

JAVA Installation

Open command prompt and see the java version

java -version

If you get below error then install java.

'java' is not recognized as an internal or external command, operable program or batch file.

After Installation set the path for java, windows start → type environment variables → select environment variables.

In **USER Variables** section:

Click New → give variable names as JAVA_HOME, value as installed path

JAVA_HOME = C:\Program Files\Java\jdk1.8.0_281

In **SYSTEM Variables** section:

Double Click Path variable → click New → give value as **C:\Program Files\Java\jdk1.8.0_281\bin**

Now close the command prompt and check again with **java -version**, you can see below output.

java version "1.8.0_281"

Java(TM) SE Runtime Environment (build 1.8.0_281-b09)

Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)

Python Installation

Install the downloaded Anaconda software.

During Installation it will prompt for various options like add path etc.

Select the option as No need to add the path to environment variables, recommended option

This installation takes 30mins time, post installation check python version as below

Windows start → type anaconda prompt → type the cmd **python -V**

Output will be **Python 3.8.5**

Spark Installation

Use 7-Zip to extract the downloaded spark **TGZ** file.

Again 7-Zip to extract the spark **TAR** file.

Create a folder named as **Spark** in the directory **C:\Users\DBREDDY**

Copy the contents of extracted spark folder to **C:\Users\DBREDDY\Spark**

Go to the below path and rename the 6 conf files present in **C:\Users\DBREDDY\Spark\conf**

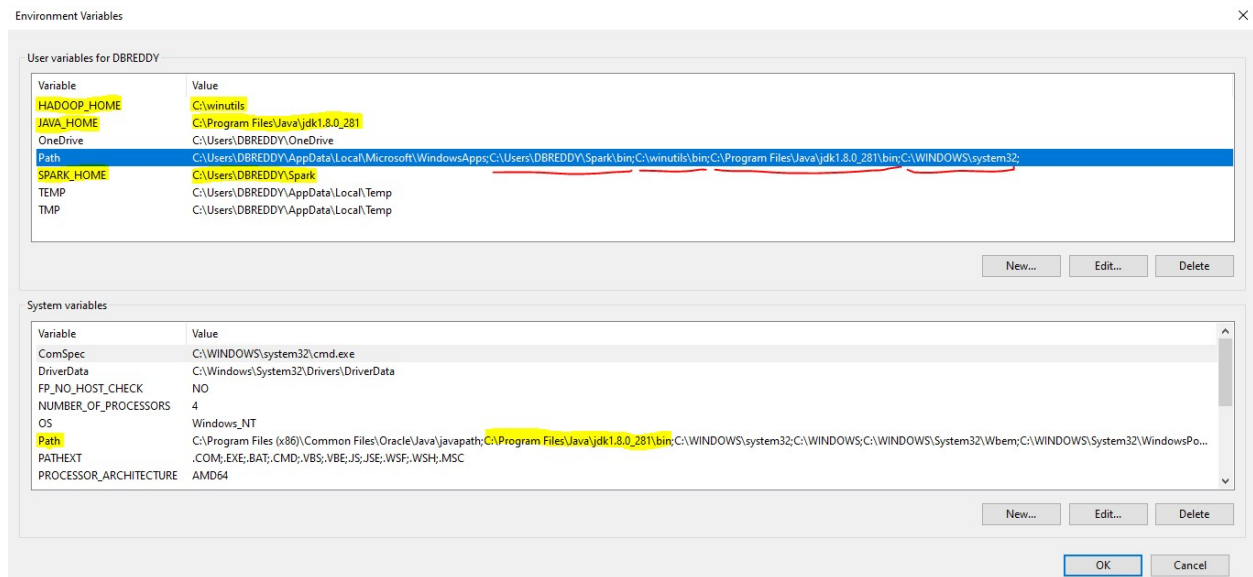
Just rename 6 files by removing the **.template** from their name and you can see the original file type.

