

Roll NO. - 31404

K4

Problem Statement - Write a simple program in SCALA using Apache Spark framework

```
te@sel-a1-216-18:~$ sudo apt update
sudo apt install openjdk-11-jdk
Hit:1 https://packages.microsoft.com/repos/code stable InRelease
Hit:2 http://oem.archive.canonical.com jammy InRelease
Get:3 https://dl.google.com/linux/chrome/deb stable InRelease [1,025 B]
Hit:4 http://dell.archive.canonical.com jammy InRelease
Hit:5 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Hit:6 http://security.ubuntu.com/ubuntu jammy-security InRelease
Err:3 https://dl.google.com/linux/chrome/deb stable InRelease
  The following signatures couldn't be verified because the public key is not available: NO_PUBKEY 32EE5355A6BC6E42
Hit:7 http://in.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:9 http://in.archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:8 https://hub-dist.unity3d.com/artifactory/hub-debian-prod-local stable InRelease
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
165 packages can be upgraded. Run 'apt list --upgradable' to see them.
W: An error occurred during the signature verification. The repository is not updated and the previous index files will be used. GPG error: https://dl.google.com/linux/chrome/deb stable InRelease: The following signatures couldn't be verified because the public key is not available: NO_PUBKEY 32EE5355A6BC6E42
W: Failed to fetch https://dl.google.com/linux/chrome/deb/dists/stable/InRelease The following signatures couldn't be verified because the public key is not available: NO_PUBKEY 32EE5355A6BC6E42
W: Some index files failed to download. They have been ignored, or old ones used instead.
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openjdk-11-jdk is already the newest version (11.0.26+4-1ubuntu1-22.04).
The following packages were automatically installed and are no longer required:
  cifs-utils dctrl-tools dmraid genders gir1.2-timzone-1.0 gir1.2-xml-1.0
  keyutils kpartx kpartx-boot libdebian-installer4 libdmraid1.0.0.rc16
  libgenders0 libtimzone-1.0-data libtimzone-1.0 python3-icu python3-pam rdate
  user-setup
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 165 not upgraded.
```

```
te@sel-a1-216-18:~$ wget https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
--2025-04-07 09:20:39-- https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.208.237, 135.181.214.104, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.208.237|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400724056 (382M) [application/x-gzip]
Saving to: 'spark-3.5.5-bin-hadoop3.tgz.1'
```

```
spark-3.5.5-bin-hado 100%[=====] 382.16M  2.84MB/s   in 2m 19s

2025-04-07 09:22:58 (2.76 MB/s) - 'spark-3.5.5-bin-hadoop3.tgz.1' saved [400724056/400724056]
```

```
te@sel-a1-216-18:~$ tar -xvzf spark-3.5.5-bin-hadoop3.tgz
spark-3.5.5-bin-hadoop3/
spark-3.5.5-bin-hadoop3/jars/
spark-3.5.5-bin-hadoop3/jars/HikariCP-2.5.1.jar
