TUGAS INDIVIDU

BIG DATA

PERTEMUAN 5

Disusun oleh:

Rijal Ammar Irsyadul Ibad

TI-3B/2041720184

D4 TEKNIK INFORMATIKA

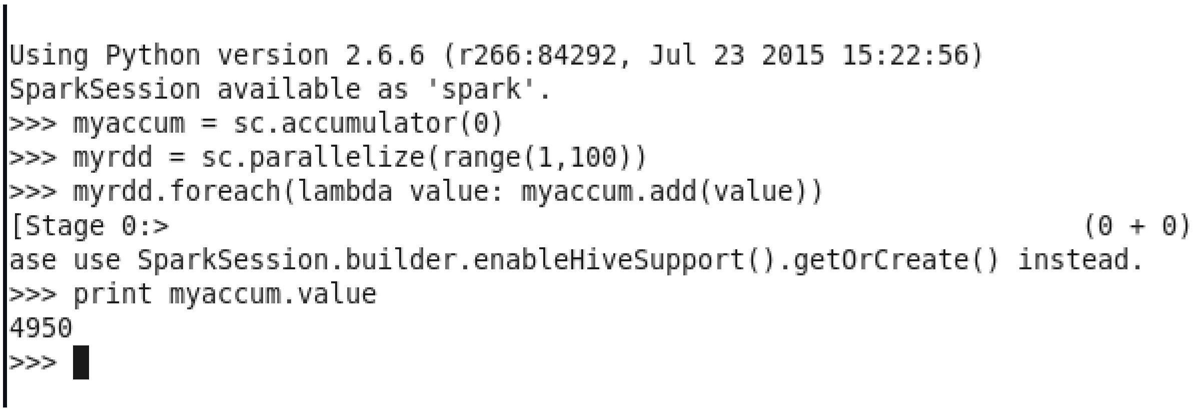
TEKNOLOGI INFORMASI

POLITEKNIK NEGERI MALANG

2023

A. Code Accumulator

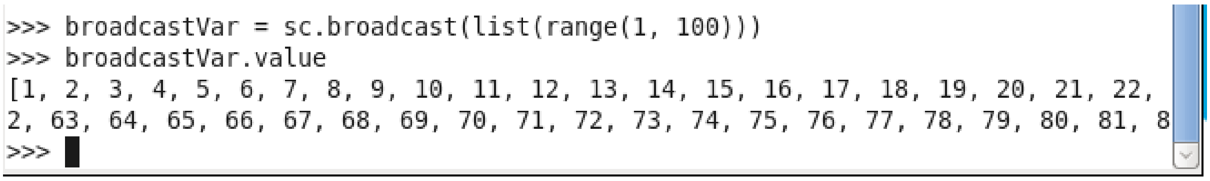
Saat dicoba di Cloudera



```
Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkSession available as 'spark'.
>>> myaccum = sc.accumulator(0)
>>> myrdd = sc.parallelize(range(1,100))
>>> myrdd.foreach(lambda value: myaccum.add(value))
[Stage 0:>                                                    (0 + 0)
ase use SparkSession.builder.enableHiveSupport().getOrCreate() instead.
>>> print myaccum.value
4950
>>> 
```

