BUAN 6346 Big Data Analytics
PHASE 3

SHIVA KUMAR REDDY KOPPULA

Table of Contents

LAB CHAPTER 10: View the SPARK DOCUMENTATION 3

Lab Chapter 10: Explore RDD's using Spark Shell 4

Start the python Spark Shell 4

Start and Use the Scala Spark Shell 6

LAB CHAPTER 10: USE RDDS TO TRANSFORM A DATASET 10

Explore the Loudacre Web Log Files 10

LAB CHAPTER 11: PROCESS DATA FILES WITH SPARK 17

Review the API Documentation for RDD Operations 17

The Data 17

The Task: 18

LAB CHAPTER 12: USE PAIR RDDS TO JOIN TWO DATASETS 21

Explore Web Log Files 21

LAB CHAPTER 13: WRITE AND RUN A SPARK APPLICATION 25

Write a Spark application in Python 25

Submit a Spark Application to the Cluster 26

LAB CHAPTER 13: CONFIGURE A SPARK APPLICATION 29

Set configuration options at the command line 29

LAB CHAPTER 14: VIEW JOBS AND STAGES IN THE SPARK APPLICATION UI 35

Explore Partitioning of File-Based RDDs 35

LAB CHAPTER 15: PERSIST AN RDD 43

Explore the Performance effect of caching an RDD 43

LAB CHAPTER 16: IMPLEMENT AN ITERATIVE ALGORITHM WITH SPARK 46

Review the Data 46

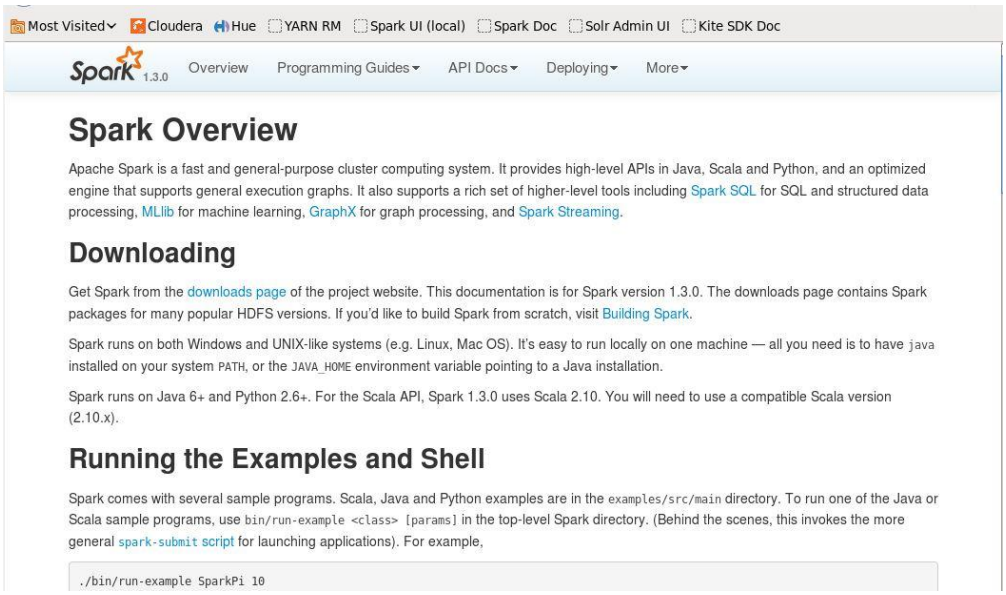
LAB CHAPTER 17: USE SPARK SQL FOR ETL 53

Review the Data in MySQL 53

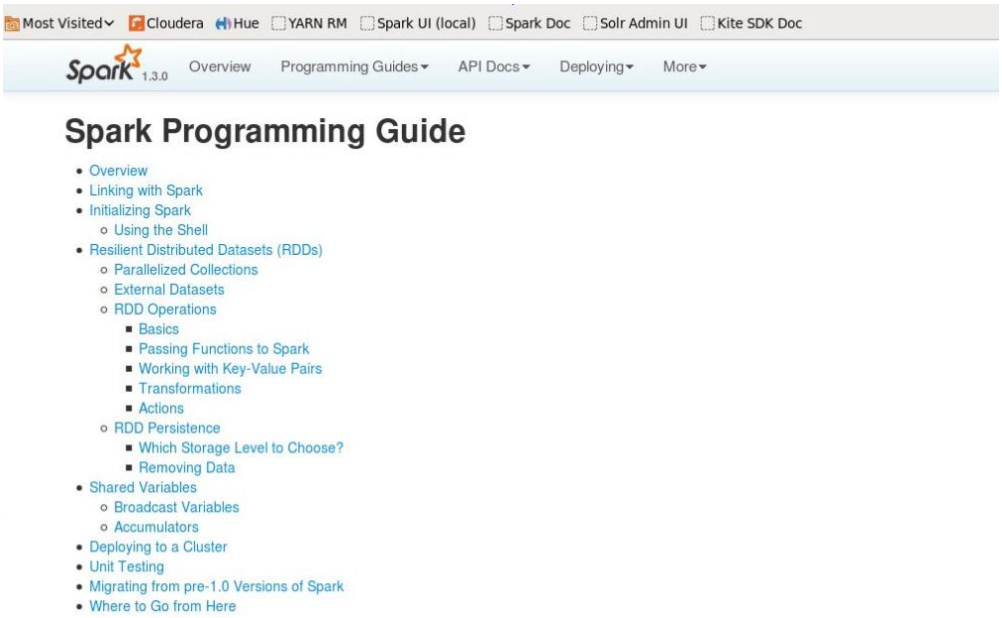
Load the Data from MySQL 53

LAB CHAPTER 10: View the SPARK DOCUMENTATION

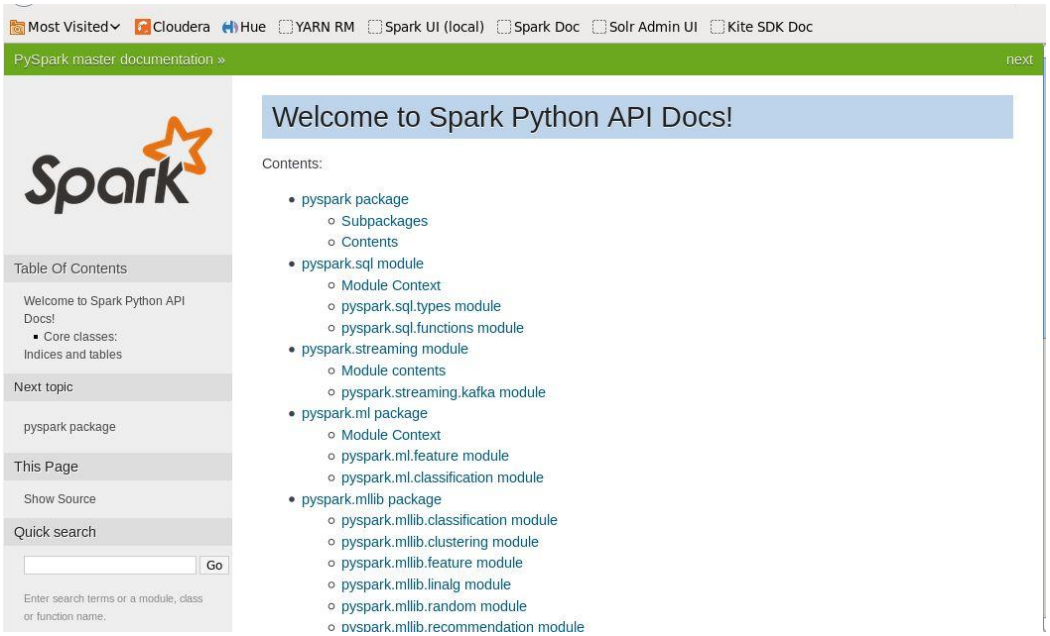
1. We can familiarize ourselves with the Spark Documentation by accessing it through the Firefox browser using the URL: file:/usr/lib/spark/docs/_site/index.html.



2. Click on the Spark Programming Guide to gain further insights into Spark programming.



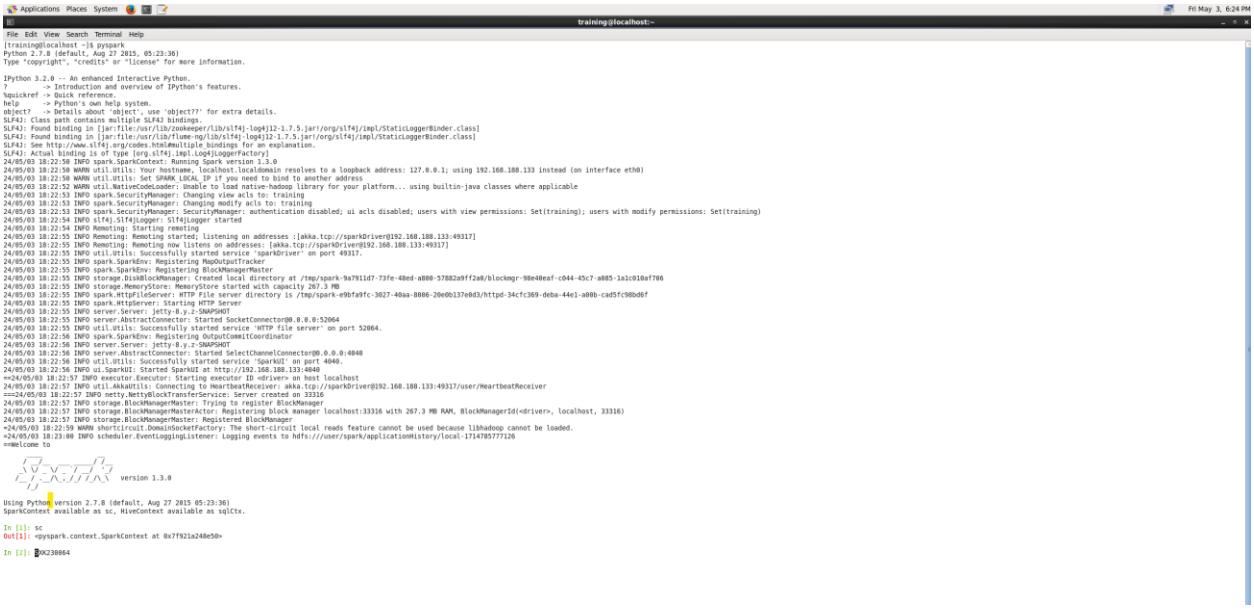
3. After selecting the Python API from the API Docs menu, I bookmarked the page.



Lab Chapter 10: Explore RDD's using Spark Shell

Start the python Spark Shell

1. We will start the PySpark shell in the terminal window with the following Command: $ pyspark.



```
Applications  Places  System                                   Fri May 3, 6:55 PM
training@localhost:~
File  Edit  View  Search  Terminal  Help
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/03 18:52:39 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/03 18:52:39 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.188.133 instead (on interface eth0)
24/05/03 18:52:39 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/03 18:52:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/05/03 18:52:42 INFO spark.SecurityManager: Changing view acls to: training
24/05/03 18:52:42 INFO spark.SecurityManager: Changing modify acls to: training
24/05/03 18:52:42 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify permissions: Set(training)
24/05/03 18:52:43 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/03 18:52:43 INFO Remoting: Starting remoting
24/05/03 18:52:44 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:50272]
24/05/03 18:52:44 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:50272]
24/05/03 18:52:44 INFO util.Utils: Successfully started service 'sparkDriver' on port 50272.
24/05/03 18:52:44 INFO spark.SparkEnv: Registering MapOutputTracker
24/05/03 18:52:44 INFO spark.SparkEnv: Registering BlockManagerMaster
24/05/03 18:52:44 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-d7c90f70-906b-4a32-9b9f-0ab3003c859b/blockmgr-29a269e8-7f56-4f45-ae96-33465ebd0982
24/05/03 18:52:44 INFO storage.MemoryStore: MemoryStore started with capacity 267.3 MB
24/05/03 18:52:44 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-555a7a54-a0f6-4c72-bec4-ca73bcae8a80/httpd-efe528c2-20ae-470f-bee1-aecb67c5a174
24/05/03 18:52:44 INFO spark.HttpServer: Starting HTTP Server
24/05/03 18:52:45 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:52:45 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:43955
24/05/03 18:52:45 INFO util.Utils: Successfully started service 'HTTP file server' on port 43955.
24/05/03 18:52:45 INFO spark.SparkEnv: Registering OutputCommitCoordinator
24/05/03 18:52:45 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:52:45 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
24/05/03 18:52:45 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
24/05/03 18:52:45 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4040
24/05/03 18:52:46 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/03 18:52:46 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:50272/user/HeartbeatReceiver
24/05/03 18:52:47 INFO netty.NettyBlockTransferService: Server created on 42444
24/05/03 18:52:47 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/03 18:52:47 INFO storage.BlockManagerMasterActor: Registering block manager localhost:42444 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 42444)
24/05/03 18:52:47 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/03 18:52:49 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/03 18:52:50 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714787566403
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: sc
Out[1]: <pyspark.context.SparkContext at 0x7f2a24f39e50>

In [2]: sc.
sc.PACKAGE_EXTENSIONS    sc.binaryRecords         sc.defaultParallelism    sc.master                sc.pythonExec            sc.setLocalProperty      sc.stop
sc.accumulator           sc.broadcast             sc.dump_profiles         sc.newAPIHadoopFile      sc.runJob                sc.setSystemProperty     sc.textFile
sc.addFile               sc.cancelAllJobs         sc.environment           sc.newAPIHadoopRDD       sc.sequenceFile          sc.show_profiles         sc.union
sc.addPyFile             sc.cancelJobGroup        sc.getLocalProperty      sc.parallelize           sc.serializer            sc.sparkHome             sc.version
sc.appName               sc.clearFiles            sc.hadoopFile            sc.pickleFile            sc.setCheckpointDir      sc.sparkUser             sc.wholeTextFiles
sc.binaryFiles           sc.defaultMinPartitions  sc.hadoopRDD             sc.profiler_collector    sc.setJobGroup           sc.statusTracker

In [2]: sc.sxk230064
training@localhost:~
```

2. To ensure that the Spark has created a SparkContext object called 'sc', use the command 'sc.' Then press the [TAB] key to view all available SparkContext methods. Exit by pressing Ctrl-D.

```
Applications  Places  System                                   Fri May 3, 6:56 PM
training@localhost:~
File  Edit  View  Search  Terminal  Help
In [2]: sc.
sc.PACKAGE_EXTENSIONS    sc.binaryRecords         sc.defaultParallelism    sc.master                sc.pythonExec            sc.setLocalProperty      sc.stop
sc.accumulator           sc.broadcast             sc.dump_profiles         sc.newAPIHadoopFile      sc.runJob                sc.setSystemProperty     sc.textFile
sc.addFile               sc.cancelAllJobs         sc.environment           sc.newAPIHadoopRDD       sc.sequenceFile          sc.show_profiles         sc.union
sc.addPyFile             sc.cancelJobGroup        sc.getLocalProperty      sc.parallelize           sc.serializer            sc.sparkHome             sc.version
sc.appName               sc.clearFiles            sc.hadoopFile            sc.pickleFile            sc.setCheckpointDir      sc.sparkUser             sc.wholeTextFiles
sc.binaryFiles           sc.defaultMinPartitions  sc.hadoopRDD             sc.profiler_collector    sc.setJobGroup           sc.statusTracker

In [2]:
Do you really want to exit ([y]/n)? y
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/metrics/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/03 18:55:53 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/03 18:55:53 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.188.133:4040
24/05/03 18:55:53 INFO scheduler.DAGScheduler: Stopping DAGScheduler
24/05/03 18:55:53 INFO spark.MapOutputTrackerMasterActor: MapOutputTrackerActor stopped!
24/05/03 18:55:53 INFO storage.MemoryStore: MemoryStore cleared
24/05/03 18:55:53 INFO storage.BlockManager: BlockManager stopped
24/05/03 18:55:53 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/03 18:55:53 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorActor: OutputCommitCoordinator stopped!
24/05/03 18:55:53 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/03 18:55:53 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
24/05/03 18:55:53 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/03 18:55:53 INFO Remoting: Remoting shut down
24/05/03 18:55:53 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost ~]$ sxk230064
```

Start and Use the Scala Spark Shell

3. I started the Scala Spark Shell in a terminal window using the command below in the snip.

```
[training@localhost ~]$ clear

[training@localhost ~]$ spark-shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/03 18:58:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/05/03 18:58:24 INFO spark.SecurityManager: Changing view acls to: training
24/05/03 18:58:24 INFO spark.SecurityManager: Changing modify acls to: training
24/05/03 18:58:24 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify permissions:
Set(training)
24/05/03 18:58:24 INFO spark.HttpServer: Starting HTTP Server
24/05/03 18:58:25 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:58:25 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:44591
24/05/03 18:58:25 INFO util.Utils: Successfully started service 'HTTP class server' on port 44591.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type :help for more information.
24/05/03 18:58:42 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.188.133 instead (on interface eth0)
24/05/03 18:58:42 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/03 18:58:42 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/03 18:58:42 INFO spark.SecurityManager: Changing view acls to: training
24/05/03 18:58:42 INFO spark.SecurityManager: Changing modify acls to: training
24/05/03 18:58:42 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify permissions:
Set(training)
24/05/03 18:58:43 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/03 18:58:44 INFO Remoting: Starting remoting
24/05/03 18:58:44 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:38341]
24/05/03 18:58:44 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:38341]
24/05/03 18:58:44 INFO util.Utils: Successfully started service 'sparkDriver' on port 38341.
24/05/03 18:58:44 INFO spark.SparkEnv: Registering MapOutputTracker
```

```
Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type :help for more information.
24/05/03 18:58:42 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.188.133 instead (on interface eth0)
24/05/03 18:58:42 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/03 18:58:42 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/03 18:58:42 INFO spark.SecurityManager: Changing view acls to: training
24/05/03 18:58:42 INFO spark.SecurityManager: Changing modify acls to: training
24/05/03 18:58:42 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify permissions:
Set(training)
24/05/03 18:58:43 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/03 18:58:44 INFO Remoting: Starting remoting
24/05/03 18:58:44 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:38341]
24/05/03 18:58:44 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:38341]
24/05/03 18:58:44 INFO util.Utils: Successfully started service 'sparkDriver' on port 38341.
24/05/03 18:58:44 INFO spark.SparkEnv: Registering MapOutputTracker
24/05/03 18:58:44 INFO spark.SparkEnv: Registering BlockManagerMaster
24/05/03 18:58:44 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-dc46d815-e0b0-4eb8-8e4d-b195dcea108c/blockmgr-d581030f-e09f-41d4-9a25-ab63ea8aed76
24/05/03 18:58:44 INFO storage.MemoryStore: MemoryStore started with capacity 267.3 MB
24/05/03 18:58:44 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-d2b4db76-1c56-4fcb-8b4d-78b79ee2ed68/httpd-c3ad0729-ac68-4522-a16d-854c52777666
24/05/03 18:58:44 INFO spark.HttpServer: Starting HTTP Server
24/05/03 18:58:44 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:58:44 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:45952
24/05/03 18:58:44 INFO util.Utils: Successfully started service 'HTTP file server' on port 45952.
24/05/03 18:58:44 INFO spark.SparkEnv: Registering OutputCommitCoordinator
24/05/03 18:58:45 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:58:45 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
24/05/03 18:58:45 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
24/05/03 18:58:45 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4040
24/05/03 18:58:46 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/03 18:58:46 INFO executor.Executor: Using REPL class URI: http://192.168.188.133:44591
24/05/03 18:58:46 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:38341/user/HeartbeatReceiver
24/05/03 18:58:46 INFO netty.NettyBlockTransferService: Server created on 60677
24/05/03 18:58:46 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/03 18:58:46 INFO storage.BlockManagerMasterActor: Registering block manager localhost:60677 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 60677)
24/05/03 18:58:46 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/03 18:58:49 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/03 18:58:50 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714787925632
24/05/03 18:58:50 INFO repl.SparkILoop: Created spark context..
Spark context available as sc.
24/05/03 18:58:52 INFO repl.SparkILoop: Created sql context (with Hive support)..
SQL context available as sqlContext.

scala> sc
res0: org.apache.spark.SparkContext = org.apache.spark.SparkContext@940dad1

scala> SXK230064
```