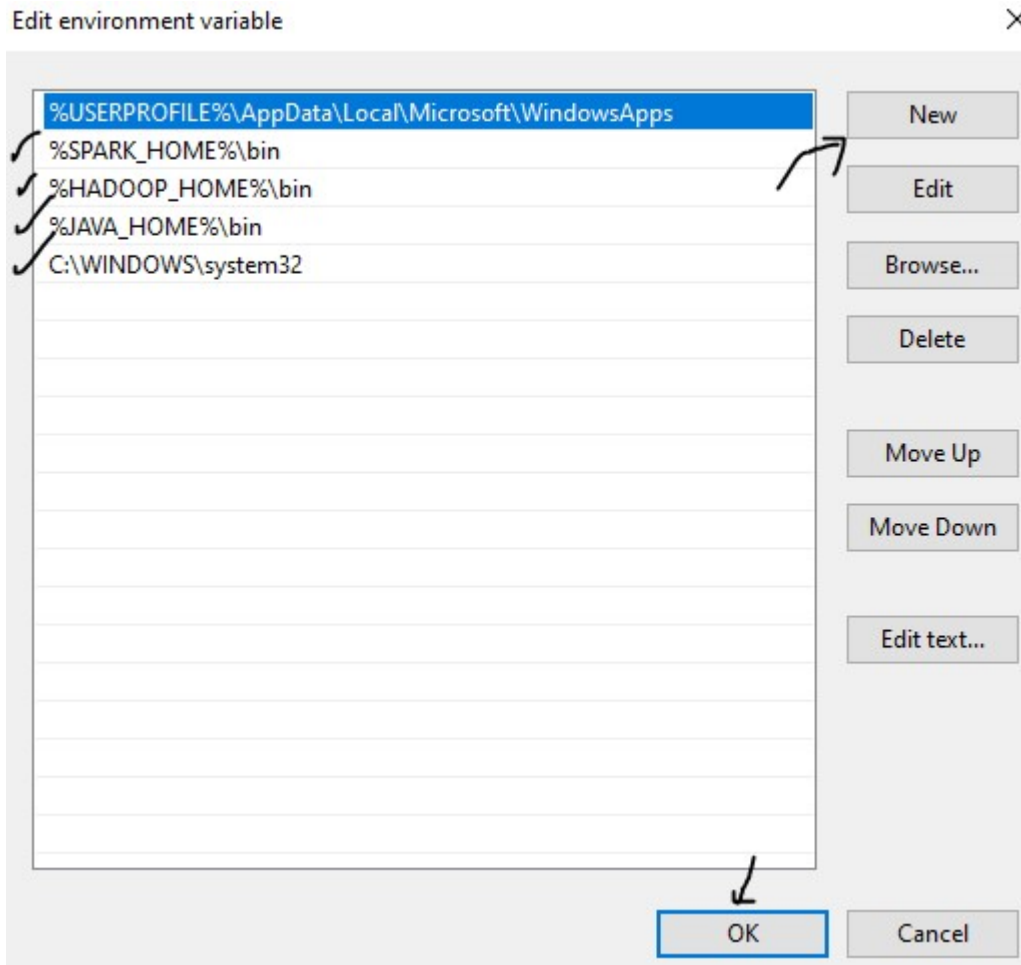
Download the winutils.exe and place in the below path

Create **winutils** folder in **drive C**, further create **bin** folder inside **winutils** and place the **winutils.exe** in the **bin** folder as below

C:\winutils\bin\winutils.exe

Add path for hadoop & spark as below, include the variables(java, hadoop & spark) in user variable **path**(blue color highlighted)





Windows start → cmd (right click run as admin) , it will open cmd and show the path as **C:\WINDOWS\system32**

Create a folder and name it as **tmp** in **drive C**

Inside **tmp** create a subfolder **hive** so the path looks as **C:\tmp\hive**

Execute the following commands in cmd(Run as administrator) option.

