01 Spark Installation

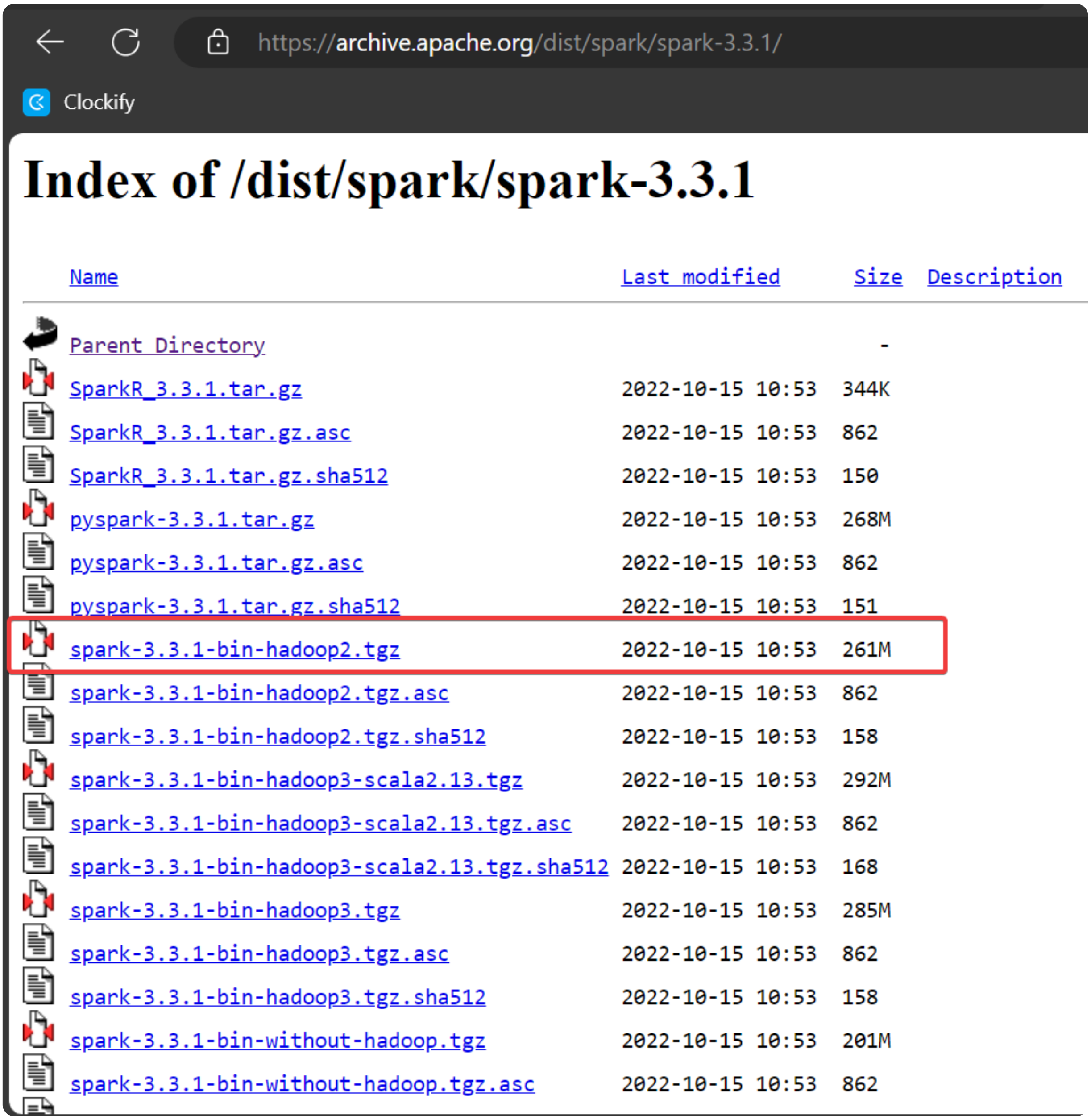
Instruction: Read this document at least two times and then assess yourself whether you can follow these installation steps.

