**Lab7 Spark**

* First I have downloaded the spark
* Unzipped and setuped the environment variables.

cd /home/hadoop/spark

cp ./conf/spark-env.sh.template ./conf/spark-env.sh

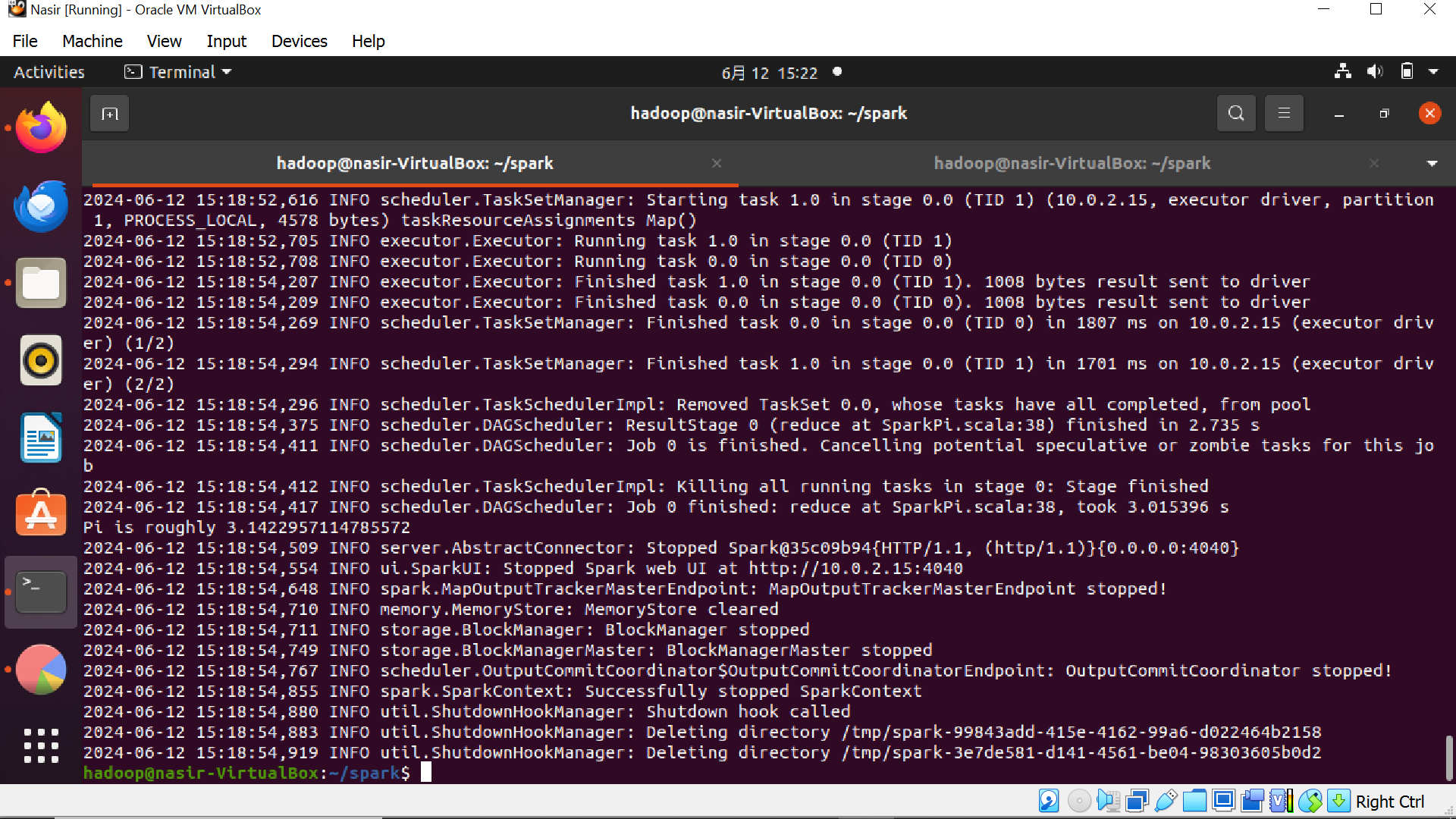
* Append a line to the file: spark-env.sh:

export SPARK\_DIST\_CLASSPATH=$(/home/hadoop/hadoop-3.2.4/bin/hadoop classpath)