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

```
winutils.exe ls -F C:\tmp\hive
```

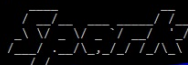
The output is something of the sort:

```
drwxrwxrwx|1|LAPTOP-.....
```

Open Anaconda prompt to check spark installation by cmd **spark-shell**

```
Anaconda Prompt (anaconda3) - spark-shell

(base) C:\Users\DBREDDY>spark-shell
21/05/02 17:48:46 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://DBREDDY-PC:4040
Spark context available as 'sc' (master = local[*], app id = local-1619957949211).
Spark session available as 'spark'.
Welcome to

 version 3.0.2

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_281)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 21/05/02 17:49:29 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped

scala>
```

Spark shell - Environment

Not secure | dbreddy-pc4040/environment/

Spark 3.0.2 | Jobs | Stages | Storage | **Environment** | Executors

Spark shell application UI

Environment

Runtime Information

Name	Value
Java Home	C:\Program Files\Java\jdk1.8.0_281\jre
Java Version	1.8.0_281 (Oracle Corporation)
Scala Version	version 2.12.10

Spark Properties

Name	Value
spark.app.id	local-1619957949211
spark.app.name	Spark shell
spark.app.startTime	1619957942917
spark.driver.host	DBREDDY-PC
spark.driver.port	56095
spark.executor.id	driver
spark.home	C:\Users\DBREDDY\Spark
spark.jars	
spark.master	local[*]
spark.repl.class.outputDir	C:\Users\DBREDDY\AppData\Local\Temp\spark-6a36f8f7-342f-4959-b8b5-8d8407ef2753\repl-479b68be-6fe9-4b77-9fe9-c4d61a62d4da
spark.repl.class.uri	spark://DBREDDY-PC:56095/classes
spark.scheduler.mode	FIFO
spark.sql.catalogImplementation	hive
spark.submit.deployMode	client
spark.submit.pyFiles	

```
Command Prompt

Microsoft Windows [Version 10.0.18363.1500]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\DBREDDY>java -version
java version "1.8.0_281"
Java(TM) SE Runtime Environment (build 1.8.0_281-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)
```

```
Anaconda Prompt (anaconda3)

(base) C:\Users\DBREDDY>python -V
Python 3.8.5

(base) C:\Users\DBREDDY>
```

How to open jupyter notebook ?