Requirements

1. Java 19
2. Python ~~latest~~ 3.10.1 (Yes, it is little old but you will not miss anything. Don't Worry!)
3. spark 3.3.1 for hadoop 2.7

---

1. Java
   Java Archive Downloads - Java SE 19 (oracle.com)
   - Download the windows x64 installer file
   - install it when asked to choose path - choose `C:\Java\jdk` (Go to your C Drive create a folder called Java and inside it create another folder jdk)
   - Done!
2. Python
   https://www.python.org/downloads/
   - ~~Just install the latest python exe and enable path.~~
   - Install Python 3.10.1 because spark 3.3.1 is not compatible with later versions
   - https://www.python.org/downloads/release/python-3101/
3. spark
   Index of /dist/spark/spark-3.3.1 (apache.org)

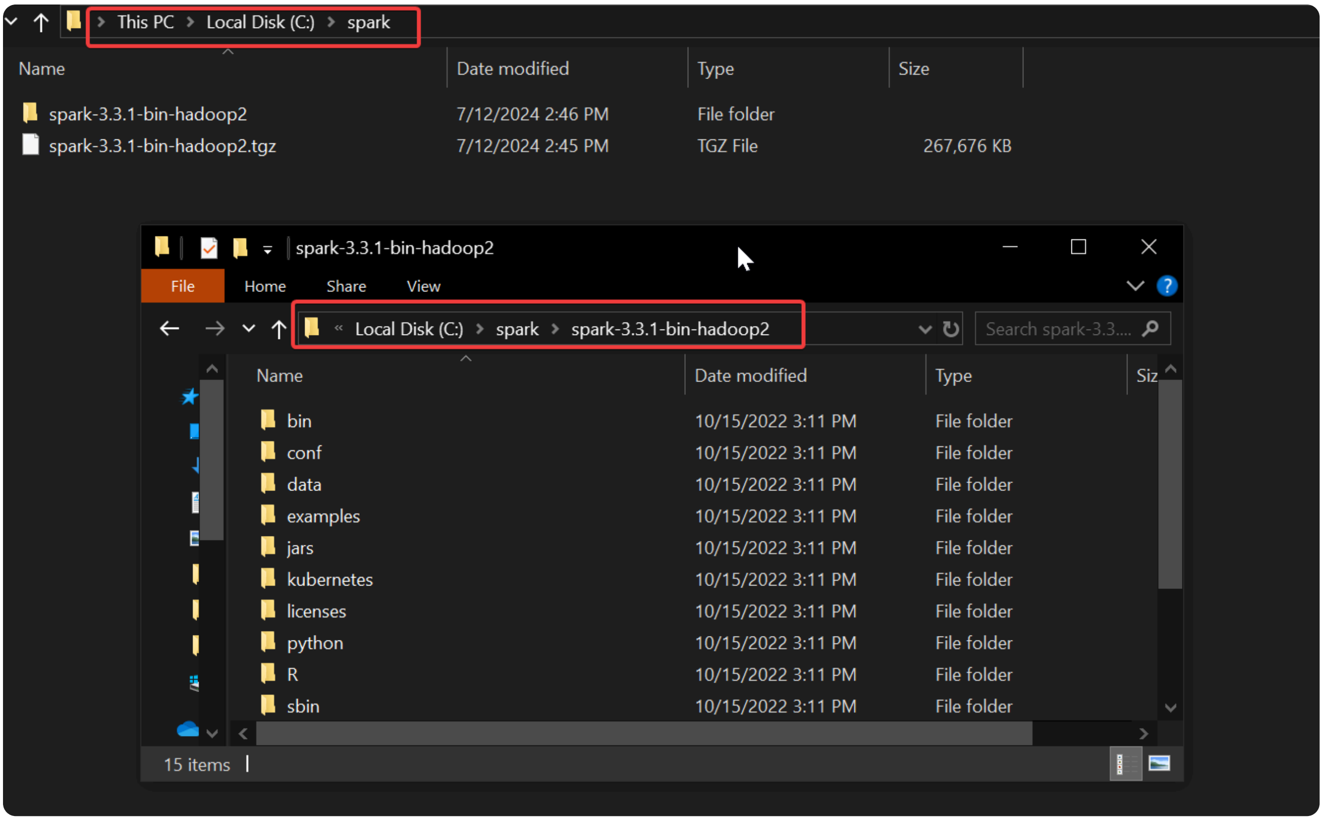


https://archive.apache.org/dist/spark/spark-3.3.1/

Index of /dist/spark/spark-3.3.1

Name	Last modified	Size	Description
Parent Directory		-	
SparkR_3.3.1.tar.gz	2022-10-15 10:53	344K	
SparkR_3.3.1.tar.gz.asc	2022-10-15 10:53	862	
SparkR_3.3.1.tar.gz.sha512	2022-10-15 10:53	150	
pyspark-3.3.1.tar.gz	2022-10-15 10:53	268M	
pyspark-3.3.1.tar.gz.asc	2022-10-15 10:53	862	
pyspark-3.3.1.tar.gz.sha512	2022-10-15 10:53	151	
spark-3.3.1-bin-hadoop2.tgz	2022-10-15 10:53	261M	
spark-3.3.1-bin-hadoop2.tgz.asc	2022-10-15 10:53	862	
spark-3.3.1-bin-hadoop2.tgz.sha512	2022-10-15 10:53	158	
spark-3.3.1-bin-hadoop3-scala2.13.tgz	2022-10-15 10:53	292M	
spark-3.3.1-bin-hadoop3-scala2.13.tgz.asc	2022-10-15 10:53	862	