B. Code BoardCast

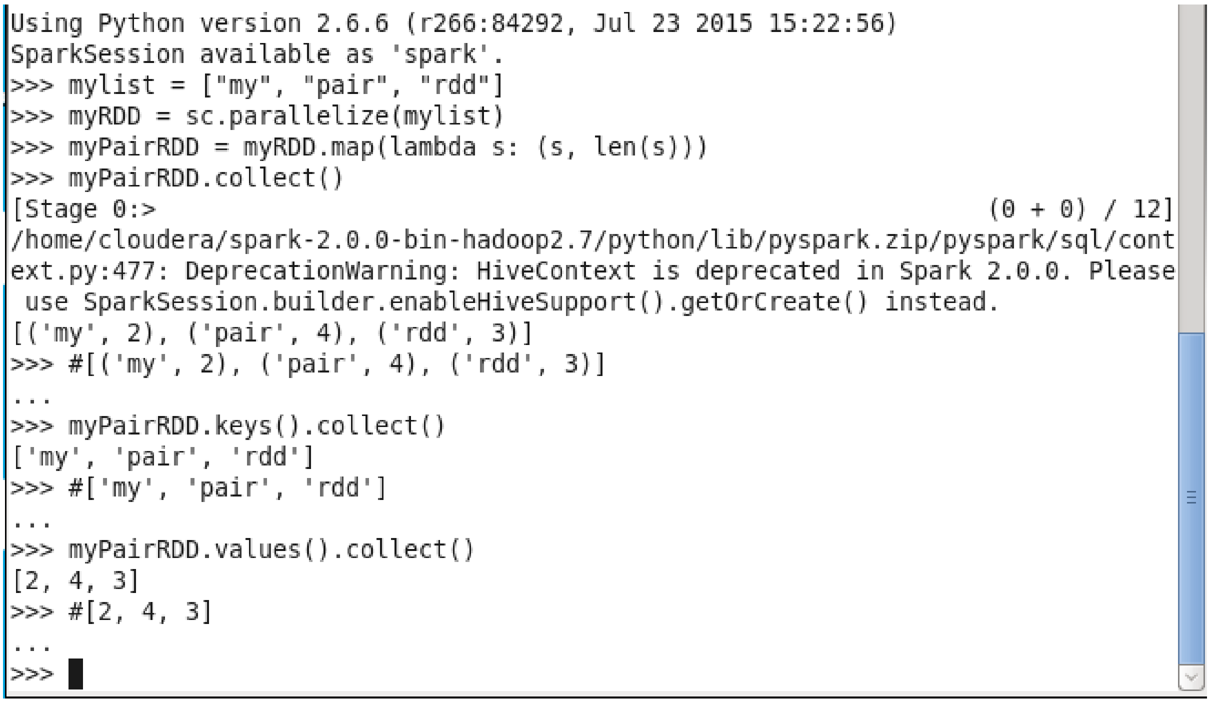
Saat dicoba di Cloudera



```
>>> broadcastVar = sc.broadcast(list(range(1, 100)))
>>> broadcastVar.value
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,
2, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 8
>>> 
```

C. PairRRD

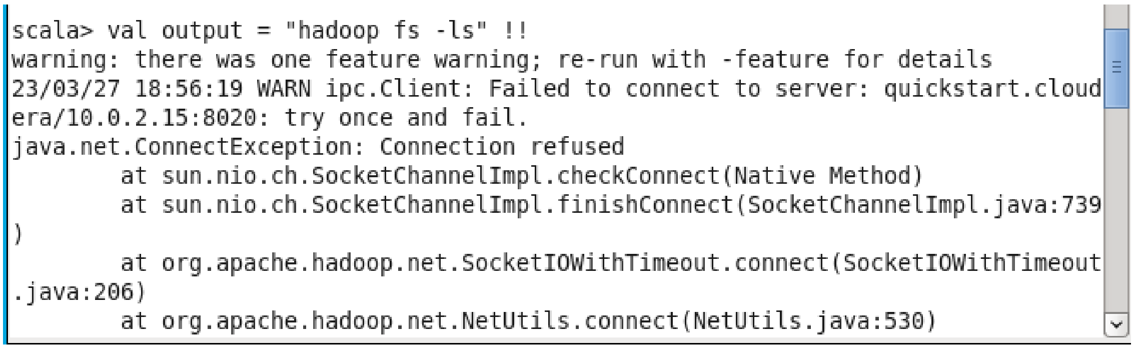
Saat dicoba di Cloudera



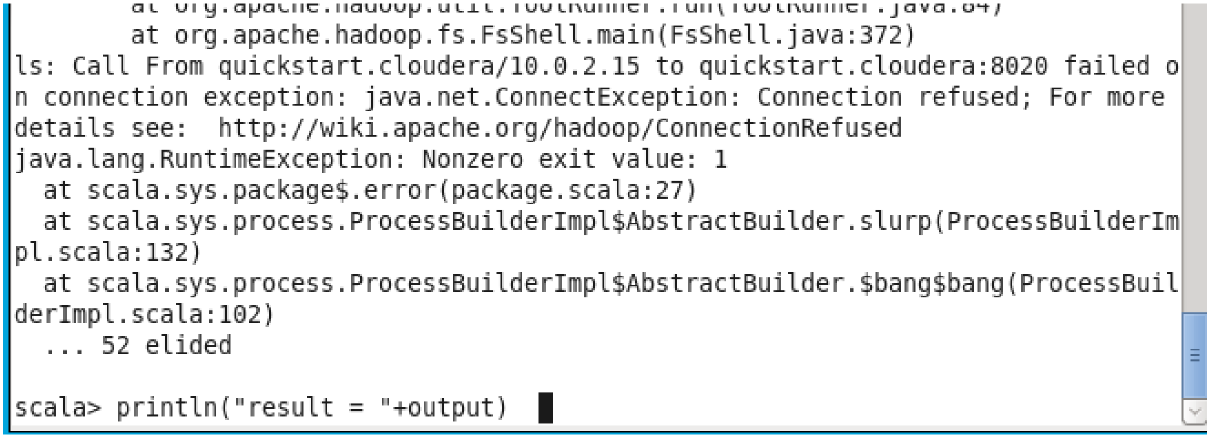
```
Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkSession available as 'spark'.
>>> mylist = ["my", "pair", "rdd"]
>>> myRDD = sc.parallelize(mylist)
>>> myPairRDD = myRDD.map(lambda s: (s, len(s)))
>>> myPairRDD.collect()
[Stage 0:>                                                   (0 + 0) / 12]
/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/lib/pyspark.zip/pyspark/sql/cont
ext.py:477: DeprecationWarning: HiveContext is deprecated in Spark 2.0.0. Please
 use SparkSession.builder.enableHiveSupport().getOrCreate() instead.
[('my', 2), ('pair', 4), ('rdd', 3)]
>>> #[('my', 2), ('pair', 4), ('rdd', 3)]
...
>>> myPairRDD.keys().collect()
['my', 'pair', 'rdd']
>>> #['my', 'pair', 'rdd']
...
>>> myPairRDD.values().collect()
[2, 4, 3]
>>> #[2, 4, 3]
...
>>> 
```