4. To verify the existence of the SparkContext object, I typed the command 'sc.' To view all available SparkContext methods, type 'sc.' followed by a dot and then the [TAB] key.

```
24/05/03 18:58:44 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:58:44 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:45952
24/05/03 18:58:44 INFO util.Utils: Successfully started service 'HTTP file server' on port 45952.
24/05/03 18:58:44 INFO spark.SparkEnv: Registering OutputCommitCoordinator
24/05/03 18:58:45 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/03 18:58:45 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
24/05/03 18:58:45 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
24/05/03 18:58:45 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4040
24/05/03 18:58:46 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/03 18:58:46 INFO executor.Executor: Using REPL class URI: http://192.168.188.133:44591
24/05/03 18:58:46 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:38341/user/HeartbeatReceiver
24/05/03 18:58:46 INFO netty.NettyBlockTransferService: Server created on 60677
24/05/03 18:58:46 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/03 18:58:46 INFO storage.BlockManagerMasterActor: Registering block manager localhost:60677 with 267.3 MB RAM, BlockManagerId(<driver>, localho
24/05/03 18:58:46 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/03 18:58:49 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/03 18:58:50 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714787925632
24/05/03 18:58:50 INFO repl.SparkILoop: Created spark context..
Spark context available as sc.
24/05/03 18:58:52 INFO repl.SparkILoop: Created sql context (with Hive support)..
SQL context available as sqlContext.

scala> sc
res0: org.apache.spark.SparkContext = org.apache.spark.SparkContext@940dad1

scala> sc.
accumulable                accumulableCollection      accumulator                addFile                    addJar                     addSparkListen
appName                    applicationId              asInstanceOf               binaryFiles                binaryRecords              broadcast
cancelAllJobs              cancelJobGroup             clearCallSite              clearFiles                 clearJars                  clearJobGroup
defaultMinPartitions       defaultMinSplits           defaultParallelism         emptyRDD                   files                      getAllPools
getCheckpointDir           getConf                    getExecutorMemoryStatus    getExecutorStorageStatus   getLocalProperty           getPersistentR
getPoolForName             getRDDStorageInfo          getSchedulingMode          hadoopConfiguration        hadoopFile                 hadoopRDD
initLocalProperties        isInstanceOf               isLocal                    jars                       killExecutor               killExecutors
makeRDD                    master                     metricsSystem              newAPIHadoopFile           newAPIHadoopRDD            objectFile
parallelize                requestExecutors           runApproximateJob          runJob                     sequenceFile               setCallSite
setCheckpointDir           setJobDescription          setJobGroup                setLocalProperty           sparkUser                  startTime
statusTracker              stop                       submitJob                  tachyonFolderName          textFile                   toString
union                      version                    wholeTextFiles

scala> sc.sxk230064
```

5. I opened a text file in a text editor in a separate window using the path below in VM.

```
Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I--
I took the one less traveled by,
And that has made all the difference.
```

6. In the PySpark terminal, we defined an RDD by reading the test file from the local system using the command: $ mydata = sc.textFile("file:/home/training/training_materials/data/frostroad.txt").

```
24/05/03 19:28:41 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
24/05/03 19:28:41 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
24/05/03 19:28:41 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4040
24/05/03 19:28:42 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/03 19:28:42 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:55638/user/Heart
24/05/03 19:28:42 INFO netty.NettyBlockTransferService: Server created on 50895
24/05/03 19:28:42 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/03 19:28:42 INFO storage.BlockManagerMasterActor: Registering block manager localhost:50895 with 267.3 MB RAM, BlockManag
st, 50895)
24/05/03 19:28:42 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/03 19:28:45 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop
24/05/03 19:28:45 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714789722
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: mydata=sc.textFile(\
   ...: "file:/home/training/training_materials/\
   ...: data/frostroad.txt")
24/05/03 19:29:57 INFO storage.MemoryStore: ensureFreeSpace(280171) called with curMem=0, maxMem=280248975
24/05/03 19:29:57 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 273.6 KB, free 267.0 M
24/05/03 19:29:57 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=280171, maxMem=280248975
24/05/03 19:29:57 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 20.7 KB, free 26
24/05/03 19:29:57 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:50895 (size: 20.7 KB, free: 26
24/05/03 19:29:57 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/03 19:29:57 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:-2

In [2]: sxk230064
```

7. Spark has not read the file yet; this happens only when you perform an operation on the RDD. Let's start by counting the number of rows in the dataset using the command: mydata.count(). The output shows 23, which includes the empty lines in the dataset.

```
24/05/03 19:32:51 INFO scheduler.DAGScheduler: Final stage: Stage 0(count at <ipython-input-2-b5419efaf7a8>:1)
24/05/03 19:32:51 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 19:32:51 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 19:32:51 INFO scheduler.DAGScheduler: Submitting Stage 0 (PythonRDD[2] at count at <ipython-input-2-b5419efaf7a8>:1),
which has no missing parents
24/05/03 19:32:51 INFO storage.MemoryStore: ensureFreeSpace(5480) called with curMem=301375, maxMem=280248975
24/05/03 19:32:51 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 5.4 KB, free 267.0 MB)
24/05/03 19:32:51 INFO storage.MemoryStore: ensureFreeSpace(3454) called with curMem=306855, maxMem=280248975
24/05/03 19:32:51 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 3.4 KB, free 267
.0 MB)
24/05/03 19:32:51 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:50895 (size: 3.4 KB, free: 267
.2 MB)
24/05/03 19:32:51 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/03 19:32:51 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:839
24/05/03 19:32:51 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 0 (PythonRDD[2] at count at <ipython-input
-2-b5419efaf7a8>:1)
24/05/03 19:32:51 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
24/05/03 19:32:51 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1321 bytes)
24/05/03 19:32:51 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/03 19:32:52 INFO rdd.HadoopRDD: Input split: file:/home/training/training_materials/data/frostroad.txt:0+730
24/05/03 19:32:52 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/05/03 19:32:52 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/05/03 19:32:52 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/05/03 19:32:52 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/05/03 19:32:52 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/05/03 19:32:52 INFO python.PythonRDD: Times: total = 722, boot = 600, init = 95, finish = 27
24/05/03 19:32:52 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 1828 bytes result sent to driver
24/05/03 19:32:52 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1187 ms on localhost (1/1)
24/05/03 19:32:52 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
24/05/03 19:32:52 INFO scheduler.DAGScheduler: Stage 0 (count at <ipython-input-2-b5419efaf7a8>:1) finished in 1.285 s
24/05/03 19:32:52 INFO scheduler.DAGScheduler: Job 0 finished: count at <ipython-input-2-b5419efaf7a8>:1, took 1.590437 s
Out[2]: 23

In [3]: sxk230064
```

8. We will now execute the collect operation to display the data in the RDD using the command mydata.collect().

```
24/05/03 19:32:52 INFO scheduler.DAGScheduler: Job 0 finished: count at <ipython-input-2-b5419efaf7a8>:1, took 1.590437 s
Out[2]: 23

In [3]: mydata.collect()
24/05/03 19:36:46 INFO spark.SparkContext: Starting job: collect at <ipython-input-3-0a9f21ece71b>:1
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Got job 1 (collect at <ipython-input-3-0a9f21ece71b>:1) with 1 output partitions
 (allowLocal=false)
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Final stage: Stage 1(collect at <ipython-input-3-0a9f21ece71b>:1)
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Submitting Stage 1 (file:/home/training/training_materials/data/frostroad.txt Ma
pPartitionsRDD[1] at textFile at NativeMethodAccessorImpl.java:-2), which has no missing parents
24/05/03 19:36:46 INFO storage.MemoryStore: ensureFreeSpace(2696) called with curMem=310309, maxMem=280248975
24/05/03 19:36:46 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 2.6 KB, free 267.0 MB)
24/05/03 19:36:46 INFO storage.MemoryStore: ensureFreeSpace(1659) called with curMem=313005, maxMem=280248975
24/05/03 19:36:46 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 1659.0 B, free 2
67.0 MB)
24/05/03 19:36:46 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:50895 (size: 1659.0 B, free: 2
67.2 MB)
24/05/03 19:36:46 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/03 19:36:46 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:839
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (file:/home/training/training_materials/
data/frostroad.txt MapPartitionsRDD[1] at textFile at NativeMethodAccessorImpl.java:-2)
24/05/03 19:36:46 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
24/05/03 19:36:46 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS_LOCAL, 1321 bytes)
24/05/03 19:36:46 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
24/05/03 19:36:46 INFO rdd.HadoopRDD: Input split: file:/home/training/training_materials/data/frostroad.txt:0+730
24/05/03 19:36:46 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 2540 bytes result sent to driver
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Stage 1 (collect at <ipython-input-3-0a9f21ece71b>:1) finished in 0.111 s
24/05/03 19:36:46 INFO scheduler.DAGScheduler: Job 1 finished: collect at <ipython-input-3-0a9f21ece71b>:1, took 0.215436 s
24/05/03 19:36:46 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 84 ms on localhost (1/1)
24/05/03 19:36:46 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
Out[3]:
[u'Two roads diverged in a yellow wood,',
 u'And sorry I could not travel both',
 u'And be one traveler, long I stood',
 u'And looked down one as far as I could',
 u'To where it bent in the undergrowth;',
 u'',
 u'Then took the other, as just as fair,',
 u'And having perhaps the better claim,',
 u'Because it was grassy and wanted wear;',
 u'Though as for that the passing there',
 u'Had worn them really about the same,',
 u'',
 u'And both that morning equally lay',
 u'In leaves no step had trodden black.',
 u'Oh, I kept the first for another day!',
 u'Yet knowing how way leads on to way,',
 u'I doubted if I should ever come back.',
 u'',
 u'I shall be telling this with a sigh',
 u'Somewhere ages and ages hence:',
 u'Two roads diverged in a wood, and I--',
 u'I took the one less traveled by,',
 u'And that has made all the difference.']

In [4]: sxk230064
```

9. Using command completion, I can view all available transformations and operations you can perform on an RDD. We used the command 'mydata. + [TAB] key.

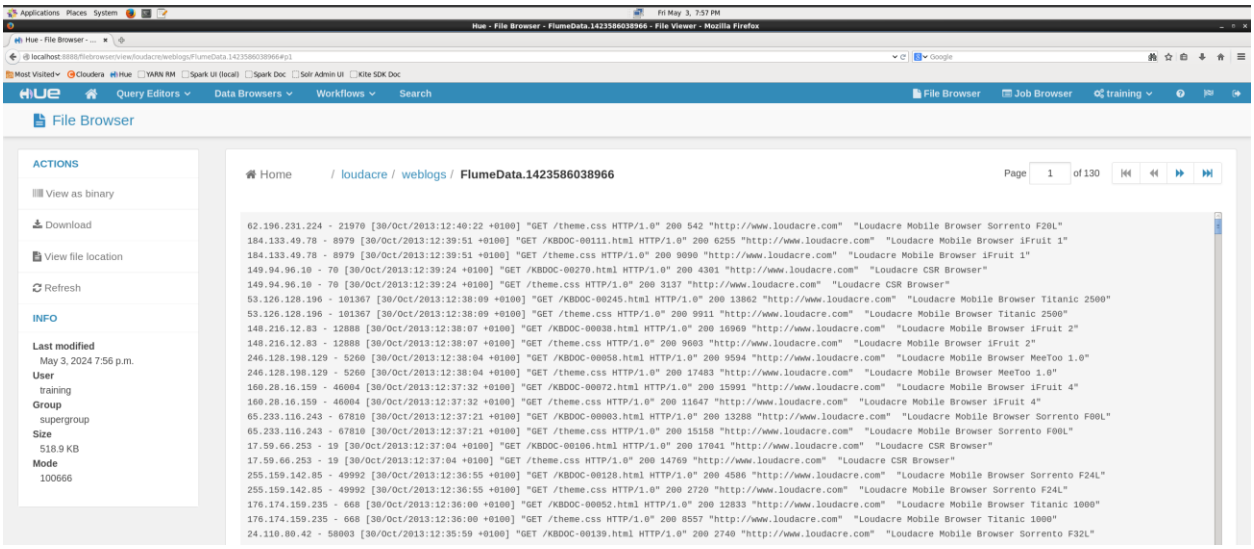
```
SyntaxError: invalid syntax


In [5]: mydata.
Display all 103 possibilities? (y or n)
mydata.aggregate                        mydata.fullOuterJoin                    mydata.mean
mydata.aggregateByKey                   mydata.getCheckpointFile                mydata.meanApprox
mydata.cache                            mydata.getNumPartitions                 mydata.min
mydata.cartesian                        mydata.getStorageLevel                  mydata.name
mydata.checkpoint                       mydata.glom                             mydata.partitionBy
mydata.coalesce                         mydata.groupBy                          mydata.partitioner
mydata.cogroup                          mydata.groupByKey                       mydata.persist
mydata.collect                          mydata.groupWith                        mydata.pipe
mydata.collectAsMap                     mydata.histogram                        mydata.randomSplit
mydata.combineByKey                     mydata.id                               mydata.reduce
mydata.context                          mydata.intersection                     mydata.reduceByKey
mydata.count                            mydata.isCheckpointed                   mydata.reduceByKeyLocally
mydata.countApprox                      mydata.isEmpty                          mydata.repartition
mydata.countApproxDistinct              mydata.is_cached                        mydata.repartitionAndSortWithinPartitions
mydata.countByKey                       mydata.is_checkpointed                  mydata.rightOuterJoin
mydata.countByValue                     mydata.join                             mydata.sample
mydata.ctx                              mydata.keyBy                            mydata.sampleByKey
mydata.distinct                         mydata.keys                             mydata.sampleStdev
mydata.filter                           mydata.leftOuterJoin                    mydata.sampleVariance
mydata.first                            mydata.lookup                           mydata.saveAsHadoopDataset
mydata.flatMap                          mydata.map                              mydata.saveAsHadoopFile
mydata.flatMapValues                    mydata.mapPartitions                    mydata.saveAsNewAPIHadoopDataset
mydata.fold                             mydata.mapPartitionsWithIndex           mydata.saveAsNewAPIHadoopFile
mydata.foldByKey                        mydata.mapPartitionsWithSplit           mydata.saveAsPickleFile
mydata.foreach                          mydata.mapValues                        mydata.saveAsSequenceFile
mydata.foreachPartition                 mydata.max                              mydata.saveAsTextFile

In [5]: mydata.sxk230064
```

LAB CHAPTER 10: USE RDDS TO TRANSFORM A DATASET

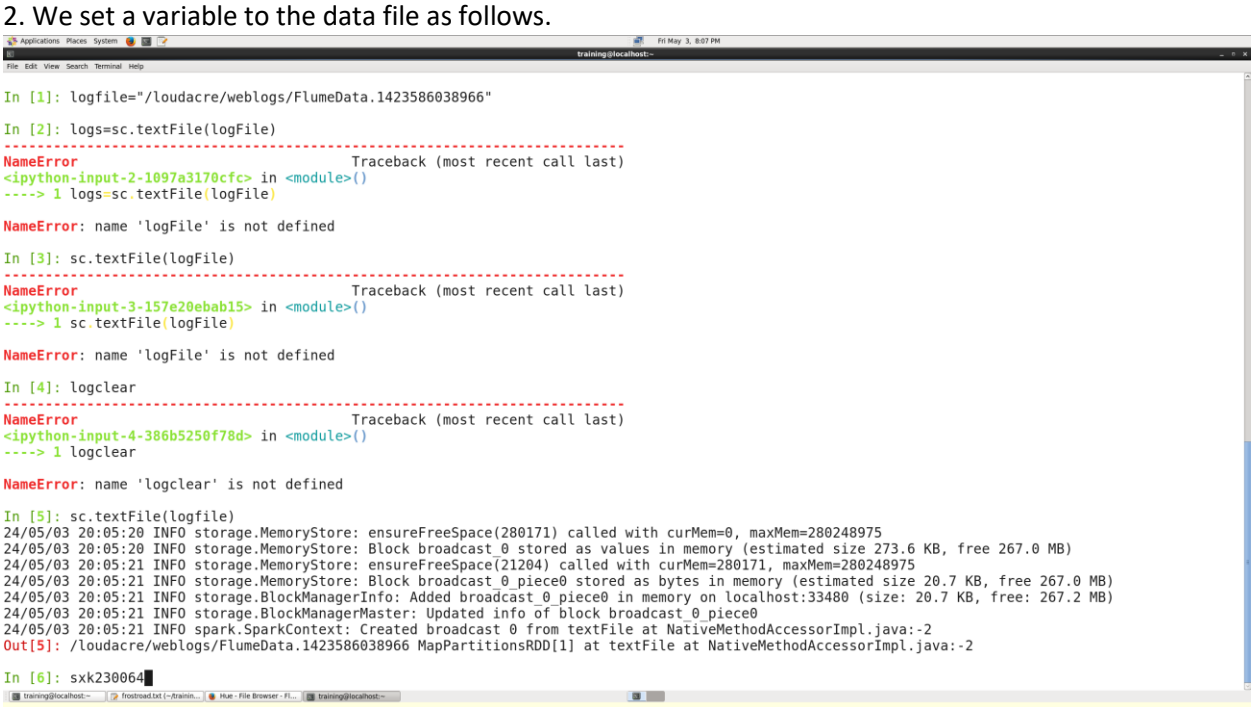
Explore the Loudacre Web Log Files

1. From the Hue Browser, I reviewed loudacre/weblogs/FlumeData.1424275921226.

- IP Address: 119.182.134.196
- User Id: 46
- Request: "GET /KBDOC-00186.html HTTP/1.0



2. We set a variable to the data file as follows.



```
In [1]: logfile="/loudacre/weblogs/FlumeData.1423586038966"

In [2]: logs=sc.textFile(logFile)
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-1097a3170cfc> in <module>()
----> 1 logs=sc.textFile(logFile)

NameError: name 'logFile' is not defined

In [3]: sc.textFile(logFile)
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-3-157e20ebab15> in <module>()
----> 1 sc.textFile(logFile)

NameError: name 'logFile' is not defined

In [4]: logclear
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-4-386b5250f78d> in <module>()
----> 1 logclear

NameError: name 'logclear' is not defined

In [5]: sc.textFile(logfile)
24/05/03 20:05:20 INFO storage.MemoryStore: ensureFreeSpace(280171) called with curMem=0, maxMem=280248975
24/05/03 20:05:20 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 273.6 KB, free 267.0 MB)
24/05/03 20:05:21 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=280171, maxMem=280248975
24/05/03 20:05:21 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 20.7 KB, free 267.0 MB)
24/05/03 20:05:21 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:33480 (size: 20.7 KB, free: 267.2 MB)
24/05/03 20:05:21 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/03 20:05:21 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:-2
Out[5]: /loudacre/weblogs/FlumeData.1423586038966 MapPartitionsRDD[1] at textFile at NativeMethodAccessorImpl.java:-2

In [6]: sxk230064
```

3. I have created an RDD from the data file.

```
In [8]:

In [8]: logs=sc.textFile(logfile)
24/05/03 20:13:51 INFO storage.MemoryStore: ensureFreeSpace(280219) called with curMem=301375, maxMem=280248975
24/05/03 20:13:51 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 273.7 KB, free 266.7 MB)
24/05/03 20:13:51 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=581594, maxMem=280248975
24/05/03 20:13:51 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.7 MB)
24/05/03 20:13:51 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:33480 (size: 20.7 KB, free: 267.2 MB)
24/05/03 20:13:51 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/03 20:13:51 INFO spark.SparkContext: Created broadcast 1 from textFile at NativeMethodAccessorImpl.java:-2

In [9]: sxk230064
```