spark-3.5.5-bin-hadoop3/jars/JLargeArrays-1.5.jar
spark-3.5.5-bin-hadoop3/jars/JTransforms-3.1.jar
spark-3.5.5-bin-hadoop3/jars/RoaringBitmap-0.9.45.jar
spark-3.5.5-bin-hadoop3/jars/ST4-4.0.4.jar
spark-3.5.5-bin-hadoop3/jars/activation-1.1.1.jar
spark-3.5.5-bin-hadoop3/jars/aircompressor-0.27.jar
spark-3.5.5-bin-hadoop3/jars/algebra_2.12-2.0.1.jar
spark-3.5.5-bin-hadoop3/jars/annotations-17.0.0.jar
spark-3.5.5-bin-hadoop3/jars/antlr-runtime-3.5.2.jar
spark-3.5.5-bin-hadoop3/jars/antlr4-runtime-4.9.3.jar
spark-3.5.5-bin-hadoop3/jars/aopalliance-repackaged-2.6.1.jar
spark-3.5.5-bin-hadoop3/jars/arpack-3.0.3.jar
spark-3.5.5-bin-hadoop3/jars/arpack_combined_all-0.1.jar
spark-3.5.5-bin-hadoop3/jars/arrow-format-12.0.1.jar
spark-3.5.5-bin-hadoop3/jars/arrow-memory-core-12.0.1.jar
spark-3.5.5-bin-hadoop3/jars/arrow-memory-netty-12.0.1.jar
spark-3.5.5-bin-hadoop3/jars/arrow-vector-12.0.1.jar
spark-3.5.5-bin-hadoop3/jars/audience-annotations-0.5.0.jar
```

```
te@sel-a1-216-18:~$ mv spark-3.5.5-bin-hadoop3 spark
export SPARK_HOME=~/.spark
export PATH=$SPARK_HOME/bin:$PATH
```

```
te@sel-a1-216-18:~$ readlink -f $(which java)
/usr/lib/jvm/java-11-openjdk-amd64/bin/java
te@sel-a1-216-18:~$ readlink -f $(which java)
/usr/lib/jvm/java-11-openjdk-amd64/bin/java
te@sel-a1-216-18:~$ export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export PATH=$JAVA_HOME/bin:$PATH
te@sel-a1-216-18:~$ echo 'export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64' >> ~/.bashrc
echo 'export PATH=$JAVA_HOME/bin:$PATH' >> ~/.bashrc
source ~/.bashrc
te@sel-a1-216-18:~$ spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/07 09:28:07 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://sel-a1-216-18:4040
Spark context available as 'sc' (master = local[*], app id = local-1743998288054).
Spark session available as 'spark'.
Welcome to

  ____  __
 / ___/  / /
/ /   /  / /
/ /___/  / /
\____/___/_/

 version 3.5.5

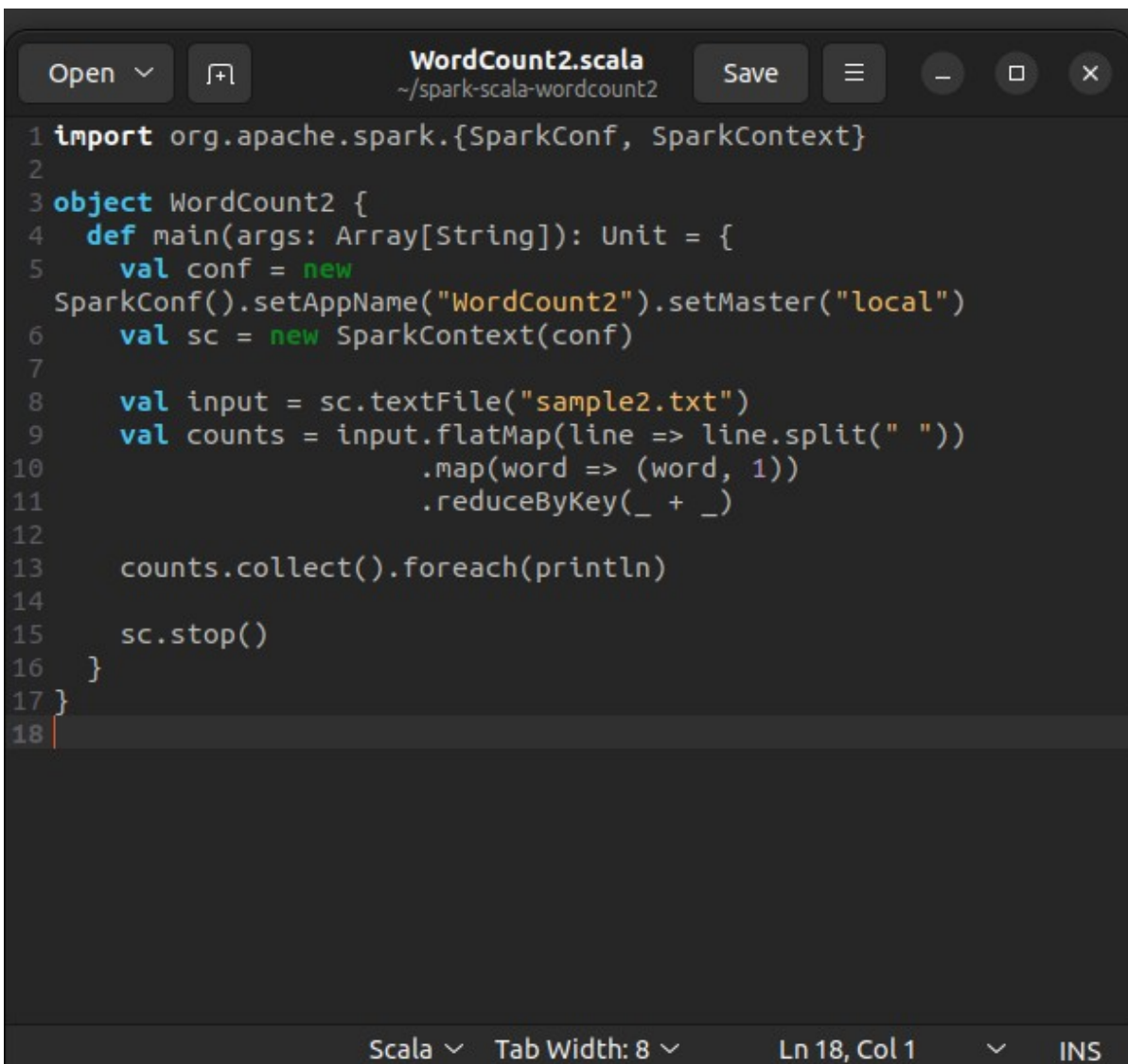
Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.26)
Type in expressions to have them evaluated.
Type :help for more information.
```

```
te@sel-a1-216-18:~/spark-scala-wordcount$ mkdir ~/spark-scala-wordcount2
cd ~/spark-scala-wordcount2
nano WordCount2.scala
te@sel-a1-216-18:~/spark-scala-wordcount2$ echo "hello world hello spark word count example" > sample2.txt
```

```
te@sel-a1-216-18:~/spark-scala-wordcount2$ scalac -classpath "$SPARK_HOME/jars/*" WordCount2.scala
te@sel-a1-216-18:~/spark-scala-wordcount2$ jar cf WordCount2.jar WordCount2*.class
te@sel-a1-216-18:~/spark-scala-wordcount2$ spark-submit \
--class WordCount2 \
--master local \
WordCount2.jar
```

```
25/04/07 10:00:29 INFO DAGScheduler: Job 0 finished: collect at WordCount2.scala:13, took 0.307149 s
(example,1)
(spark,1)
(word,1)
(hello,2)
(count,1)
(world,1)
25/04/07 10:00:29 INFO SparkContext: SparkContext is stopping with exitCode 0.
```

