1. Install OpenJDK

```
$ sudo apt-get install openjdk-8-jdk
```

2. Identify the location of java

```
$ which java
```

3. Download Apache Spark (tgz file)

4. Extract the tgz file, rename (optinal) and save it in Home directory (optional)

5. Set PATH and JAVA_HOME environment variables in ".profile"

```
$ cd ~
$ gedit .profile
```

Add the following lines at the end of the file (assuming folder is renamed as “spark” in Home dir):

```
PATH=$PATH:~/spark/bin
JAVA_HOME=/usr/bin/java
```

Logout and Login for changes to take effect

6. Write a Spark program

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[1]").appName("test").getOrCreate()
spark.sparkContext.setLogLevel("ERROR")
dataList = [("Java", 20000), ("Python", 100000), ("Scala", 3000)]
rdd=spark.sparkContext.parallelize(dataList)
print(rdd.count())
```

7. Save the file with .py extension

8. Execute the file with spark-submit

```
$ spark-submit spark1.py
```

9. Download Anaconda for linux

10. Install Anaconda

```
$ bash Anaconda3-2022.05-Linux-x86_64.sh
```

11. Start Anaconda

```
$ anaconda-navigator
```

12. Link pyspark with jupyter

```
$ cd ~
$ gedit .bashrc
```

Add the following lines at the end

```
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
```

13. Run pyspark in terminal to open jupyter notebook.

```
$ pyspark
```

14. Run code in jupyter notebook.

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[1]").appName("test").getOrCreate()
dataList = [("Java", 20000), ("Python", 100000), ("Scala", 3000)]
rdd=spark.sparkContext.parallelize(dataList)
print(rdd.count())
```