4. I then created an RDD containing only those lines that are requests for JPG files. We can view the first 10 lines of this data using the command jpglogs.take(10).

```
In [9]: jpglogs=\
   ...: logs.filter(lambda line: ".jpg" in line)

In [10]: jpglogs.take(10)
24/05/03 20:17:53 INFO mapred.FileInputFormat: Total input paths to process : 1
24/05/03 20:17:54 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/03 20:17:54 INFO scheduler.DAGScheduler: Got job 0 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/03 20:17:54 INFO scheduler.DAGScheduler: Final stage: Stage 0(runJob at PythonRDD.scala:356)
24/05/03 20:17:54 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:17:54 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:17:54 INFO scheduler.DAGScheduler: Submitting Stage 0 (PythonRDD[5] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/03 20:17:54 INFO storage.MemoryStore: ensureFreeSpace(5248) called with curMem=602798, maxMem=280248975
24/05/03 20:17:54 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 5.1 KB, free 266.7 MB)
24/05/03 20:17:54 INFO storage.MemoryStore: ensureFreeSpace(3354) called with curMem=608046, maxMem=280248975
24/05/03 20:17:54 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 3.3 KB, free 266.7 MB)
24/05/03 20:17:54 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:33480 (size: 3.3 KB, free: 267.2 MB)
24/05/03 20:17:54 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/03 20:17:54 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:839
24/05/03 20:17:54 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 0 (PythonRDD[5] at RDD at PythonRDD.scala:42)
24/05/03 20:17:54 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
24/05/03 20:17:54 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:17:54 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/03 20:17:55 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/05/03 20:17:55 INFO python.PythonRDD: Times: total = 1259, boot = 1007, init = 194, finish = 58
24/05/03 20:17:56 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 3516 bytes result sent to driver
24/05/03 20:17:56 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1688 ms on localhost (1/1)
24/05/03 20:17:56 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
24/05/03 20:17:56 INFO scheduler.DAGScheduler: Stage 0 (runJob at PythonRDD.scala:356) finished in 1.767 s
24/05/03 20:17:56 INFO scheduler.DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:356, took 2.190781 s
Out[10]:
[u'136.76.29.92 - 104751 [30/Oct/2013:12:31:19 +0100] "GET /sorrento_f21l.jpg HTTP/1.0" 200 11049 "http://www.loudacre.com"  "Loudacre Mobile Browser
 Titanic 2000"',
```

```
24/05/03 20:17:54 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
24/05/03 20:17:54 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:17:54 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/03 20:17:55 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/05/03 20:17:55 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/05/03 20:17:55 INFO python.PythonRDD: Times: total = 1259, boot = 1007, init = 194, finish = 58
24/05/03 20:17:56 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 3516 bytes result sent to driver
24/05/03 20:17:56 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1688 ms on localhost (1/1)
24/05/03 20:17:56 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
24/05/03 20:17:56 INFO scheduler.DAGScheduler: Stage 0 (runJob at PythonRDD.scala:356) finished in 1.767 s
24/05/03 20:17:56 INFO scheduler.DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:356, took 2.190781 s
Out[10]:
[u'136.76.29.92 - 104751 [30/Oct/2013:12:31:19 +0100] "GET /sorrento_f21l.jpg HTTP/1.0" 200 11049 "http://www.loudacre.com"  "Loudacre Mobile Browser
 Titanic 2000"',
 u'24.168.246.104 - 53083 [30/Oct/2013:12:30:45 +0100] "GET /map.jpg HTTP/1.0" 200 8046 "http://www.loudacre.com"  "Loudacre Mobile Browser Sorrento
F10L"',
 u'249.196.186.161 - 28033 [30/Oct/2013:12:12:41 +0100] "GET /sorrento_f22l.jpg HTTP/1.0" 200 7864 "http://www.loudacre.com"  "Loudacre Mobile Browse
r Sorrento F20L"',
 u'29.178.162.189 - 40516 [30/Oct/2013:12:11:55 +0100] "GET /meetoo_5.0.jpg HTTP/1.0" 200 2833 "http://www.loudacre.com"  "Loudacre Mobile Browser So
rrento F11L"',
 u'252.17.152.64 - 31147 [30/Oct/2013:12:04:29 +0100] "GET /ifruit_5.jpg HTTP/1.0" 200 6517 "http://www.loudacre.com"  "Loudacre Mobile Browser Sorre
nto F21L"',
 u'241.129.9.248 - 47821 [30/Oct/2013:11:55:51 +0100] "GET /meetoo_3.1.jpg HTTP/1.0" 200 11501 "http://www.loudacre.com"  "Loudacre Mobile Browser So
rrento F10L"',
 u'73.31.66.132 - 40231 [30/Oct/2013:11:53:09 +0100] "GET /sorrento_f21l.jpg HTTP/1.0" 200 4698 "http://www.loudacre.com"  "Loudacre Mobile Browser R
onin Novelty Note 2"',
 u'110.144.41.164 - 18769 [30/Oct/2013:11:36:01 +0100] "GET /ronin_s3.jpg HTTP/1.0" 200 6980 "http://www.loudacre.com"  "Loudacre Mobile Browser Sorr
ento F11L"',
 u'98.125.208.32 - 100245 [30/Oct/2013:11:30:09 +0100] "GET /ifruit_4a.jpg HTTP/1.0" 200 8130 "http://www.loudacre.com"  "Loudacre Mobile Browser Mee
Too 3.0"',
 u'146.151.22.18 - 3439 [30/Oct/2013:11:22:54 +0100] "GET /titanic_2100.jpg HTTP/1.0" 200 10157 "http://www.loudacre.com"  "Loudacre Mobile Browser T
itanic 1000"']

In [11]: sxb230064
```

5. Every time we don't need to store intermediate objects in a variable, we can combine steps into a single line of code. For example, to count the number of JPG requests, we can execute this in a single command.

```
In [16]: sc.textFile(logfile).filter(lambda line: \
   ....: ".jpg" in line).count()
24/05/03 20:27:26 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=611400, maxMem=280248975
24/05/03 20:27:26 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
24/05/03 20:27:26 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=891643, maxMem=280248975
24/05/03 20:27:26 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
24/05/03 20:27:26 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:33480 (size: 20.7 KB, free: 267.2 MB)
24/05/03 20:27:26 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
24/05/03 20:27:26 INFO spark.SparkContext: Created broadcast 3 from textFile at NativeMethodAccessorImpl.java:-2
24/05/03 20:27:26 INFO mapred.FileInputFormat: Total input paths to process : 1
24/05/03 20:27:26 INFO spark.SparkContext: Starting job: count at <ipython-input-16-a079c423cba4>:1
24/05/03 20:27:26 INFO scheduler.DAGScheduler: Got job 1 (count at <ipython-input-16-a079c423cba4>:1) with 1 output partitions (allowLocal=false)
24/05/03 20:27:26 INFO scheduler.DAGScheduler: Final stage: Stage 1(count at <ipython-input-16-a079c423cba4>:1)
24/05/03 20:27:26 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:27:26 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:27:26 INFO scheduler.DAGScheduler: Submitting Stage 1 (PythonRDD[8] at count at <ipython-input-16-a079c423cba4>:1), which has no missing
parents
24/05/03 20:27:26 INFO storage.MemoryStore: ensureFreeSpace(5848) called with curMem=912847, maxMem=280248975
24/05/03 20:27:26 INFO storage.MemoryStore: Block broadcast_4 stored as values in memory (estimated size 5.7 KB, free 266.4 MB)
24/05/03 20:27:27 INFO storage.MemoryStore: ensureFreeSpace(3739) called with curMem=918695, maxMem=280248975
24/05/03 20:27:27 INFO storage.MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 3.7 KB, free 266.4 MB)
24/05/03 20:27:27 INFO storage.BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:33480 (size: 3.7 KB, free: 267.2 MB)
24/05/03 20:27:27 INFO storage.BlockManagerMaster: Updated info of block broadcast_4_piece0
24/05/03 20:27:27 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:839
24/05/03 20:27:27 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (PythonRDD[8] at count at <ipython-input-16-a079c423cba4>:1)
24/05/03 20:27:27 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
24/05/03 20:27:27 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:27:27 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
24/05/03 20:27:27 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:27:27 INFO python.PythonRDD: Times: total = 253, boot = 13, init = 64, finish = 176
24/05/03 20:27:27 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 1883 bytes result sent to driver
24/05/03 20:27:27 INFO scheduler.DAGScheduler: Stage 1 (count at <ipython-input-16-a079c423cba4>:1) finished in 0.378 s
24/05/03 20:27:27 INFO scheduler.DAGScheduler: Job 1 finished: count at <ipython-input-16-a079c423cba4>:1, took 0.484367 s
24/05/03 20:27:27 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 341 ms on localhost (1/1)
24/05/03 20:27:27 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
Out[16]: 218

In [17]: SXK230064
```

6. Now we will use the map function to define a new RDD. We will start with a simple map that returns the length of each line in the log file. This prints out an array of five integers corresponding to the first five lines in the file.

```
24/05/03 20:27:27 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
24/05/03 20:27:27 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:27:27 INFO python.PythonRDD: Times: total = 253, boot = 13, init = 64, finish = 176
24/05/03 20:27:27 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 1883 bytes result sent to driver
24/05/03 20:27:27 INFO scheduler.DAGScheduler: Stage 1 (count at <ipython-input-16-a079c423cba4>:1) finished in 0.378 s
24/05/03 20:27:27 INFO scheduler.DAGScheduler: Job 1 finished: count at <ipython-input-16-a079c423cba4>:1, took 0.484367 s
24/05/03 20:27:27 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 341 ms on localhost (1/1)
24/05/03 20:27:27 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
Out[16]: 218

In [17]: logs.map(lambda line: len(line)).take(5)
24/05/03 20:32:28 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/03 20:32:28 INFO scheduler.DAGScheduler: Got job 2 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/03 20:32:28 INFO scheduler.DAGScheduler: Final stage: Stage 2(runJob at PythonRDD.scala:356)
24/05/03 20:32:28 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:32:28 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:32:28 INFO scheduler.DAGScheduler: Submitting Stage 2 (PythonRDD[10] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/03 20:32:28 INFO storage.MemoryStore: ensureFreeSpace(5208) called with curMem=922434, maxMem=280248975
24/05/03 20:32:28 INFO storage.MemoryStore: Block broadcast_5 stored as values in memory (estimated size 5.1 KB, free 266.4 MB)
24/05/03 20:32:28 INFO storage.MemoryStore: ensureFreeSpace(3321) called with curMem=927642, maxMem=280248975
24/05/03 20:32:28 INFO storage.MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 3.2 KB, free 266.4 MB)
24/05/03 20:32:28 INFO storage.BlockManagerInfo: Added broadcast_5_piece0 in memory on localhost:33480 (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:32:28 INFO storage.BlockManagerMaster: Updated info of block broadcast_5_piece0
24/05/03 20:32:28 INFO spark.SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:839
24/05/03 20:32:28 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 2 (PythonRDD[10] at RDD at PythonRDD.scala:42)
24/05/03 20:32:28 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
24/05/03 20:32:28 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:32:28 INFO executor.Executor: Running task 0.0 in stage 2.0 (TID 2)
24/05/03 20:32:28 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:32:29 INFO python.PythonRDD: Times: total = 77, boot = 6, init = 69, finish = 2
24/05/03 20:32:29 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 1911 bytes result sent to driver
24/05/03 20:32:29 INFO scheduler.DAGScheduler: Stage 2 (runJob at PythonRDD.scala:356) finished in 0.219 s
24/05/03 20:32:29 INFO scheduler.DAGScheduler: Job 2 finished: runJob at PythonRDD.scala:356, took 0.302328 s
Out[17]: [152, 153, 146, 138, 131]

In [18]: 24/05/03 20:32:29 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 201 ms on localhost (1/1)
24/05/03 20:32:29 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
SXK230064
```

```
Out[17]: [152, 153, 146, 138, 131]

In [18]: 24/05/03 20:32:29 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 201 ms on localhost (1/1)
24/05/03 20:32:29 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool


In [18]: logs.map(lambda line: line.split()).take(5)
24/05/03 20:37:49 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Got job 3 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Final stage: Stage 3(runJob at PythonRDD.scala:356)
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Submitting Stage 3 (PythonRDD[12] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/03 20:37:49 INFO storage.MemoryStore: ensureFreeSpace(5200) called with curMem=930963, maxMem=280248975
24/05/03 20:37:49 INFO storage.MemoryStore: Block broadcast_6 stored as values in memory (estimated size 5.1 KB, free 266.4 MB)
24/05/03 20:37:49 INFO storage.MemoryStore: ensureFreeSpace(3321) called with curMem=936163, maxMem=280248975
24/05/03 20:37:49 INFO storage.MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 3.2 KB, free 266.4 MB)
24/05/03 20:37:49 INFO storage.BlockManagerInfo: Added broadcast_6_piece0 in memory on localhost:33480 (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:37:49 INFO storage.BlockManagerMaster: Updated info of block broadcast_6_piece0
24/05/03 20:37:49 INFO spark.SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:839
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 3 (PythonRDD[12] at RDD at PythonRDD.scala:42)
24/05/03 20:37:49 INFO scheduler.TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
24/05/03 20:37:49 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:37:49 INFO executor.Executor: Running task 0.0 in stage 3.0 (TID 3)
24/05/03 20:37:49 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:37:49 INFO python.PythonRDD: Times: total = 109, boot = 10, init = 92, finish = 7
24/05/03 20:37:49 INFO executor.Executor: Finished task 0.0 in stage 3.0 (TID 3). 3100 bytes result sent to driver
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Stage 3 (runJob at PythonRDD.scala:356) finished in 0.248 s
24/05/03 20:37:49 INFO scheduler.DAGScheduler: Job 3 finished: runJob at PythonRDD.scala:356, took 0.346040 s
24/05/03 20:37:49 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 234 ms on localhost (1/1)
24/05/03 20:37:49 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
Out[18]:
[u'62.196.231.224',
 u'-',
 u'21970',
 u'[30/Oct/2013:12:40:22',
 u'+0100]',
 u'"GET',
```

7. In the previous step, we printed numbers which are not very useful. Instead, we will try mapping to an array of words for each line. This time it prints out five arrays, each containing the words from the corresponding log file line.

```
 u'200',
 u'9090',
 u'"http://www.loudacre.com"',
 u'"Loudacre',
 u'Mobile',
 u'Browser',
 u'iFruit',
 u'1"'],
[u'149.94.96.10',
 u'-',
 u'70',
 u'[30/Oct/2013:12:39:24',
 u'+0100]',
 u'"GET',
 u'/KBDOC-00270.html',
 u'HTTP/1.0"',
 u'200',
 u'4301',
 u'"http://www.loudacre.com"',
 u'"Loudacre',
 u'CSR',
 u'Browser"'],
[u'149.94.96.10',
 u'-',
 u'70',
 u'[30/Oct/2013:12:39:24',
 u'+0100]',
 u'"GET',
 u'/theme.css',
 u'HTTP/1.0"',
 u'200',
 u'3137',
 u'"http://www.loudacre.com"',
 u'"Loudacre',
 u'CSR',
 u'Browser"']]

In [19]: SXK230064
```

8. From the above steps, we learned how map works; now we will write code that splits all the lines by space. Here, I created a new RDD by splitting lines.

```
 u'CSR',
 u'Browser"']]

In [19]: ips = logs.map(lambda line: line.split()[0])

In [20]: ips.take(5)
24/05/03 20:42:42 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/03 20:42:42 INFO scheduler.DAGScheduler: Got job 4 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/03 20:42:42 INFO scheduler.DAGScheduler: Final stage: Stage 4(runJob at PythonRDD.scala:356)
24/05/03 20:42:42 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:42:42 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:42:42 INFO scheduler.DAGScheduler: Submitting Stage 4 (PythonRDD[14] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/03 20:42:42 INFO storage.MemoryStore: ensureFreeSpace(5208) called with curMem=939484, maxMem=280248975
24/05/03 20:42:42 INFO storage.MemoryStore: Block broadcast_7 stored as values in memory (estimated size 5.1 KB, free 266.4 MB)
24/05/03 20:42:42 INFO storage.MemoryStore: ensureFreeSpace(3326) called with curMem=944692, maxMem=280248975
24/05/03 20:42:42 INFO storage.MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 3.2 KB, free 266.4 MB)
24/05/03 20:42:42 INFO storage.BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:33480 (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:42:42 INFO storage.BlockManagerMaster: Updated info of block broadcast_7_piece0
24/05/03 20:42:42 INFO spark.SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:839
24/05/03 20:42:42 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 4 (PythonRDD[14] at RDD at PythonRDD.scala:42)
24/05/03 20:42:42 INFO scheduler.TaskSchedulerImpl: Adding task set 4.0 with 1 tasks
24/05/03 20:42:42 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 4.0 (TID 4, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:42:42 INFO executor.Executor: Running task 0.0 in stage 4.0 (TID 4)
24/05/03 20:42:42 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:42:42 INFO python.PythonRDD: Times: total = 83, boot = 10, init = 71, finish = 2
24/05/03 20:42:43 INFO executor.Executor: Finished task 0.0 in stage 4.0 (TID 4). 2000 bytes result sent to driver
24/05/03 20:42:43 INFO scheduler.DAGScheduler: Stage 4 (runJob at PythonRDD.scala:356) finished in 0.188 s
24/05/03 20:42:43 INFO scheduler.DAGScheduler: Job 4 finished: runJob at PythonRDD.scala:356, took 0.287183 s
Out[20]:
[u'62.196.231.224',
 u'184.133.49.78',
 u'184.133.49.78',
 u'149.94.96.10',
 u'149.94.96.10']
24/05/03 20:42:43 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 4.0 (TID 4) in 162 ms on localhost (1/1)
24/05/03 20:42:43 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool

In [21]: SXK230064
```

9. From the above, we can see that take and collect return outputs in arrays, which allows us to iterate through them.

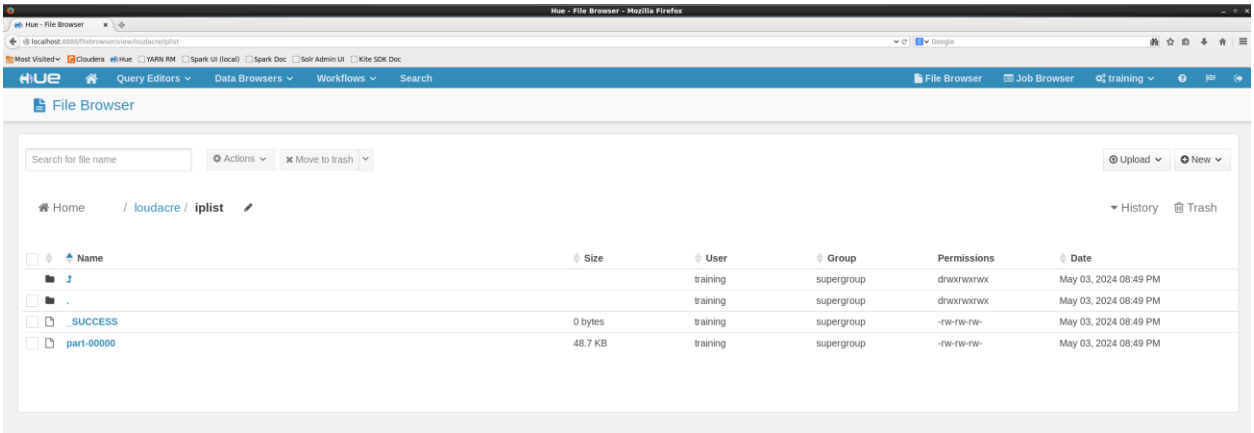
```
In [21]: for ip in ips.take(10): print ip
24/05/03 20:46:13 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/03 20:46:13 INFO scheduler.DAGScheduler: Got job 5 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/03 20:46:13 INFO scheduler.DAGScheduler: Final stage: Stage 5(runJob at PythonRDD.scala:356)
24/05/03 20:46:13 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:46:13 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:46:13 INFO scheduler.DAGScheduler: Submitting Stage 5 (PythonRDD[15] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/03 20:46:13 INFO storage.MemoryStore: ensureFreeSpace(5208) called with curMem=948018, maxMem=280248975
24/05/03 20:46:13 INFO storage.MemoryStore: Block broadcast_8 stored as values in memory (estimated size 5.1 KB, free 266.4 MB)
24/05/03 20:46:13 INFO storage.MemoryStore: ensureFreeSpace(3327) called with curMem=953226, maxMem=280248975
24/05/03 20:46:13 INFO storage.MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 3.2 KB, free 266.4 MB)
24/05/03 20:46:13 INFO storage.BlockManagerInfo: Added broadcast_8_piece0 in memory on localhost:33480 (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:46:13 INFO storage.BlockManagerMaster: Updated info of block broadcast_8_piece0
24/05/03 20:46:13 INFO spark.SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:839
24/05/03 20:46:13 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 5 (PythonRDD[15] at RDD at PythonRDD.scala:42)
24/05/03 20:46:13 INFO scheduler.TaskSchedulerImpl: Adding task set 5.0 with 1 tasks
24/05/03 20:46:13 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 5, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:46:13 INFO executor.Executor: Running task 0.0 in stage 5.0 (TID 5)
24/05/03 20:46:14 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:46:14 INFO python.PythonRDD: Times: total = 86, boot = 9, init = 70, finish = 7
24/05/03 20:46:14 INFO executor.Executor: Finished task 0.0 in stage 5.0 (TID 5). 2117 bytes result sent to driver
24/05/03 20:46:14 INFO scheduler.DAGScheduler: Stage 5 (runJob at PythonRDD.scala:356) finished in 0.256 s
24/05/03 20:46:14 INFO scheduler.DAGScheduler: Job 5 finished: runJob at PythonRDD.scala:356, took 0.347268 s
62.196.231.224
184.133.49.78
184.133.49.78
149.94.96.10
149.94.96.10
53.126.128.196
53.126.128.196
148.216.12.83
148.216.12.83
246.128.198.129

In [22]: 24/05/03 20:46:14 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 206 ms on localhost (1/1)
24/05/03 20:46:14 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
SXK230064
```