D. System Commands Output (Scala)

Saat dicoba di Cloudera



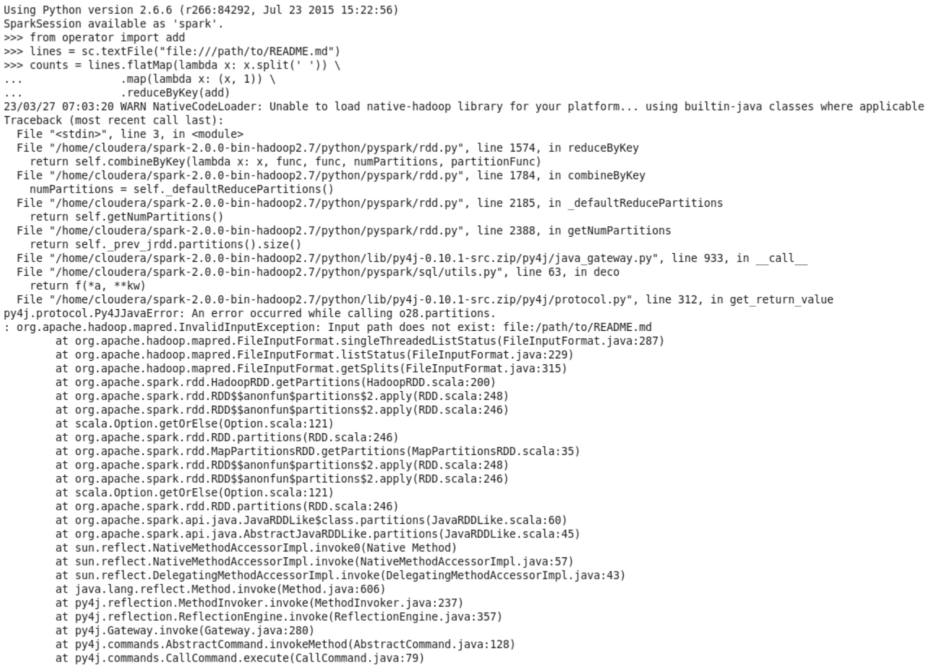
```
scala> val output = "hadoop fs -ls" !!
warning: there was one feature warning; re-run with -feature for details
23/03/27 18:56:19 WARN ipc.Client: Failed to connect to server: quickstart.cloud
era/10.0.2.15:8020: try once and fail.
java.net.ConnectException: Connection refused
        at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
        at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:739
)
        at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout
.java:206)
        at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
```



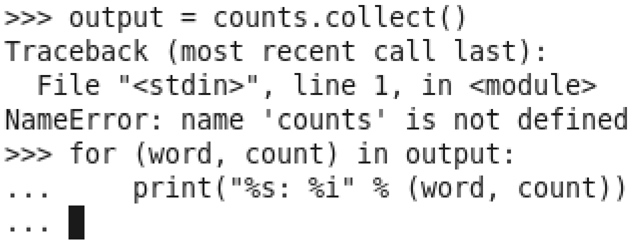
```
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
        at org.apache.hadoop.fs.FsShell.main(FsShell.java:372)
ls: Call From quickstart.cloudera/10.0.2.15 to quickstart.cloudera:8020 failed o
n connection exception: java.net.ConnectException: Connection refused; For more
details see:  http://wiki.apache.org/hadoop/ConnectionRefused
java.lang.RuntimeException: Nonzero exit value: 1
  at scala.sys.package$.error(package.scala:27)
  at scala.sys.process.ProcessBuilderImpl$AbstractBuilder.slurp(ProcessBuilderIm
pl.scala:132)
  at scala.sys.process.ProcessBuilderImpl$AbstractBuilder.$bang$bang(ProcessBuil
derImpl.scala:102)
  ... 52 elided

scala> println("result = "+output)
```