Windows start → type Anaconda prompt → use cmd **jupyter-notebook**

Redirects the request via browser

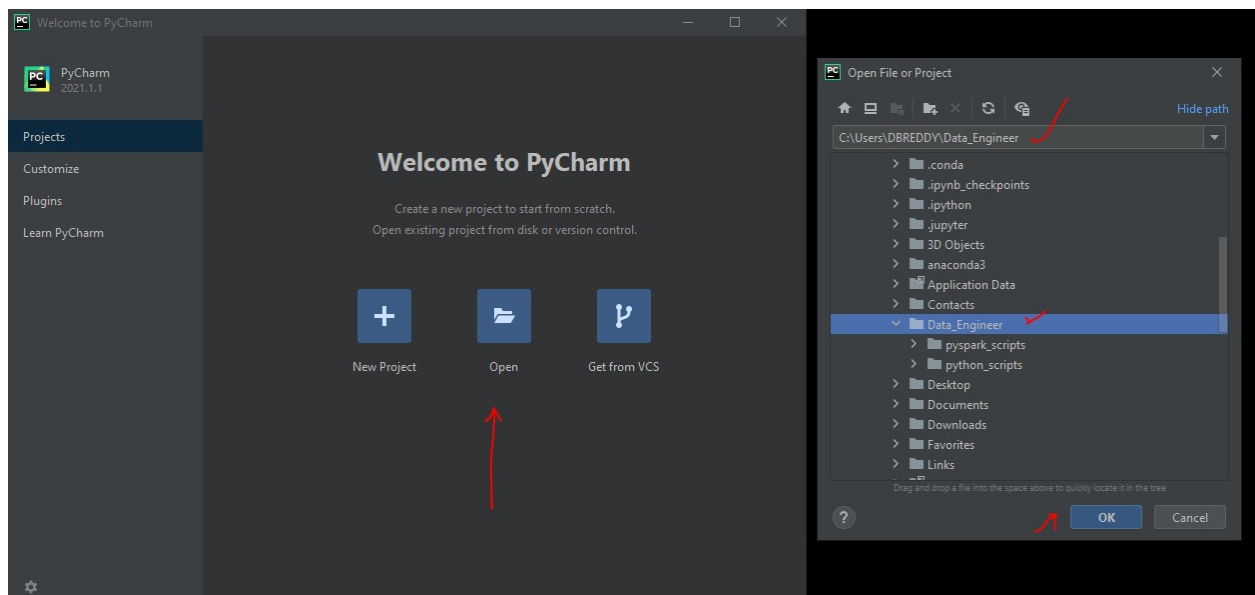
```
Anaconda Prompt (anaconda3) - jupyter-notebook
(base) C:\Users\DBREDDY>jupyter-notebook ✓
[I 17:56:41.035 NotebookApp] JupyterLab extension loaded from C:\Users\DBREDDY\anaconda3\lib\site-packages\jupyterlab
[I 17:56:41.035 NotebookApp] JupyterLab application directory is C:\Users\DBREDDY\anaconda3\share\jupyter\lab
[I 17:56:41.040 NotebookApp] Serving notebooks from local directory: C:\Users\DBREDDY
[I 17:56:41.042 NotebookApp] Jupyter Notebook 6.1.4 is running at:
[I 17:56:41.042 NotebookApp] http://localhost:8888/?token=0e3a9f6a8f6ad53070c8e32704ac73180d8c5422008a1d4c
[I 17:56:41.042 NotebookApp] or http://127.0.0.1:8888/?token=0e3a9f6a8f6ad53070c8e32704ac73180d8c5422008a1d4c
[I 17:56:41.042 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 17:56:41.200 NotebookApp]

✓ To access the notebook, open this file in a browser:
  file:///C:/Users/DBREDDY/AppData/Roaming/jupyter/runtime/nbserver-2916-open.html
Or copy and paste one of these URLs:
  http://localhost:8888/?token=0e3a9f6a8f6ad53070c8e32704ac73180d8c5422008a1d4c
  or http://127.0.0.1:8888/?token=0e3a9f6a8f6ad53070c8e32704ac73180d8c5422008a1d4c
```

Other way to open is Windows start → type **jupyter notebook** and open it

Pycharm IDE Setup

Install Pycharm, open it and you can see below options, select open existing project



C:\Users\DBREDDY\anaconda3\python.exe

Pycharm File → Settings → Project → Project Interpreter

File Edit View Navigate Code Refactor Run Tools VCS Window Help Data_Engineer

Data_Engineer pyspark_scripts

Project

- Project
- Data_Engineer C:\Users\DBREDDY\Data_Engineer
 - pyspark_scripts
 - python_scripts
 - External Libraries
 - Scratches and Consoles

Structure

Python Package

Python 3.8 has been configured as a project interpreter // Config

Settings

Appearance & Behavior

Keymap

Editor

Plugins

Version Control

Project: Data_Engineer

Python Interpreter

Project Structure

Build, Execution, Deployment

Languages & Frameworks

Tools

Project: Data_Engineer > Python Interpreter

Python Interpreter: Python 3.8 C:\Users\DBREDDY\anaconda3\python.exe

Package	Version	Latest version
_ipyw_jlab_nb_ext_conf	0.1.0	
alabaster	0.7.12	
anaconda	2020.11	
anaconda-client	1.7.2	
anaconda-navigator	1.10.0	
anaconda-project	0.8.4	
argh	0.26.2	
argon2-cffi	20.1.0	
asn1crypto	1.4.0	
astroid	2.4.2	
astropy	4.0.2	
async_generator	1.10	
atomicwrites	1.4.0	
attrs	20.3.0	
autopep8	1.5.4	
babel	2.8.1	
backcall	0.2.0	
backports	1.0	
backports.functools_lru_cache	1.6.1	
backports.shutil_get_terminal_size	1.0.0	
backports.tempfile	1.0	
backports.weakref	1.0.post1	
bcrypt	3.2.0	

Event Log

Python 3.8