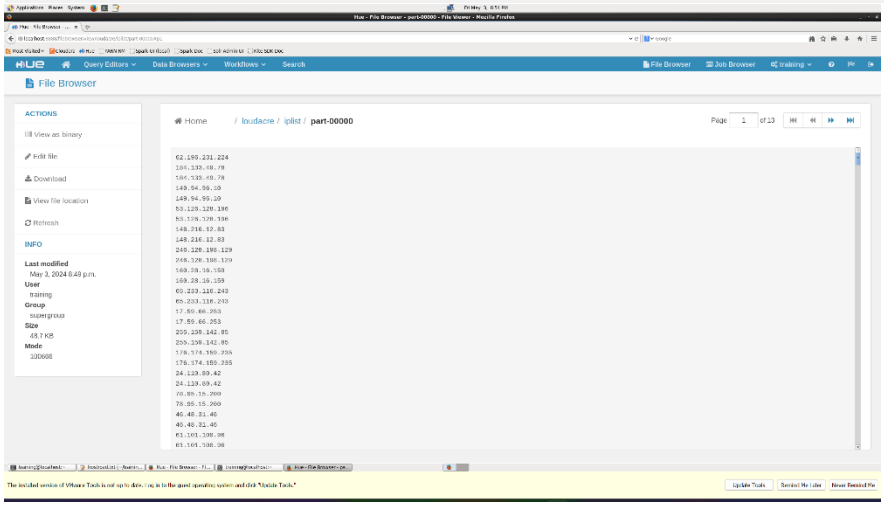
10. Lastly, I saved the list of IP addresses as a text file.

```
24/05/03 20:46:14 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool


In [22]: ips.saveAsTextFile("/loudacre/iplist")
24/05/03 20:49:01 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/03 20:49:02 INFO spark.SparkContext: Starting job: saveAsTextFile at NativeMethodAccessorImpl.java:-2
24/05/03 20:49:02 INFO scheduler.DAGScheduler: Got job 6 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) with 1 output partitions (allowLocal=fa
lse)
24/05/03 20:49:02 INFO scheduler.DAGScheduler: Final stage: Stage 6(saveAsTextFile at NativeMethodAccessorImpl.java:-2)
24/05/03 20:49:02 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/03 20:49:02 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/03 20:49:02 INFO scheduler.DAGScheduler: Submitting Stage 6 (MapPartitionsRDD[18] at saveAsTextFile at NativeMethodAccessorImpl.java:-2), which
 has no missing parents
24/05/03 20:49:02 INFO storage.MemoryStore: ensureFreeSpace(137448) called with curMem=956553, maxMem=280248975
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_9 stored as values in memory (estimated size 134.2 KB, free 266.2 MB)
24/05/03 20:49:02 INFO storage.MemoryStore: ensureFreeSpace(48024) called with curMem=1094001, maxMem=280248975
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 46.9 KB, free 266.2 MB)
24/05/03 20:49:02 INFO storage.BlockManagerInfo: Added broadcast_9_piece0 in memory on localhost:33480 (size: 46.9 KB, free: 267.1 MB)
24/05/03 20:49:02 INFO storage.BlockManagerMaster: Updated info of block broadcast_9_piece0
24/05/03 20:49:02 INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:839
24/05/03 20:49:02 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 6 (MapPartitionsRDD[18] at saveAsTextFile at NativeMethodAccesso
rImpl.java:-2)
24/05/03 20:49:02 INFO scheduler.TaskSchedulerImpl: Adding task set 6.0 with 1 tasks
24/05/03 20:49:02 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 6.0 (TID 6, localhost, PROCESS_LOCAL, 1326 bytes)
24/05/03 20:49:02 INFO executor.Executor: Running task 0.0 in stage 6.0 (TID 6)
24/05/03 20:49:02 INFO storage.BlockManager: Removing broadcast 8
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_8
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_8 of size 5208 dropped from memory (free 279112158)
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_8_piece0
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_8_piece0 of size 3327 dropped from memory (free 279115485)
24/05/03 20:49:02 INFO storage.BlockManagerInfo: Removed broadcast_8_piece0 on localhost:33480 in memory (size: 3.2 KB, free: 267.1 MB)
24/05/03 20:49:02 INFO storage.BlockManagerMaster: Updated info of block broadcast_8_piece0
24/05/03 20:49:02 INFO spark.ContextCleaner: Cleaned broadcast 8
24/05/03 20:49:02 INFO storage.BlockManager: Removing broadcast 3
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_3_piece0
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_3_piece0 of size 21204 dropped from memory (free 279136689)
24/05/03 20:49:02 INFO storage.BlockManagerInfo: Removed broadcast_3_piece0 on localhost:33480 in memory (size: 20.7 KB, free: 267.2 MB)
24/05/03 20:49:02 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
```

```
24/05/03 20:49:02 INFO spark.ContextCleaner: Cleaned broadcast 4
24/05/03 20:49:02 INFO storage.BlockManager: Removing broadcast 5
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_5_piece0
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_5_piece0 of size 3321 dropped from memory (free 279429840)
24/05/03 20:49:02 INFO storage.BlockManagerInfo: Removed broadcast_5_piece0 on localhost:33480 in memory (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:49:02 INFO storage.BlockManagerMaster: Updated info of block broadcast_5_piece0
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_5
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_5 of size 5208 dropped from memory (free 279435048)
24/05/03 20:49:02 INFO spark.ContextCleaner: Cleaned broadcast 5
24/05/03 20:49:02 INFO storage.BlockManager: Removing broadcast 6
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_6_piece0
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_6_piece0 of size 3321 dropped from memory (free 279438369)
24/05/03 20:49:02 INFO storage.BlockManagerInfo: Removed broadcast_6_piece0 on localhost:33480 in memory (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:49:02 INFO storage.BlockManagerMaster: Updated info of block broadcast_6_piece0
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_6
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_6 of size 5200 dropped from memory (free 279443569)
24/05/03 20:49:02 INFO spark.ContextCleaner: Cleaned broadcast 6
24/05/03 20:49:02 INFO storage.BlockManager: Removing broadcast 7
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_7
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_7 of size 5208 dropped from memory (free 279448777)
24/05/03 20:49:02 INFO storage.BlockManager: Removing block broadcast_7_piece0
24/05/03 20:49:02 INFO storage.MemoryStore: Block broadcast_7_piece0 of size 3326 dropped from memory (free 279452103)
24/05/03 20:49:02 INFO storage.BlockManagerInfo: Removed broadcast_7_piece0 on localhost:33480 in memory (size: 3.2 KB, free: 267.2 MB)
24/05/03 20:49:02 INFO storage.BlockManagerMaster: Updated info of block broadcast_7_piece0
24/05/03 20:49:02 INFO spark.ContextCleaner: Cleaned broadcast 7
24/05/03 20:49:02 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/03 20:49:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/03 20:49:03 INFO python.PythonRDD: Times: total = 260, boot = 6, init = 55, finish = 199
24/05/03 20:49:03 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405032049_0006_m_000000_6' to hdfs://localhost:8020/loudacre/ipli
st/_temporary/0/task_202405032049_0006_m_000000
24/05/03 20:49:03 INFO spark.SparkHadoopWriter: attempt_202405032049_0006_m_000000_6: Committed
24/05/03 20:49:03 INFO executor.Executor: Finished task 0.0 in stage 6.0 (TID 6). 2160 bytes result sent to driver
24/05/03 20:49:03 INFO scheduler.DAGScheduler: Stage 6 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 1.204 s
24/05/03 20:49:03 INFO scheduler.DAGScheduler: Job 6 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 1.380751 s
24/05/03 20:49:03 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 6.0 (TID 6) in 1189 ms on localhost (1/1)
24/05/03 20:49:03 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool

In [23]: sxk230064
```

11. Now using the Hue File Browser, I displayed the contents of the IP list.

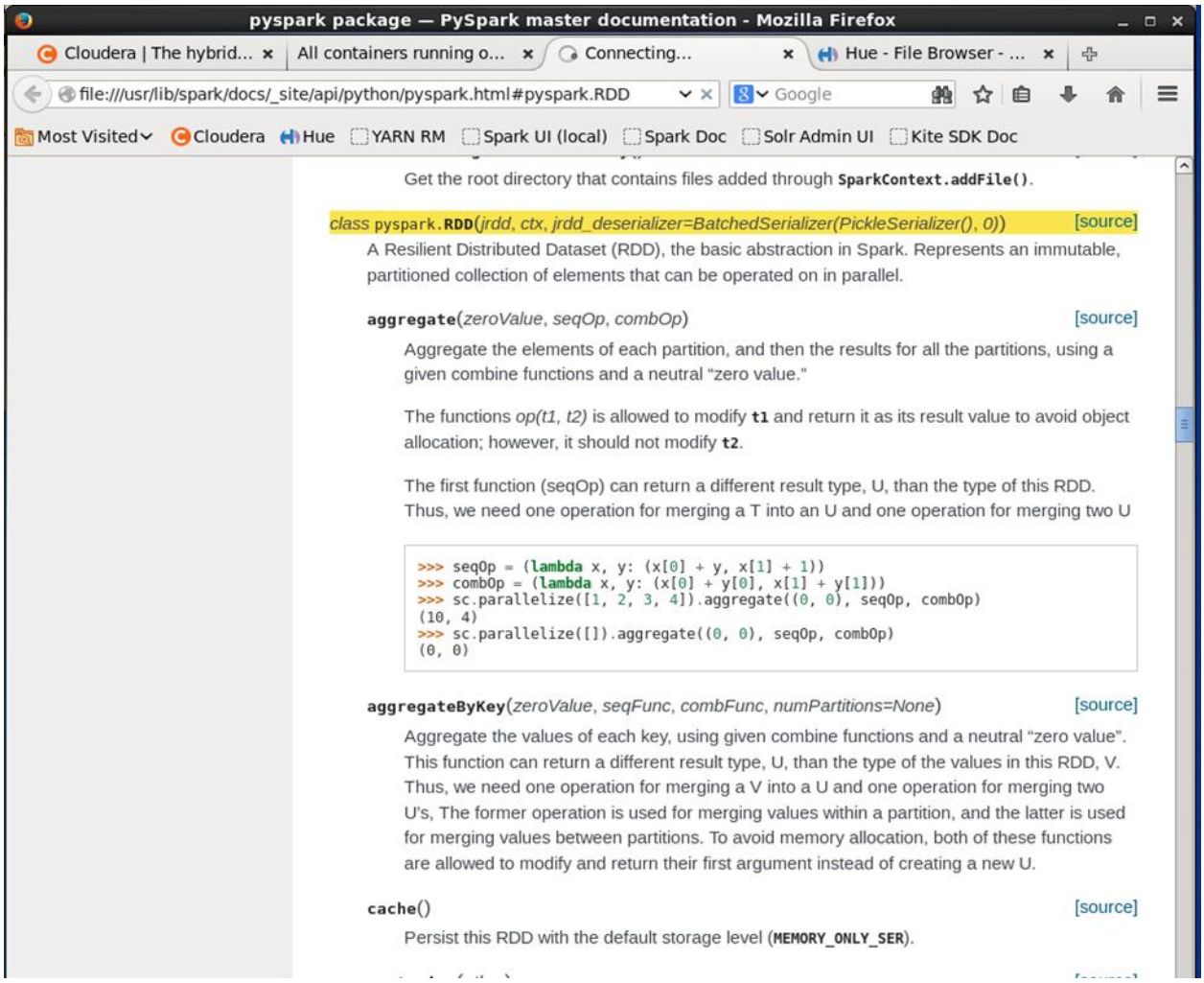




LAB CHAPTER 11: PROCESS DATA FILES WITH SPARK

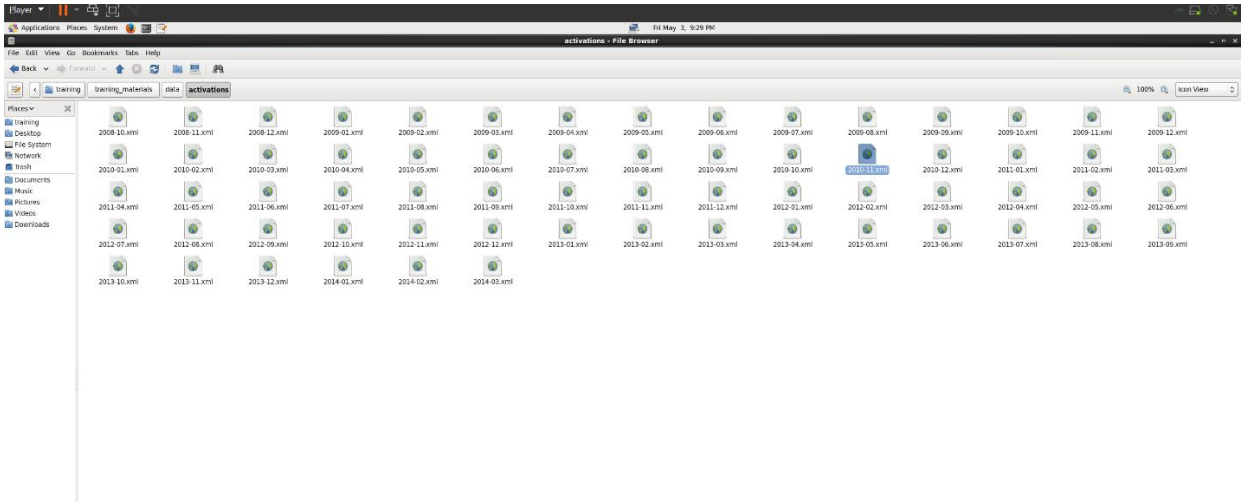
Review the API Documentation for RDD Operations

1. Reviewed methods available for the RDD class



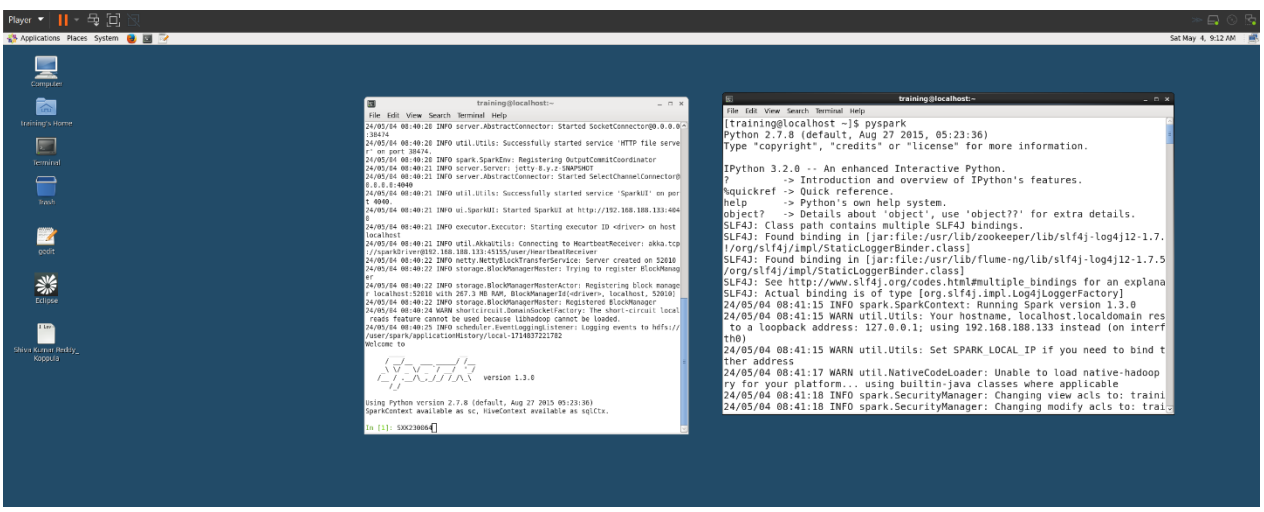
The Data

1. Each XML file contains data for all the devices activated by customers during a specific month.



2. Copied this data to /loudacre in HDFS.

    1) cd /home/training/training_materials/data
    2) hdfs dfs -put activations /loudacre
    3) hdfs dfs -cat /loudacre/activations/2008-10.xml | tail -n 20



The Task:

1. In this step we run the code (ActivationModels) provided in the pyspark terminal.

    Code used:

    import xml.etree.ElementTree as ElementTree

    # Given a string containing XML, parse the string, and

    # return an iterator of activation XML records (Elements) contained in the string

    def getactivations(s):

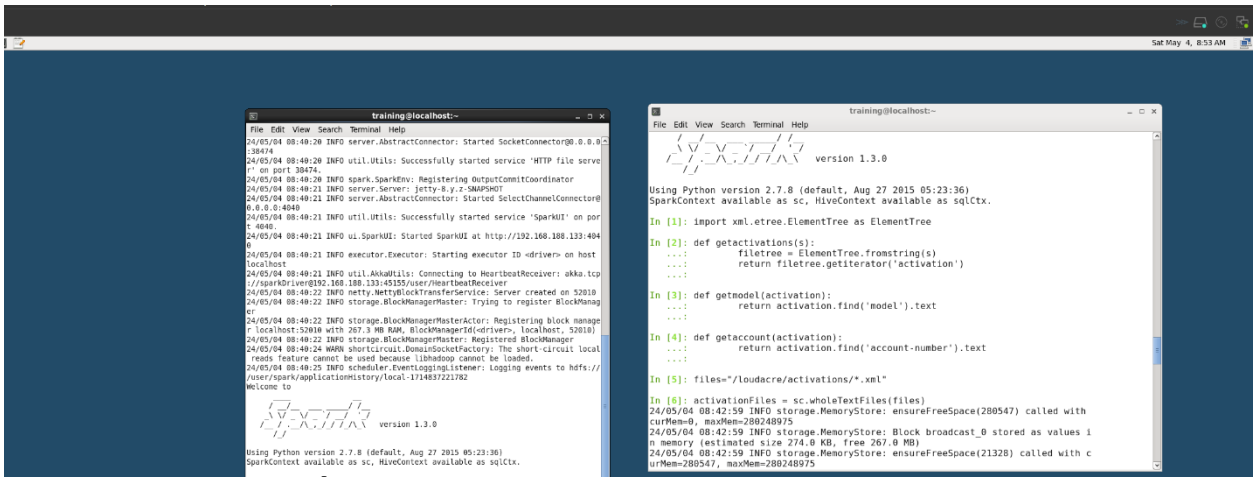
    filetree = ElementTree.fromstring(s)

```
    return filetree.getiterator('activation')
# Given an activation record (XML Element), return the model name
def getmodel(activation):
    return activation.find('model').text
# Given an activation record (XML Element), return the account number
def getaccount(activation):
    return activation.find('account-number').text
```

2. By using wholeTextFiles I have created an RDD from the activations dataset.

   Code used:

   ```
   files="/loudacre/activations/*.xml"
   activationFiles = sc.wholeTextFiles(files)
   ```

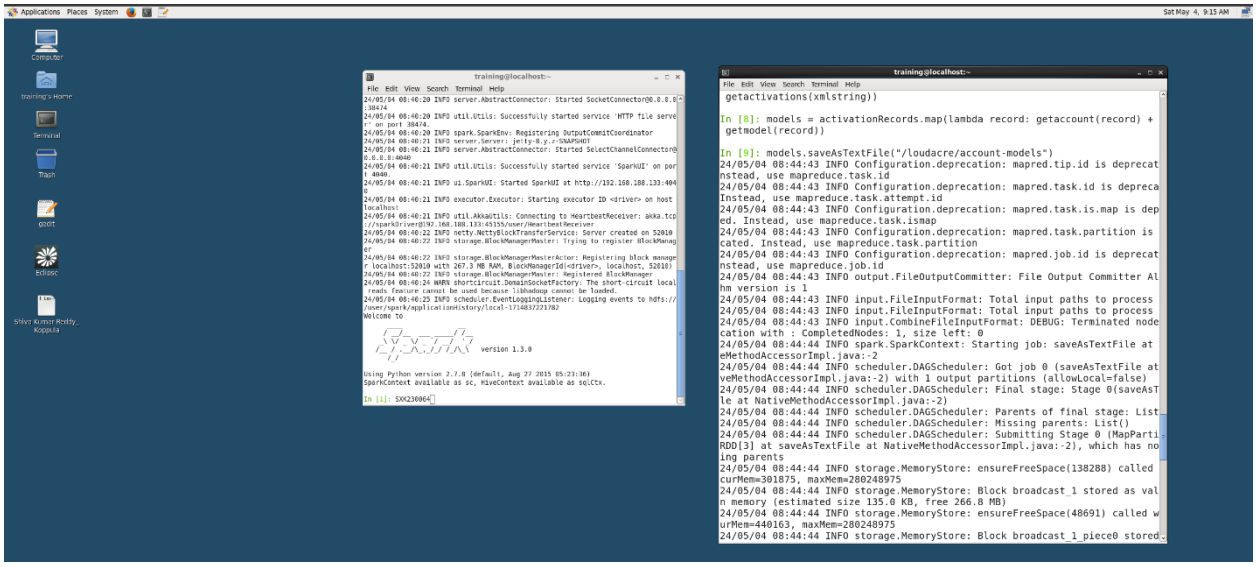


3. Using flatMap I defined getActivations which returns a collection of XML records. Where flatMap will map each record to a separate RDD element.

