

ORACLE®

Certified Professional

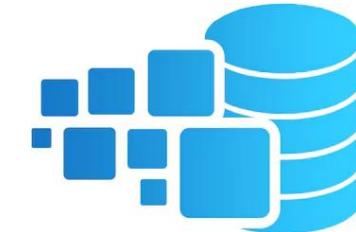
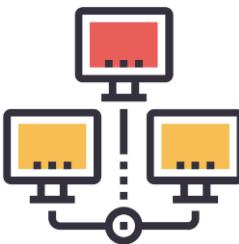


Oracle Certified
Professional Java
Programmer

Microsoft Certified
Professional



Computer
SCIENCE



Dr. SOHAIL IMRAN



APACHE
Spark™
Install



Login and Download



1

Cloud Account

[Sign in to Cloud](#)

[Sign Up for Free Cloud Tier](#)

Oracle Account

Sohail IMRAN

Help

[Sign Out](#)

2

<https://www.oracle.com/java/technologies/downloads/#java17-windows>

ORACLE Products Industries Resources Customers Partners Developers Company

Java downloads Tools and resources Java archive

Java SE Development Kit 17.0.16 downloads

Java SE subscribers will receive JDK 17 updates until at least **September 2029**.

The Oracle JDK 17 license changed in October 2024

The [Oracle Technology Network License Agreement for Oracle Java SE](#) used for JDK 17 updates 17.0.13 and greater is substantially different from prior Oracle Java licenses. It permits personal use and development use, such as personal use and development use, at no cost -- but other uses authorized under prior Oracle JDK licenses may not be permitted. FAQs are available [here](#).

Commercial license and support are available for a low cost with [Java SE Universal Subscription](#).

JDK 17.0.16 [checksums](#) and [OL 9 GPG Keys](#) for RPMs

Linux macOS Windows

Product/file description	File size	Download
x64 Compressed Archive	172.93 MB	 jdk-17.0.16_windows-x64_bin.zip
x64 Installer	154.10 MB	 jdk-17.0.16_windows-x64_bin.exe
x64 MSI Installer	152.85 MB	 jdk-17.0.16_windows-x64_bin.msi

3

Recent download history

	jdk-17.0.16_windows-x64_bin.zip
	173 MB • 6 minutes ago



APACHE
Spark

big
data
ANALYTICS

Setup JAVA_HOME

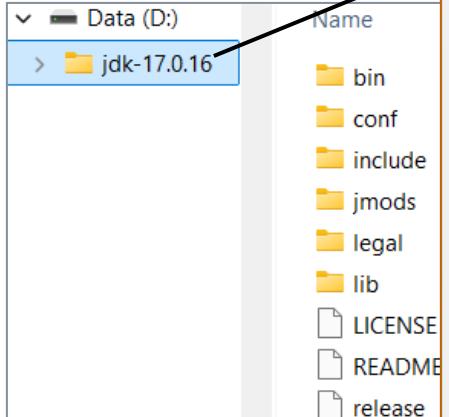


Recent download history

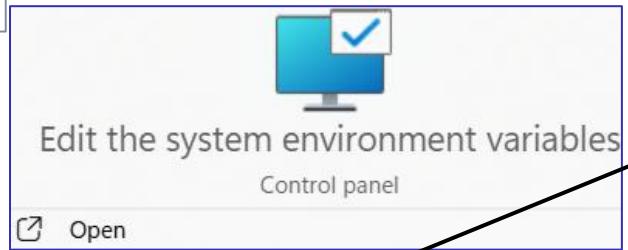
jdk-17.0.16_windows-x64_bin.zip
173 MB • 6 minutes ago

