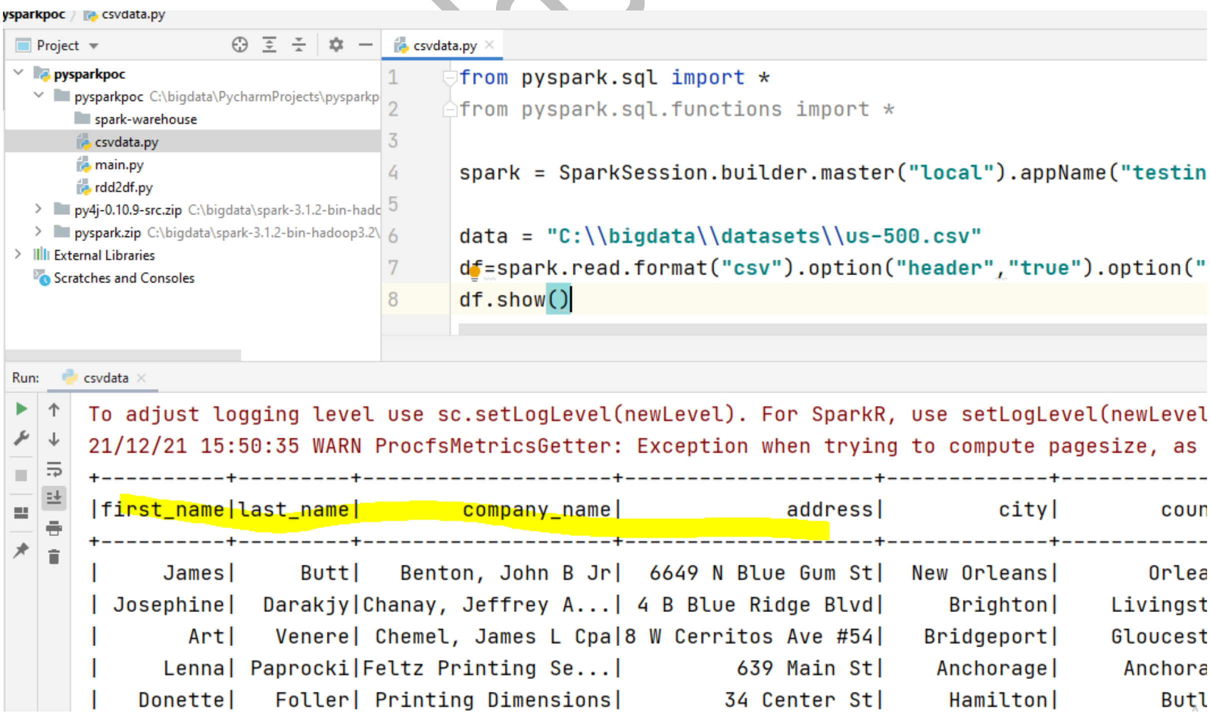
☑ Private networks, such as my home or work network

□ Public networks, such as those in airports and coffee shops (not recommended because these networks often have little or no security)

What are the risks of allowing an app through a firewall?

It means spark want to access java, you are granting permission .. Click Allow access.

