Sean McLean

ALY 6110

Module 5

Individual Lab #2

JDK, Spark Shell and Scala are installed and imported using the Command Prompt:

```
C:\Users\seanm>spark-shell
24/06/25 16:56:30 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP_HOME an
d hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/25 16:56:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where a
pplicable
Spark context Web UI available at http://mcleanse.myfiosgateway.com:4040
Spark context available as 'sc' (master = local[*], app id = local-1719348998648).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.1
      /_/

Using Scala version 2.12.18 (Java HotSpot(TM) Client VM, Java 1.8.0_411)
Type in expressions to have them evaluated.
Type :help for more information.
```

```
scala> // Import SparkContext

scala> import org.apache.spark.SparkContext
import org.apache.spark.SparkContext

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

Define the Path to the Text File:

```
scala> val txtFile = "C:/Users/seanm/Downloads/spark-3.5.1-bin-hadoop3/spark-3.5.1-bin-hadoop3/README.md"
txtFile: String = C:/Users/seanm/Downloads/spark-3.5.1-bin-hadoop3/spark-3.5.1-bin-hadoop3/README.md
```

```
scala> // Read the text file into an RDD

scala> val txtData = sc.textFile(txtFile)
txtData: org.apache.spark.rdd.RDD[String] = C:/Users/seanm/Downloads/spark-3.5.1-bin-hadoop3/spark-3.5.1-bin-hadoop3/README.md MapPar
titionsRDD[5] at textFile at <console>:28
```

```
scala> // Cache the RDD

scala> txtData.cache()
res4: txtData.type = C:/Users/seanm/Downloads/spark-3.5.1-bin-hadoop3/spark-3.5.1-bin-hadoop3/README.md MapPartitionsRDD[5] at textFi
le at <console>:28
```

```
scala> // Count the number of lines in the text file

scala> val lineCount = txtData.count()
lineCount: Long = 125

scala> println(s"Number of lines in the file: $lineCount")
Number of lines in the file: 125
```

```
scala> // Perform the word count

scala> val wcData = txtData.flatMap(line => line.split(" "))
wcData: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at flatMap at <console>:27

scala>                      .map(word => (word, 1))
res6: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[7] at map at <console>:28

scala>                      .reduceByKey(_ + _)
res7: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[8] at reduceByKey at <console>:28
```

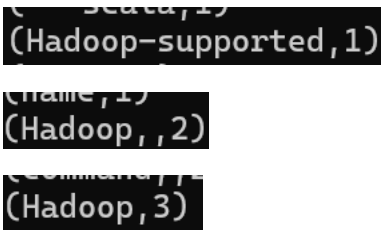
Run the following commands to perform the word count. The count shows up next to each word in the text file:

```
scala> val wcData = txtData.flatMap(l => l.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
wcData: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[11] at reduceByKey at <console>:27

scala> wcData.collect().foreach(println)
(package,1)
(this,1)
(integration,1)
(Python,2)
(cluster.,1)
(its,1)
([run,1)
(There,1)
(general,2)
(have,1)
(pre-built,1)
(Because,1)
(YARN,,1)
(locally,2)
(changed,1)
(locally.,1)
(several,1)
(only,1)
(Configuration,1)
(This,2)
(basic,1)
(first,1)
(learning,,1)
(documentation,3)
(graph,1)
(Hive,2)
(info,1)
(["Specifying,1)
```

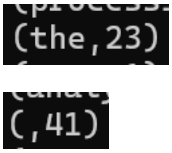
Count the words in the README.md file and answer the set of questions below:

- How many times is the word "Hadoop" counted when the tutorial has printed out all the word counts?



The word ‘Hadoop’ was mentioned in the file six times total after the tutorial printed out all of the word counts.

- Which is the most common word used in the file? How many times does the word occur?



The most common word in the file is the word ‘the’ which appeared 23 times, and the most common symbol in the file was a comma which was evident 41 times.

- Which word occurs the fewest times? How many times does the word occur?

There were nearly 200 words that appeared the fewest in the file, with each word occurring only one time. The code generated came from Open AI.

```
scala> // Filter the words that appear only once

scala> val wordsAppearedOnce = wordCounts.filter { case (word, count) => count == 1 }
wordsAppearedOnce: Array[(String, Int)] = Array((package,1), (this,1), (integration,1), (cluster.,1), (its,1), ([run,1), (There,1), (
have,1), (pre-built,1), (Because,1), (YARN,,1), (changed,1), (locally.,1), (several,1), (only,1), (Configuration,1), (basic,1), (firs
t,1), (learning,,1), (graph,1), (info,1), (["Specifying,1), ("yarn",1), ([params]`.,1), (Downloads](https://static.pepy.tech/personal
ized-badge/pyspark?period=month&units=international_system&left_color=black&right_color=orange&left_text=PyPI%20downloads)](https://p
ypi.org/project/pyspark/),1), ([project,1), (Build](https://github.com/apache/spark/actions/workflows/build_main.yml/badge.svg)](http
s://github.com/apache/spark/actions/workflows/build_main.yml),1), (prefer,1), (version,1), (file,1), (documentation,,1), (MASTER,1),.
..

scala> // Print the words that appear only once

scala> println("Words that appear only once:")
Words that appear only once:

scala> wordsAppearedOnce.foreach { case (word, count) => println(word) }
package
this
integration
cluster.
its
[run
There
have
pre-built
Because
YARN,
changed
locally.
several
only
```

