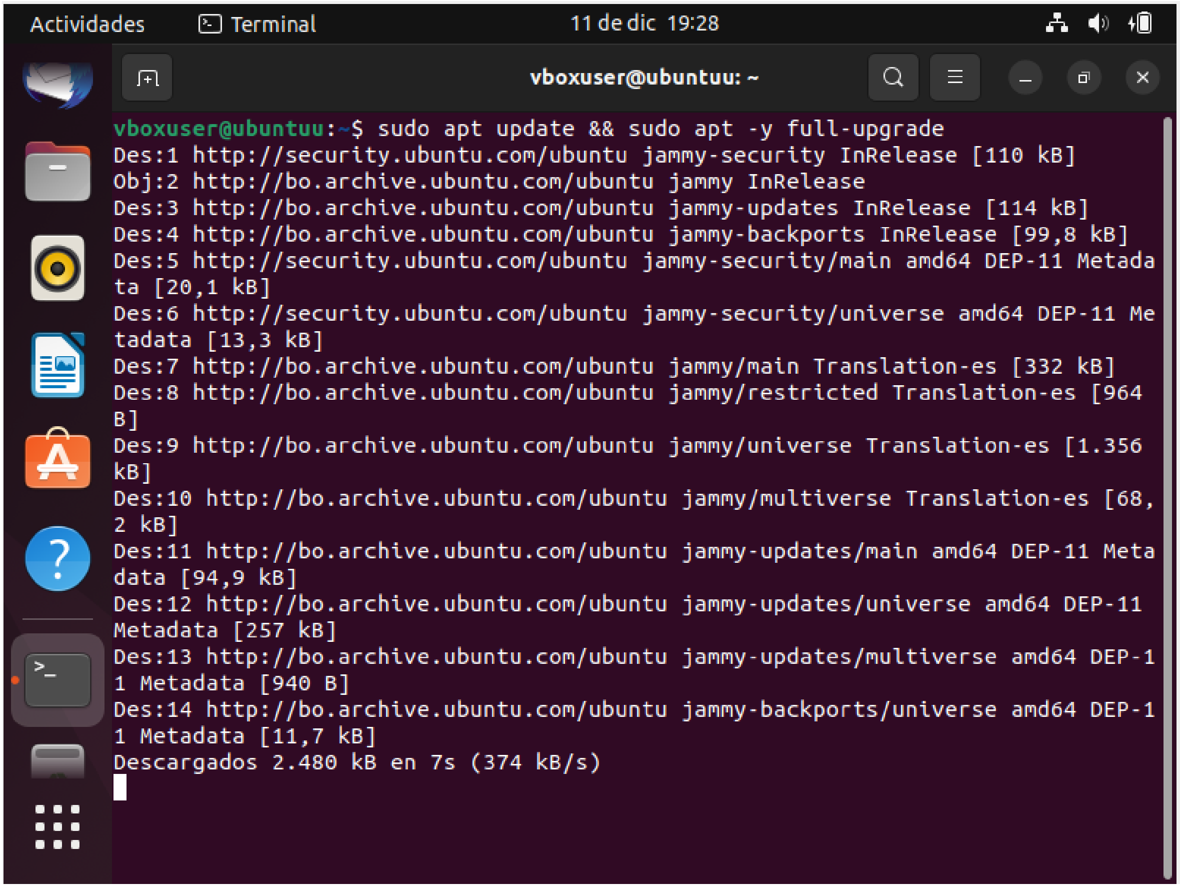
Examen Final

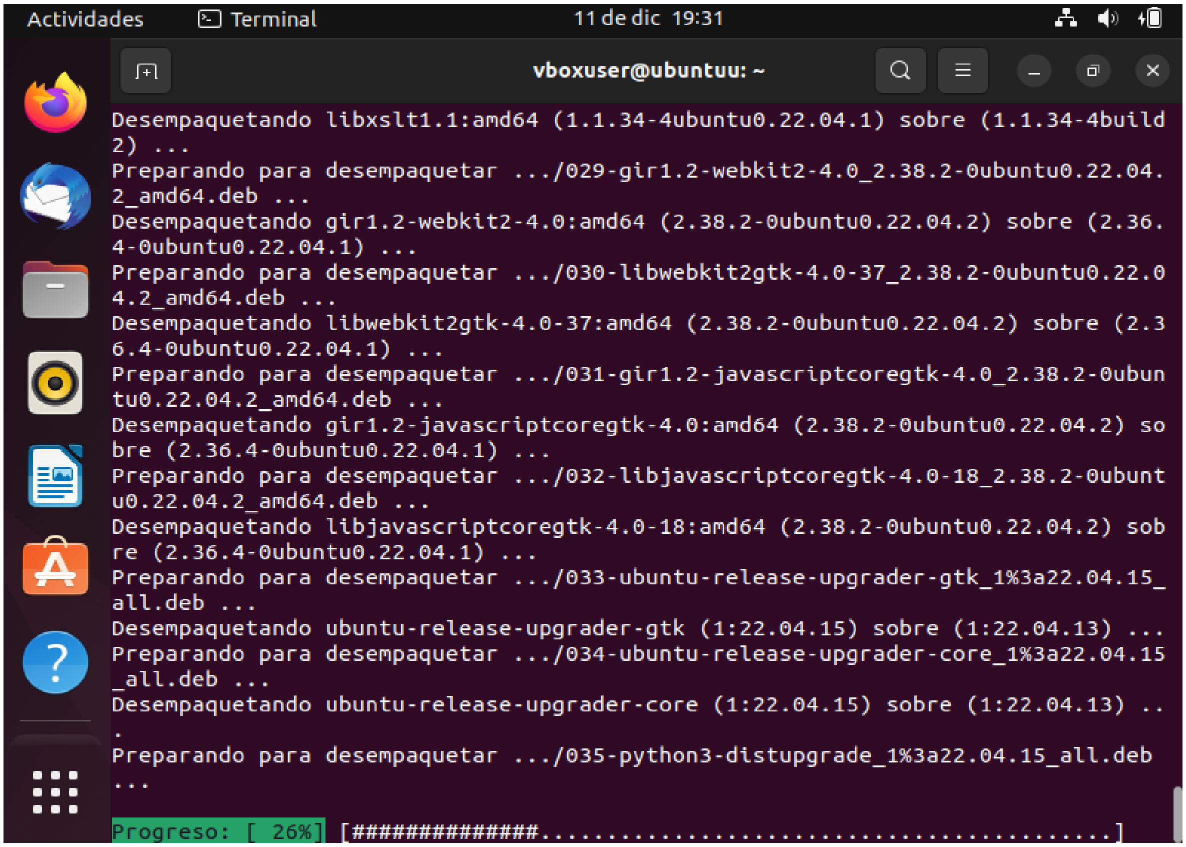
Nombre: Miguel Angel Quispe Mamani

C.I: 9971512 LP

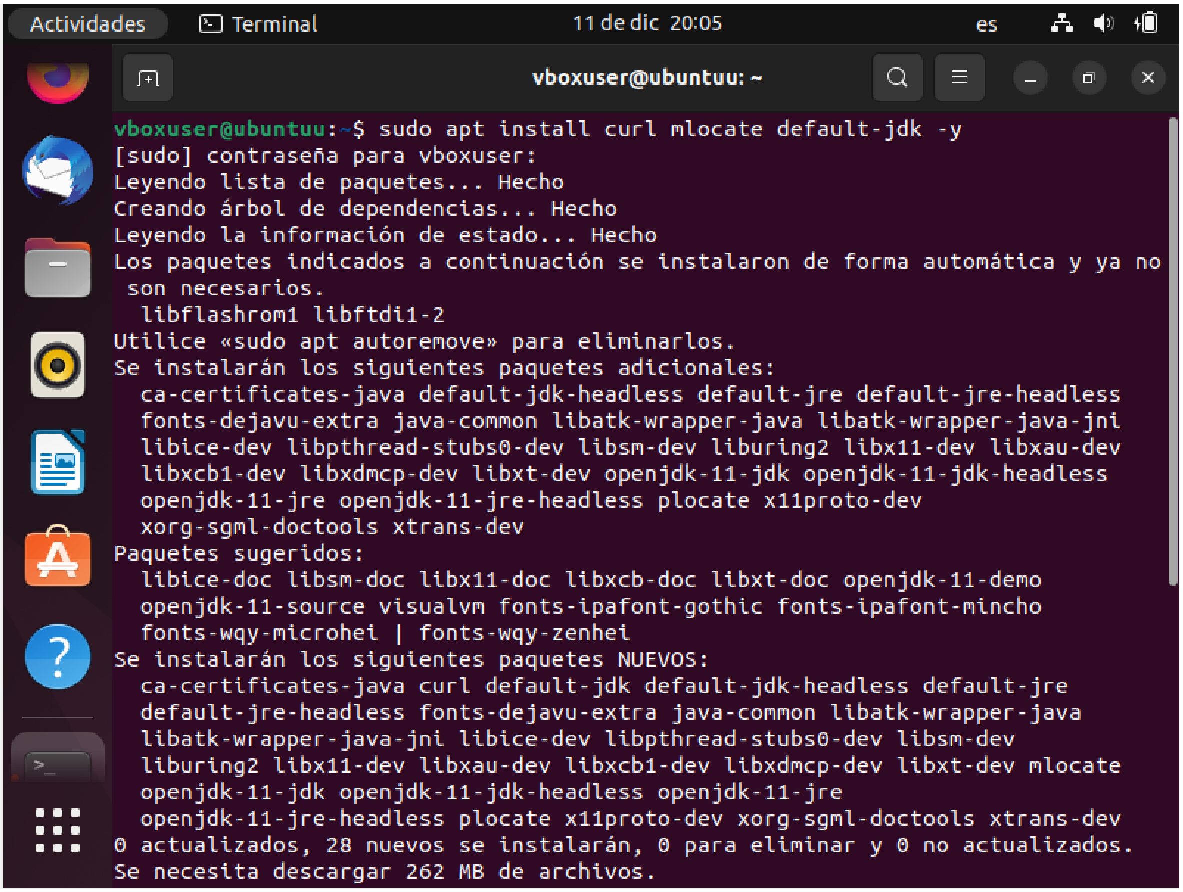
1. En una máquina virtual realice la configuración de apache spark, puede guiarse en cualquier tutorial o el proporcionado por el docente.
   url: https://computingforgeeks.com/how-to-install-apache-spark-on-ubuntu-debian/
   Con el shell podra ejecutar scala por defecto
   Instale Python para spark



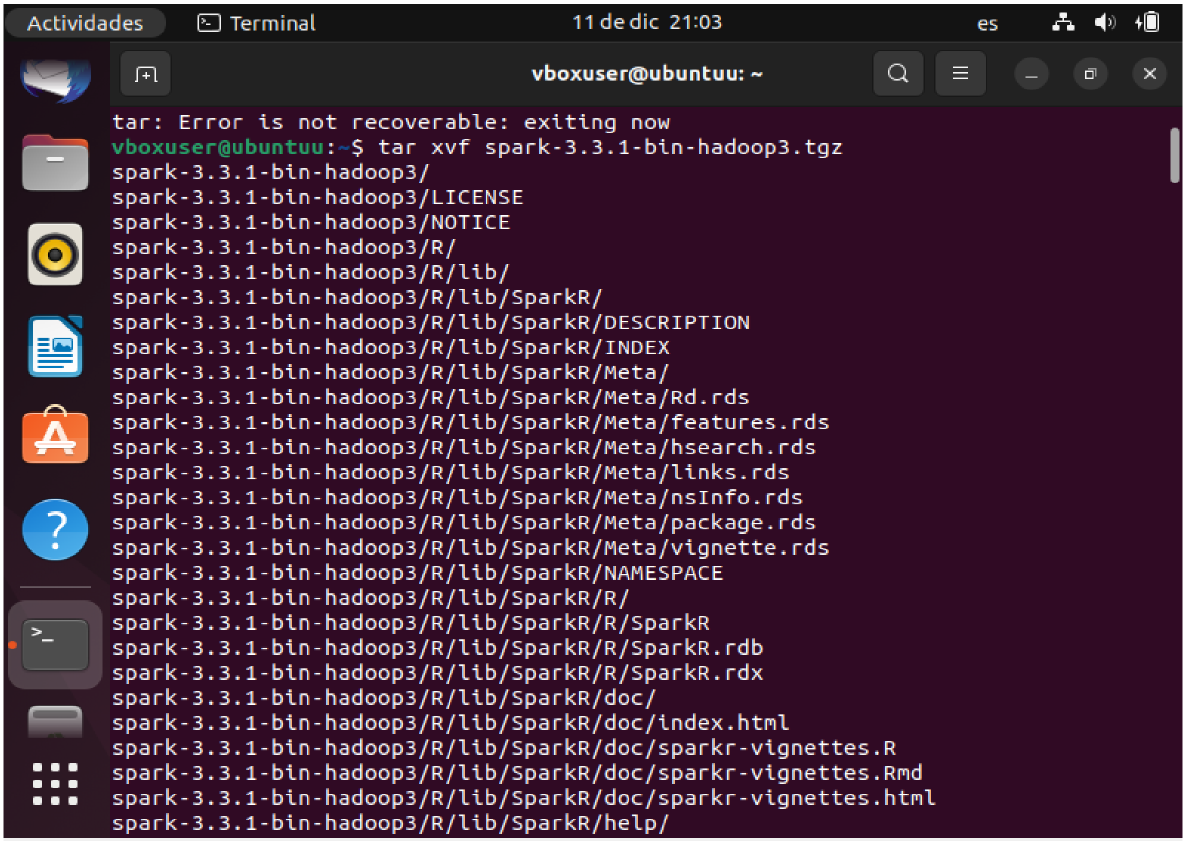
```
vboxuser@ubuntuu:~$ sudo apt update && sudo apt -y full-upgrade
Des:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Obj:2 http://bo.archive.ubuntu.com/ubuntu jammy InRelease
Des:3 http://bo.archive.ubuntu.com/ubuntu jammy-updates InRelease [114 kB]
Des:4 http://bo.archive.ubuntu.com/ubuntu jammy-backports InRelease [99,8 kB]
Des:5 http://security.ubuntu.com/ubuntu jammy-security/main amd64 DEP-11 Metada
ta [20,1 kB]
Des:6 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 DEP-11 Me
tadata [13,3 kB]
Des:7 http://bo.archive.ubuntu.com/ubuntu jammy/main Translation-es [332 kB]
Des:8 http://bo.archive.ubuntu.com/ubuntu jammy/restricted Translation-es [964
B]
Des:9 http://bo.archive.ubuntu.com/ubuntu jammy/universe Translation-es [1.356
kB]
Des:10 http://bo.archive.ubuntu.com/ubuntu jammy/multiverse Translation-es [68,
2 kB]
Des:11 http://bo.archive.ubuntu.com/ubuntu jammy-updates/main amd64 DEP-11 Meta
data [94,9 kB]
Des:12 http://bo.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 DEP-11
Metadata [257 kB]
Des:13 http://bo.archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 DEP-1
1 Metadata [940 B]
Des:14 http://bo.archive.ubuntu.com/ubuntu jammy-backports/universe amd64 DEP-1
1 Metadata [11,7 kB]
Descargados 2.480 kB en 7s (374 kB/s)
```



