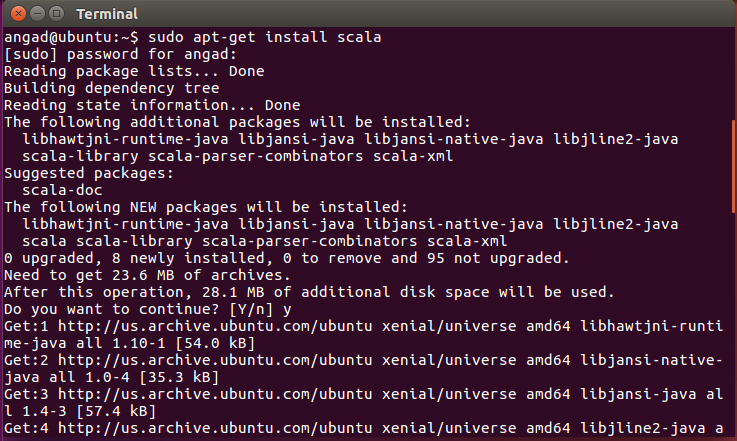
***Installing Scala***

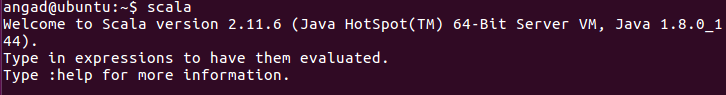
1. First, we have to download the Scala by using this command:

sudo apt-get install scala



1. Now, we should check whether Scala is properly installed or not. So, we are simply going to type the following command:

Scala



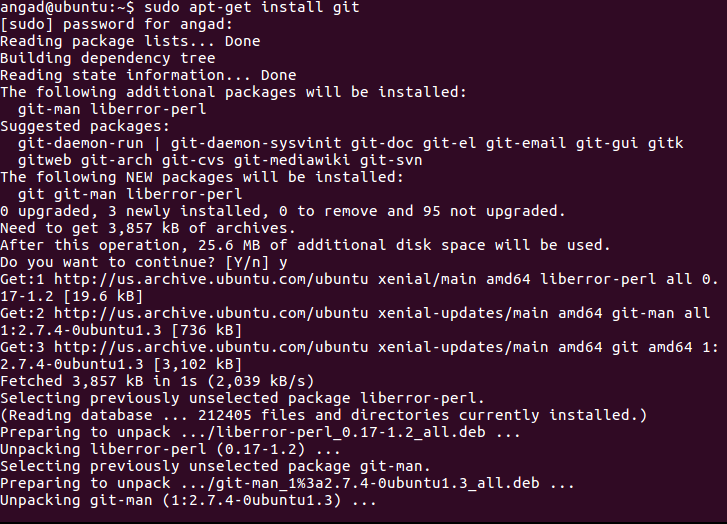
1. You should see the Scala REPL running. Test it with:

println(“Hello World”)



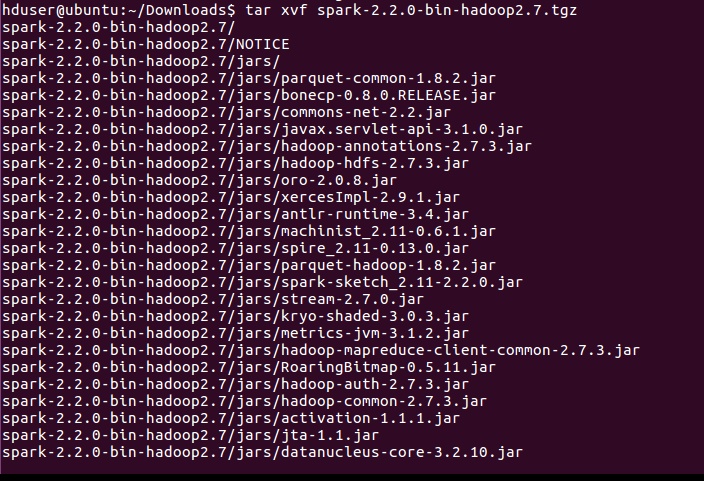
1. Next it is time to install Spark. We need git for this, so in your terminal type:

sudo apt-get install git



1. Now, we have to extract the spark package by

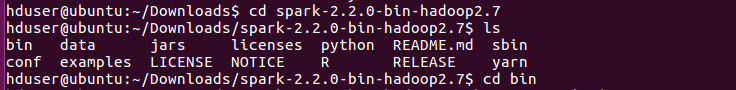
tar xvf spark-2.0.2-bin-hadoop2.7.tgz



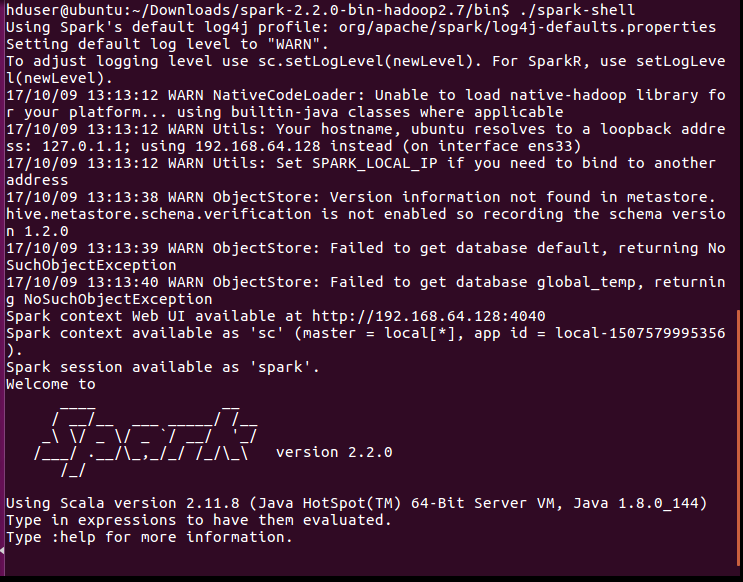
1. Then once its done extracting the Spark folder, use:

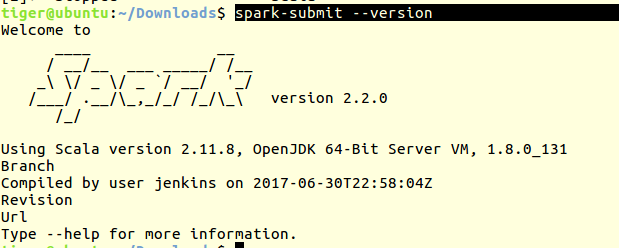
cd spark-2.0.2-bin-hadoop2.7

1. Then we should go in the bin folder of Scala

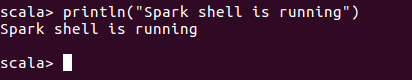


1. ./spark-shell





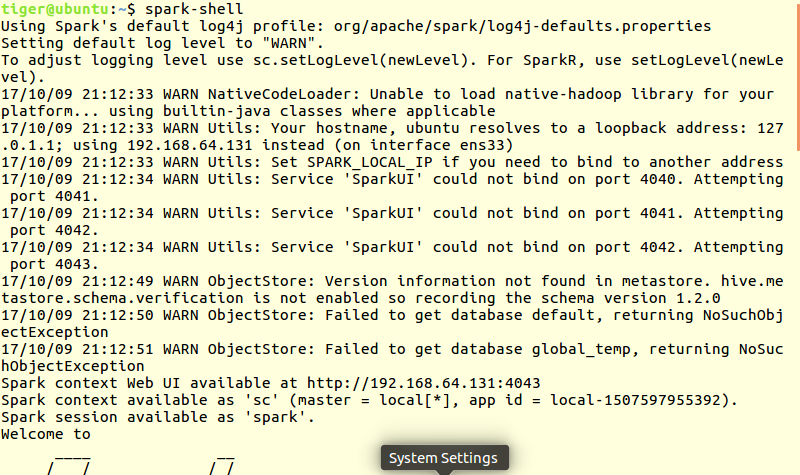
1. To verify the Scala is working:



Wordcount Program:

1. Run the command

spark-shell

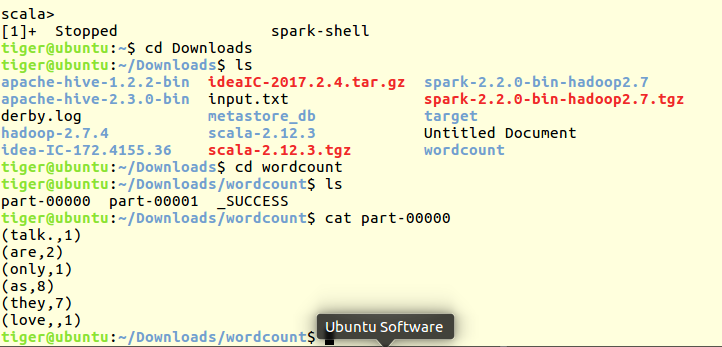


1. Create a simple RDD from the text file

val infile = sc.textFile(“/home/tiger/Downloads/input.txt”)

val counts = infile.flatMap(line => line.split(“ ”)).map(word => (word, 1)).reduceByKey(\_ + \_);

counts.saveAsTextFile(“/usr/tiger/Downloads/wordcount”)



***Import CSV Files into HIVE Using Spark (Python)***

>>> from pyspark.sql import HiveContext

>>> from pyspark.sql.types import \*

>>> from pyspark.sql import Row

>>> csv\_data = sc.textFile("/home/tiger/Downloads/chronic.csv")

>>> type(csv\_data)

<class 'pyspark.rdd.RDD'>

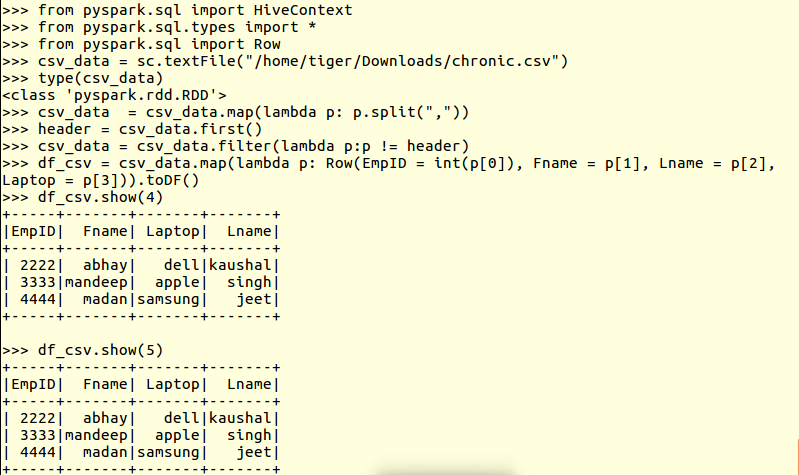
>>> csv\_data = csv\_data.map(lambda p: p.split(","))

>>> header = csv\_data.first()

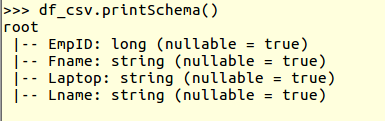
>>> csv\_data = csv\_data.filter(lambda p:p != header)

>>> df\_csv = csv\_data.map(lambda p: Row(EmpID = int(p[0]), Fname = p[1], Lname = p[2], Laptop = p[3])).toDF()

>>> df\_csv.show(4)



>>> df\_csv.printSchema()



>>> from pyspark.sql import HiveContext

>>> hc = HiveContext(sc)

>>> df\_csv.write.format("orc").saveAsTable("employees")