   Code Used**: activationRecos = activationFiles.flatMap(lambda (filename, xmlstring): getactivations(xmlstring))**

4. I mapped each activation record to a string which first displays the account number.

   Code used: **models = activationRecos.map(lambda record: getaccount (record) + ":" + getmodel(record))**



5. Then, I saved the formatted strings to a text file in loudacre directory.
models.saveAsTextFile("/loudacre/account-models")

LAB CHAPTER 12: USE PAIR RDDS TO JOIN TWO DATASETS

Explore Web Log Files

1. In this we use smaller dataset, consisting of only a few of the web log files, rather than all of them. I have specified a wildcard textFile("/loudacre/weblogs/*6") which would include only filenames ending with the digit 6.
   logs = sc.textFile("/loudacre/weblogs/*6")

2. Using map-reduce, we will count the number of requests from each user
   Created a pair RDD where the user ID is the key, and the number 1 is the value.
   Reduced to the sum values for each user ID.
   user_request = logs.map(lambda line: line.split()).map(lambda words:(words[2],1))
   ...: .reduceByKey(lambda count1, count2: count1 + count2)

3. In this step by using countByKey we determine how many users visited the site for each frequency. That is, how many users visited once, twice, three times and so on.

   user_freq = user_request.map(lambda(userid,freq): (freq,userid)).countByKey()

   print user_freq

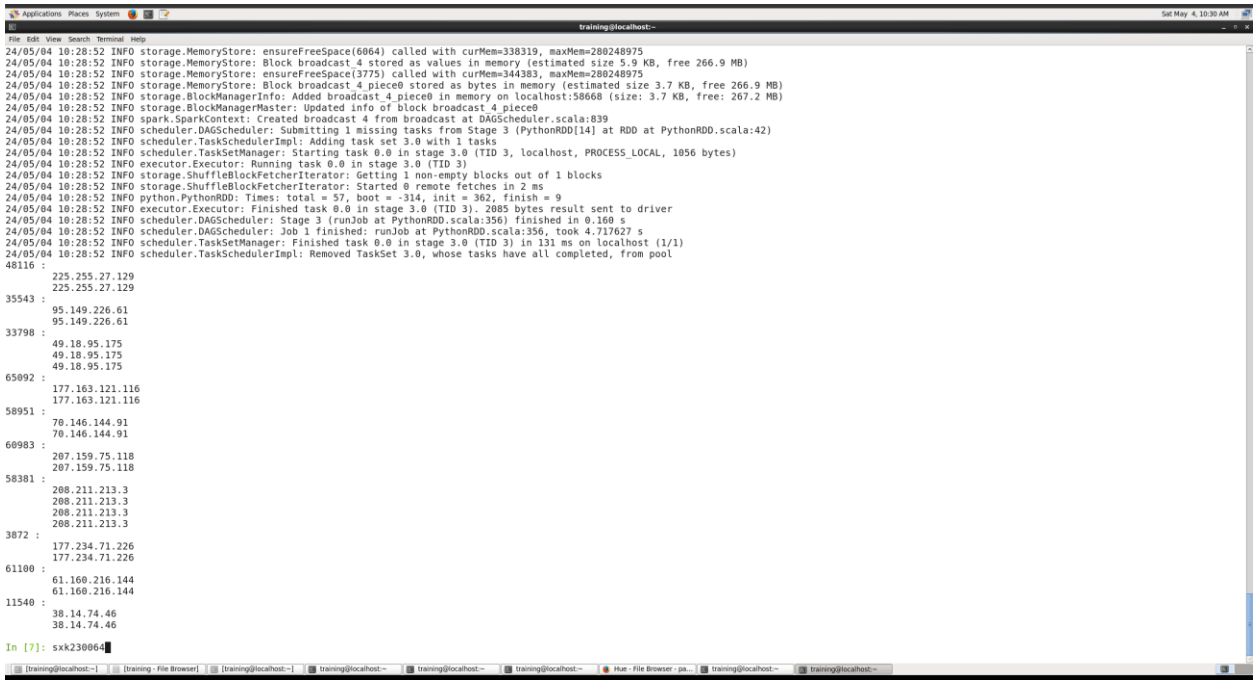
```
24/05/04 09:51:49 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=280171, maxMem=280248975
24/05/04 09:51:49 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 20.7 KB, free 267.0 MB)
24/05/04 09:51:49 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:36744 (size: 20.7 KB, free: 267.2 MB)
24/05/04 09:51:49 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/04 09:51:49 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:-2

In [2]: userReqs = logs.map(lambda line: line.split()).map(lambda words: (words[2],1)).reduceByKey(lambda count1, count2: count1 + count2)
24/05/04 09:53:56 INFO mapred.FileInputFormat: Total input paths to process : 1

In [3]: freqcnt=userReqs.map(lambda (x,y): (y,x)).countByKey()
24/05/04 09:59:22 INFO spark.SparkContext: Starting job: countByKey at <ipython-input-3-d26148b1d718>:1
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Registering RDD 4 (reduceByKey at <ipython-input-2-c1d4266cd9ac>:1)
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Got job 0 (countByKey at <ipython-input-3-d26148b1d718>:1) with 1 output partitions (allowLocal=false)
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Final stage: Stage 1(countByKey at <ipython-input-3-d26148b1d718>:1)
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 0)
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Missing parents: List(Stage 0)
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Submitting Stage 0 (PairwiseRDD[4] at reduceByKey at <ipython-input-2-c1d4266cd9ac>:1), which has no missing parents
24/05/04 09:59:22 INFO storage.MemoryStore: ensureFreeSpace(7904) called with curMem=301375, maxMem=280248975
24/05/04 09:59:22 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 7.7 KB, free 267.0 MB)
24/05/04 09:59:22 INFO storage.MemoryStore: ensureFreeSpace(5070) called with curMem=309279, maxMem=280248975
24/05/04 09:59:22 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.0 KB, free 267.0 MB)
24/05/04 09:59:22 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:36744 (size: 5.0 KB, free: 267.2 MB)
24/05/04 09:59:22 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 09:59:22 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:839
24/05/04 09:59:22 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 0 (PairwiseRDD[4] at reduceByKey at <ipython-input-2-c1d4266cd9ac>:1)
24/05/04 09:59:22 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
24/05/04 09:59:22 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1315 bytes)
24/05/04 09:59:22 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/04 09:59:23 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/04 09:59:23 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/05/04 09:59:23 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/05/04 09:59:23 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/05/04 09:59:23 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/05/04 09:59:23 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/05/04 09:59:24 INFO python.PythonRDD: Times: total = 916, boot = 542, init = 214, finish = 160
24/05/04 09:59:24 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 1994 bytes result sent to driver
24/05/04 09:59:24 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1229 ms on localhost (1/1)
24/05/04 09:59:24 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
24/05/04 09:59:24 INFO scheduler.DAGScheduler: Stage 0 (reduceByKey at <ipython-input-2-c1d4266cd9ac>:1) finished in 1.470 s
24/05/04 09:59:24 INFO scheduler.DAGScheduler: looking for newly runnable stages
24/05/04 09:59:24 INFO scheduler.DAGScheduler: running: Set()
24/05/04 09:59:24 INFO scheduler.DAGScheduler: waiting: Set(Stage 1)
24/05/04 09:59:24 INFO scheduler.DAGScheduler: failed: Set()
24/05/04 09:59:24 INFO scheduler.DAGScheduler: Missing parents for Stage 1: List()
24/05/04 09:59:24 INFO scheduler.DAGScheduler: Submitting Stage 1 (PythonRDD[7] at countByKey at <ipython-input-3-d26148b1d718>:1), which is now runnable
24/05/04 09:59:24 INFO storage.MemoryStore: ensureFreeSpace(6480) called with curMem=314349, maxMem=280248975
24/05/04 09:59:24 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 6.3 KB, free 267.0 MB)
24/05/04 09:59:24 INFO storage.MemoryStore: ensureFreeSpace(4145) called with curMem=320829, maxMem=280248975
24/05/04 09:59:24 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 4.0 KB, free 267.0 MB)
24/05/04 09:59:24 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:36744 (size: 4.0 KB, free: 267.2 MB)
24/05/04 09:59:24 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/04 09:59:24 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:839
24/05/04 09:59:24 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (PythonRDD[7] at countByKey at <ipython-input-3-d26148b1d718>:1)
24/05/04 09:59:24 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
24/05/04 09:59:24 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 09:59:24 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
24/05/04 09:59:24 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
24/05/04 09:59:24 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 31 ms
24/05/04 09:59:24 INFO python.PythonRDD: Times: total = 141, boot = -394, init = 532, finish = 3
24/05/04 09:59:24 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 1018 bytes result sent to driver
24/05/04 09:59:24 INFO scheduler.DAGScheduler: Stage 1 (countByKey at <ipython-input-3-d26148b1d718>:1) finished in 0.244 s
24/05/04 09:59:24 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 227 ms on localhost (1/1)
24/05/04 09:59:24 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
24/05/04 09:59:24 INFO scheduler.DAGScheduler: Job 0 finished: countByKey at <ipython-input-3-d26148b1d718>:1, took 2.178353 s

In [4]: sxk230064
```

```
In [4]: print freqcount
defaultdict(<type 'int'>, {1: 1, 2: 757, 3: 5, 4: 292, 5: 1, 6: 58, 8: 31, 10: 9, 12: 3, 14: 3, 16: 1})

In [5]: userips = logs \
   ...:     .map(lambda line: line.split()) \
   ...:     .map(lambda words: (words[2],words[0])) \
   ...:     .groupByKey()
```

4. Then, I created another RDD where user ID is the key, and the value list is the list of all the IP addresses that user has connected from.

user_ip_addr = logs.map(lambda line: line.split()).map(lambda words:(words[2],words[0])).groupByKey()

```
In [4]: print freqcount
defaultdict(<type 'int'>, {1: 1, 2: 757, 3: 5, 4: 292, 5: 1, 6: 58, 8: 31, 10: 9, 12: 3, 14: 3, 16: 1})

In [5]: userips = logs \
   ...:     .map(lambda line: line.split()) \
   ...:     .map(lambda words: (words[2],words[0])) \
   ...:     .groupByKey()

In [6]: for (userid,ips) in userips.take(10):
   ...:         print userid, ":"
   ...:         for ip in ips: print "\t",ip
   ...:
24/05/04 10:28:47 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/04 10:28:47 INFO scheduler.DAGScheduler: Registering RDD 10 (groupByKey at <ipython-input-5-783ca6d831cd>:1)
24/05/04 10:28:47 INFO scheduler.DAGScheduler: Got job 1 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/04 10:28:47 INFO scheduler.DAGScheduler: Final stage: Stage 3(runJob at PythonRDD.scala:356)
24/05/04 10:28:47 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 2)
24/05/04 10:28:47 INFO scheduler.DAGScheduler: Missing parents: List(Stage 2)
24/05/04 10:28:48 INFO scheduler.DAGScheduler: Submitting Stage 2 (PairwiseRDD[10] at groupByKey at <ipython-input-5-783ca6d831cd>:1), which has no missing parents
24/05/04 10:28:48 INFO storage.MemoryStore: ensureFreeSpace(8120) called with curMem=324998, maxMem=280248975
24/05/04 10:28:48 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 7.9 KB, free 266.9 MB)
24/05/04 10:28:48 INFO storage.MemoryStore: ensureFreeSpace(5201) called with curMem=333118, maxMem=280248975
24/05/04 10:28:48 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 5.1 KB, free 266.9 MB)
24/05/04 10:28:48 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:58668 (size: 5.1 KB, free: 267.2 MB)
24/05/04 10:28:48 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
24/05/04 10:28:48 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:839
24/05/04 10:28:48 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 2 (PairwiseRDD[10] at groupByKey at <ipython-input-5-783ca6d831cd>:1)
24/05/04 10:28:48 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
24/05/04 10:28:48 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, PROCESS_LOCAL, 1315 bytes)
24/05/04 10:28:48 INFO executor.Executor: Running task 0.0 in stage 2.0 (TID 2)
24/05/04 10:28:48 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+531346
24/05/04 10:28:52 INFO python.PythonRDD: Times: total = 2892, boot = 64, init = 367, finish = 2461
24/05/04 10:28:52 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 2049 bytes result sent to driver
24/05/04 10:28:52 INFO scheduler.DAGScheduler: Stage 2 (groupByKey at <ipython-input-5-783ca6d831cd>:1) finished in 4.088 s
24/05/04 10:28:52 INFO scheduler.DAGScheduler: looking for newly runnable stages
24/05/04 10:28:52 INFO scheduler.DAGScheduler: running: Set()
24/05/04 10:28:52 INFO scheduler.DAGScheduler: waiting: Set(Stage 3)
24/05/04 10:28:52 INFO scheduler.DAGScheduler: failed: Set()
24/05/04 10:28:52 INFO scheduler.DAGScheduler: Missing parents for Stage 3: List()
24/05/04 10:28:52 INFO scheduler.DAGScheduler: Submitting Stage 3 (PythonRDD[14] at RDD at PythonRDD.scala:42), which is now runnable
24/05/04 10:28:52 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 4031 ms on localhost (1/1)
24/05/04 10:28:52 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
```



```
24/05/04 10:28:52 INFO storage.MemoryStore: ensureFreeSpace(6064) called with curMem=338319, maxMem=280248975
24/05/04 10:28:52 INFO storage.MemoryStore: Block broadcast_4 stored as values in memory (estimated size 5.9 KB, free 266.9 MB)
24/05/04 10:28:52 INFO storage.MemoryStore: ensureFreeSpace(3775) called with curMem=344383, maxMem=280248975
24/05/04 10:28:52 INFO storage.MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 3.7 KB, free 266.9 MB)
24/05/04 10:28:52 INFO storage.BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:58668 (size: 3.7 KB, free: 267.2 MB)
24/05/04 10:28:52 INFO storage.BlockManagerMaster: Updated info of block broadcast_4_piece0
24/05/04 10:28:52 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:839
24/05/04 10:28:52 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 3 (PythonRDD[14] at RDD at PythonRDD.scala:42)
24/05/04 10:28:52 INFO scheduler.TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
24/05/04 10:28:52 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 10:28:52 INFO executor.Executor: Running task 0.0 in stage 3.0 (TID 3)
24/05/04 10:28:52 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
24/05/04 10:28:52 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 2 ms
24/05/04 10:28:52 INFO python.PythonRDD: Times: total = 57, boot = -314, init = 362, finish = 9
24/05/04 10:28:52 INFO executor.Executor: Finished task 0.0 in stage 3.0 (TID 3). 2085 bytes result sent to driver
24/05/04 10:28:52 INFO scheduler.DAGScheduler: Stage 3 (runJob at PythonRDD.scala:356) finished in 0.160 s
24/05/04 10:28:52 INFO scheduler.DAGScheduler: Job 1 finished: runJob at PythonRDD.scala:356, took 4.717627 s
24/05/04 10:28:52 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 131 ms on localhost (1/1)
24/05/04 10:28:52 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
48116 :
        225.255.27.129
        225.255.27.129
35543 :
        95.149.226.61
        95.149.226.61
33798 :
        49.18.95.175
        49.18.95.175
        49.18.95.175
65092 :
        177.163.121.116
        177.163.121.116
58951 :
        70.146.144.91
        70.146.144.91
60983 :
        207.159.75.118
        207.159.75.118
58381 :
        208.211.213.3
        208.211.213.3
        208.211.213.3
        208.211.213.3
3872 :
        177.234.71.226
        177.234.71.226
61100 :
        61.160.216.144
        61.160.216.144
11540 :
        38.14.74.46
        38.14.74.46

In [7]: sxk230064
```

Join Web Log Data with Account Data

5. Join the accounts data with the weblog data to produce a dataset keyed by user ID which contains the user account information and the number of website hits for that user

```
acc = sc.textFile("/loudacre/accounts").map(lambda s: s.split(',')).map(lambda account:(account[0],account))

account = acc.join(user_request)

for (userid, (values,count)) in account.take(5):
print userid, count, values[3], values[4]
```

```
In [7]: accounts = sc.textFile("/loudacre/accounts") \
   ...:     .map(lambda s: s.split(',')) \
   ...:     .map(lambda account: (account[0],account))
24/05/04 10:31:32 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=348158, maxMem=280248975
24/05/04 10:31:32 INFO storage.MemoryStore: Block broadcast_5 stored as values in memory (estimated size 273.7 KB, free 266.7 MB)
24/05/04 10:31:33 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=628401, maxMem=280248975
24/05/04 10:31:33 INFO storage.MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.6 MB)
24/05/04 10:31:33 INFO storage.BlockManagerInfo: Added broadcast_5_piece0 in memory on localhost:58668 (size: 20.7 KB, free: 267.2 MB)
24/05/04 10:31:33 INFO storage.BlockManagerMaster: Updated info of block broadcast_5_piece0
24/05/04 10:31:33 INFO spark.SparkContext: Created broadcast 5 from textFile at NativeMethodAccessorImpl.java:-2

In [8]: accounthits = accounts.join(userreqs)
24/05/04 10:32:02 INFO mapred.FileInputFormat: Total input paths to process : 7

In [9]: for (userid,(values,count)) in accounthits.take(5) :
   ...:         print  userid, count, values[3],values[4]
   ...:
24/05/04 10:32:16 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/04 10:32:16 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 145 bytes
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Registering RDD 21 (join at <ipython-input-8-b8d0361ddb3f>:1)
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Got job 2 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Final stage: Stage 6(runJob at PythonRDD.scala:356)
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 5)
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Missing parents: List(Stage 5)
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Submitting Stage 5 (PairwiseRDD[21] at join at <ipython-input-8-b8d0361ddb3f>:1), which has no missing parents
24/05/04 10:32:16 INFO storage.MemoryStore: ensureFreeSpace(12264) called with curMem=649605, maxMem=280248975
24/05/04 10:32:16 INFO storage.MemoryStore: Block broadcast_6 stored as values in memory (estimated size 12.0 KB, free 266.6 MB)
24/05/04 10:32:16 INFO storage.MemoryStore: ensureFreeSpace(6608) called with curMem=661869, maxMem=280248975
24/05/04 10:32:16 INFO storage.MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 6.5 KB, free 266.6 MB)
24/05/04 10:32:16 INFO storage.BlockManagerInfo: Added broadcast_6_piece0 in memory on localhost:58668 (size: 6.5 KB, free: 267.2 MB)
24/05/04 10:32:16 INFO storage.BlockManagerMaster: Updated info of block broadcast_6_piece0
24/05/04 10:32:16 INFO spark.SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:839
24/05/04 10:32:16 INFO scheduler.DAGScheduler: Submitting 8 missing tasks from Stage 5 (PairwiseRDD[21] at join at <ipython-input-8-b8d0361ddb3f>:1)
24/05/04 10:32:16 INFO scheduler.TaskSchedulerImpl: Adding task set 5.0 with 8 tasks
24/05/04 10:32:16 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 4, localhost, PROCESS_LOCAL, 1414 bytes)
24/05/04 10:32:16 INFO executor.Executor: Running task 0.0 in stage 5.0 (TID 4)
24/05/04 10:32:16 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/accounts/part-m-00000:0+4706617
24/05/04 10:32:18 INFO storage.BlockManager: Removing broadcast 4
24/05/04 10:32:18 INFO storage.BlockManager: Removing block broadcast_4
24/05/04 10:32:18 INFO storage.MemoryStore: Block broadcast_4 of size 6064 dropped from memory (free 279586562)
24/05/04 10:32:18 INFO storage.BlockManager: Removing block broadcast_4_piece0
24/05/04 10:32:18 INFO storage.MemoryStore: Block broadcast_4_piece0 of size 3775 dropped from memory (free 279590337)
24/05/04 10:32:18 INFO storage.BlockManagerInfo: Removed broadcast_4_piece0 on localhost:58668 in memory (size: 3.7 KB, free: 267.2 MB)
24/05/04 10:32:18 INFO storage.BlockManagerMaster: Updated info of block broadcast_4_piece0
24/05/04 10:32:18 INFO spark.ContextCleaner: Cleaned broadcast 4
```