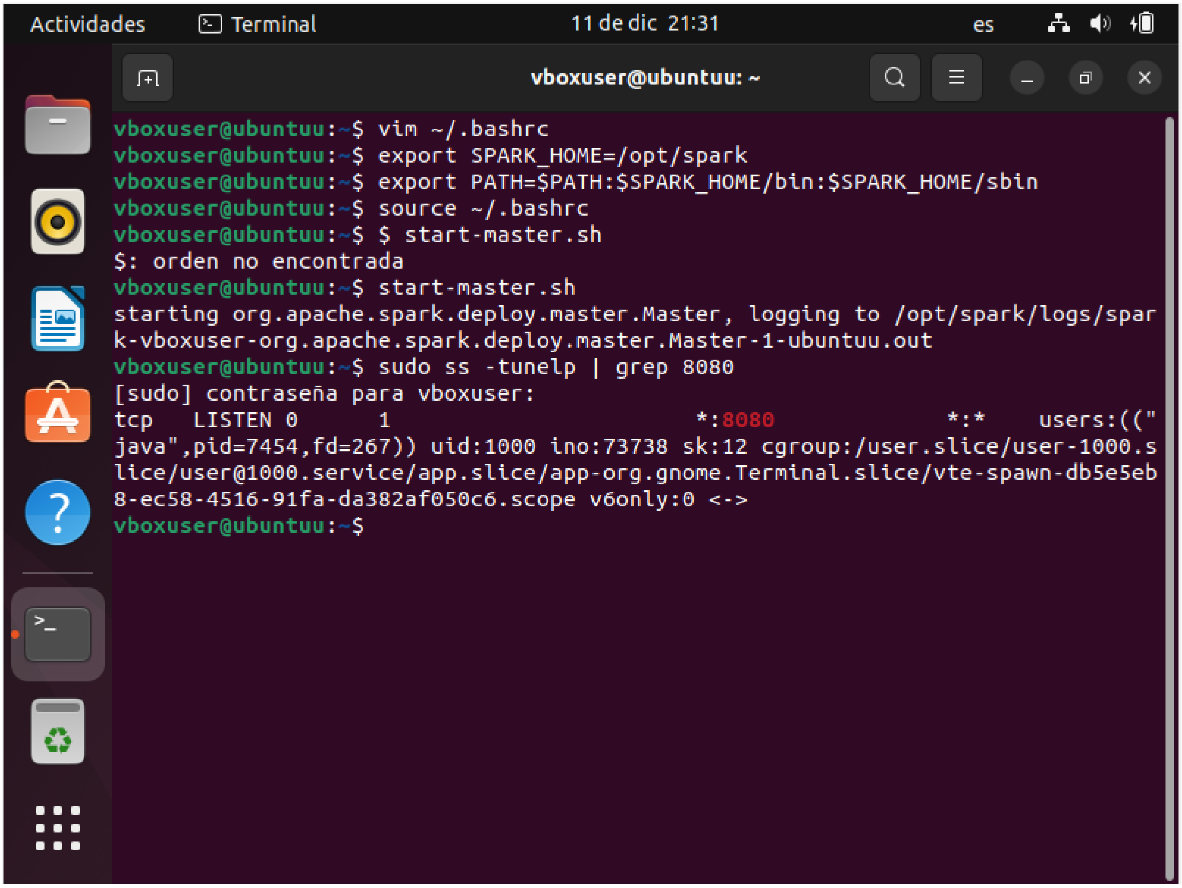
```
Desempaquetando libxslt1.1:amd64 (1.1.34-4ubuntu0.22.04.1) sobre (1.1.34-4build
2) ...
Preparando para desempaquetar .../029-gir1.2-webkit2-4.0_2.38.2-0ubuntu0.22.04.
2_amd64.deb ...
Desempaquetando gir1.2-webkit2-4.0:amd64 (2.38.2-0ubuntu0.22.04.2) sobre (2.36.
4-0ubuntu0.22.04.1) ...
Preparando para desempaquetar .../030-libwebkit2gtk-4.0-37_2.38.2-0ubuntu0.22.0
4.2_amd64.deb ...
Desempaquetando libwebkit2gtk-4.0-37:amd64 (2.38.2-0ubuntu0.22.04.2) sobre (2.3
6.4-0ubuntu0.22.04.1) ...
Preparando para desempaquetar .../031-gir1.2-javascriptcoregtk-4.0_2.38.2-0ubun
tu0.22.04.2_amd64.deb ...
Desempaquetando gir1.2-javascriptcoregtk-4.0:amd64 (2.38.2-0ubuntu0.22.04.2) so
bre (2.36.4-0ubuntu0.22.04.1) ...
Preparando para desempaquetar .../032-libjavascriptcoregtk-4.0-18_2.38.2-0ubunt
u0.22.04.2_amd64.deb ...
Desempaquetando libjavascriptcoregtk-4.0-18:amd64 (2.38.2-0ubuntu0.22.04.2) sob
re (2.36.4-0ubuntu0.22.04.1) ...
Preparando para desempaquetar .../033-ubuntu-release-upgrader-gtk_1%3a22.04.15_
all.deb ...
Desempaquetando ubuntu-release-upgrader-gtk (1:22.04.15) sobre (1:22.04.13) ...
Preparando para desempaquetar .../034-ubuntu-release-upgrader-core_1%3a22.04.15
_all.deb ...
Desempaquetando ubuntu-release-upgrader-core (1:22.04.15) sobre (1:22.04.13) ..
.
Preparando para desempaquetar .../035-python3-distupgrade_1%3a22.04.15_all.deb
...

Progreso: [ 26%] [##############.........................................]
```