Now results you will get like this



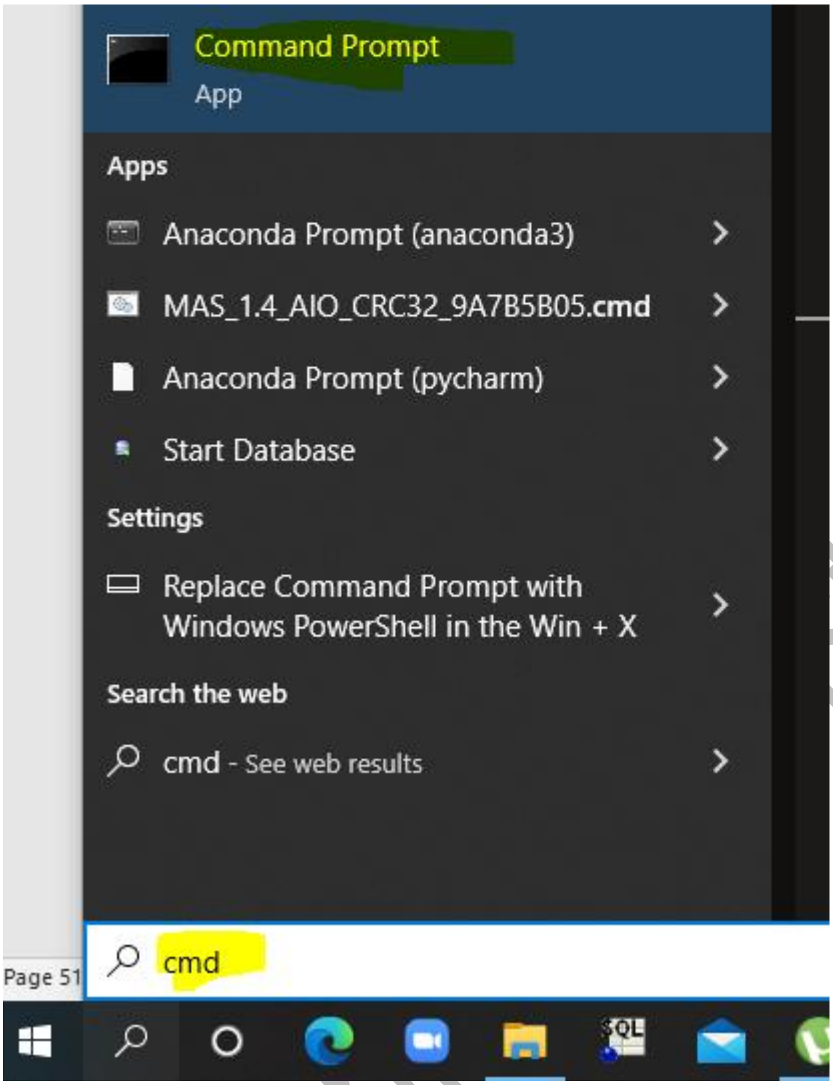
ysparkpoc
csvdata.py
Project
pysparkpoc
pysparkpoc C:\bigdata\PycharmProjects\pysparkp
spark-warehouse
csvdata.py
main.py
rdd2df.py
py4j-0.10.9-src.zip C:\bigdata\spark-3.1.2-bin-hadc
pyspark.zip C:\bigdata\spark-3.1.2-bin-hadoop3.2\
External Libraries
Scratches and Consoles
Run: csvdata

```
from pyspark.sql import *
from pyspark.sql.functions import *

spark = SparkSession.builder.master("local").appName("testin

data = "C:\\bigdata\\datasets\\us-500.csv"
df=spark.read.format("csv").option("header","true").option("
df.show()
```

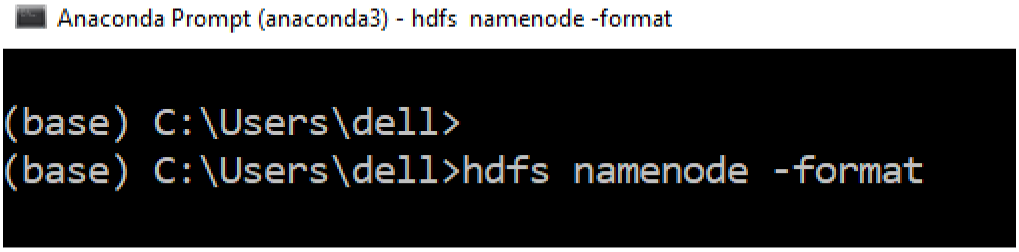
```
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel
21/12/21 15:50:35 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as
+----------+---------+--------------------+--------------------+-------------+------------
|first_name|last_name|        company_name|             address|         city|        coun
+----------+---------+--------------------+--------------------+-------------+------------
|     James|     Butt|   Benton, John B Jr|  6649 N Blue Gum St|  New Orleans|      Orlea
| Josephine|  Darakjy|Chanay, Jeffrey A...|  4 B Blue Ridge Blvd|     Brighton|   Livingst
|       Art|   Venere| Chemel, James L Cpa|8 W Cerritos Ave #54|   Bridgeport|   Gloucest
|     Lenna| Paprocki|Feltz Printing Se...|        639 Main St|    Anchorage|     Anchora
|   Donette|   Foller| Printing Dimensions|        34 Center St|     Hamilton|       Butl
```

It means Spark 100% working fine.