* In shell, run the command:

cd /home/hadoop/spark

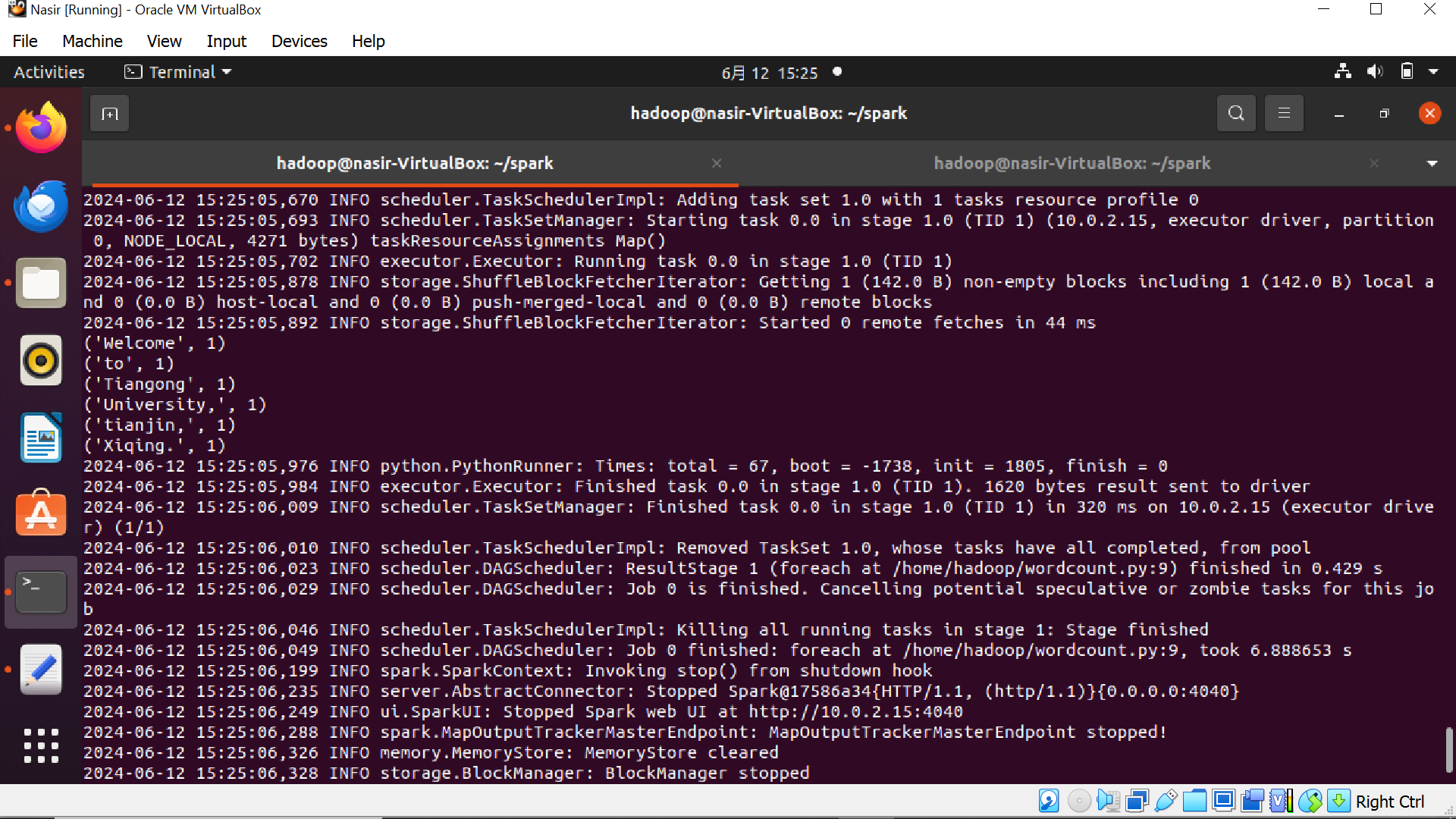
bin/run-example SparkPi



* Then I Create and edit wordcount.py in /home/hadoop/
* Then In shell, run :