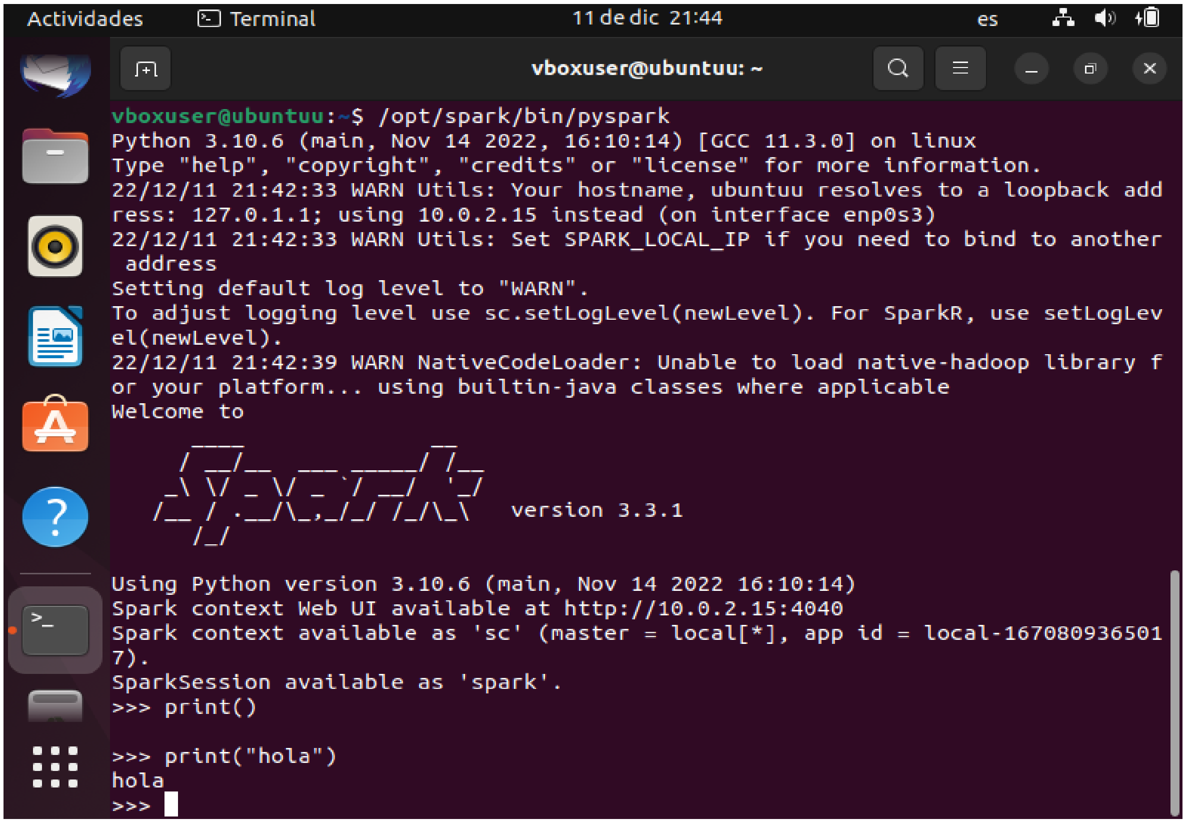
```
vboxuser@ubuntuu:~$ sudo apt install curl mlocate default-jdk -y
[sudo] contraseña para vboxuser:
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
Los paquetes indicados a continuación se instalaron de forma automática y ya no
 son necesarios.
  libflashrom1 libftdi1-2
Utilice «sudo apt autoremove» para eliminarlos.
Se instalarán los siguientes paquetes adicionales:
  ca-certificates-java default-jdk-headless default-jre default-jre-headless
  fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni
  libice-dev libpthread-stubs0-dev libsm-dev liburing2 libx11-dev libxau-dev
  libxcb1-dev libxdmcp-dev libxt-dev openjdk-11-jdk openjdk-11-jdk-headless
  openjdk-11-jre openjdk-11-jre-headless plocate x11proto-dev
  xorg-sgml-doctools xtrans-dev
Paquetes sugeridos:
  libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo
  openjdk-11-source visualvm fonts-ipafont-gothic fonts-ipafont-mincho
  fonts-wqy-microhei | fonts-wqy-zenhei
Se instalarán los siguientes paquetes NUEVOS:
  ca-certificates-java curl default-jdk default-jdk-headless default-jre
  default-jre-headless fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev
  liburing2 libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev mlocate
  openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre
  openjdk-11-jre-headless plocate x11proto-dev xorg-sgml-doctools xtrans-dev
0 actualizados, 28 nuevos se instalarán, 0 para eliminar y 0 no actualizados.
Se necesita descargar 262 MB de archivos.
```



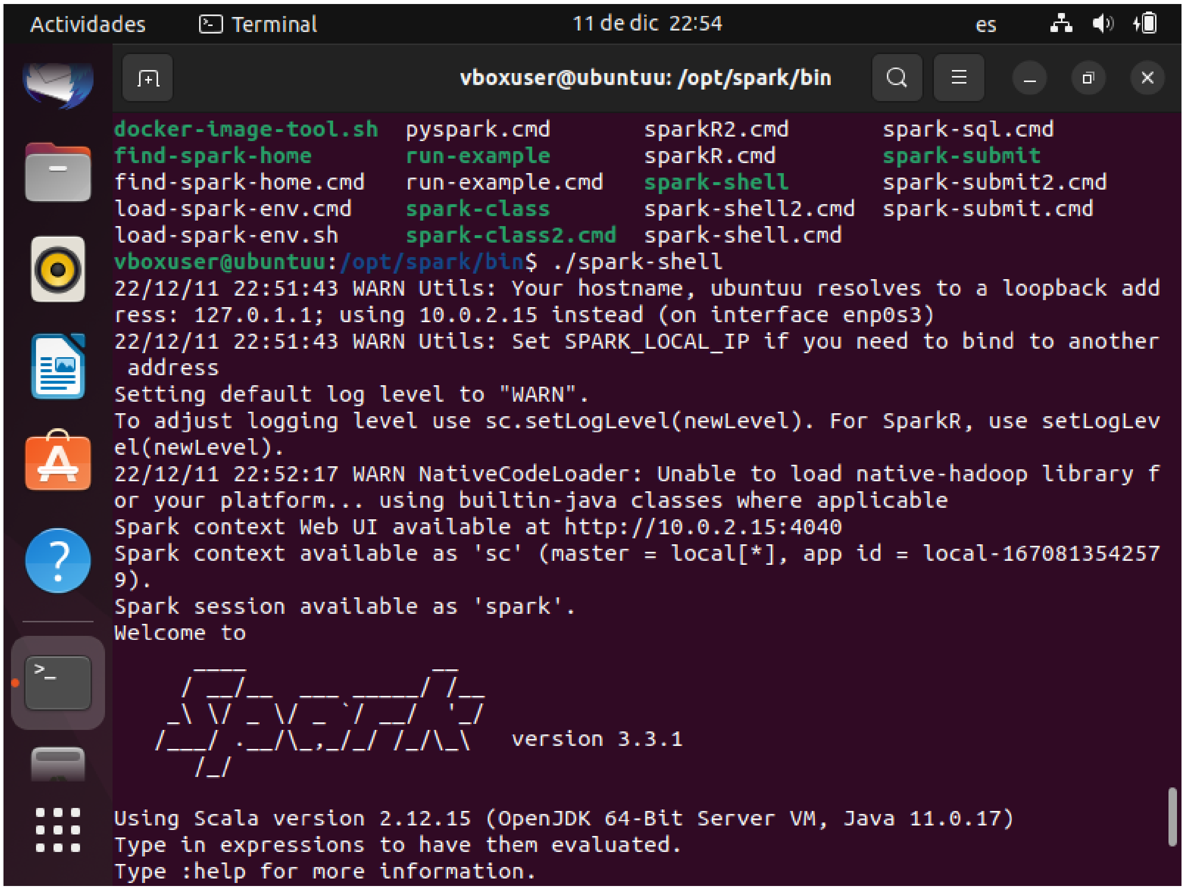
```
tar: Error is not recoverable: exiting now
vboxuser@ubuntuu:~$ tar xvf spark-3.3.1-bin-hadoop3.tgz
spark-3.3.1-bin-hadoop3/
spark-3.3.1-bin-hadoop3/LICENSE
spark-3.3.1-bin-hadoop3/NOTICE
spark-3.3.1-bin-hadoop3/R/
spark-3.3.1-bin-hadoop3/R/lib/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/DESCRIPTION
spark-3.3.1-bin-hadoop3/R/lib/SparkR/INDEX
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/Rd.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/features.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/hsearch.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/links.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/nsInfo.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/package.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/vignette.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/NAMESPACE
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/SparkR
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdb
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdx
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/index.html
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.R
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.Rmd
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.html
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/
```



