

# BigData Analytics

Dr. SOHAIL IMRAN

APACHE  
**Spark**  
Install

# Login and Download



ORACLE®

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 substantial, such as personal use and development use, at no cost -- but other uses authorized under prior Oracle JDK licenses may and using this product. 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	<a href="#">jdk-17.0.16_windows-x64_bin.zip</a>
x64 Installer	154.10 MB	<a href="#">jdk-17.0.16_windows-x64_bin.exe</a>
x64 MSI Installer	152.85 MB	<a href="#">jdk-17.0.16_windows-x64_bin.msi</a>

## Recent download history



jdk-17.0.16\_windows-x64\_bin.zip

173 MB • 6 minutes ago

3



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

Data (D:)

> jdk-17.0.16

- bin
- conf
- include
- jmods
- legal
- lib
- LICENSE
- README
- release

2

Edit the system environment variables

Control panel

Open

System Properties

Computer Name Hardware Advanced System Protection Remote

You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage, and virtual memory

Settings...

User Profiles

Desktop settings related to your sign-in

Settings...

Startup and Recovery

System startup, system failure, and debugging information

Settings...

Environment Variables...

OK Cancel Apply

3

4

New System Variable

Variable name: JAVA\_HOME

Variable value: D:\jdk-17.0.16

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

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
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_ARCHITECTU...	AMD64
PROCESSOR_IDENTIFIER	...