Extract

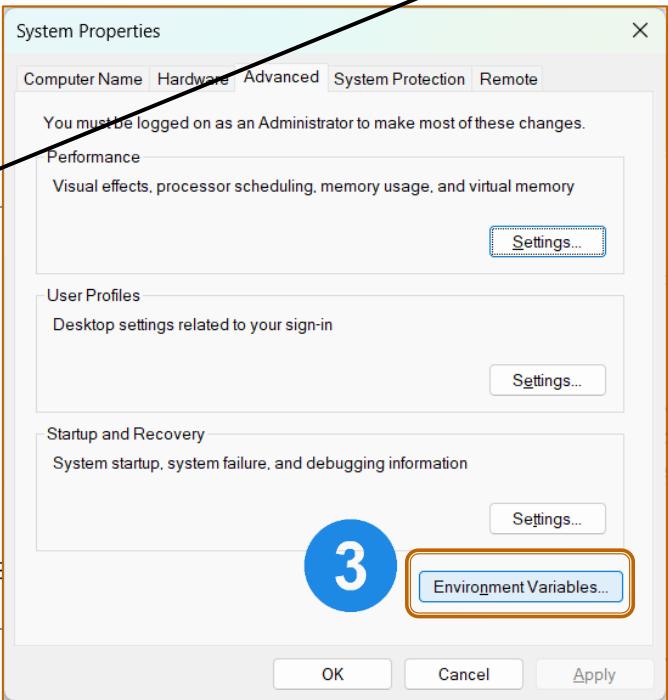
1



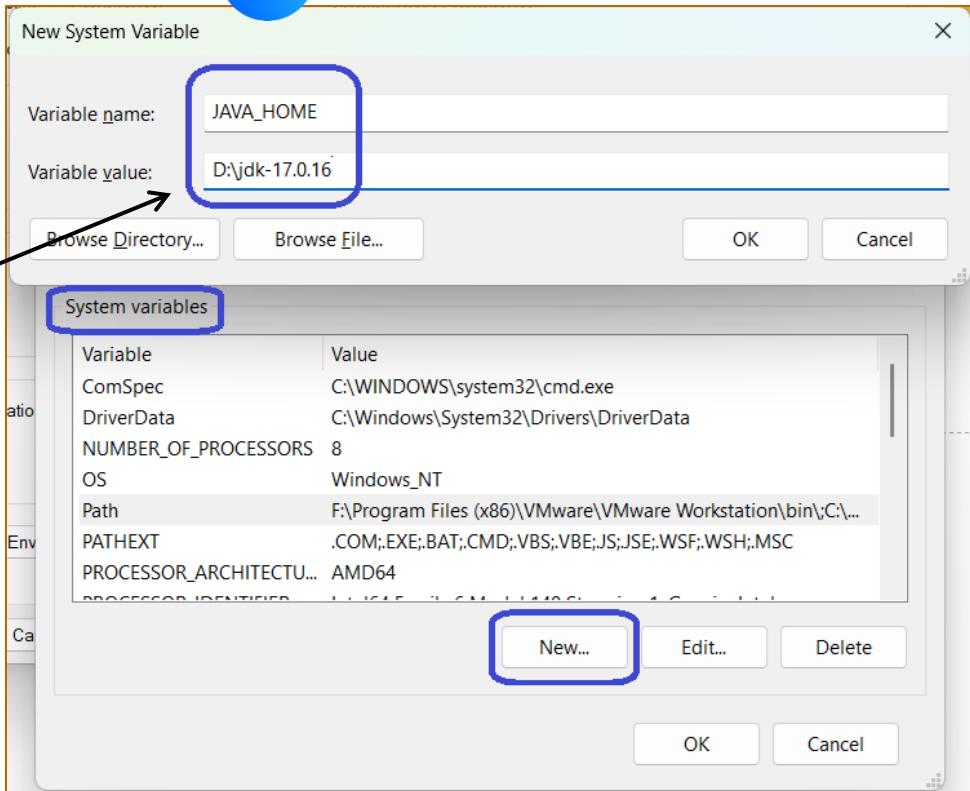
2



3



4



APACHE
Spark
big
data
ANALYTICS

Setup - PATH



1

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
JAVA_HOME	D:\jdk-17.0.16
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	F:\Program Files (x86)\VMware\VMware Workstation\bin;C:\...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