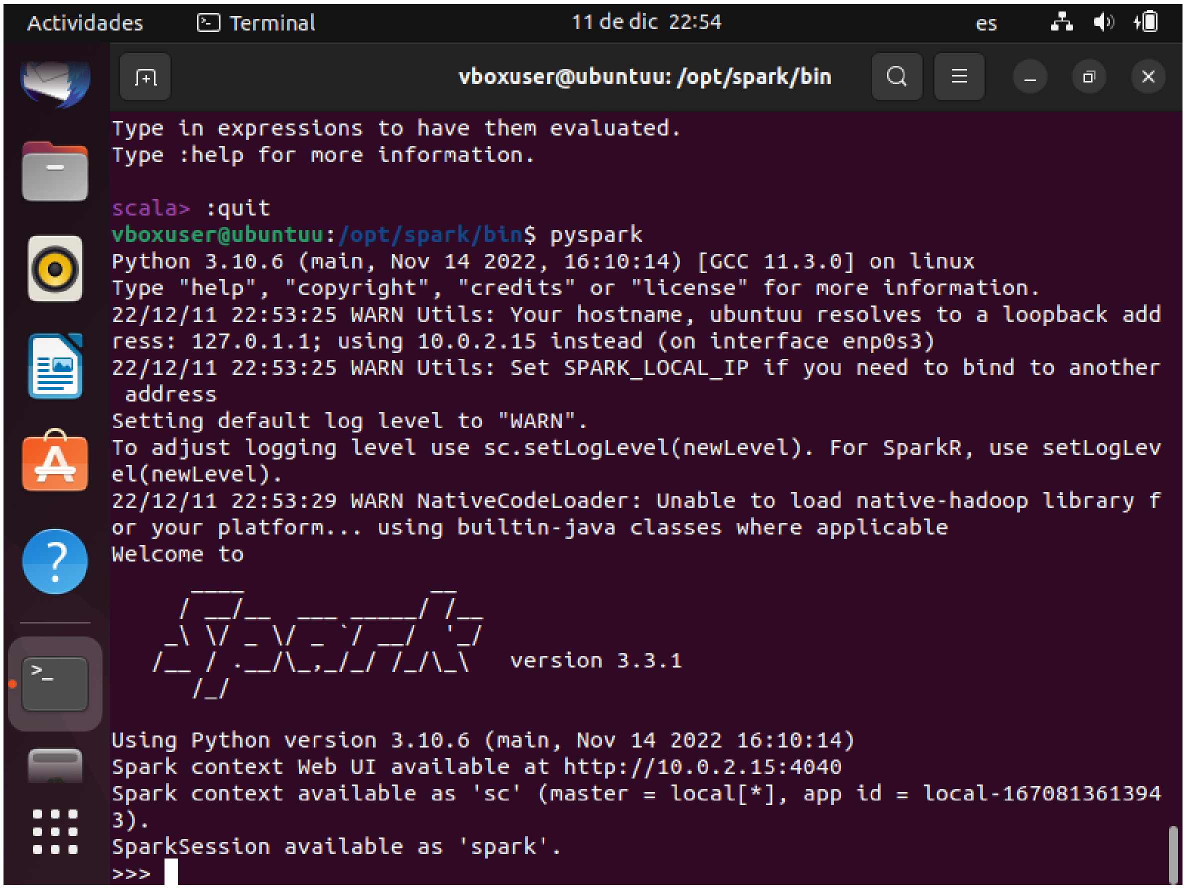
```
vboxuser@ubuntuu:~$ vim ~/.bashrc
vboxuser@ubuntuu:~$ export SPARK_HOME=/opt/spark
vboxuser@ubuntuu:~$ export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
vboxuser@ubuntuu:~$ source ~/.bashrc
vboxuser@ubuntuu:~$ $ start-master.sh
$: orden no encontrada
vboxuser@ubuntuu:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spar
k-vboxuser-org.apache.spark.deploy.master.Master-1-ubuntuu.out
vboxuser@ubuntuu:~$ sudo ss -tunelp | grep 8080
[sudo] contraseña para vboxuser:
tcp   LISTEN 0      1                   *:8080            *:*    users:(("
java",pid=7454,fd=267)) uid:1000 ino:73738 sk:12 cgroup:/user.slice/user-1000.s
lice/user@1000.service/app.slice/app-org.gnome.Terminal.slice/vte-spawn-db5e5eb
8-ec58-4516-91fa-da382af050c6.scope v6only:0 <->
vboxuser@ubuntuu:~$
```



```
vboxuser@ubuntuu:~$ /opt/spark/bin/pyspark
Python 3.10.6 (main, Nov 14 2022, 16:10:14) [GCC 11.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
22/12/11 21:42:33 WARN Utils: Your hostname, ubuntuu resolves to a loopback add
ress: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
22/12/11 21:42:33 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
 address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
22/12/11 21:42:39 WARN NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Python version 3.10.6 (main, Nov 14 2022 16:10:14)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-167080936501
7).
SparkSession available as 'spark'.
>>> print()

>>> print("hola")
hola
>>> 
```