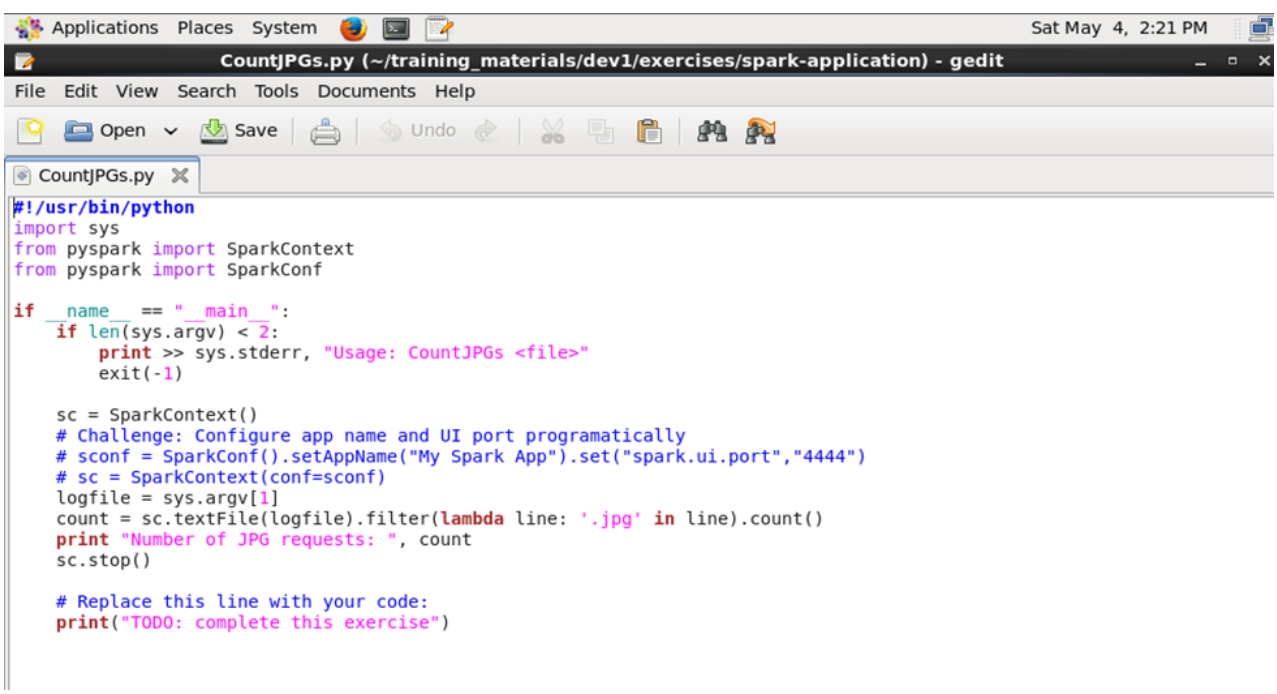
```
24/05/04 10:32:39 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/accounts/part-m-00005:0+131
24/05/04 10:32:39 INFO python.PythonRDD: Times: total = 46, boot = -207, init = 253, finish = 0
24/05/04 10:32:39 INFO python.PythonRDD: Times: total = 54, boot = -199, init = 252, finish = 1
24/05/04 10:32:39 INFO executor.Executor: Finished task 5.0 in stage 5.0 (TID 9). 2056 bytes result sent to driver
24/05/04 10:32:39 INFO scheduler.TaskSetManager: Starting task 6.0 in stage 5.0 (TID 10, localhost, PROCESS_LOCAL, 1414 bytes)
24/05/04 10:32:39 INFO executor.Executor: Running task 6.0 in stage 5.0 (TID 10)
24/05/04 10:32:39 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 5.0 (TID 9) in 262 ms on localhost (6/8)
24/05/04 10:32:39 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/accounts/part-m-00006:0+131
24/05/04 10:32:39 INFO python.PythonRDD: Times: total = 40, boot = -149, init = 189, finish = 0
24/05/04 10:32:39 INFO python.PythonRDD: Times: total = 44, boot = -140, init = 183, finish = 1
24/05/04 10:32:39 INFO executor.Executor: Finished task 6.0 in stage 5.0 (TID 10). 2056 bytes result sent to driver
24/05/04 10:32:39 INFO scheduler.TaskSetManager: Starting task 7.0 in stage 5.0 (TID 11, localhost, PROCESS_LOCAL, 1154 bytes)
24/05/04 10:32:39 INFO executor.Executor: Running task 7.0 in stage 5.0 (TID 11)
24/05/04 10:32:39 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 5.0 (TID 10) in 204 ms on localhost (7/8)
24/05/04 10:32:39 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
24/05/04 10:32:39 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/05/04 10:32:39 INFO python.PythonRDD: Times: total = 52, boot = -160, init = 207, finish = 5
24/05/04 10:32:40 INFO python.PythonRDD: Times: total = 82, boot = -150, init = 212, finish = 20
24/05/04 10:32:40 INFO executor.Executor: Finished task 7.0 in stage 5.0 (TID 11). 1095 bytes result sent to driver
24/05/04 10:32:40 INFO scheduler.DAGScheduler: Stage 5 (join at <ipython-input-8-b8d0361ddb3f>:1) finished in 23.150 s
24/05/04 10:32:40 INFO scheduler.DAGScheduler: looking for newly runnable stages
24/05/04 10:32:40 INFO scheduler.DAGScheduler: running: Set()
24/05/04 10:32:40 INFO scheduler.DAGScheduler: waiting: Set(Stage 6)
24/05/04 10:32:40 INFO scheduler.DAGScheduler: failed: Set()
24/05/04 10:32:40 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 5.0 (TID 11) in 281 ms on localhost (8/8)
24/05/04 10:32:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
24/05/04 10:32:40 INFO scheduler.DAGScheduler: Missing parents for Stage 6: List()
24/05/04 10:32:40 INFO scheduler.DAGScheduler: Submitting Stage 6 (PythonRDD[25] at RDD at PythonRDD.scala:42), which is now runnable
24/05/04 10:32:40 INFO storage.MemoryStore: ensureFreeSpace(7264) called with curMem=634668, maxMem=280248975
24/05/04 10:32:40 INFO storage.MemoryStore: Block broadcast_7 stored as values in memory (estimated size 7.1 KB, free 266.7 MB)
24/05/04 10:32:40 INFO storage.MemoryStore: ensureFreeSpace(4579) called with curMem=641932, maxMem=280248975
24/05/04 10:32:40 INFO storage.MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 4.5 KB, free 266.6 MB)
24/05/04 10:32:40 INFO storage.BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:58668 (size: 4.5 KB, free: 267.2 MB)
24/05/04 10:32:40 INFO storage.BlockManagerMaster: Updated info of block broadcast_7_piece0
24/05/04 10:32:40 INFO spark.SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:839
24/05/04 10:32:40 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 6 (PythonRDD[25] at RDD at PythonRDD.scala:42)
24/05/04 10:32:40 INFO scheduler.TaskSchedulerImpl: Adding task set 6.0 with 1 tasks
24/05/04 10:32:40 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 6.0 (TID 12, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 10:32:40 INFO executor.Executor: Running task 0.0 in stage 6.0 (TID 12)
24/05/04 10:32:40 INFO storage.ShuffleBlockFetcherIterator: Getting 8 non-empty blocks out of 8 blocks
24/05/04 10:32:40 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/05/04 10:32:40 INFO python.PythonRDD: Times: total = 313, boot = -249, init = 286, finish = 276
24/05/04 10:32:40 INFO executor.Executor: Finished task 0.0 in stage 6.0 (TID 12). 2143 bytes result sent to driver
24/05/04 10:32:40 INFO scheduler.DAGScheduler: Stage 6 (runJob at PythonRDD.scala:356) finished in 0.356 s
24/05/04 10:32:40 INFO scheduler.DAGScheduler: Job 2 finished: runJob at PythonRDD.scala:356, took 23.792299 s
99 4 Gerald Bellomy
72369 2 Evelyn Simmons
69259 2 Judith Higgins
177 2 Daryl Obrien
44489 2 Cheryl Reedy
24/05/04 10:32:40 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 6.0 (TID 12) in 353 ms on localhost (1/1)
24/05/04 10:32:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool

In [10]: sxk230064
```

```
44489 2 Cheryl Reedy
24/05/04 10:32:40 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 6.0 (TID 12) in 353 ms on localhost (1/1)
24/05/04 10:32:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool

In [10]: accountsByPCode = sc.textFile("/loudacre/accounts") \
   ...:     .map(lambda s: s.split(','))\
   ...:     .keyBy(lambda account: account[8])
24/05/04 10:36:45 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=646511, maxMem=280248975
24/05/04 10:36:45 INFO storage.MemoryStore: Block broadcast_8 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
24/05/04 10:36:45 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=926754, maxMem=280248975
24/05/04 10:36:45 INFO storage.MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
24/05/04 10:36:45 INFO storage.BlockManagerInfo: Added broadcast_8_piece0 in memory on localhost:58668 (size: 20.7 KB, free: 267.2 MB)
24/05/04 10:36:45 INFO storage.BlockManagerMaster: Updated info of block broadcast_8_piece0
24/05/04 10:36:45 INFO spark.SparkContext: Created broadcast 8 from textFile at NativeMethodAccessorImpl.java:-2

In [11]: namesByPCode = accountsByPCode\
   ...:     .mapValues(lambda account: account[4] + ',' + account[3]) \
   ...:     .groupByKey()
24/05/04 10:37:09 INFO mapred.FileInputFormat: Total input paths to process : 7

In [12]: for (pcode,names) in namesByPCode.sortByKey().take(5):
   ...:         print "---" ,pcode
   ...:         for name in names: print name
   ...:
24/05/04 10:37:40 INFO spark.SparkContext: Starting job: sortByKey at <ipython-input-12-876fc4764298>:1
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Registering RDD 30 (groupByKey at <ipython-input-11-886eefc832d6>:1)
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Got job 3 (sortByKey at <ipython-input-12-876fc4764298>:1) with 7 output partitions (allowLocal=false)
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Final stage: Stage 8(sortByKey at <ipython-input-12-876fc4764298>:1)
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 7)
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Missing parents: List(Stage 7)
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Submitting Stage 7 (PairwiseRDD[30] at groupByKey at <ipython-input-11-886eefc832d6>:1), which has no missing parents
24/05/04 10:37:40 INFO storage.MemoryStore: ensureFreeSpace(8768) called with curMem=947958, maxMem=280248975
24/05/04 10:37:40 INFO storage.MemoryStore: Block broadcast_9 stored as values in memory (estimated size 8.6 KB, free 266.4 MB)
24/05/04 10:37:40 INFO storage.MemoryStore: ensureFreeSpace(5590) called with curMem=956726, maxMem=280248975
24/05/04 10:37:40 INFO storage.MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 5.5 KB, free 266.3 MB)
24/05/04 10:37:40 INFO storage.BlockManagerInfo: Added broadcast_9_piece0 in memory on localhost:58668 (size: 5.5 KB, free: 267.2 MB)
24/05/04 10:37:40 INFO storage.BlockManagerMaster: Updated info of block broadcast_9_piece0
24/05/04 10:37:40 INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:839
24/05/04 10:37:40 INFO scheduler.DAGScheduler: Submitting 7 missing tasks from Stage 7 (PairwiseRDD[30] at groupByKey at <ipython-input-11-886eefc832d6>:1)
24/05/04 10:37:40 INFO scheduler.TaskSchedulerImpl: Adding task set 7.0 with 7 tasks
24/05/04 10:37:40 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 7.0 (TID 13, localhost, PROCESS_LOCAL, 1305 bytes)
24/05/04 10:37:40 INFO executor.Executor: Running task 0.0 in stage 7.0 (TID 13)
```

LAB CHAPTER 13: WRITE AND RUN A SPARK APPLICATION

Write a Spark application in Python

1. I load the file passed into the program, count the number of JPG requests, and display the count using 'SparkContext.' As shown below, I save this file as 'CountJPGs.py'



2. Now I ran the program, passing the name of the log file to process

   cd $DEV1/exercises/spark-application
   spark-submit CountJPGs.py /loudacre/weblogs/*

```
[training@localhost ~]$ DEV1/exercises/spark-application
bash: DEV1/exercises/spark-application: No such file or directory
[training@localhost ~]$ cd $DEV1/exercises/spark-application
[training@localhost spark-application]$ spark-submit CountJPGs.py /loudacre/weblogs/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/04 14:18:45 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/04 14:18:45 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.16
8.188.133 instead (on interface eth0)
24/05/04 14:18:45 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/04 14:18:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
24/05/04 14:18:49 INFO spark.SecurityManager: Changing view acls to: training
24/05/04 14:18:49 INFO spark.SecurityManager: Changing modify acls to: training
24/05/04 14:18:49 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permi
ssions: Set(training); users with modify permissions: Set(training)
24/05/04 14:18:51 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/04 14:18:51 INFO Remoting: Starting remoting
24/05/04 14:18:52 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:41283]
24/05/04 14:18:52 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:41283]
24/05/04 14:18:52 INFO util.Utils: Successfully started service 'sparkDriver' on port 41283.
24/05/04 14:18:52 INFO spark.SparkEnv: Registering MapOutputTracker
24/05/04 14:18:52 INFO spark.SparkEnv: Registering BlockManagerMaster
24/05/04 14:18:52 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-e1112872-db93-435e-b107-09d7713ec663/blo
```

```
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 14:20:31 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.188.133:4044
24/05/04 14:20:31 INFO scheduler.DAGScheduler: Stopping DAGScheduler
24/05/04 14:20:31 INFO spark.MapOutputTrackerMasterActor: MapOutputTrackerActor stopped!
24/05/04 14:20:31 INFO storage.MemoryStore: MemoryStore cleared
24/05/04 14:20:31 INFO storage.BlockManager: BlockManager stopped
24/05/04 14:20:31 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/04 14:20:31 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorActor: OutputCommitCoordinator stopped!
24/05/04 14:20:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/04 14:20:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remo
te transports.
24/05/04 14:20:31 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/04 14:20:31 INFO Remoting: Remoting shut down
24/05/04 14:20:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost spark-application]$ spark-submit \
>  --master yarn-client \
> sxk230064
```

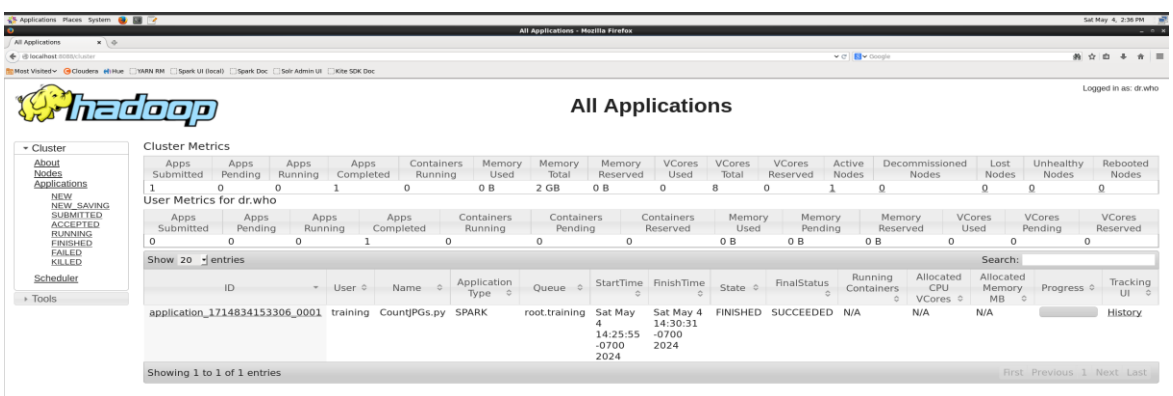
Submit a Spark Application to the Cluster

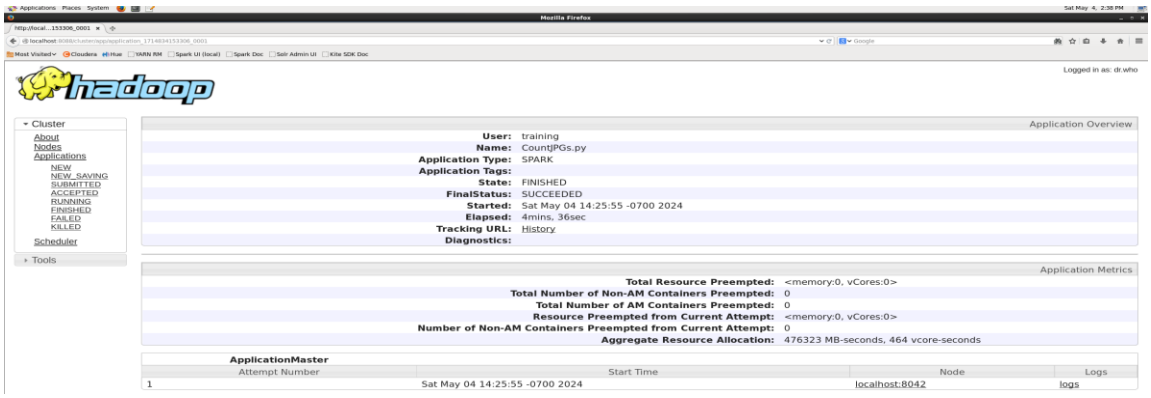
1. I re-ran the program, this time specifying the cluster master to run it on the cluster, as in the below screenshot.
   spark-submit --master yarn-client CountJPGs.py /loudacre/weblogs/*

```
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 14:20:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 14:20:31 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.188.133:4044
24/05/04 14:20:31 INFO scheduler.DAGScheduler: Stopping DAGScheduler
24/05/04 14:20:31 INFO spark.MapOutputTrackerMasterActor: MapOutputTrackerActor stopped!
24/05/04 14:20:31 INFO storage.MemoryStore: MemoryStore cleared
24/05/04 14:20:31 INFO storage.BlockManager: BlockManager stopped
24/05/04 14:20:31 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/04 14:20:31 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorActor: OutputCommitCoordinator stopped!
24/05/04 14:20:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/04 14:20:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remo
te transports.
24/05/04 14:20:31 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/04 14:20:31 INFO Remoting: Remoting shut down
24/05/04 14:20:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost spark-application]$ spark-submit \
>  --master yarn-client \
> CountJPGs.py /loudacre/weblogs/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
```

```
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 14:30:31 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 14:30:31 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.188.133:4044
24/05/04 14:30:31 INFO scheduler.DAGScheduler: Stopping DAGScheduler
24/05/04 14:30:31 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
24/05/04 14:30:31 INFO cluster.YarnClientSchedulerBackend: Asking each executor to shut down
24/05/04 14:30:31 INFO cluster.YarnClientSchedulerBackend: Stopped
24/05/04 14:30:31 INFO spark.MapOutputTrackerMasterActor: MapOutputTrackerActor stopped!
24/05/04 14:30:31 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorActor: OutputCommitCoordinator stopped!
24/05/04 14:30:31 INFO storage.MemoryStore: MemoryStore cleared
24/05/04 14:30:31 INFO storage.BlockManager: BlockManager stopped
24/05/04 14:30:31 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/04 14:30:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/04 14:30:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remo
te transports.
24/05/04 14:30:31 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/04 14:30:31 INFO Remoting: Remoting shut down
24/05/04 14:30:31 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost spark-application]$ sxk230064
```

2. Then I opened the YARN Resource Manager UI and look at the completed application:





3. My application ID is application_1659476506168_0001, by using this I ran the command **yarn logs -applicationId application_1659476506168_0001**

```
[training@localhost spark-application]$ ^C
[training@localhost spark-application]$ yarn logs -applicationId application_1714834153306_0001
24/05/04 14:43:59 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032


Container: container_1714834153306_0001_01_000002 on localhost_51315
=====================================================================
LogType:stderr
Log Upload Time:Sat May 04 14:30:33 -0700 2024
LogLength:7746
Log Contents:
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/04 14:28:00 INFO executor.CoarseGrainedExecutorBackend: Registered signal handlers for [TERM, HUP, INT]
24/05/04 14:29:40 INFO spark.SecurityManager: Changing view acls to: yarn,training
24/05/04 14:29:40 INFO spark.SecurityManager: Changing modify acls to: yarn,training
24/05/04 14:29:40 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permi
ssions: Set(yarn, training); users with modify permissions: Set(yarn, training)
24/05/04 14:29:44 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/04 14:29:44 INFO Remoting: Starting remoting
24/05/04 14:29:49 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://driverPropsFetcher@localhost:57267]
24/05/04 14:29:49 INFO Remoting: Remoting now listens on addresses: [akka.tcp://driverPropsFetcher@localhost:57267]
24/05/04 14:29:49 INFO util.Utils: Successfully started service 'driverPropsFetcher' on port 57267.
24/05/04 14:29:56 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/04 14:29:56 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remo
te transports.
24/05/04 14:29:57 INFO spark.SecurityManager: Changing view acls to: yarn,training
24/05/04 14:29:57 INFO spark.SecurityManager: Changing modify acls to: yarn,training
24/05/04 14:29:57 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permi
ssions: Set(yarn, training); users with modify permissions: Set(yarn, training)
24/05/04 14:29:57 INFO slf4j.Slf4jLogger: Slf4jLogger started
```

```
k__.jar<CPS>$HADOOP_CONF_DIR<CPS>$HADOOP_COMMON_HOME/*<CPS>$HADOOP_COMMON_HOME/lib/*<CPS>$HADOOP_HDFS_HOME/*<CPS>$HADOOP_HDFS_H
OME/lib/*<CPS>$HADOOP_MAPRED_HOME/*<CPS>$HADOOP_MAPRED_HOME/lib/*<CPS>$HADOOP_YARN_HOME/*<CPS>$HADOOP_YARN_HOME/lib/*<CPS>$HADO
OP_MAPRED_HOME/share/hadoop/mapreduce/*<CPS>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*<CPS>:/usr/lib/spark/lib/spark-asse
mbly.jar::/usr/lib/hadoop/lib/*:/usr/lib/hadoop/*:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/*:/usr/lib/hadoop-mapreduce/l
ib/*:/usr/lib/hadoop-mapreduce/*:/usr/lib/hadoop-yarn/lib/*:/usr/lib/hadoop-yarn/*:/usr/lib/hive/lib/*:/usr/lib/flume-ng/lib/*:
/usr/lib/paquet/lib/*:/usr/lib/avro/lib/*, SPARK_LOG_URL_STDERR -> http://localhost:8042/node/containerlogs/container_171483415
3306_0001_01_000002/training/stderr?start=0, SPARK_DIST_CLASSPATH -> :/usr/lib/spark/lib/spark-assembly.jar::/usr/lib/hadoop/li
b/*:/usr/lib/hadoop/*:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/*:/usr/lib/hadoop-mapreduce/lib/*:/usr/lib/hadoop-mapredu
ce/*:/usr/lib/hadoop-yarn/lib/*:/usr/lib/hadoop-yarn/*:/usr/lib/hive/lib/*:/usr/lib/flume-ng/lib/*:/usr/lib/paquet/lib/*:/usr/l
ib/avro/lib/*, SPARK_YARN_STAGING_DIR -> .sparkStaging/application_1714834153306_0001, SPARK_YARN_CACHE_FILES_FILE_SIZES -> 976
48667,575, SPARK_USER -> training, SPARK_YARN_CACHE_FILES_VISIBILITIES -> PRIVATE,PRIVATE, SPARK_YARN_MODE -> true, SPARK_YARN_
CACHE_FILES_TIME_STAMPS -> 1714857954099,1714857954827, SPARK_LOG_URL_STDOUT -> http://localhost:8042/node/containerlogs/contai
ner_1714834153306_0001_01_000002/training/stdout?start=0, SPARK_YARN_CACHE_FILES -> hdfs://localhost:8020/user/training/.sparkS
taging/application_1714834153306_0001/spark-assembly-1.3.0-cdh5.4.3-hadoop2.6.0-cdh5.4.3.jar#__spark__.jar,hdfs://localhost:802
0/user/training/.sparkStaging/application_1714834153306_0001/CountJPGs.py#CountJPGs.py)
24/05/04 14:27:32 INFO yarn.ExecutorRunnable: Setting up executor with commands: List({{JAVA_HOME}}/bin/java, -server, -XX:OnOu
tOfMemoryError='kill %p', -Xms400m, -Xmx400m, -Djava.io.tmpdir={{PWD}}/tmp, '-Dspark.driver.port=43793', -Dspark.yarn.app.conta
iner.log.dir=<LOG_DIR>, org.apache.spark.executor.CoarseGrainedExecutorBackend, --driver-url, akka.tcp://sparkDriver@192.168.18
8.133:43793/user/CoarseGrainedScheduler, --executor-id, 1, --hostname, localhost, --cores, 1, --app-id, application_17148341533
06_0001, --user-class-path, file:$PWD/__app__.jar, 1>, <LOG_DIR>/stdout, 2>, <LOG_DIR>/stderr)
24/05/04 14:27:32 INFO impl.ContainerManagementProtocolProxy: Opening proxy : localhost:51315
24/05/04 14:30:31 INFO yarn.ApplicationMaster: Driver terminated or disconnected! Shutting down. Disassociated [akka.tcp://spar
kYarnAM@192.168.188.133:37454] -> [akka.tcp://sparkDriver@192.168.188.133:43793]
24/05/04 14:30:31 INFO yarn.ApplicationMaster: Final app status: SUCCEEDED, exitCode: 0
24/05/04 14:30:31 INFO yarn.ApplicationMaster: Unregistering ApplicationMaster with SUCCEEDED
24/05/04 14:30:31 INFO impl.AMRMClientImpl: Waiting for application to be successfully unregistered.
24/05/04 14:30:31 INFO yarn.ApplicationMaster: Deleting staging directory .sparkStaging/application_1714834153306_0001

LogType:stdout
Log Upload Time:Sat May 04 14:30:33 -0700 2024
LogLength:0
Log Contents:

[training@localhost spark-application]$ sxk230064
```