New... Edit... Delete

2

Edit environment variable

F:\Program Files (x86)\VMware\VMware Workstation\bin\
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
%USERPROFILE%\dnx\bin
C:\Program Files\Microsoft DNX\Dnvm\
C:\Program Files (x86)\Windows Kits\8.1\Windows Performance ...
C:\Program Files\Microsoft SQL Server\120\Tools\Binn\
C:\Program Files (x86)\Microsoft SQL Server\150\DT\Binn\
C:\Program Files\Azure Data Studio\bin
C:\Program Files (x86)\nodejs\
D:\jdk-17.0.16\bin

New Edit... Browse... Delete Move Up Move Down Edit text... OK Cancel

3
`java -version`

Windows PowerShell

```
PS C:\Users\DELL> java -version
java version "1.7.0_80"
Java(TM) SE Runtime Environment (build 1.7.0_80-b15)
Java HotSpot(TM) 64-Bit Server VM (build 24.80-b11, mixed mode)
PS C:\Users\DELL>
```



APACHE
Spark
big
data
ANALYTICS



Download & Extract



- 1 <https://dlcdn.apache.org/spark/spark-4.0.1/spark-4.0.1-bin-hadoop3.tgz>
- 2 <https://code.load.github.com/cdarlint/winutils/zip/refs/heads/master>

A screenshot of a GitHub repository page. The URL is `github.com/cdarlint/winutils/tree/master`. The repository name is `cdarlint / winutils` and it is public. There are 26 issues and 3 pull requests. The master branch is selected. Two commits are listed:

- `hadoop-2.6.1/bin` - add 261 (4 years ago)
- `hadoop-2.6.5/bin` - fixed exe and lib... (6 years ago)

A screenshot of a Windows File Explorer window titled "Downloads". It shows three files:

- `spark-4.0.1-bin-hadoop3.tgz`
- `winutils-master.zip`
- `jdk-17.0.16_windows-x64_bin.zip`

A large blue circle containing the number "3" is positioned above the first file, with a curved arrow pointing down towards it. To the right of the "Downloads" window, there is a watermark-like graphic for "APACHE Spark big data ANALYTICS" featuring the Apache logo and a star.

System Variables & Path



Windows



1

Edit System Variable

Variable name: HADOOP_HOME

Variable value: D:\hadoop-3.3.6

Browse Directory... Browse File... OK Cancel

2

System variables	
Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
JAVA_HOME	D:\jdk-17.0.16
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	F:\Program Files (x86)\VMware\VMware Workstation\bin;C:\...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMDX86

Edit...

Delete

3

Edit System Variable

Variable name: SPARK_HOME

Variable value: D:\spark-4.0.1-bin-hadoop3

Browse Directory... Browse File... OK Cancel

Edit environment variable

F:\Program Files (x86)\VMware\VMware Workstation\bin
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
%USERPROFILE%\dnx\bin
C:\Program Files\Microsoft DNX\Dnvm\
C:\Program Files (x86)\Windows Kits\8.1\Windows Performance ...
C:\Program Files\Microsoft SQL Server\120\Tools\Binn\
C:\Program Files (x86)\Microsoft SQL Server\150\DTS\Binn\
C:\Program Files\Azure Data Studio\bin
C:\Program Files (x86)\nodejs\
D:\jdk-17.0.16\bin
D:\hadoop-3.3.6\bin
D:\spark-4.0.1-bin-hadoop3\bin

New Edit Browse... Delete Move Up Move Down Edit text... OK Cancel

New System Variable

Variable name: PYSPARK_PYTHON

Variable value: python

Browse Directory... Browse File... OK Cancel



spark-shell



spark-shell

```
PS C:\Users\DELL> spark-shell
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      /---\
     / \ \ - \ \ - \ \ - \ \ - \
    /_ _/ . __ \_ , _ / / _ / \_ \
   /_ /           version 4.0.1

Using Scala version 2.13.16 (Java HotSpot(TM) 64-Bit Server VM, Java 17.0.16)
Type in expressions to have them evaluated.
Type :help for more information.
Spark context Web UI available at http://DESKTOP-9VA0BA3:4040
Spark context available as 'sc' (master = local[*], app id = local-176113995553).
Spark session available as 'spark'.

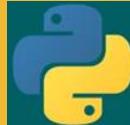
scala> |
```