E. WordCount

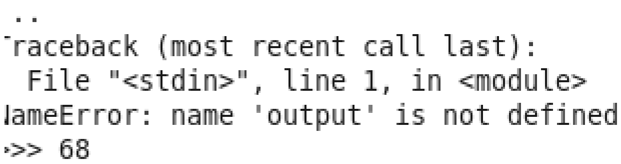
Saat dicoba di Cloudera



```
Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkSession available as 'spark'.
>>> from operator import add
>>> lines = sc.textFile("file:///path/to/README.md")
>>> counts = lines.flatMap(lambda x: x.split(' ')) \
...               .map(lambda x: (x, 1)) \
...               .reduceByKey(add)
23/03/27 07:03:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Traceback (most recent call last):
  File "<stdin>", line 3, in <module>
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/pyspark/rdd.py", line 1574, in reduceByKey
    return self.combineByKey(lambda x: x, func, func, numPartitions, partitionFunc)
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/pyspark/rdd.py", line 1784, in combineByKey
    numPartitions = self._defaultReducePartitions()
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/pyspark/rdd.py", line 2185, in _defaultReducePartitions
    return self.getNumPartitions()
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/pyspark/rdd.py", line 2388, in getNumPartitions
    return self._prev_jrdd.partitions().size()
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/lib/py4j-0.10.1-src.zip/py4j/java_gateway.py", line 933, in __call__
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/pyspark/sql/utils.py", line 63, in deco
    return f(*a, **kw)
  File "/home/cloudera/spark-2.0.0-bin-hadoop2.7/python/lib/py4j-0.10.1-src.zip/py4j/protocol.py", line 312, in get_return_value
py4j.protocol.Py4JJavaError: An error occurred while calling o28.partitions.
: org.apache.hadoop.mapred.InvalidInputException: Input path does not exist: file:/path/to/README.md
	at org.apache.hadoop.mapred.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:287)
	at org.apache.hadoop.mapred.FileInputFormat.listStatus(FileInputFormat.java:229)
	at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:315)
	at org.apache.spark.rdd.HadoopRDD.getPartitions(HadoopRDD.scala:200)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:248)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:246)
	at scala.Option.getOrElse(Option.scala:121)
	at org.apache.spark.rdd.RDD.partitions(RDD.scala:246)
	at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:35)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:248)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:246)
	at scala.Option.getOrElse(Option.scala:121)
	at org.apache.spark.rdd.RDD.partitions(RDD.scala:246)
	at org.apache.spark.api.java.JavaRDDLike$class.partitions(JavaRDDLike.scala:60)
	at org.apache.spark.api.java.AbstractJavaRDDLike.partitions(JavaRDDLike.scala:45)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
	at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
	at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
	at java.lang.reflect.Method.invoke(Method.java:606)
	at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:237)
	at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
	at py4j.Gateway.invoke(Gateway.java:280)
	at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:128)
	at py4j.commands.CallCommand.execute(CallCommand.java:79)
```