LAB CHAPTER 13: CONFIGURE A SPARK APPLICATION

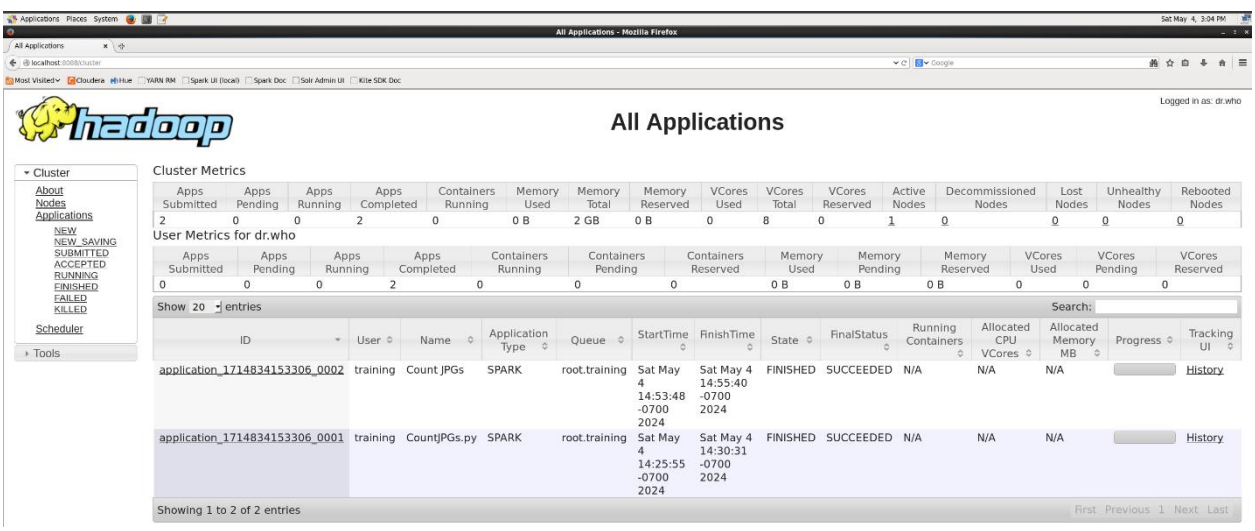
Set configuration options at the command line

1. I'm re-running the CountJPGs Python program I wrote in the previous exercise, but this time with a different application name that is “JPGsCount”
   spark-submit --master yarn-client --name 'JPGsCount' CountJPGs.py /loudacre/weblogs/*

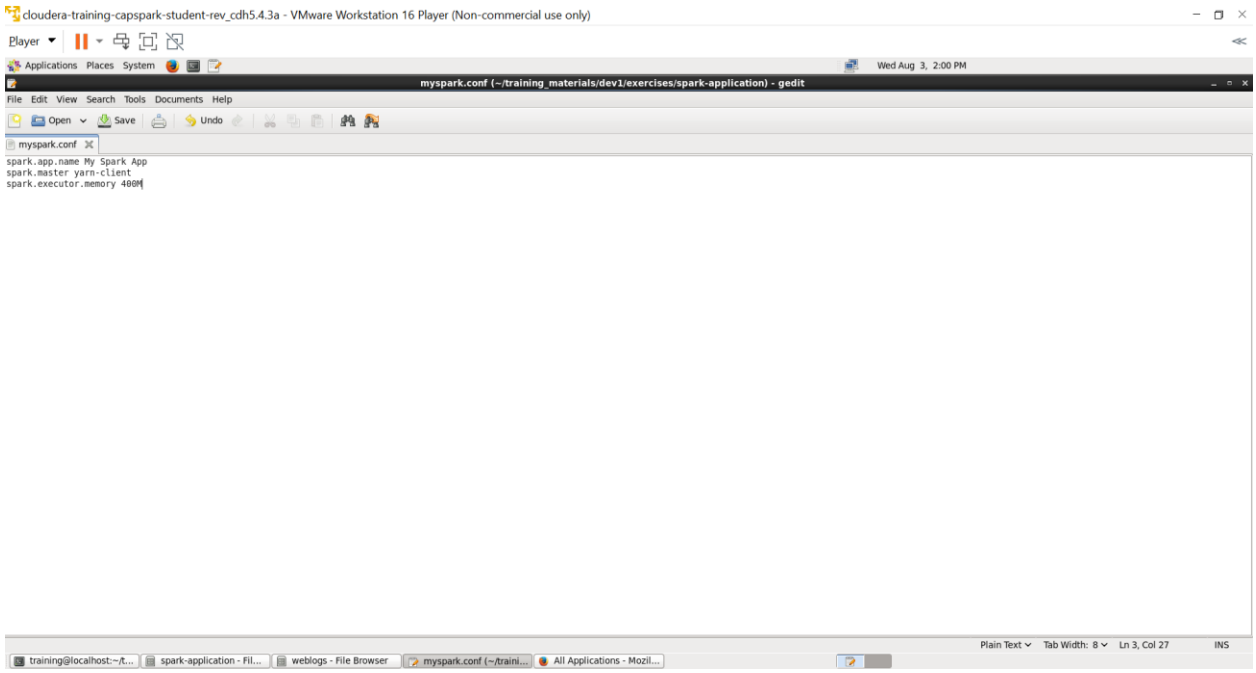
```
[training@localhost spark-application]$ spark-submit --master yarn-client \
>  --name 'Count JPGs' \
>  CountJPGs.py /loudacre/weblogs/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/04 14:53:32 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/04 14:53:32 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.16
8.188.133 instead (on interface eth0)
24/05/04 14:53:32 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/04 14:53:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
24/05/04 14:53:34 INFO spark.SecurityManager: Changing view acls to: training
24/05/04 14:53:34 INFO spark.SecurityManager: Changing modify acls to: training
24/05/04 14:53:34 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permi
ssions: Set(training); users with modify permissions: Set(training)
24/05/04 14:53:35 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/04 14:53:36 INFO Remoting: Starting remoting
24/05/04 14:53:36 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:37633]
24/05/04 14:53:36 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:37633]
24/05/04 14:53:36 INFO util.Utils: Successfully started service 'sparkDriver' on port 37633.
24/05/04 14:53:36 INFO spark.SparkEnv: Registering MapOutputTracker
24/05/04 14:53:36 INFO spark.SparkEnv: Registering BlockManagerMaster
24/05/04 14:53:36 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-9894132d-4b15-403c-b053-428e19f72b21/blo
```

```
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 14:55:39 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 14:55:39 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.188.133:4044
24/05/04 14:55:39 INFO scheduler.DAGScheduler: Stopping DAGScheduler
24/05/04 14:55:39 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
24/05/04 14:55:39 INFO cluster.YarnClientSchedulerBackend: Asking each executor to shut down
24/05/04 14:55:39 INFO cluster.YarnClientSchedulerBackend: Stopped
24/05/04 14:55:39 INFO spark.MapOutputTrackerMasterActor: MapOutputTrackerActor stopped!
24/05/04 14:55:40 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorActor: OutputCommitCoordinator stopped!
24/05/04 14:55:40 INFO storage.MemoryStore: MemoryStore cleared
24/05/04 14:55:40 INFO storage.BlockManager: BlockManager stopped
24/05/04 14:55:40 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/04 14:55:40 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/04 14:55:40 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remo
te transports.
24/05/04 14:55:40 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/04 14:55:40 INFO Remoting: Remoting shut down
24/05/04 14:55:40 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost spark-application]$ sxk230064
```

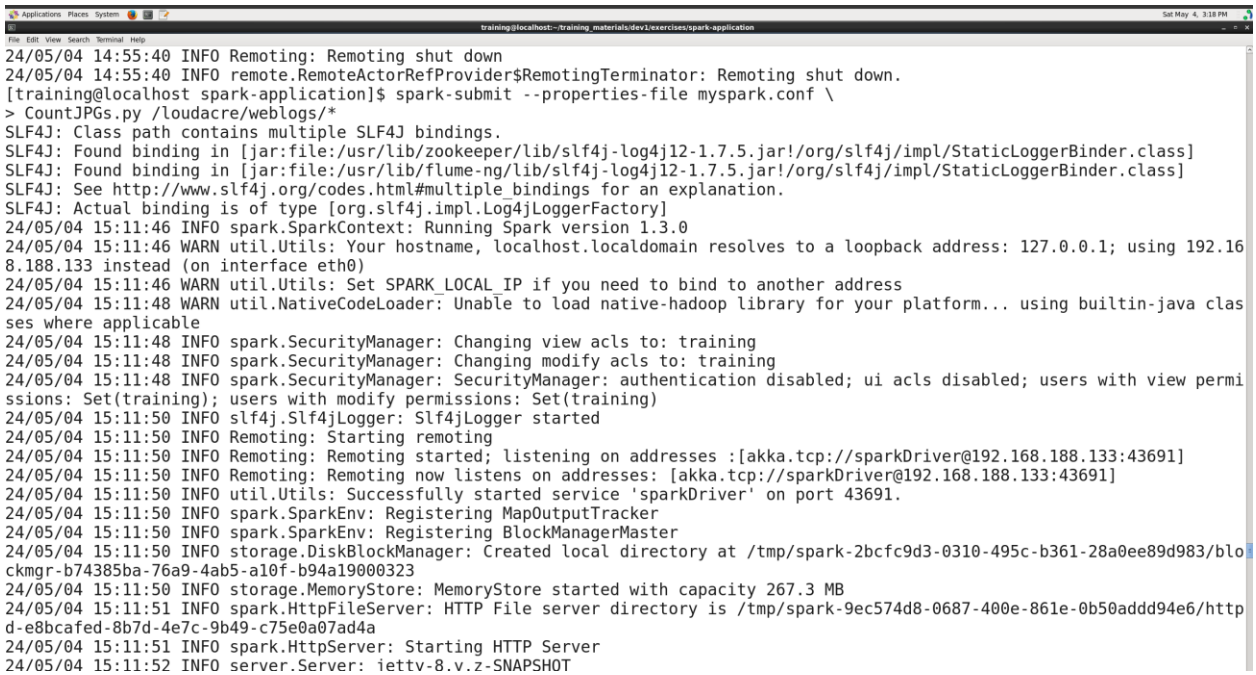
2. I visited the Resource Manager UI to see the application name listed as the one specified in the command line.



3. I created a configuration file named'myspark.conf' in the working directory using a text editor such as gedit, as shown below



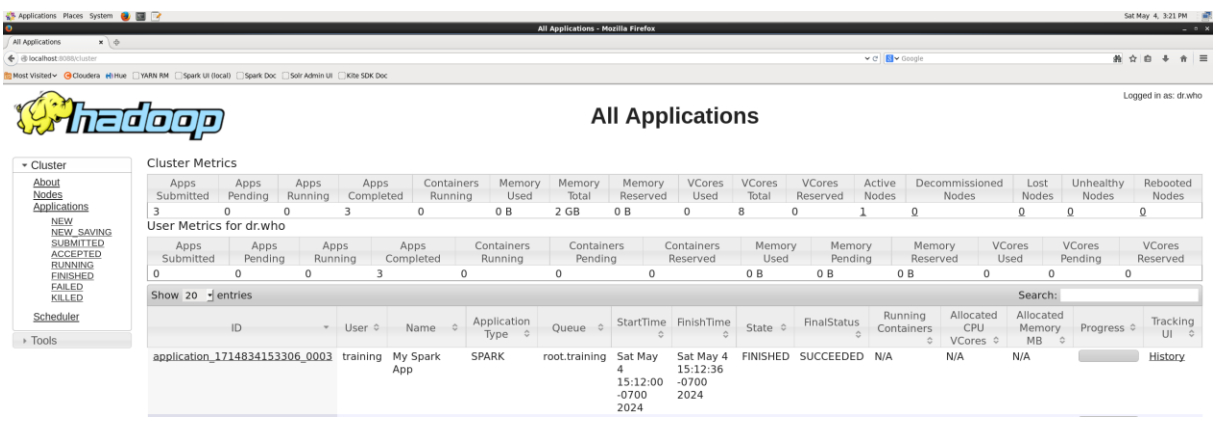
4. I have re-run the application, this time using the properties file instead of using the script options to configure Spark properties:
spark-submit --properties-file myspark.conf \ CountJPGs.py /loudacre/weblogs/*



```
24/05/04 14:55:40 INFO Remoting: Remoting shut down
24/05/04 14:55:40 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost spark-application]$ spark-submit --properties-file myspark.conf \
> CountJPGs.py /loudacre/weblogs/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/04 15:11:46 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/04 15:11:46 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.16
8.188.133 instead (on interface eth0)
24/05/04 15:11:46 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/04 15:11:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
24/05/04 15:11:48 INFO spark.SecurityManager: Changing view acls to: training
24/05/04 15:11:48 INFO spark.SecurityManager: Changing modify acls to: training
24/05/04 15:11:48 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permi
ssions: Set(training); users with modify permissions: Set(training)
24/05/04 15:11:50 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/04 15:11:50 INFO Remoting: Starting remoting
24/05/04 15:11:50 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:43691]
24/05/04 15:11:50 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:43691]
24/05/04 15:11:50 INFO util.Utils: Successfully started service 'sparkDriver' on port 43691.
24/05/04 15:11:50 INFO spark.SparkEnv: Registering MapOutputTracker
24/05/04 15:11:50 INFO spark.SparkEnv: Registering BlockManagerMaster
24/05/04 15:11:50 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-2bcfc9d3-0310-495c-b361-28a0ee89d983/blo
ckmgr-b74385ba-76a9-4ab5-a10f-b94a19000323
24/05/04 15:11:50 INFO storage.MemoryStore: MemoryStore started with capacity 267.3 MB
24/05/04 15:11:51 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-9ec574d8-0687-400e-861e-0b50addd94e6/http
d-e8bcafed-8b7d-4e7c-9b49-c75e0a07ad4a
24/05/04 15:11:51 INFO spark.HttpServer: Starting HTTP Server
24/05/04 15:11:52 INFO server.Server: jetty-8.y.z-SNAPSHOT
```

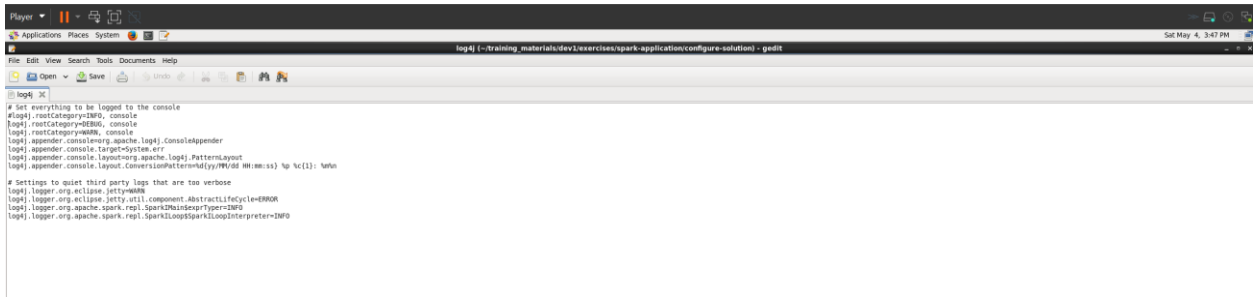
```
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 15:12:35 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 15:12:35 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.188.133:4044
24/05/04 15:12:35 INFO scheduler.DAGScheduler: Stopping DAGScheduler
24/05/04 15:12:35 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
24/05/04 15:12:35 INFO cluster.YarnClientSchedulerBackend: Asking each executor to shut down
24/05/04 15:12:35 INFO cluster.YarnClientSchedulerBackend: Stopped
24/05/04 15:12:35 INFO spark.MapOutputTrackerMasterActor: MapOutputTrackerActor stopped!
24/05/04 15:12:35 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorActor: OutputCommitCoordinator stopped!
24/05/04 15:12:35 INFO storage.MemoryStore: MemoryStore cleared
24/05/04 15:12:35 INFO storage.BlockManager: BlockManager stopped
24/05/04 15:12:35 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/04 15:12:35 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
24/05/04 15:12:35 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remo
te transports.
24/05/04 15:12:35 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/04 15:12:36 INFO Remoting: Remoting shut down
24/05/04 15:12:36 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[training@localhost spark-application]$ sxk230064
```

5. I used the YARN UI to confirm the application name is displayed as 'My Spark App' while the application was running, as shown below.

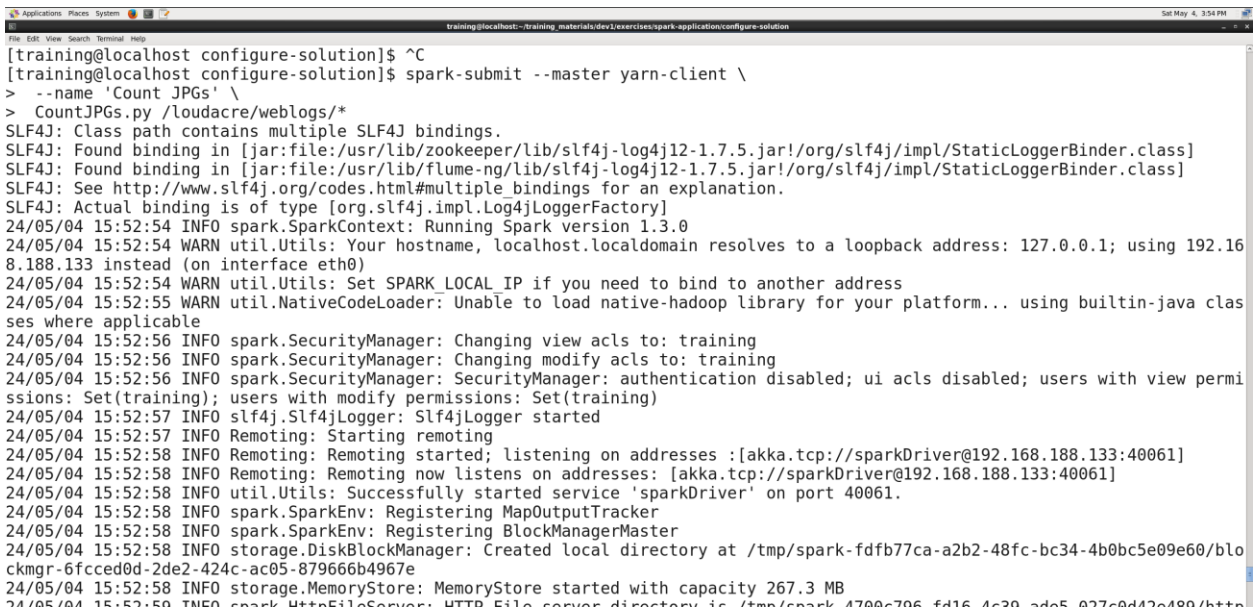


6. Copied the template file /etc/spark/conf/log4j.properties.template to log4j.properties in my exercise working directory.