cd /home/hadoop/spark

./bin/spark-submit /home/hadoop/wordcount.py



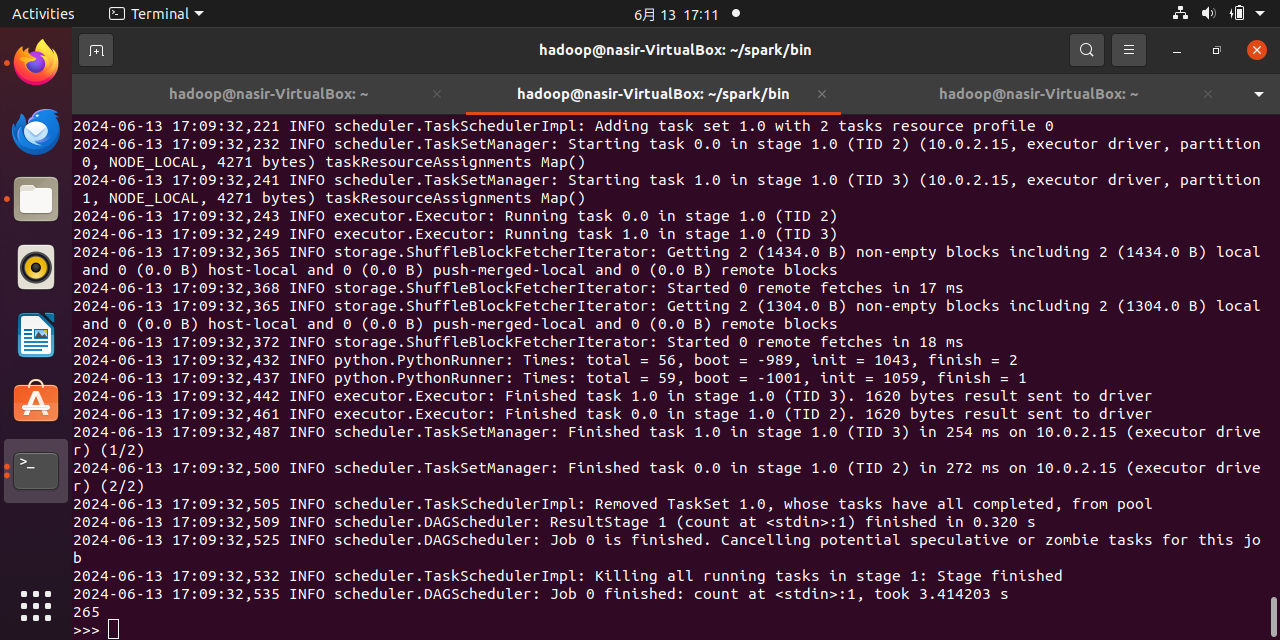
**Task1**

lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")