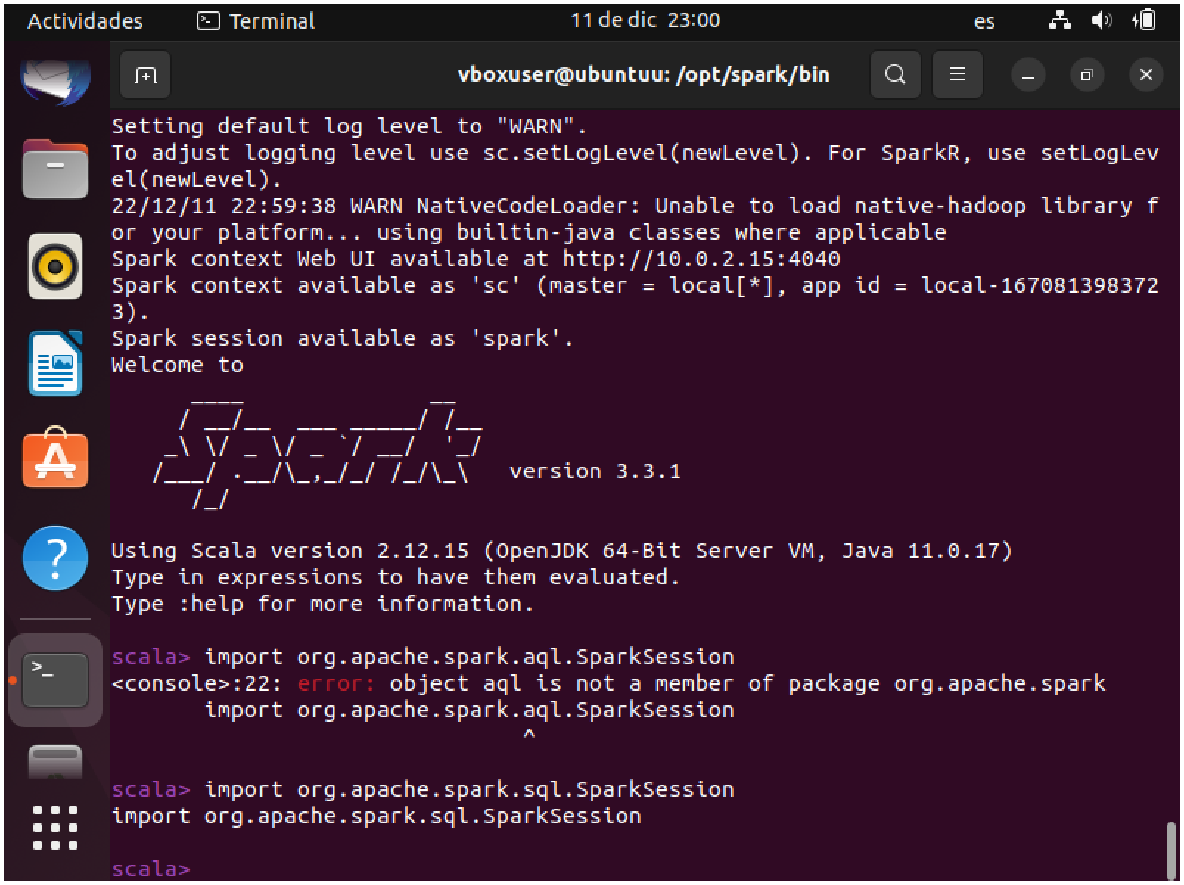
```
docker-image-tool.sh   pyspark.cmd       sparkR2.cmd        spark-sql.cmd
find-spark-home        run-example       sparkR.cmd         spark-submit
find-spark-home.cmd    run-example.cmd   spark-shell        spark-submit2.cmd
load-spark-env.cmd     spark-class       spark-shell2.cmd   spark-submit.cmd
load-spark-env.sh      spark-class2.cmd  spark-shell.cmd
vboxuser@ubuntuu:/opt/spark/bin$ ./spark-shell
22/12/11 22:51:43 WARN Utils: Your hostname, ubuntuu resolves to a loopback add
ress: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
22/12/11 22:51:43 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
 address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
22/12/11 22:52:17 WARN NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-167081354257
9).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 11.0.17)
Type in expressions to have them evaluated.
Type :help for more information.
```



```
Type in expressions to have them evaluated.
Type :help for more information.

scala> :quit
vboxuser@ubuntuu:/opt/spark/bin$ pyspark
Python 3.10.6 (main, Nov 14 2022, 16:10:14) [GCC 11.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
22/12/11 22:53:25 WARN Utils: Your hostname, ubuntuu resolves to a loopback add
ress: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
22/12/11 22:53:25 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
 address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
22/12/11 22:53:29 WARN NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Python version 3.10.6 (main, Nov 14 2022 16:10:14)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-167081361394
3).
SparkSession available as 'spark'.
>>> 
```

2. Realice el siguiente código, documente su funcionamiento en apache spark

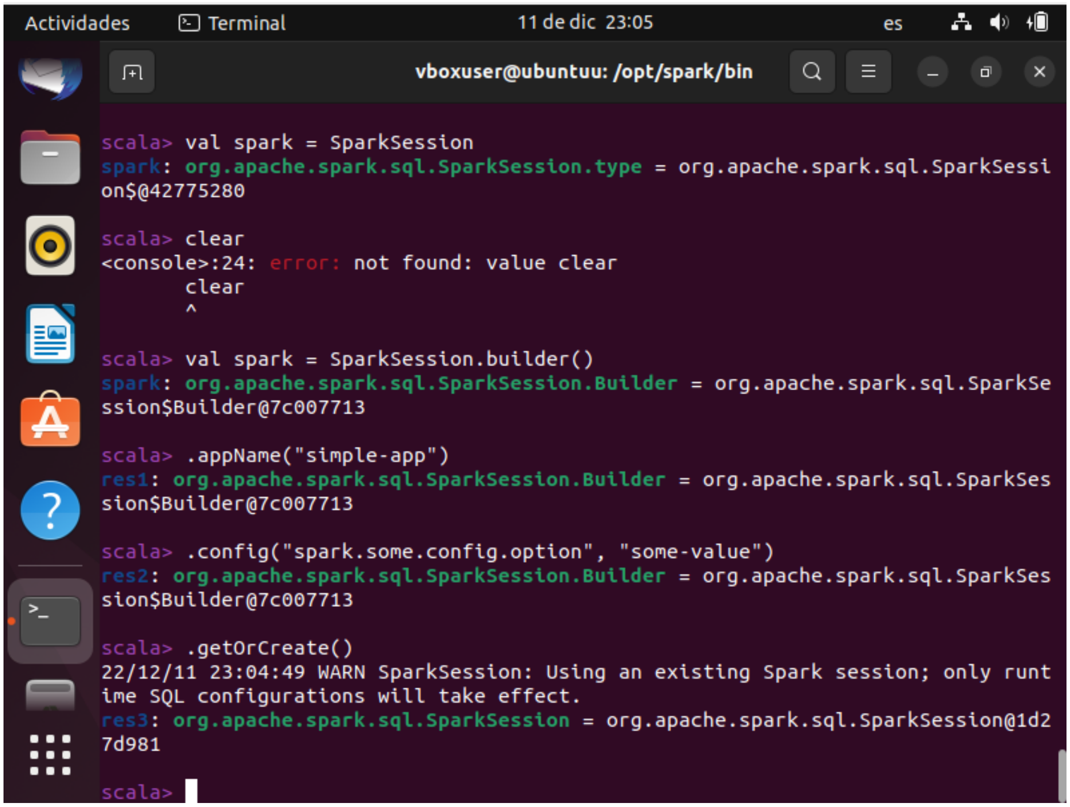
Importamos las librerías necesarias



```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
22/12/11 22:59:38 WARN NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-167081398372
3).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 11.0.17)
Type in expressions to have them evaluated.
Type :help for more information.

scala> import org.apache.spark.aql.SparkSession
<console>:22: error: object aql is not a member of package org.apache.spark
       import org.apache.spark.aql.SparkSession
                               ^

scala> import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.SparkSession

scala>
```

Copiamos el código

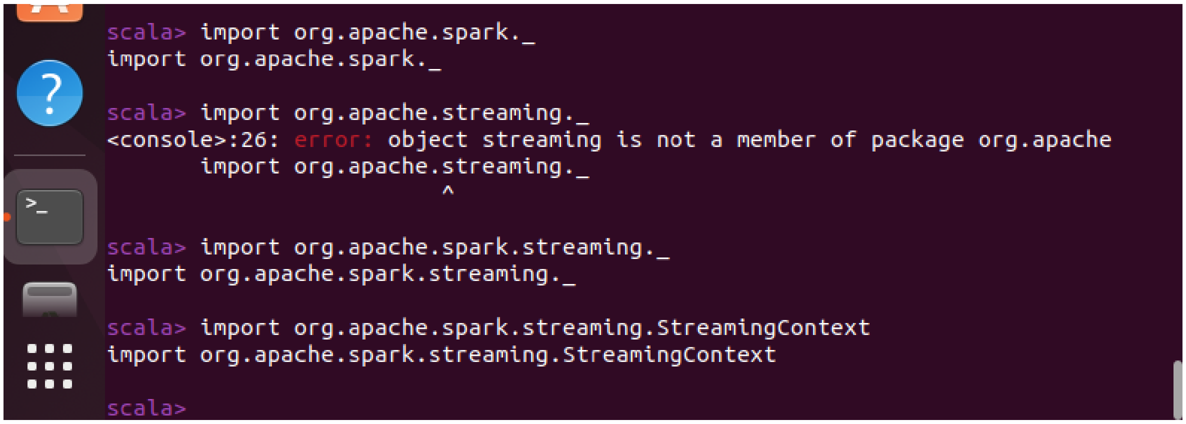
```
val spark: SparkSession = SparkSession.builder()
   .master("local[*]")
   .appName("simple-app")
   .getOrCreate()

val dataSet: Dataset[String] = spark.read.textFile("textfile.csv")
val df: DataFrame = dataSet.toDF()
```



```
scala> val spark = SparkSession
spark: org.apache.spark.sql.SparkSession.type = org.apache.spark.sql.SparkSessi
on$@42775280

scala> clear
<console>:24: error: not found: value clear
       clear
       ^

scala> val spark = SparkSession.builder()
spark: org.apache.spark.sql.SparkSession.Builder = org.apache.spark.sql.SparkSe
ssion$Builder@7c007713

scala> .appName("simple-app")
res1: org.apache.spark.sql.SparkSession.Builder = org.apache.spark.sql.SparkSes
sion$Builder@7c007713

scala> .config("spark.some.config.option", "some-value")
res2: org.apache.spark.sql.SparkSession.Builder = org.apache.spark.sql.SparkSes
sion$Builder@7c007713

scala> .getOrCreate()
22/12/11 23:04:49 WARN SparkSession: Using an existing Spark session; only runt
ime SQL configurations will take effect.
res3: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@1d2
7d981

scala> 
```