New... Edit... Delete

OK Cancel

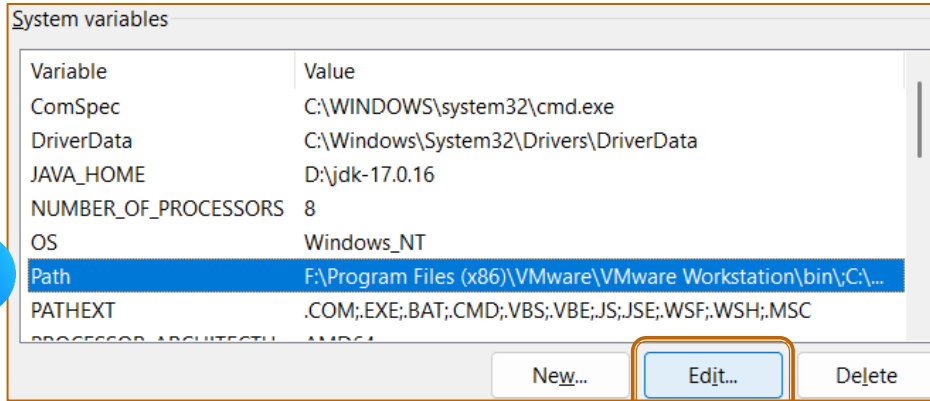


APACHE  
Spark  
big  
data  
ANALYTICS

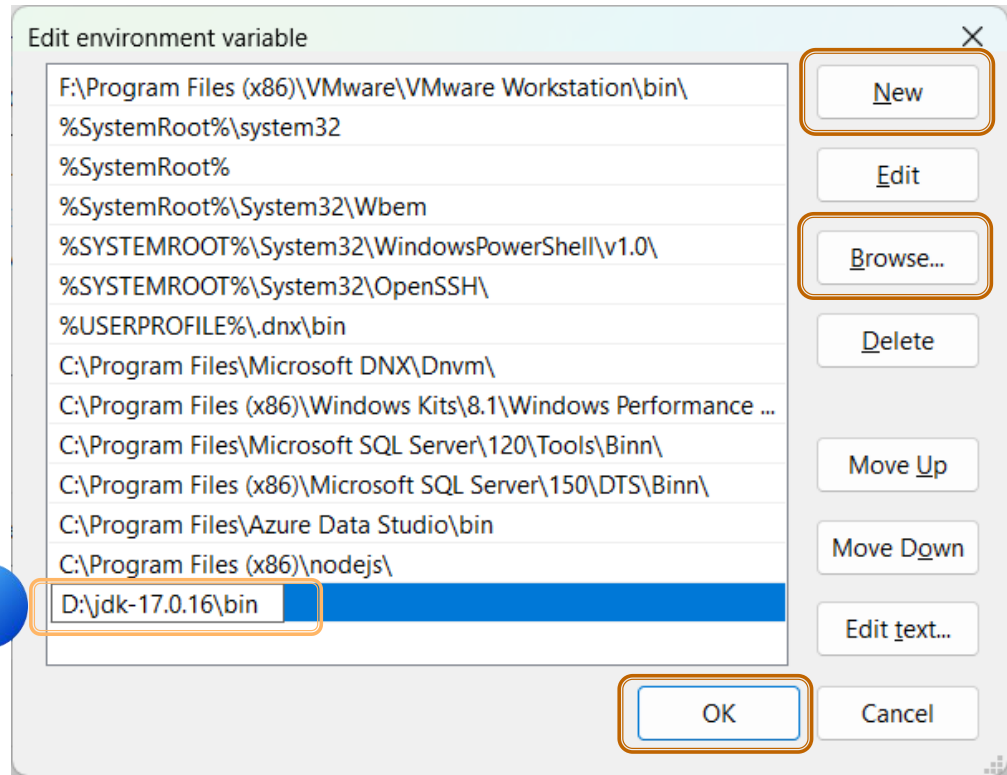
# Setup - PATH



1



2



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> |
```



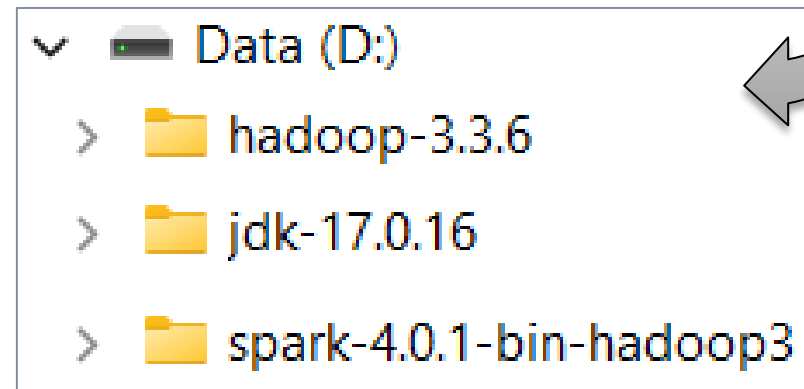
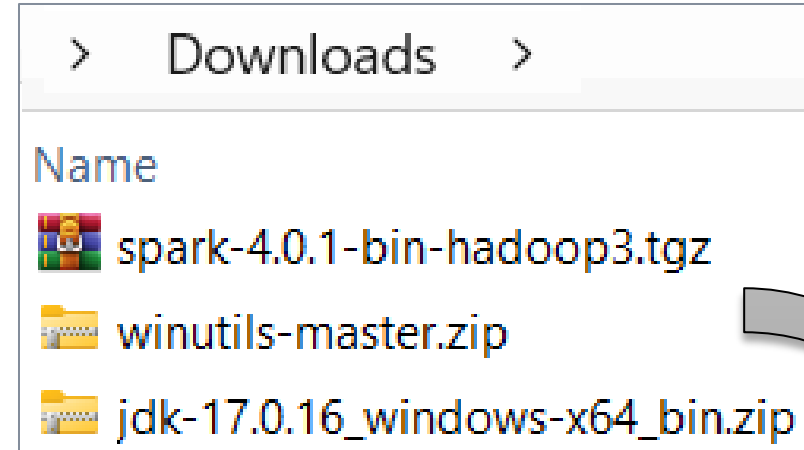
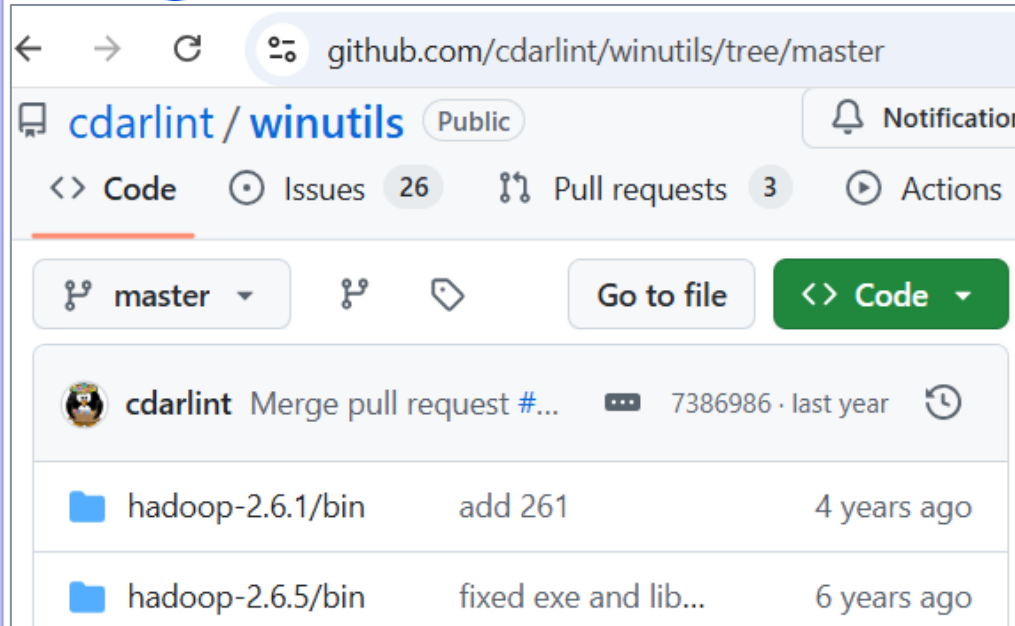


# Download & Extract

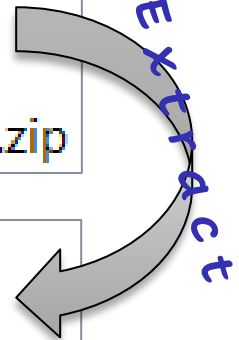


1 <https://dlcdn.apache.org/spark/spark-4.0.1/spark-4.0.1-bin-hadoop3.tgz>

2 <https://codeload.github.com/cdarlint/winutils/zip/refs/heads/master>



3



Extract



APACHE  
Spark™

big

data

ANALYTICS



# System Variables & Path



Windows



1

Edit System Variable

Variable name:

Variable value:

Edit System Variable

Variable name:

Variable value:

New System Variable

Variable name:

Variable value:

2

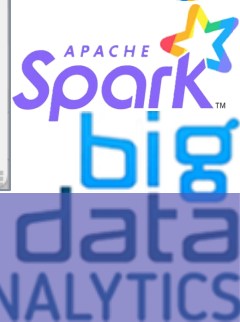
System variables

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	AMD64

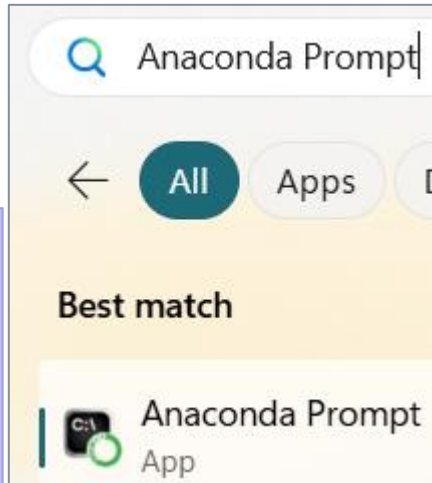
Edit environment variable

%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\DTBS\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

3







**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
    
```