spark-3.3.1-bin-hadoop3-scala2.13.tgz.sha512	2022-10-15 10:53	168	
spark-3.3.1-bin-hadoop3.tgz	2022-10-15 10:53	285M	
spark-3.3.1-bin-hadoop3.tgz.asc	2022-10-15 10:53	862	
spark-3.3.1-bin-hadoop3.tgz.sha512	2022-10-15 10:53	158	
spark-3.3.1-bin-without-hadoop.tgz	2022-10-15 10:53	201M	
spark-3.3.1-bin-without-hadoop.tgz.asc	2022-10-15 10:53	862	

4. Create a folder called `spark` in your C drive
5. Cut and paste the downloaded file in `C:\spark` and extract it there
6. Tip: If you move(cut and paste) the downloaded tgz file it will save you some time, and extract it inside the spark folder in C drive
7. If you are using 7-Zip to unzip the downloaded file, you will have to unzip it two times, and then move the content to just parent folder and delete the empty folder.



This PC > Local Disk (C:) > spark

Name	Date modified	Type	Size
spark-3.3.1-bin-hadoop2	7/12/2024 2:46 PM	File folder	
spark-3.3.1-bin-hadoop2.tgz	7/12/2024 2:45 PM	TGZ File	267,676 KB

spark-3.3.1-bin-hadoop2

Local Disk (C:) > spark > spark-3.3.1-bin-hadoop2

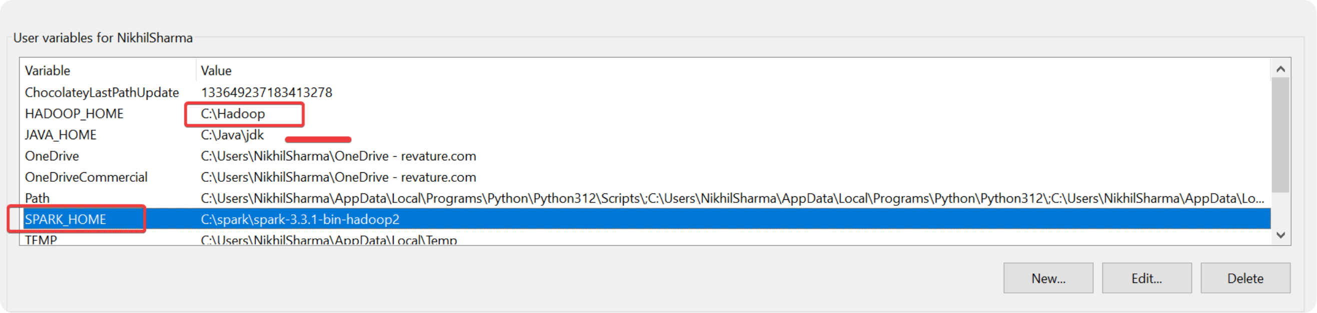
Name	Date modified	Type
bin	10/15/2022 3:11 PM	File folder
conf	10/15/2022 3:11 PM	File folder
data	10/15/2022 3:11 PM	File folder
examples	10/15/2022 3:11 PM	File folder
jars	10/15/2022 3:11 PM	File folder
kubernetes	10/15/2022 3:11 PM	File folder
licenses	10/15/2022 3:11 PM	File folder
python	10/15/2022 3:11 PM	File folder
R	10/15/2022 3:11 PM	File folder
sbin	10/15/2022 3:11 PM	File folder