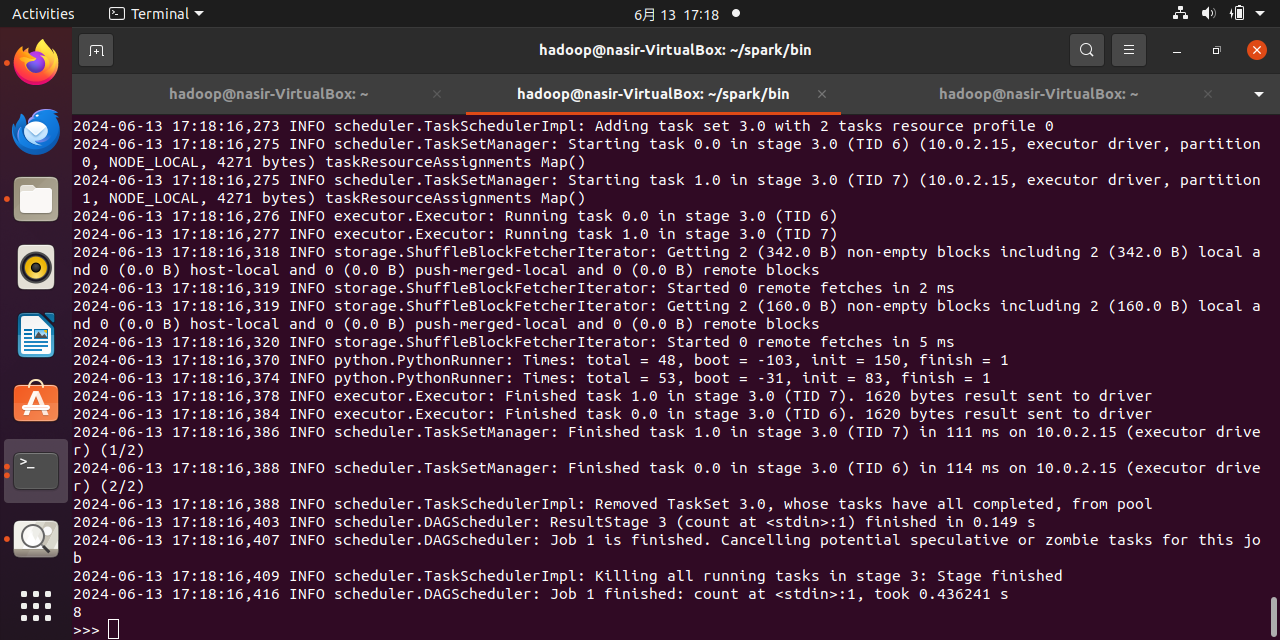
res = lines.map(lambda x: x.split(",")).map(lambda x: x[0])

distinct\_res = res.distinct() # Remove duplicates

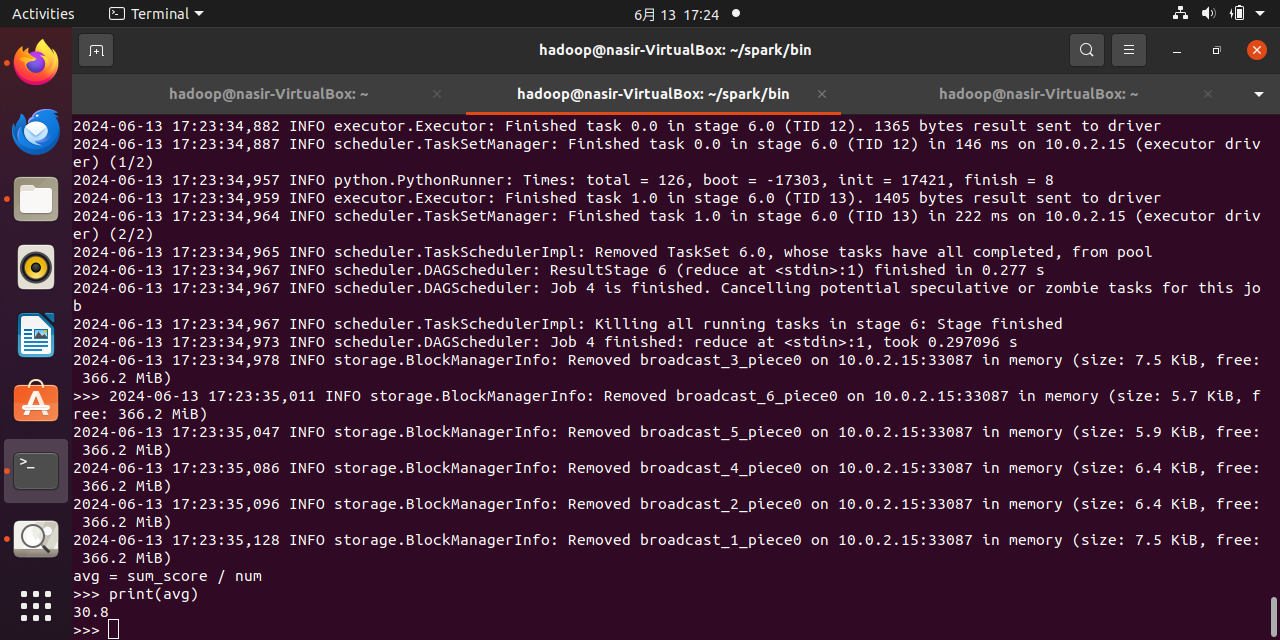
distinct\_res.count() # Get the total number of elements