```
Configuration
basic
first
learning,
graph
info
["Specifying
"yarn"
[params]`.
Downloads](https://static.pepy.tech/personalized-badge/pyspark?period=month&units=international_system&left_color=black&right_color=o
range&left_text=PyPI%20downloads)](https://pypi.org/project/pyspark/)
[project
Build](https://github.com/apache/spark/actions/workflows/build_main.yml/badge.svg)](https://github.com/apache/spark/actions/workflows
/build_main.yml)
prefer
version
file
documentation,
MASTER
are
systems.
params
scala>
DataFrames,
provides
configure
R,
when
easiest
Maven](https://maven.apache.org/).
Apache
guide](https://spark.apache.org/contributing.html)
large-scale
```

```
package.
Note
tips,
Alternatively,
>>>
variable
submit
Testing
Streaming
module,
Developer
test,
Version
thread,
rich
them,
stream
GraphX
Guide](https://spark.apache.org/docs/latest/configuration.html)
distribution
review
Build](https://img.shields.io/appveyor/ci/ApacheSoftwareFoundation/spark/master.svg?style=plastic&logo=appveyor)](https://ci.appveyor
.com/project/ApacheSoftwareFoundation/spark)
Thriftserver
API
same
start
built
Spark](#building-spark).
Kubernetes
Contributing
talk
[![GitHub
```

```
Enabling
README
example:
from
N
workloads,
```scala
Hadoop-supported
other
Example
analysis.
runs.
Building
higher-level
need
guide,
Java,
<class>
uses
will
information
IDE,
requires
get
analytics
Documentation
web
using:
cluster
```python
MLlib
contributing
Scala,
```

```
unified
built,
./dev/run-tests
sample
The
Programs
APIs
computation
Try
[Configuration
library
A
through
#
./bin/pyspark
More
storage
Once
["Useful
setup
mesos://
latest
processing,
your
not
different
distributions.
Coverage](https://codecov.io/gh/apache/spark/branch/master/graph/badge.svg)](https://codecov.io/gh/apache/spark)
given.
About
instructions.
Tests
no
```

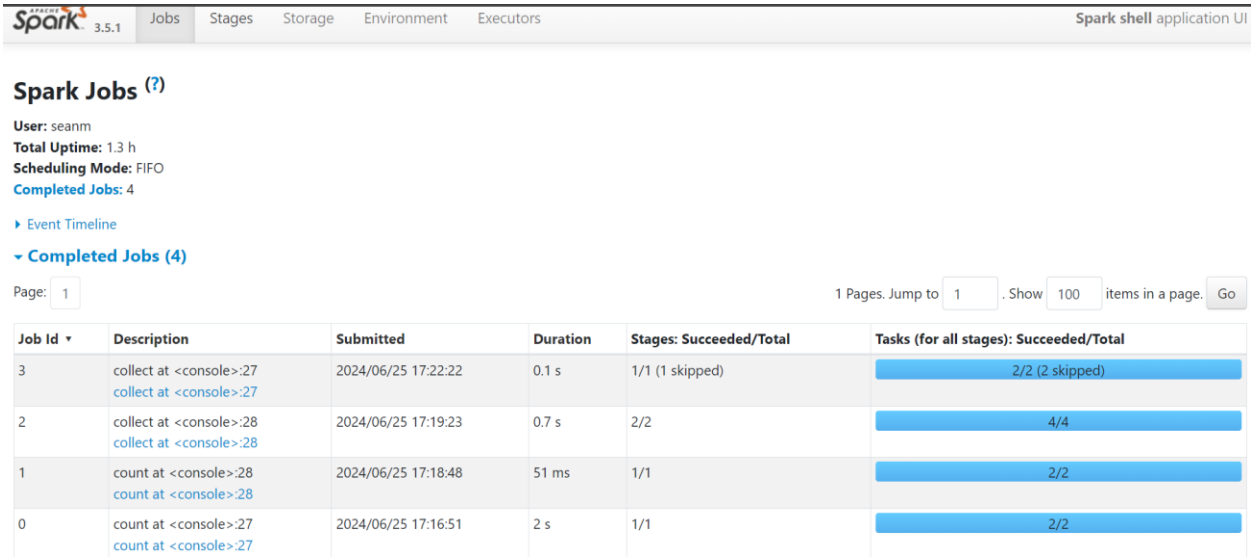
```
project.
programs,
`./bin/run-example
Spark.
./build/mvn
Versions
started
HDFS
individual
spark://
[![PyPI
programming
machine
run:
environment
clean
And
<https://spark.apache.org/>
developing
./bin/spark-shell
URL,
YARN"](https://spark.apache.org/docs/latest/building-spark.html#specifying-the-hadoop-version-and-enabling-yarn)
MASTER=spark://host:7077
threads.
against
[Apache
help
print
"local"
Structured
-DskipTests
optimized
development
```

```
Tools"](https://spark.apache.org/developer-tools.html).
graphs
downloaded
versions
usage
online
abbreviated
comes
directory.
overview
[building
Many
Running
way
Online
[![PySpark
site,
[![AppVeyor
page](https://spark.apache.org/documentation.html).
[Contribution
find
running
contains
project
tests](https://spark.apache.org/developer-tools.html#individual-tests).
Pi
protocols
high-level
Spark"](https://spark.apache.org/docs/latest/building-spark.html).
name
available
(You
core
```

```
instance:
tools
resource-managers/kubernetes/integration-tests/README.md
Actions
"local[N]"
package.)
["Building
must
```

```
scala> val currentDateTime = java.time.LocalDateTime.now()
currentDateTime: java.time.LocalDateTime = 2024-06-25T18:20:12.233
```

Then look at the web console (http://<>:4040/jobs). How many seconds did it take to complete the word count job?:



References

OpenAI. (2024). *ChatGPT (June 25 version)* [Large language model]. Retrieved from https://www.openai.com/

Penchikala, S. (2015, Jan. 30th). *Big Data Processing with Apache Spark – Part 1: Introduction.* InfoQ. https://www.infoq.com/articles/apache-spark-introduction/