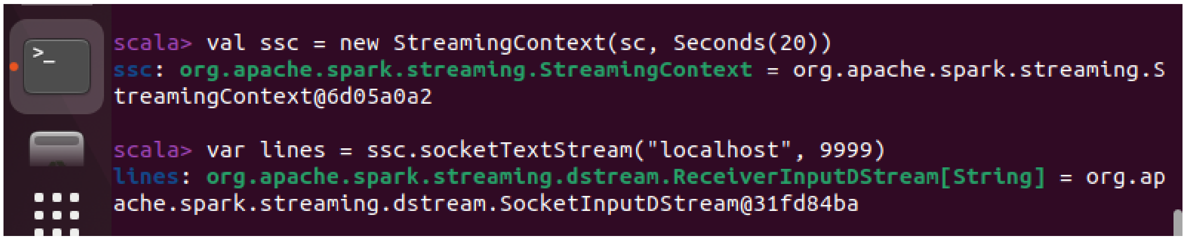
```
val streamingContext: StreamingContext = new StreamingContext(sparkContext, Seconds(20))
val lines: ReceiverInputDStream[String] = streamingContext.socketTextStream("localhost", 9999)
```

Pero antes necesitamos importar algunas librerías:



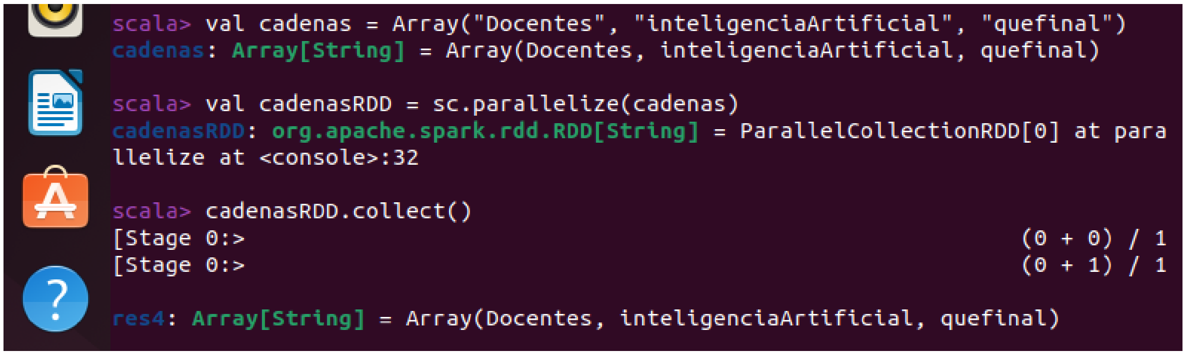
```
scala> import org.apache.spark._
import org.apache.spark._

scala> import org.apache.streaming._
<console>:26: error: object streaming is not a member of package org.apache
       import org.apache.streaming._
                         ^

scala> import org.apache.spark.streaming._
import org.apache.spark.streaming._

scala> import org.apache.spark.streaming.StreamingContext
import org.apache.spark.streaming.StreamingContext

scala>
```

Ahora si, continuando copinado el codigo:

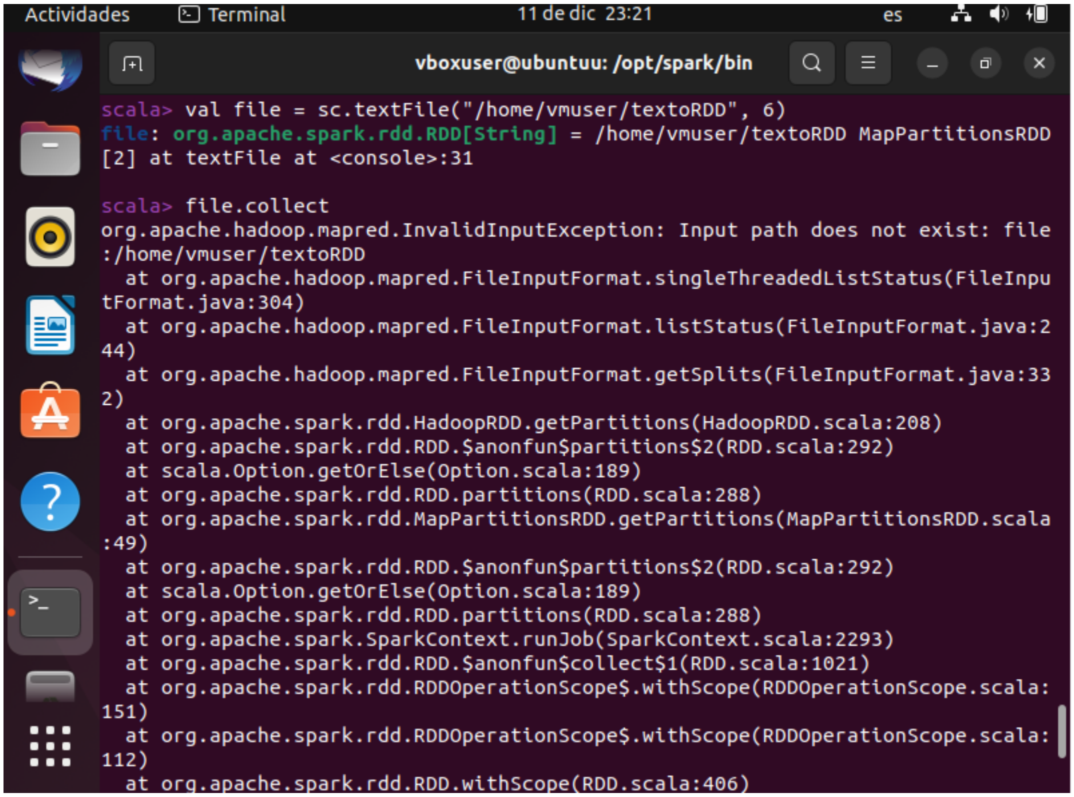


```
scala> val ssc = new StreamingContext(sc, Seconds(20))
ssc: org.apache.spark.streaming.StreamingContext = org.apache.spark.streaming.S
treamingContext@6d05a0a2

scala> var lines = ssc.socketTextStream("localhost", 9999)
lines: org.apache.spark.streaming.dstream.ReceiverInputDStream[String] = org.ap
ache.spark.streaming.dstream.SocketInputDStream@31fd84ba
```