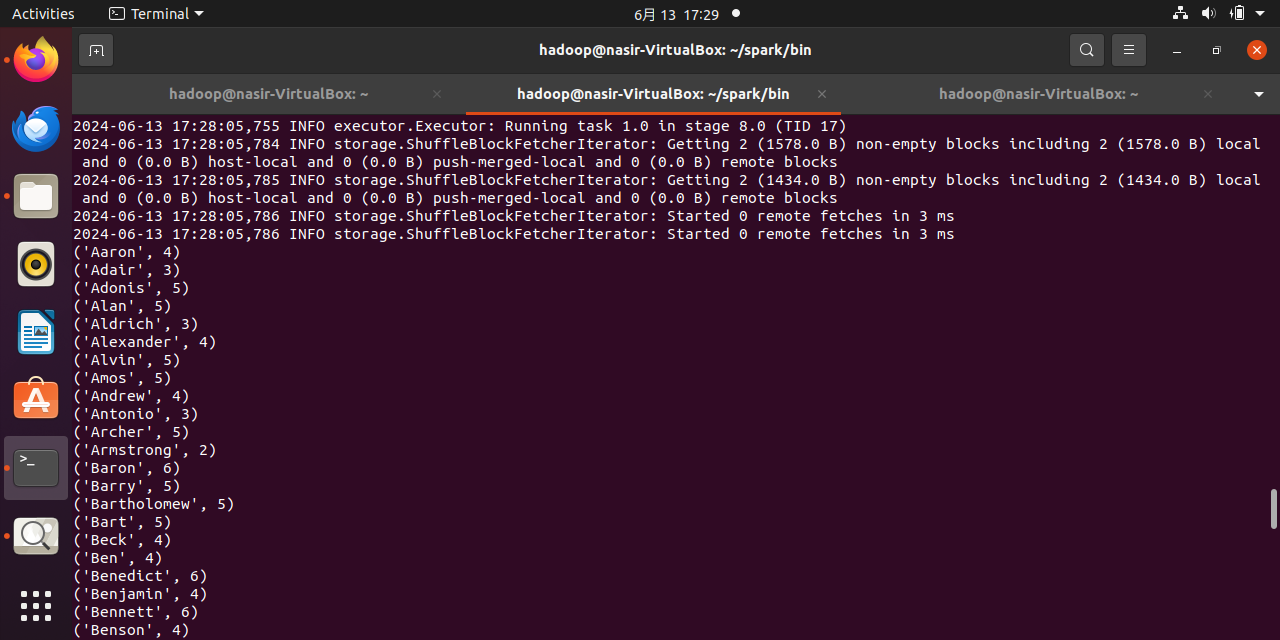
**How many courses are offered in the department?**