Similarly Hadoop working or not. Open cmd anaconda or normal cmd anything ok

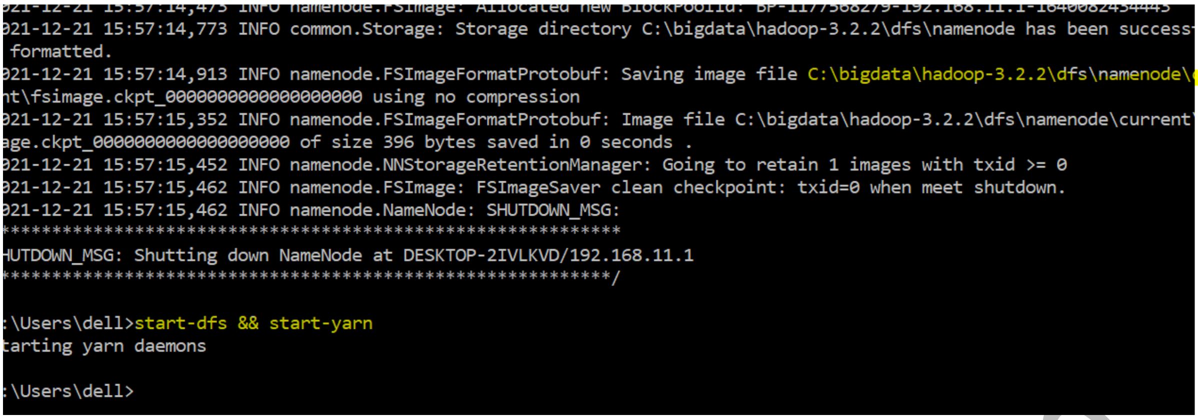

Command Prompt
App
Apps
Anaconda Prompt (anaconda3)
MAS_1.4_AIO_CRC32_9A7B5B05.cmd
Anaconda Prompt (pycharm)
Start Database
Settings
Replace Command Prompt with Windows PowerShell in the Win + X
Search the web
cmd - See web results
Page 51
cmd

Pls execute this command



```
(base) C:\Users\dell>
(base) C:\Users\dell>hdfs namenode -format
```

If you get like this Hadoop 100% formatted now Hadoop 100% working



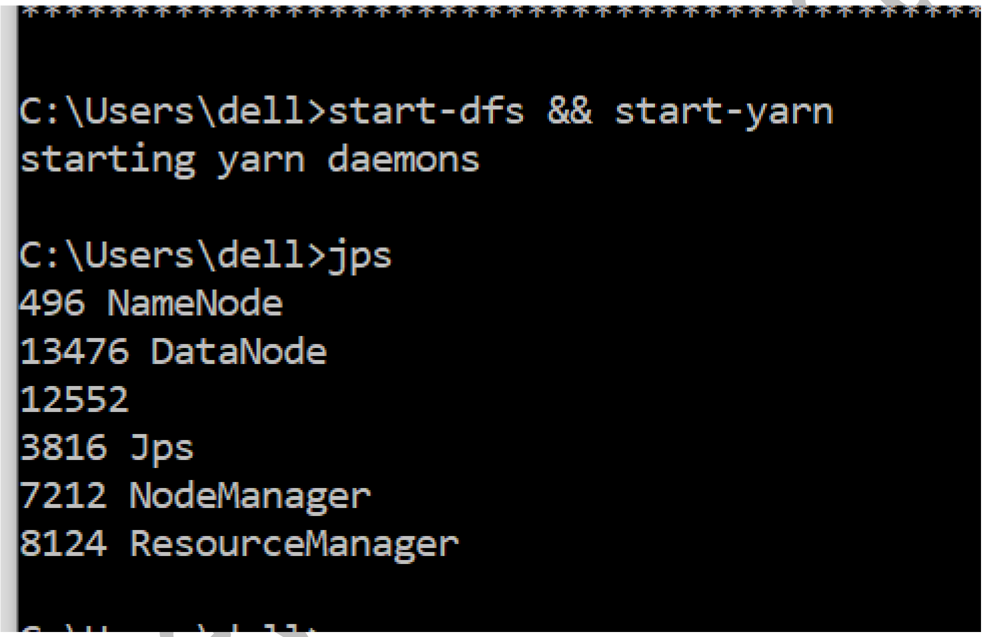
```
021-12-21 15:57:14,473 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1177368279-192.168.11.1-1640082434443
021-12-21 15:57:14,773 INFO common.Storage: Storage directory C:\bigdata\hadoop-3.2.2\dfs\namenode has been success
 formatted.
021-12-21 15:57:14,913 INFO namenode.FSImageFormatProtobuf: Saving image file C:\bigdata\hadoop-3.2.2\dfs\namenode\
nt\fsimage.ckpt_0000000000000000000 using no compression
021-12-21 15:57:15,352 INFO namenode.FSImageFormatProtobuf: Image file C:\bigdata\hadoop-3.2.2\dfs\namenode\current
age.ckpt_0000000000000000000 of size 396 bytes saved in 0 seconds .
021-12-21 15:57:15,452 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
021-12-21 15:57:15,462 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
021-12-21 15:57:15,462 INFO namenode.NameNode: SHUTDOWN_MSG:
************************************************************
HUTDOWN_MSG: Shutting down NameNode at DESKTOP-2IVLKVD/192.168.11.1
************************************************************/

:\Users\dell>start-dfs && start-yarn
tarting yarn daemons

:\Users\dell>
```

Now try any Hadoop commands its working

Now run start-dfs && start-yarn

You will get 4 popup boxes don’t worry just minimise it.(don’t close)

If you get like this



```
****************************************

C:\Users\dell>start-dfs && start-yarn
starting yarn daemons

C:\Users\dell>jps
496 NameNode
13476 DataNode
12552
3816 Jps
7212 NodeManager
8124 ResourceManager

C:\Users\dell>
```

100% Hadoop working fine.

You can start any Hadoop command now.