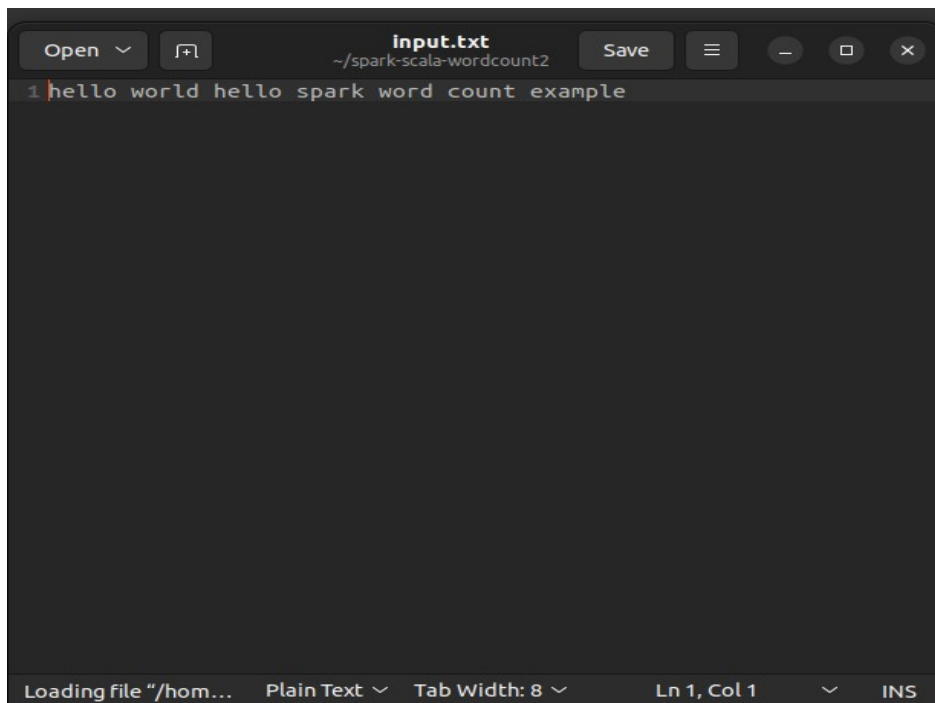
WordCount.java -



```
1 import org.apache.spark.{SparkConf, SparkContext}
2
3 object WordCount2 {
4   def main(args: Array[String]): Unit = {
5     val conf = new
      SparkConf().setAppName("WordCount2").setMaster("local")
6     val sc = new SparkContext(conf)
7
8     val input = sc.textFile("sample2.txt")
9     val counts = input.flatMap(line => line.split(" "))
10                        .map(word => (word, 1))
11                        .reduceByKey(_ + _)
12
13     counts.collect().foreach(println)
14
15     sc.stop()
16   }
17 }
18
```