APACHE
Spark
big
data
ANALYTICS



Prompt



PYTHON ENV



Anaconda Prompt

All Apps

Best match

Anaconda Prompt
App

python --version

```
Anaconda Prompt
(base) C:\Users\DELL>python --version
Python 3.13.5
```

conda env list

```
Anaconda Prompt
(base) C:\Users\DELL>conda env list
# conda environments:
#
base F:\ProgramData\anaconda3
```

conda create -n spark310 python=3.10

```
Anaconda Prompt
(base) C:\Users\DELL>conda create -n spark310 python=3.10
```



APACHE

Spark

big

data
ANALYTICS



conda activate spark310

```
Anaconda Prompt + - X
(base) C:\Users\DELL>conda activate spark310
(spark310) C:\Users\DELL>python --version
Python 3.10.19
(spark310) C:\Users\DELL>pip install pyspark
Collecting pyspark
  Using cached pyspark-4.0.1.tar.gz (434.2 MB)
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.9 (from pyspark)
  Using cached py4j-0.10.9.9-py2.py3-none-any.whl.metadata (1.3 kB)
Using cached py4j-0.10.9.9-py2.py3-none-any.whl (203 kB)
Building wheels for collected packages: pyspark
  DEPRECATION: Building 'pyspark' using the legacy setup.py bdist_wheel mechanism, which will be removed in a future version. pip 25.3 will enforce this behaviour change. A possible replacement is to use the standardized build interface by setting the '--use-pep517' option, (possibly combined with '--no-build-isolation'), or adding a 'pyproject.toml' file to the source tree of 'pyspark'. Discussion can be found at https://github.com/pypa/pip/issues/6334
Building wheel for pyspark (setup.py) ...+
```

python --version

pip install pyspark

conda install ipykernel





PYTHON ENV



Alternate Method
to Connect

New File



Next Page

Existing File →

Home

localhost:8888/tree/Documents/Code/BDA?

jupyter

File View Settings Help

Files Running

Select items to perform actions on them.

/ ... / Code / BDA /

Name

New

Python [conda env:conda-spark310]
Python [conda env:base] *
Terminal
Console
New File
New Folder

1 Kernel

Interrupt Kernel

Restart Kernel...

Restart Kernel and Clear Outputs of All Cells...

Restart Kernel and Run up to Selected Cell...

Restart Kernel and Run All Cells...

Restart Kernel and Debug...

Reconnect to Kernel

Shut Down Kernel

Shut Down All Kernels...

2 Change Kernel...

3 Select Kernel

Select kernel for: "Untitled2.ipynb"

Python [conda env:conda-spark310]

Start Kernel

Python [conda env:conda-spark310]

Python [conda env:base] *

Use No Kernel

No Kernel

Connect to Existing Kernel

Untitled2.ipynb (b1bfc5c7)