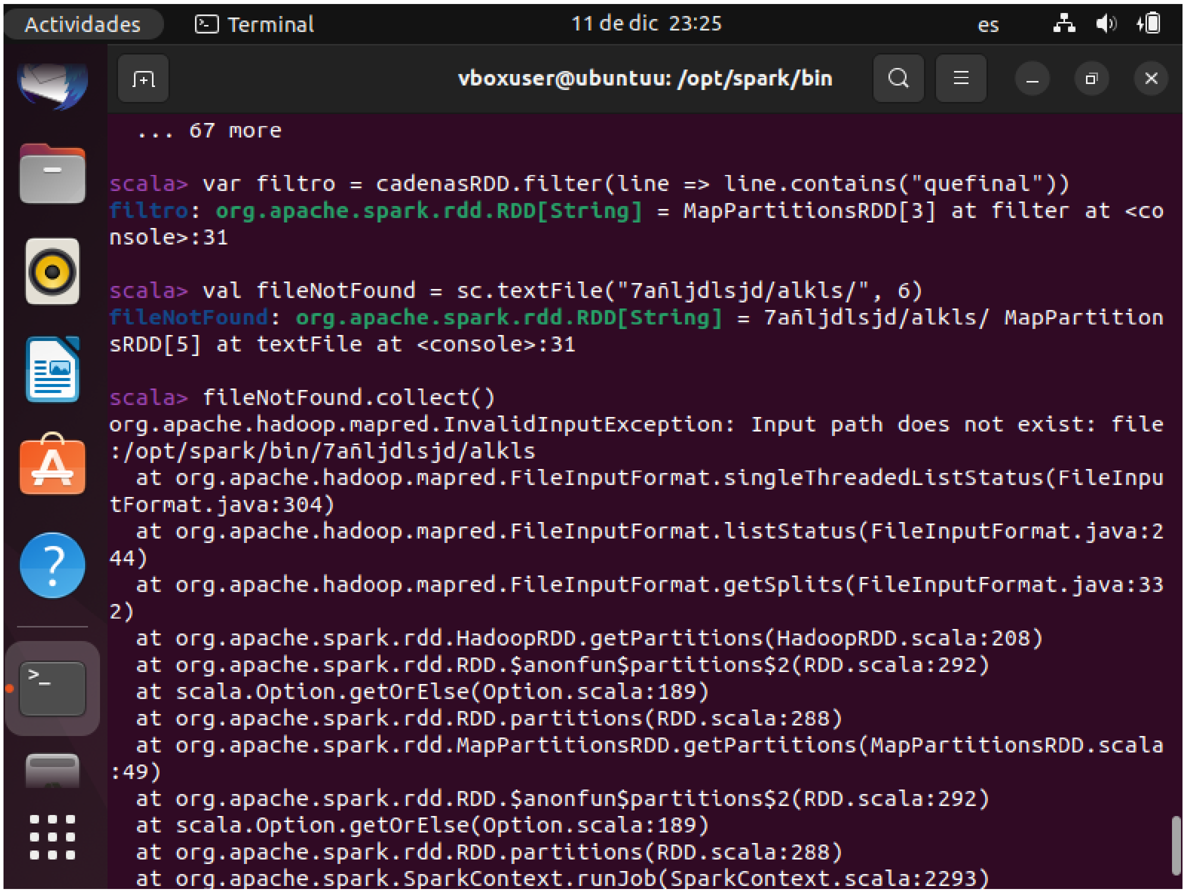
val cadenas = Array(“Docentes”, “inteligenciaArtificial”, “quefinal”)
val cadenasRDD = sc . parallelize (cadenas)
cadenasRDD.collect()
file.collect()
val filtro = cadenasRDD.filter(line => line.contains(“quefinal”))
val fileNotFound = sc.textFile(“/7añljdlsjd/alkls/”, 6)
fileNotFound.collect()



```
scala> val cadenas = Array("Docentes", "inteligenciaArtificial", "quefinal")
cadenas: Array[String] = Array(Docentes, inteligenciaArtificial, quefinal)

scala> val cadenasRDD = sc.parallelize(cadenas)
cadenasRDD: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[0] at para
llelize at <console>:32

scala> cadenasRDD.collect()
[Stage 0:>                                                          (0 + 0) / 1
[Stage 0:>                                                          (0 + 1) / 1

res4: Array[String] = Array(Docentes, inteligenciaArtificial, quefinal)
```



```
scala> val file = sc.textFile("/home/vmuser/textoRDD", 6)
file: org.apache.spark.rdd.RDD[String] = /home/vmuser/textoRDD MapPartitionsRDD
[2] at textFile at <console>:31

scala> file.collect
org.apache.hadoop.mapred.InvalidInputException: Input path does not exist: file
:/home/vmuser/textoRDD
  at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInpu
tFormat.java:304)
  at org.apache.hadoop.mapred.FileInputFormat.listStatus(FileInputFormat.java:2
44)
  at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:33
2)
  at org.apache.spark.rdd.HadoopRDD.getPartitions(HadoopRDD.scala:208)
  at org.apache.spark.rdd.RDD.$anonfun$partitions$2(RDD.scala:292)
  at scala.Option.getOrElse(Option.scala:189)
  at org.apache.spark.rdd.RDD.partitions(RDD.scala:288)
  at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala
:49)
  at org.apache.spark.rdd.RDD.$anonfun$partitions$2(RDD.scala:292)
  at scala.Option.getOrElse(Option.scala:189)
  at org.apache.spark.rdd.RDD.partitions(RDD.scala:288)
  at org.apache.spark.SparkContext.runJob(SparkContext.scala:2293)
  at org.apache.spark.rdd.RDD.$anonfun$collect$1(RDD.scala:1021)
  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:
151)
  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:
112)
  at org.apache.spark.rdd.RDD.withScope(RDD.scala:406)
```



```
  ... 67 more

scala> var filtro = cadenasRDD.filter(line => line.contains("quefinal"))
filtro: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[3] at filter at <co
nsole>:31

scala> val fileNotFound = sc.textFile("7añljdlsjd/alkls/", 6)
fileNotFound: org.apache.spark.rdd.RDD[String] = 7añljdlsjd/alkls/ MapPartition
sRDD[5] at textFile at <console>:31

scala> fileNotFound.collect()
org.apache.hadoop.mapred.InvalidInputException: Input path does not exist: file
:/opt/spark/bin/7añljdlsjd/alkls
  at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInpu
tFormat.java:304)
  at org.apache.hadoop.mapred.FileInputFormat.listStatus(FileInputFormat.java:2
44)
  at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:33
2)
  at org.apache.spark.rdd.HadoopRDD.getPartitions(HadoopRDD.scala:208)
  at org.apache.spark.rdd.RDD.$anonfun$partitions$2(RDD.scala:292)
  at scala.Option.getOrElse(Option.scala:189)
  at org.apache.spark.rdd.RDD.partitions(RDD.scala:288)
  at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala
:49)
  at org.apache.spark.rdd.RDD.$anonfun$partitions$2(RDD.scala:292)
  at scala.Option.getOrElse(Option.scala:189)
  at org.apache.spark.rdd.RDD.partitions(RDD.scala:288)
  at org.apache.spark.SparkContext.runJob(SparkContext.scala:2293)
```