15 items

5. Hadoop Home
   https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe
   Download this file and put it in `C:\Hadoop\bin`

-- Final Step is creating these Environment variables
USER Variables

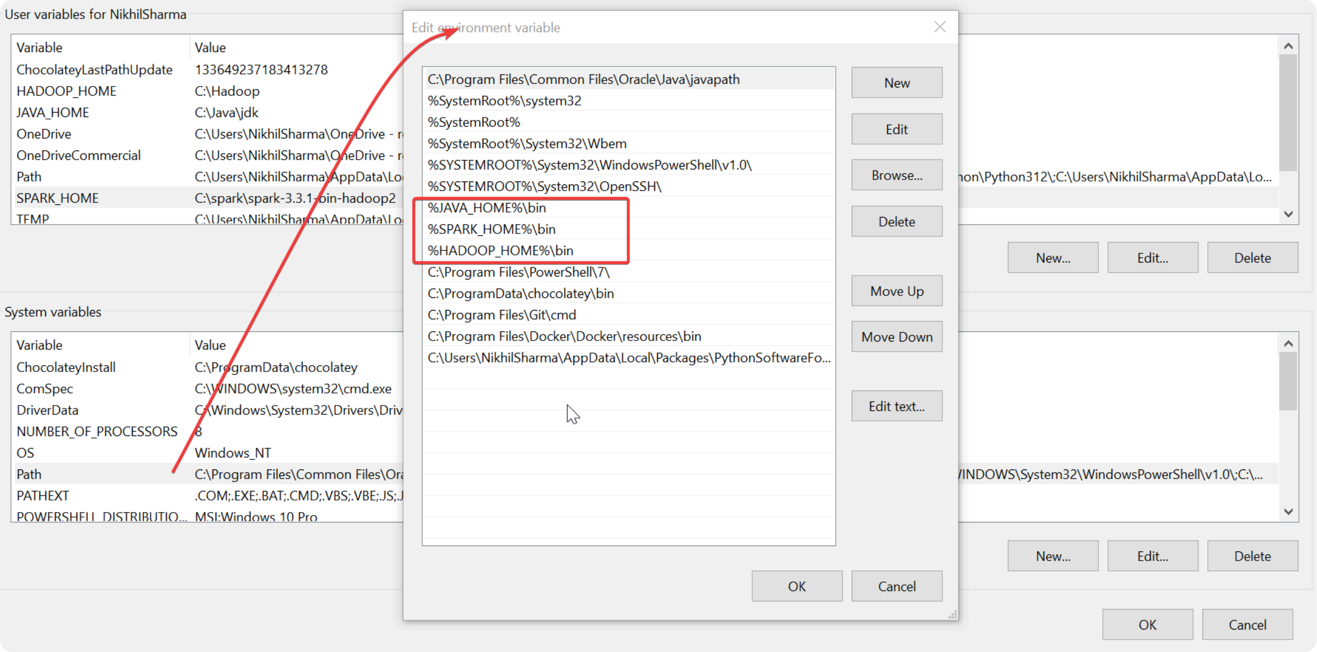
```
HADOOP_HOME- C:\hadoop
JAVA_HOME- C:\java\jdk
SPARK_HOME- C:\spark\spark-3.3.1-bin-hadoop2
```