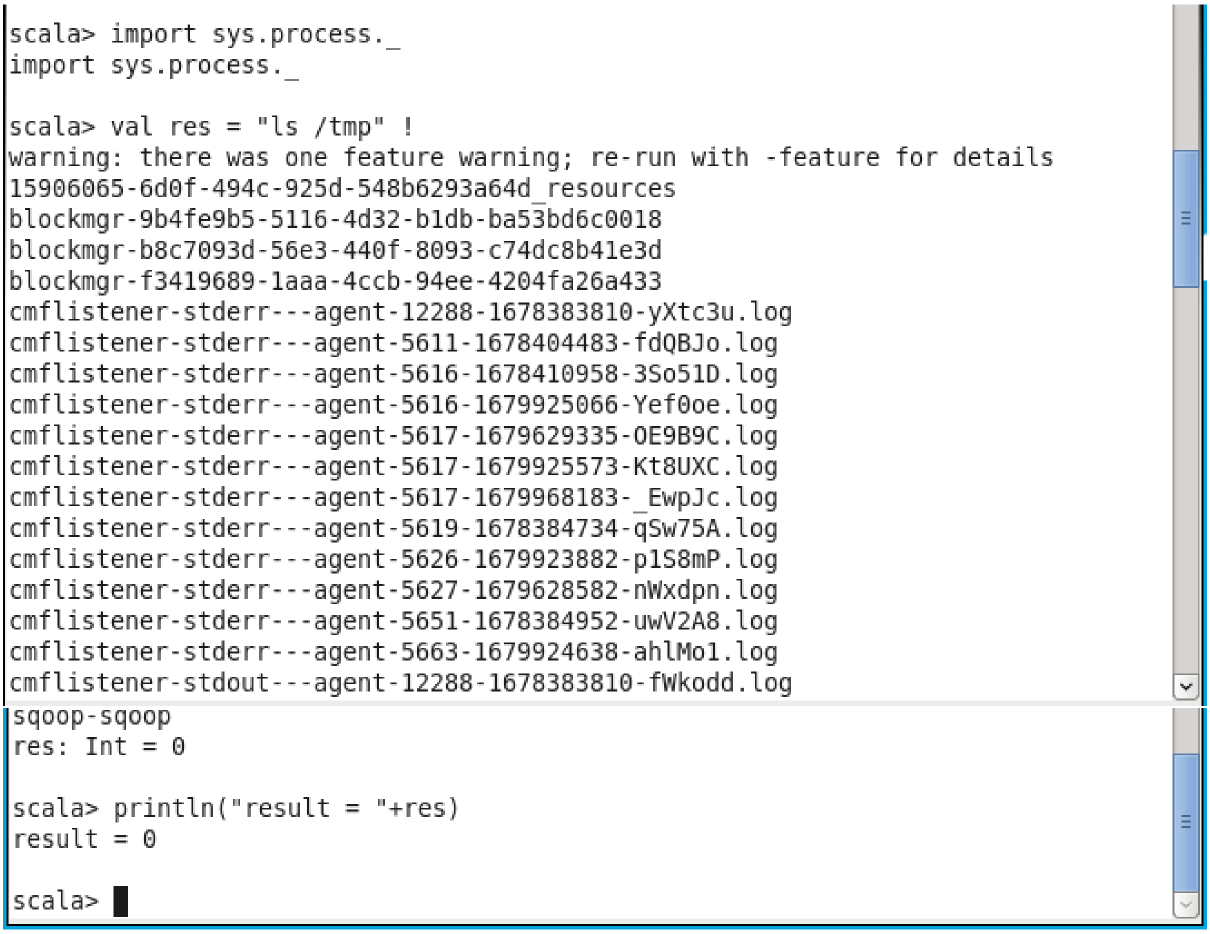
```
>>> output = counts.collect()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'counts' is not defined
>>> for (word, count) in output:
...     print("%s: %i" % (word, count))
...
```



```
 ..
raceback (most recent call last):
 File "<stdin>", line 1, in <module>
lameError: name 'output' is not defined
>> 68
```

F. System Commands Return (Scala)

Saat dicoba di Cloudera



```
scala> import sys.process._
import sys.process._

scala> val res = "ls /tmp" !
warning: there was one feature warning; re-run with -feature for details
15906065-6d0f-494c-925d-548b6293a64d_resources
blockmgr-9b4fe9b5-5116-4d32-b1db-ba53bd6c0018
blockmgr-b8c7093d-56e3-440f-8093-c74dc8b41e3d
blockmgr-f3419689-1aaa-4ccb-94ee-4204fa26a433
cmflistener-stderr---agent-12288-1678383810-yXtc3u.log
cmflistener-stderr---agent-5611-1678404483-fdQBJo.log
cmflistener-stderr---agent-5616-1678410958-3So51D.log
cmflistener-stderr---agent-5616-1679925066-Yef0oe.log
cmflistener-stderr---agent-5617-1679629335-OE9B9C.log
cmflistener-stderr---agent-5617-1679925573-Kt8UXC.log
cmflistener-stderr---agent-5617-1679968183-_EwpJc.log
cmflistener-stderr---agent-5619-1678384734-qSw75A.log
cmflistener-stderr---agent-5626-1679923882-p1S8mP.log
cmflistener-stderr---agent-5627-1679628582-nWxdpn.log
cmflistener-stderr---agent-5651-1678384952-uwV2A8.log
cmflistener-stderr---agent-5663-1679924638-ahlMo1.log
cmflistener-stdout---agent-12288-1678383810-fWkodd.log
```

```
sqoop-sqoop
res: Int = 0

scala> println("result = "+res)
result = 0

scala> 
```