****

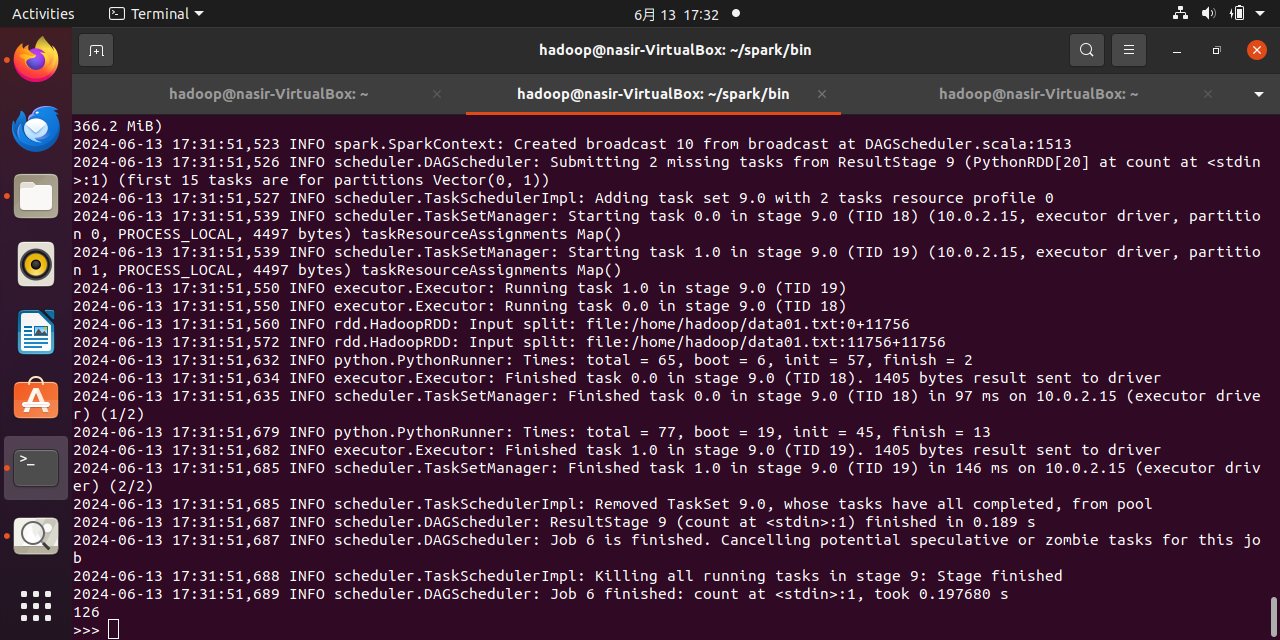
**What is the average total score of Tom?**

****

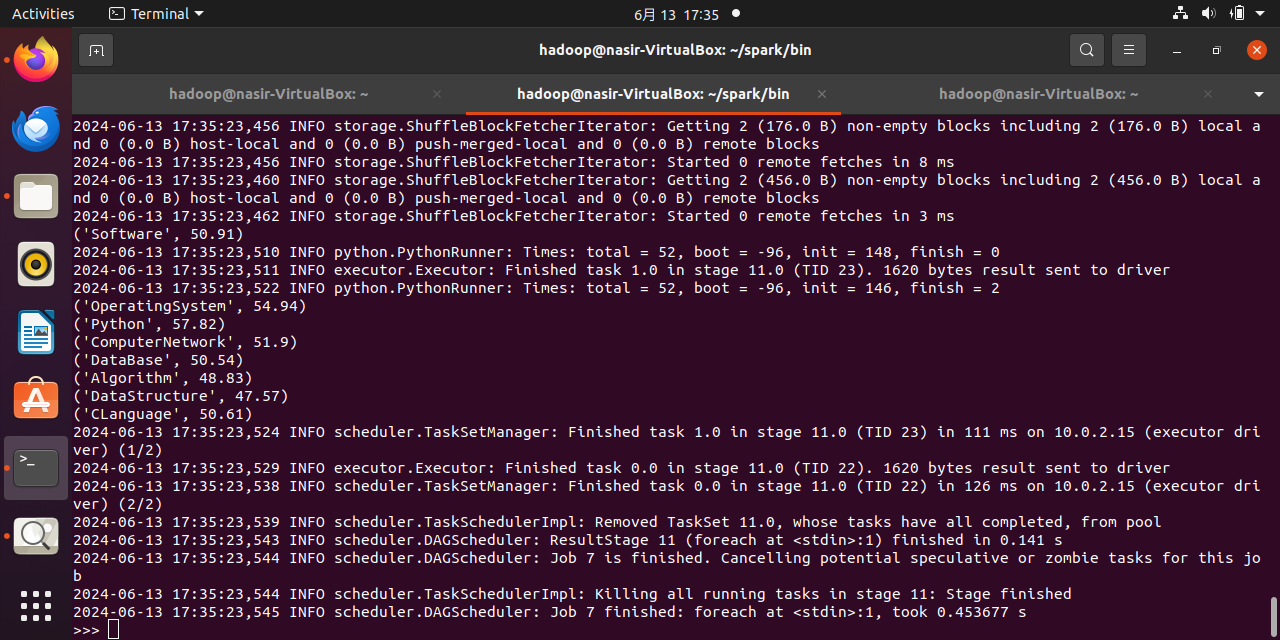
**How many courses has each student taken?**