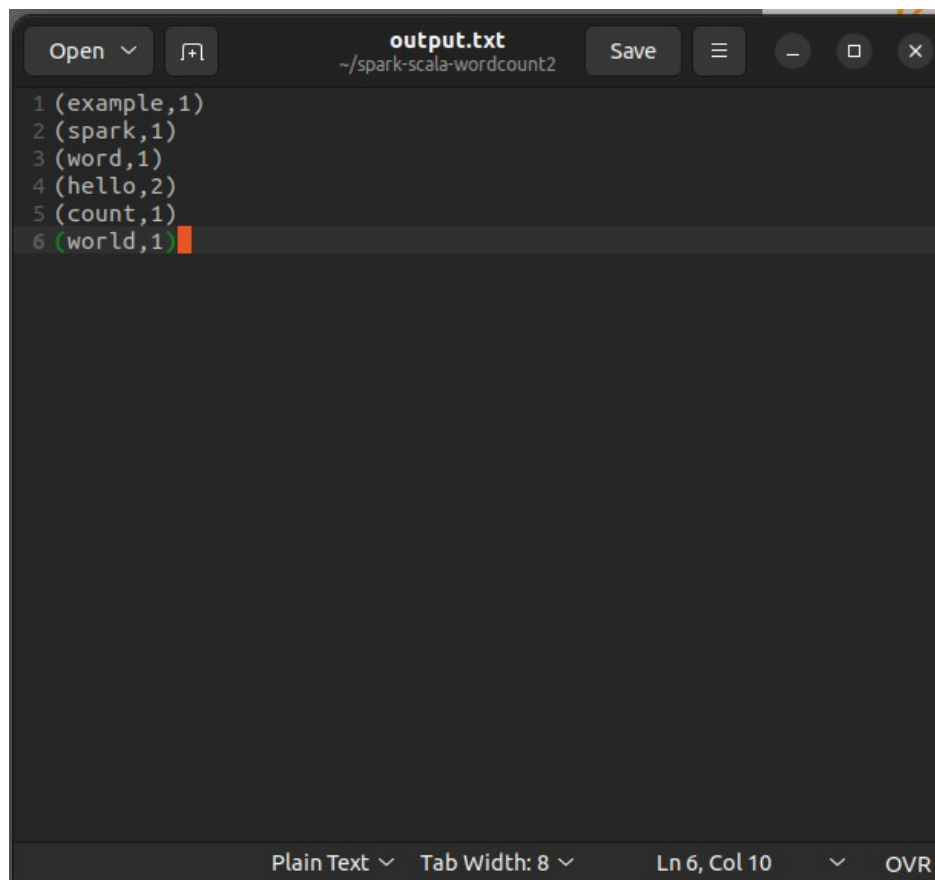
Scala ▾ Tab Width: 8 ▾ Ln 18, Col 1 ▾ INS

input.txt -



A screenshot of a text editor window titled "input.txt" with the path "~/spark-scala-wordcount2". The window contains a single line of text: "1 hello world hello spark word count example". The status bar at the bottom indicates "Loading file "/>input.txt -

output.txt -



A screenshot of a text editor window titled "output.txt" with the path "~/spark-scala-wordcount2". The window contains a list of word counts, each on a new line: "1 (example,1)", "2 (spark,1)", "3 (word,1)", "4 (hello,2)", "5 (count,1)", and "6 (world,1)". The status bar at the bottom indicates "Plain Text", "Tab Width: 8", "Ln 6, Col 10", and "OVR".

| Line | Word | Count |
|------|---------|-------|
| 1 | example | 1 |
| 2 | spark | 1 |
| 3 | word | 1 |
| 4 | hello | 2 |
| 5 | count | 1 |
| 6 | world | 1 |