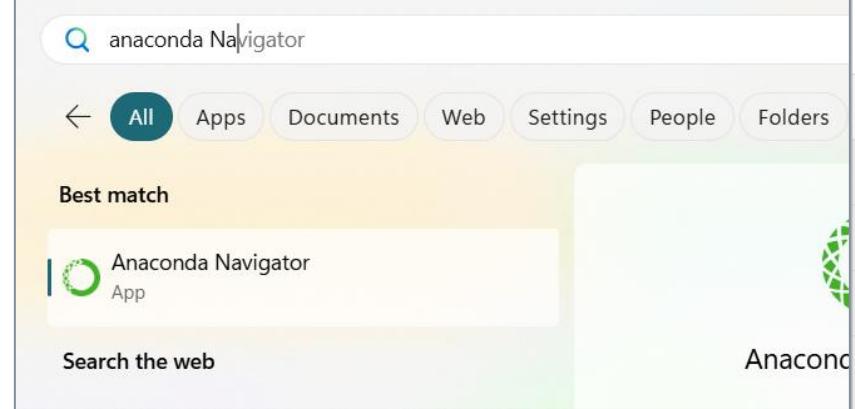


APACHE
Spark
big
data
ANALYTICS

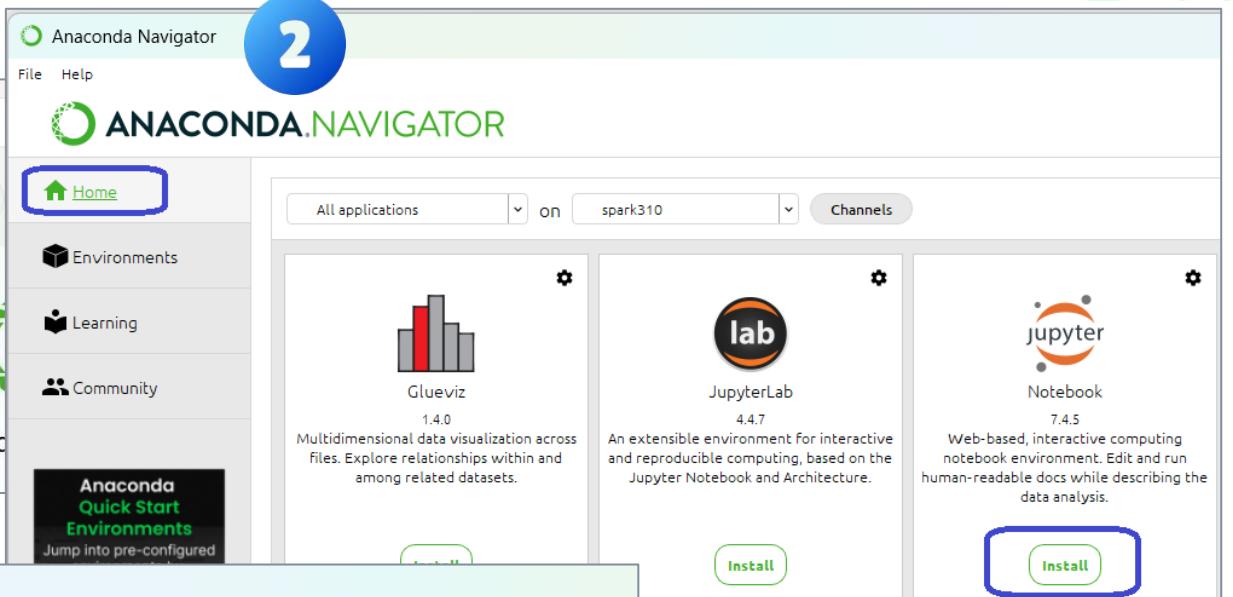
Alternate Method to Connect



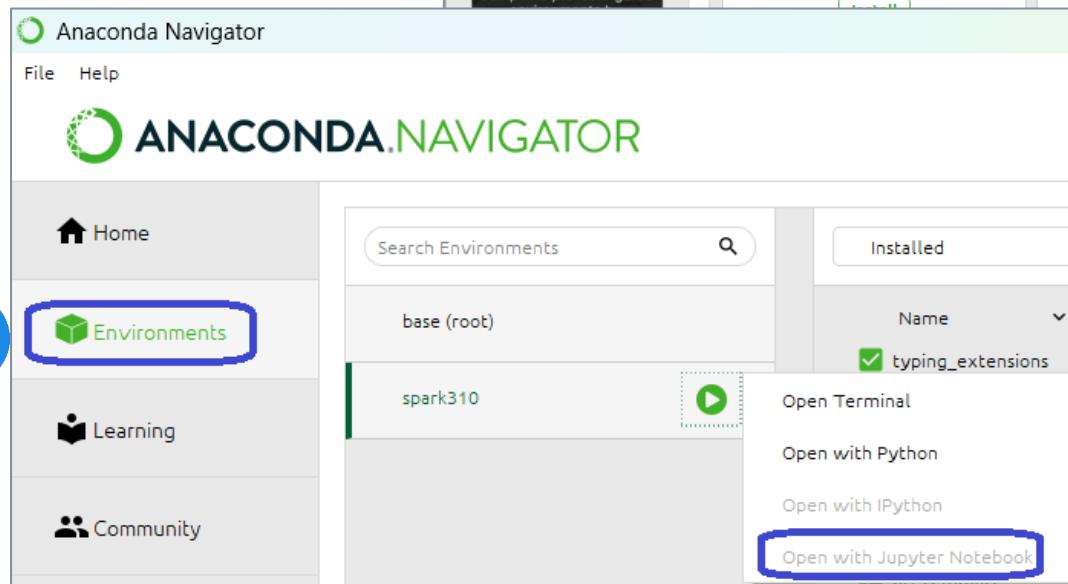
1



2



3



```
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder.appName('pySp0').getOrCreate()
```

```
[1]: from pyspark.sql import SparkSession
```

```
[2]: spark = SparkSession.builder.appName('pySp0').getOrCreate()
```

```
[3]: spark
```

```
[3]: SparkSession - in-memory
```

```
      SparkContext
```

```
      Spark UI
```

Version	v4.0.1
Master	local[*]
AppName	pySp0



install pyspark



```
pip install pyspark
```

```