****

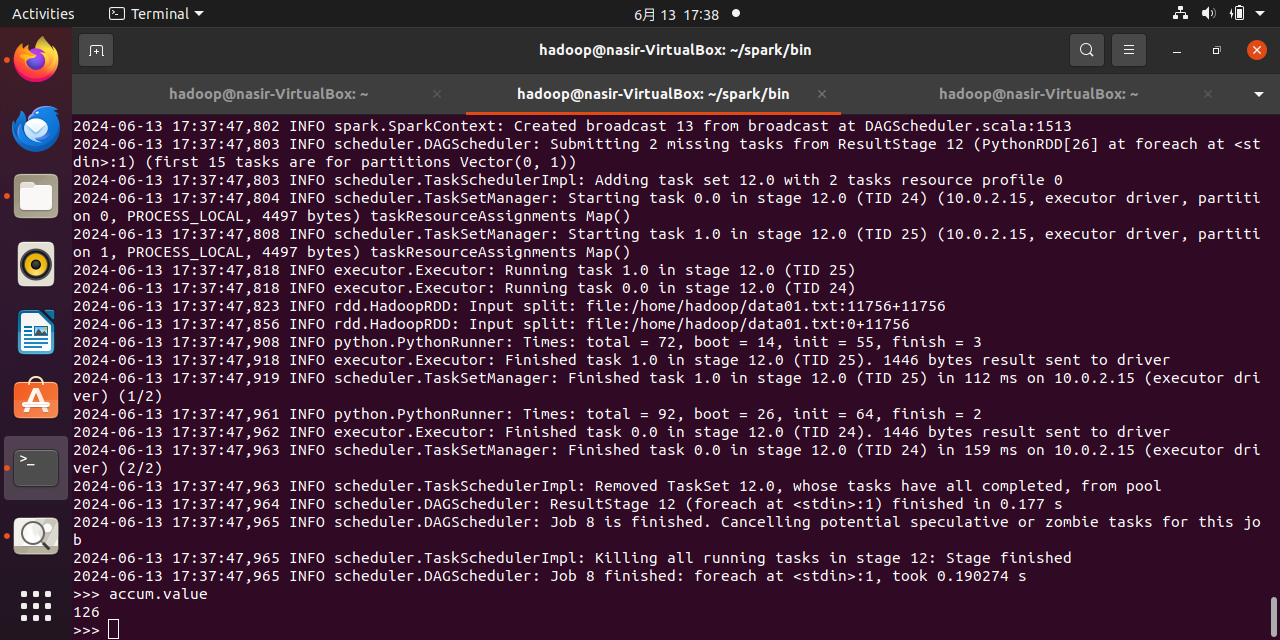
**How many students have enrolled in the DataBase course?**

****

**What is the average score for each course?**

****

**Use an accumulator to calculate how many students have taken the DataBase course**

****

**Task 2**