User variables for NikhilSharma

Variable	Value
ChocolateyLastPathUpdate	133649237183413278
HADOOP_HOME	C:\Hadoop
JAVA_HOME	C:\Java\jdk
OneDrive	C:\Users\NikhilSharma\OneDrive - revature.com
OneDriveCommercial	C:\Users\NikhilSharma\OneDrive - revature.com
Path	C:\Users\NikhilSharma\AppData\Local\Programs\Python\Python312\Scripts\;C:\Users\NikhilSharma\AppData\Local\Programs\Python\Python312\;C:\Users\NikhilSharma\AppData\Lo...
SPARK_HOME	C:\spark\spark-3.3.1-bin-hadoop2
TEMP	C:\Users\NikhilSharma\AppData\Local\Temp

System PATH



User variables for NikhilSharma

Variable	Value
ChocolateyLastPathUpdate	133649237183413278
HADOOP_HOME	C:\Hadoop
JAVA_HOME	C:\Java\jdk
OneDrive	C:\Users\NikhilSharma\OneDrive - r
OneDriveCommercial	C:\Users\NikhilSharma\OneDrive - r
Path	C:\Users\NikhilSharma\AppData\Lo
SPARK_HOME	C:\spark\spark-3.3.1-bin-hadoop2
TEMP	C:\Users\NikhilSharma\AppData\Lo

System variables

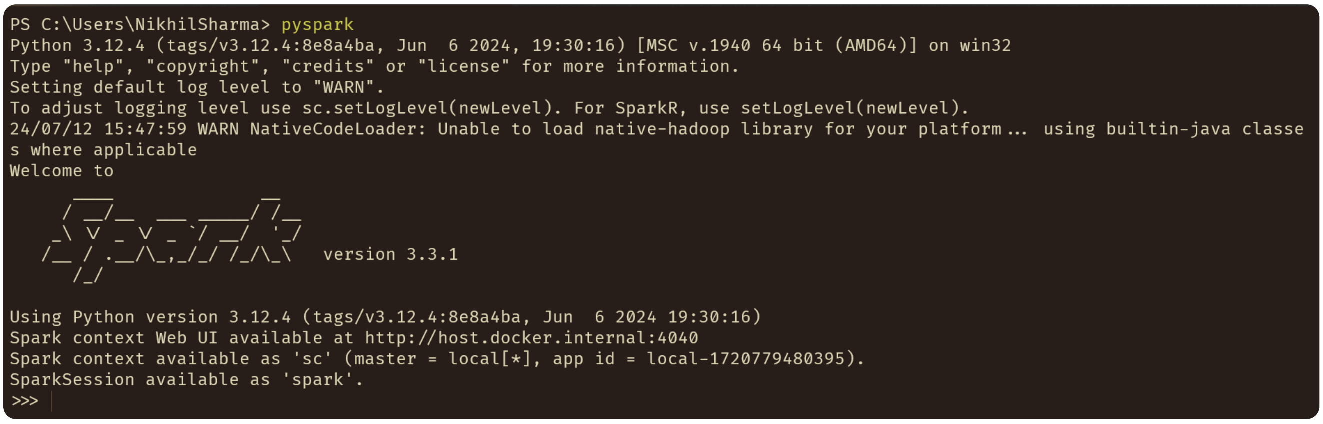
Variable	Value
ChocolateyInstall	C:\ProgramData\chocolatey
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\Driv
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Ora
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.J
POWERSHELL_DISTRIBUTIO...	MSI:Windows 10 Pro

Edit environment variable

```
C:\Program Files\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
%JAVA_HOME%\bin
%SPARK_HOME%\bin
%HADOOP_HOME%\bin
C:\Program Files\PowerShell\7\
C:\ProgramData\chocolatey\bin
C:\Program Files\Git\cmd
C:\Program Files\Docker\Docker\resources\bin
C:\Users\NikhilSharma\AppData\Local\Packages\PythonSoftwareFo...
```

Execute below commands

```
pip install py4j
pip install pyspark==3.3.1
```



```
PS C:\Users\NikhilSharma> pyspark
Python 3.12.4 (tags/v3.12.4:8e8a4ba, Jun  6 2024, 19:30:16) [MSC v.1940 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/07/12 15:47:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Python version 3.12.4 (tags/v3.12.4:8e8a4ba, Jun  6 2024 19:30:16)
Spark context Web UI available at http://host.docker.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1720779480395).
SparkSession available as 'spark'.
>>>
```

If it worked as expected you should be able to run PySpark in your machine!!