Defaulting to user installation because normal site-packages is not writeable
```

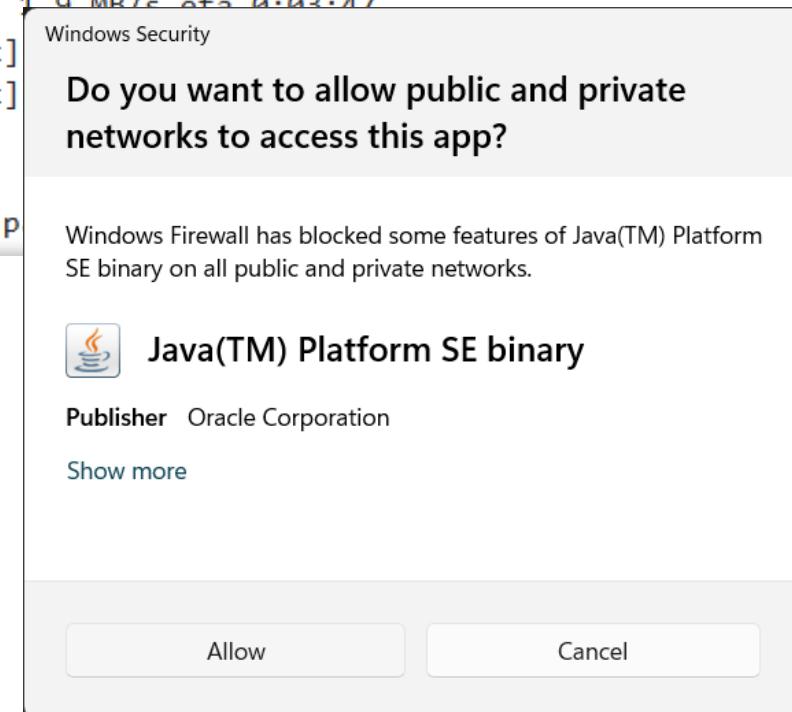
```
Collecting pyspark
```

```
  Downloading pyspark-4.0.1.tar.gz (434.2 MB)
```

```
----- 0.0/434.2 MB ? eta ----  
----- 0.3/434.2 MB ? eta ----  
----- 0.8/434.2 MB 2.3 MB/s eta 0:03:11  
----- 1.0/434.2 MB 1.9 MB/s eta 0:03:17  
----- 1/2 [pyspark]  
----- 2/2 [pyspark]
```

```
Successfully installed py4j-0.10.9.9 pyspark-4.0.1
```

```
Note: you may need to restart the kernel to use updated p
```



APACHE
Spark
big
data
ANALYTICS

delete py environment



```
conda deactivate spark310
```

```
conda remove --name spark310 --all
```



APACHE
Spark
big
data
ANALYTICS