conda activate spark310

```
(base) C:\Users\DELL>conda activate spark310
```

```
(spark310) C:\Users\DELL>python --version  
Python 3.10.19
```

python --version

```
(spark310) C:\Users\DELL>pip install pyspark  
Collecting pyspark
```

pip install pyspark

```
Using cached pyspark-4.0.1.tar.gz (434.2 MB)  
Preparing metadata (setup.py) ... done
```

conda install ipykernel

```
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 ENV



Alternate Method  
to Connect

Existing File

Next Page

New 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

jupyter Untitled1

Checkpoint: 3 minutes ago

File Edit View Run Kernel Settings Help

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...

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)



# Alternate Method to Connect



**1**

Search for Anaconda Navigator in the Windows Start menu. The search results show "Anaconda Navigator App" as the best match.

**2**

Open Anaconda Navigator. The Home screen displays the "All applications" section, which includes Glueviz, JupyterLab, and Jupyter Notebook. The "Install" button for Jupyter Notebook is highlighted.

**3**

Click on the "Environments" tab in the left sidebar. The "Environments" list shows "base (root)" and "spark310". The "spark310" environment is selected, and the "Open with Jupyter Notebook" option is highlighted in the context menu.



```
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
```

Windows Security

**Do you want to allow public and private networks to access this app?**

Windows Firewall has blocked some features of Java(TM) Platform SE binary



**Java(TM) Platform SE binary**

**Publisher** Oracle Corporation

[Show more](#)

Allow

Cancel



# delete py environment



`conda deactivate spark310`

`conda remove --name spark310 --all`



APACHE  
**Spark**  
**big**

**data**  
ANALYTICS