Now you ahead and install Jupyter Lab.

Run the below command in your terminal.

```
pip install jupyter notebook jupyterlab
```

`cd` to a folder of your choice and run `jupyter lab`

```
jupyter lab
```

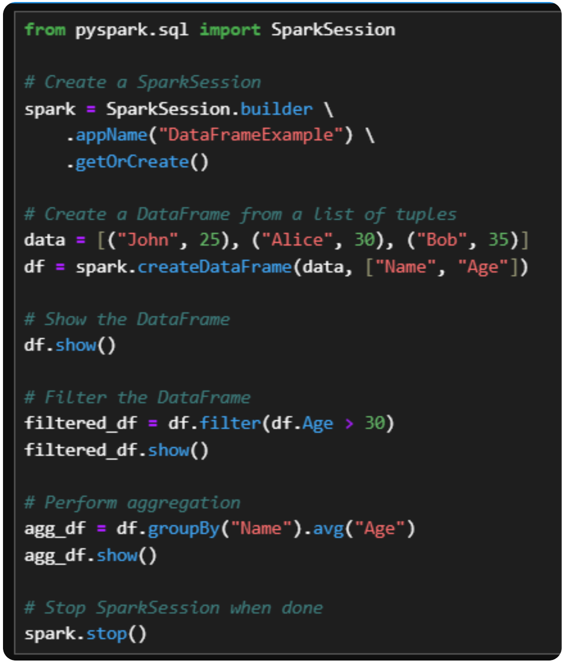
This will open Jupyter lab in a browser tab, create a new `ipynb` file and run the below code.

Spark DataFrame

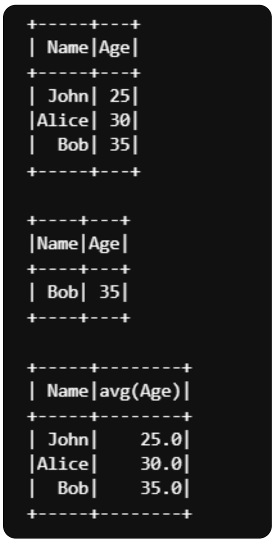
```
from pyspark.sql import SparkSession

# Create a SparkSession
spark = SparkSession.builder \
    .appName("DataFrameExample") \
    .getOrCreate()
```

```
# Create a DataFrame from a list of tuples
data = [("John", 25), ("Alice", 30), ("Bob", 35)]
df = spark.createDataFrame(data, ["Name", "Age"])

# Show the DataFrame
df.show()

# Filter the DataFrame
filtered_df = df.filter(df.Age > 30)
filtered_df.show()

# Perform aggregation
agg_df = df.groupBy("Name").avg("Age")
agg_df.show()

# Stop SparkSession when done
spark.stop()
```

Expected output:



```
from pyspark.sql import SparkSession

# Create a SparkSession
spark = SparkSession.builder \
    .appName("DataFrameExample") \
    .getOrCreate()

# Create a DataFrame from a list of tuples
data = [("John", 25), ("Alice", 30), ("Bob", 35)]
df = spark.createDataFrame(data, ["Name", "Age"])

# Show the DataFrame
df.show()

# Filter the DataFrame
filtered_df = df.filter(df.Age > 30)
filtered_df.show()

# Perform aggregation
agg_df = df.groupBy("Name").avg("Age")
agg_df.show()

# Stop SparkSession when done
spark.stop()
```



```
+-----+---+
| Name|Age|
+-----+---+
| John| 25|
|Alice| 30|
|  Bob| 35|
+-----+---+

+----+---+
|Name|Age|
+----+---+
| Bob| 35|
+----+---+

+-----+--------+
| Name|avg(Age)|
+-----+--------+
| John|    25.0|
|Alice|    30.0|
|  Bob|    35.0|
+-----+--------+
```

Give yourself a treat if you were able to run the code and get the expected output.

Thankyou for your time and patience!!