7. I edited file log4j.properties, replaced log4j.rootCategory=INFO, console with log4j.rootCategory=DEBUG, console.

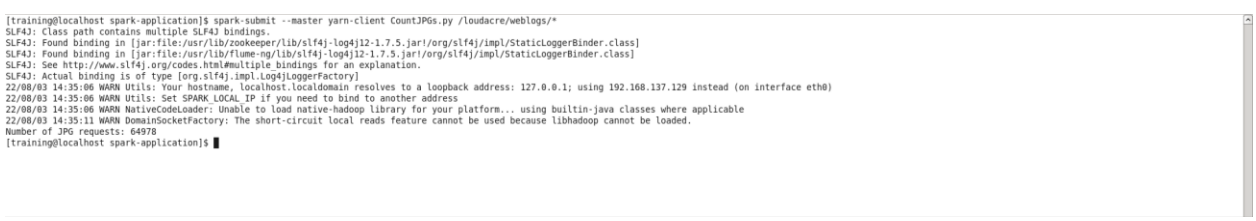


8. I re-ran the Spark application with the new logging level 'DEBUG' as shown below, spark-submit --master yarn-client CountJPGs.py /loudacre/weblogs/*



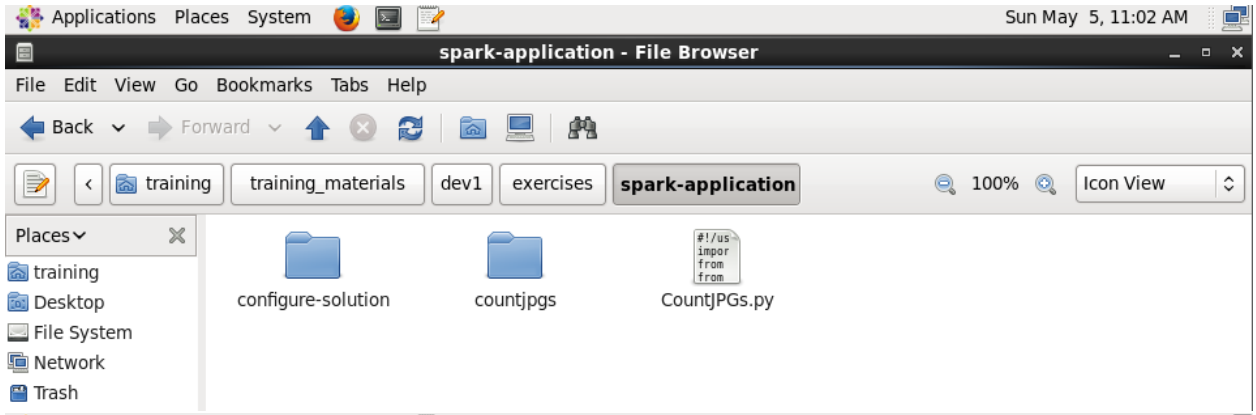
We notice that the output contains both 'info' and 'debug' messages.

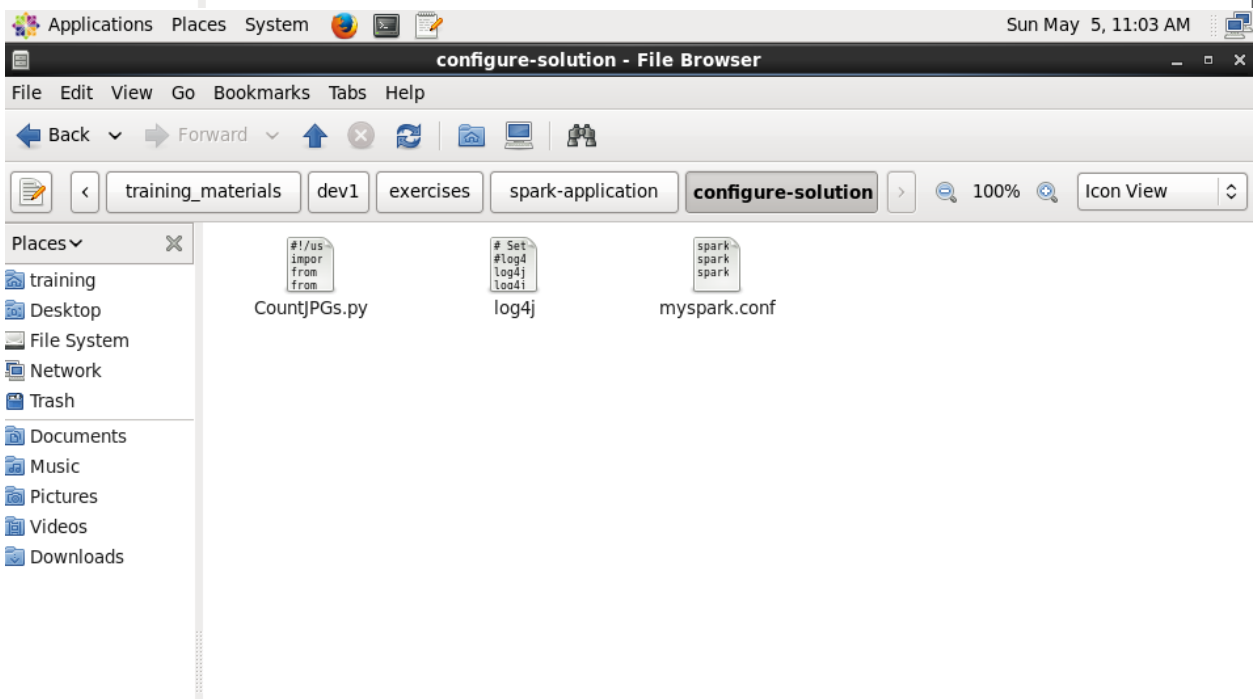
9. I then ran the application again, replacing '=DEBUG' with '=WARN'. As a result, neither INFO nor DEBUG messages were displayed in the output. As shown below, this removes all distracting messaging.

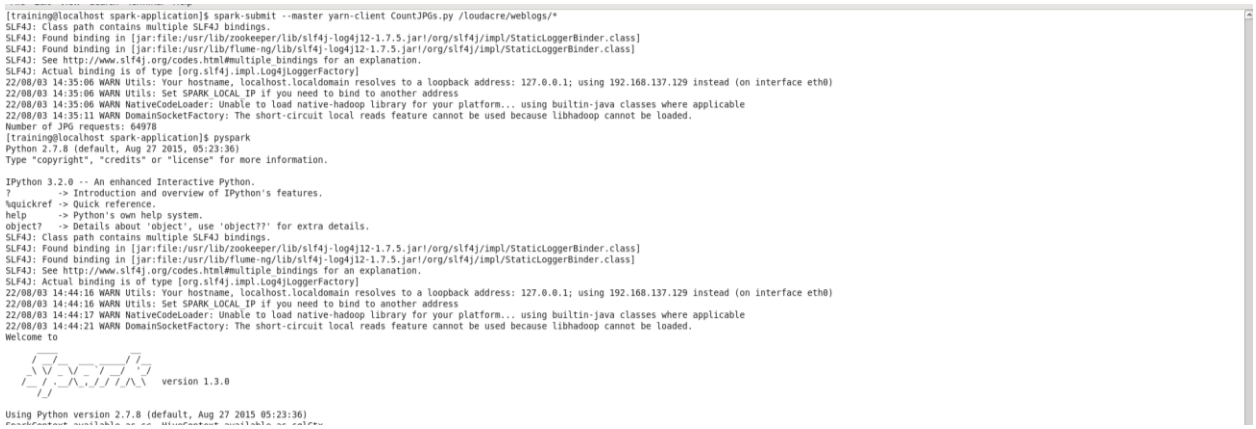


This time I noticed that there are no INFO or DEBUG messages displayed, only WARN messages.

10. Finally, In this step I saved the configuration file to my working directory and launched Spark Shell. As you can see below, this is a short output with only WARN messages.







```
[training@localhost spark-application]$ spark-submit --master yarn-client CountJPGs.py /loudacre/weblogs/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/08/03 14:35:06 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.137.129 instead (on interface eth0)
22/08/03 14:35:06 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
22/08/03 14:35:06 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/08/03 14:35:11 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
Number of JPG requests: 64970
[training@localhost spark-application]$ pyspark
Python 2.7.8 (default, Aug 27 2015, 05:23:36)
Type "copyright", "credits" or "license" for more information.

IPython 3.2.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/08/03 14:44:16 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.137.129 instead (on interface eth0)
22/08/03 14:44:16 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
22/08/03 14:44:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/08/03 14:44:21 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
```

LAB CHAPTER 14: VIEW JOBS AND STAGES IN THE SPARK APPLICATION UI

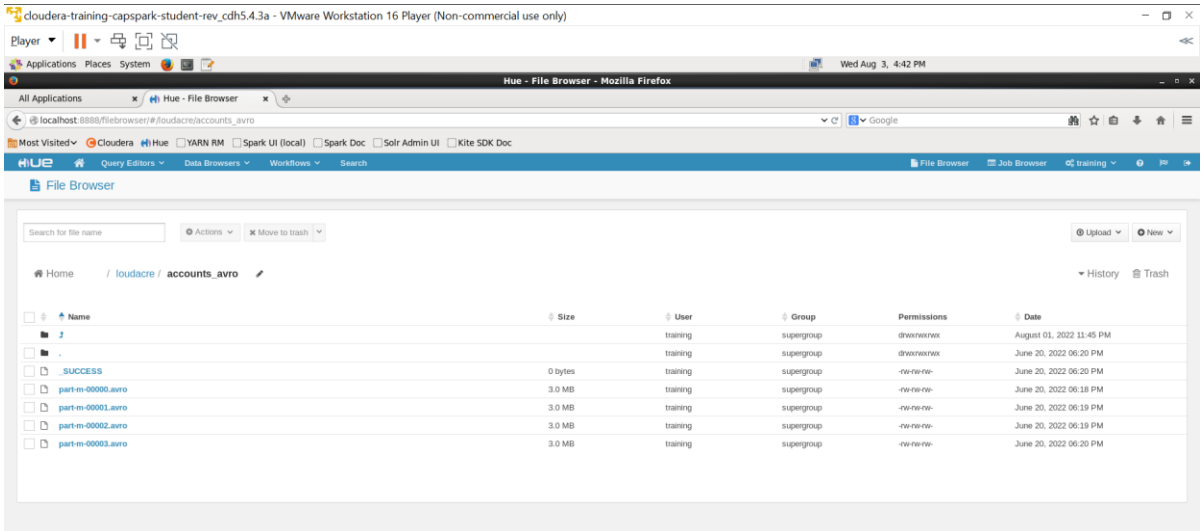
Explore Partitioning of File-Based RDDs

1. To simulate a more realistic multi-node cluster, I ran in local mode with 2 threads

```
[training@localhost configure-solution]$ pyspark --master local[2]
Python 2.7.8 (default, Aug 27 2015, 05:23:36)
Type "copyright", "credits" or "license" for more information.

IPython 3.2.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/04 16:04:16 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/04 16:04:16 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.188.133 instead (on interface eth0)
24/05/04 16:04:16 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/04 16:04:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/05/04 16:04:17 INFO spark.SecurityManager: Changing view acls to: training
24/05/04 16:04:17 INFO spark.SecurityManager: Changing modify acls to: training
24/05/04 16:04:17 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify permissions:
Set(training)
24/05/04 16:04:18 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/04 16:04:19 INFO Remoting: Starting remoting
```

```
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 16:04:21 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 16:04:21 WARN util.Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
24/05/04 16:04:21 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/04 16:04:21 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4044
24/05/04 16:04:21 INFO util.Utils: Successfully started service 'SparkUI' on port 4044.
24/05/04 16:04:21 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4044
24/05/04 16:04:21 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/04 16:04:21 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:49472/user/HeartbeatReceiver
24/05/04 16:04:22 INFO netty.NettyBlockTransferService: Server created on 59196
24/05/04 16:04:22 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/04 16:04:22 INFO storage.BlockManagerMasterActor: Registering block manager localhost:59196 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 59196)
24/05/04 16:04:22 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/04 16:04:24 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/04 16:04:24 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714863861559
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: sxk230064
```

2. Then, I reviewed the accounts dataset using Hue



3. Then I made an RDD out of a single file, which I named 'toDebugString.' As shown below, there are two partitions
   accounts=sc. \ textFile("/loudacre/accounts/part-m-00000")
   print accounts.toDebugString()

4. Repeated the above step with three partitions
   accounts=sc. \
   textFile("/loudacre/accounts/part-m-00000",3)
   print accounts.toDebugString()

```
In [1]: accounts=sc. \
   ...:  textFile("/loudacre/accounts/part-m-00000")
24/05/04 16:23:57 INFO storage.MemoryStore: ensureFreeSpace(280171) called with curMem=0, maxMem=280248975
24/05/04 16:23:57 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 273.6 KB, free 267.0 MB)
24/05/04 16:23:58 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=280171, maxMem=280248975
24/05/04 16:23:58 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 20.7 KB, free 267.0 MB)
24/05/04 16:23:58 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:23:58 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/04 16:23:58 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:-2

In [2]: print accounts.toDebugString()
24/05/04 16:24:12 INFO mapred.FileInputFormat: Total input paths to process : 1
(2) /loudacre/accounts/part-m-00000 MapPartitionsRDD[1] at textFile at NativeMethodAccessorImpl.java:-2 []
 |  /loudacre/accounts/part-m-00000 HadoopRDD[0] at textFile at NativeMethodAccessorImpl.java:-2 []

In [3]: sxk230064
```

5. In this step I created an RDD with all the files in the “accounts” dataset. The number of files is 7 same as the number of partitions.
   accountsfile= "/loudacre/accounts/*"
   accounts=sc.textFile(accountsfile)
   print(accounts.toDebugString())

```
In [3]: accountsfile= "/loudacre/accounts/*"

In [4]: accounts=sc.textFile(accountsfile)
24/05/04 16:30:16 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=301375, maxMem=280248975
24/05/04 16:30:16 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 273.7 KB, free 266.7 MB)
24/05/04 16:30:16 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=581618, maxMem=280248975
24/05/04 16:30:16 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.7 MB)
24/05/04 16:30:16 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:30:16 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 16:30:16 INFO spark.SparkContext: Created broadcast 1 from textFile at NativeMethodAccessorImpl.java:-2

In [5]: print(accounts.toDebugString())
24/05/04 16:30:40 INFO mapred.FileInputFormat: Total input paths to process : 7
(7) /loudacre/accounts/* MapPartitionsRDD[3] at textFile at NativeMethodAccessorImpl.java:-2 []
 |  /loudacre/accounts/* HadoopRDD[2] at textFile at NativeMethodAccessorImpl.java:-2 []

In [6]: accountsByID = accounts \
   ...: .map(lambda s: s.split(',')) \
   ...:  .map(lambda values: \
   ...:  (values[0],values[4] + ',' + values[3]))

In [7]: userreqs = sc \
   ...:  .textFile("/loudacre/weblogs/*6") \
   ...:  .map(lambda line: line.split()) \
   ...:  .map(lambda words: (words[2],1)) \
   ...:  .reduceByKey(lambda v1,v2: v1+v2)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=602822, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=883065, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:31:54 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/04 16:31:54 INFO spark.SparkContext: Created broadcast 2 from textFile at NativeMethodAccessorImpl.java:-2
24/05/04 16:31:54 INFO mapred.FileInputFormat: Total input paths to process : 1

In [8]: sxk230064
```

Set up the job

6. First, as shown below, I created an RDD of accounts, keyed by ID and with first and last names as values
   accountsByID = accounts \
   .map(lambda s: s.split(',')) \
   .map(lambda values: \ (values[0],values[4] + ',' + values[3]))

```
In [6]: accountsByID = accounts \
   ...: .map(lambda s: s.split(',')) \
   ...:  .map(lambda values: \
   ...:  (values[0],values[4] + ',' + values[3]))
```

7. The first step is to construct a userreqs RDD with the total number of web hits for each user ID:

```
In [5]: print(accounts.toDebugString())
24/05/04 16:30:40 INFO mapred.FileInputFormat: Total input paths to process : 7
(7) /loudacre/accounts/* MapPartitionsRDD[3] at textFile at NativeMethodAccessorImpl.java:-2 []
 |  /loudacre/accounts/* HadoopRDD[2] at textFile at NativeMethodAccessorImpl.java:-2 []

In [6]: accountsByID = accounts \
   ...: .map(lambda s: s.split(',')) \
   ...:  .map(lambda values: \
   ...:  (values[0],values[4] + ',' + values[3]))

In [7]: userreqs = sc \
   ...:  .textFile("/loudacre/weblogs/*6") \
   ...:  .map(lambda line: line.split()) \
   ...:  .map(lambda words: (words[2],1)) \
   ...:  .reduceByKey(lambda v1,v2: v1+v2)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=602822, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=883065, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:31:54 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/04 16:31:54 INFO spark.SparkContext: Created broadcast 2 from textFile at NativeMethodAccessorImpl.java:-2
24/05/04 16:31:54 INFO mapred.FileInputFormat: Total input paths to process : 1

In [8]: sxk230064
```

8. In this step I joined the two RDDs by user ID, and construct a new RDD based on first name, last name and total hits:

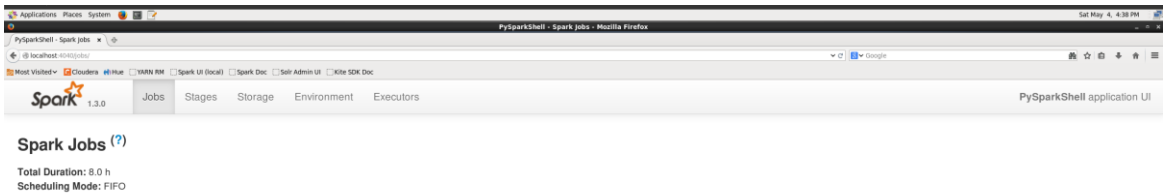
```
In [7]: userreqs = sc \
   ...:  .textFile("/loudacre/weblogs/*6") \
   ...:  .map(lambda line: line.split()) \
   ...:  .map(lambda words: (words[2],1)) \
   ...:  .reduceByKey(lambda v1,v2: v1+v2)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=602822, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=883065, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:31:54 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/04 16:31:54 INFO spark.SparkContext: Created broadcast 2 from textFile at NativeMethodAccessorImpl.java:-2
24/05/04 16:31:54 INFO mapred.FileInputFormat: Total input paths to process : 1

In [8]: sxk230064
```

9. After calling the toDebugString function, I print the results and review them as follows

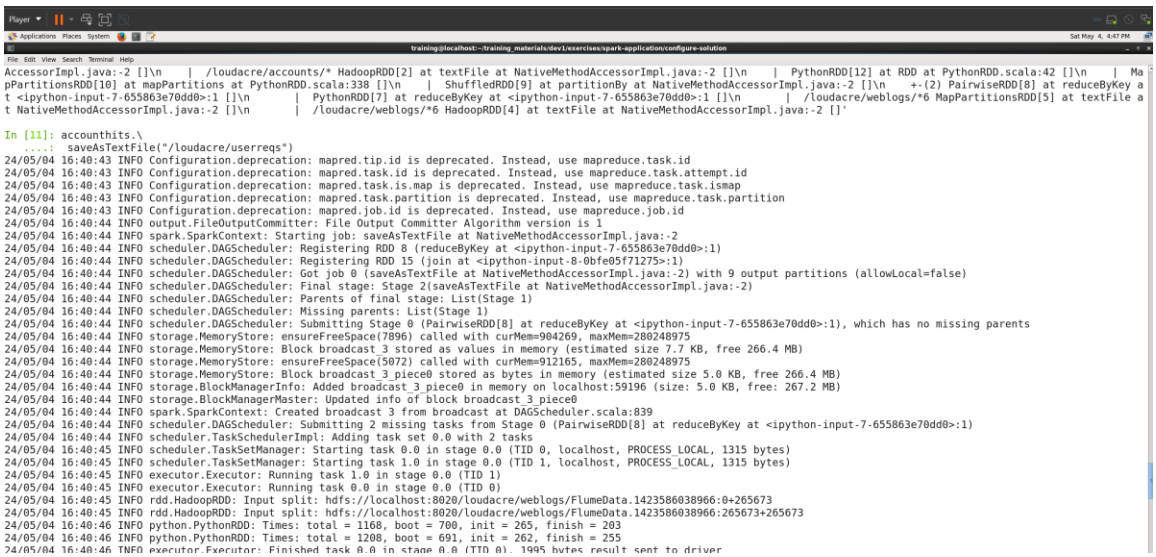
```
Applications  Places  System                                                                                    Sat May 4, 4:37 PM
training@localhost:~/training_materials/dev1/exercises/spark-application/configure-solution
File  Edit  View  Search  Terminal  Help
24/05/04 16:30:16 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=581618, maxMem=280248975
24/05/04 16:30:16 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.7 MB)
24/05/04 16:30:16 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:30:16 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 16:30:16 INFO spark.SparkContext: Created broadcast 1 from textFile at NativeMethodAccessorImpl.java:-2

In [5]: print(accounts.toDebugString())
24/05/04 16:30:40 INFO mapred.FileInputFormat: Total input paths to process : 7
(7) /loudacre/accounts/* MapPartitionsRDD[3] at textFile at NativeMethodAccessorImpl.java:-2 []
 |  /loudacre/accounts/* HadoopRDD[2] at textFile at NativeMethodAccessorImpl.java:-2 []

In [6]: accountsByID = accounts \
   ...: .map(lambda s: s.split(',')) \
   ...:  .map(lambda values: \
   ...:  (values[0],values[4] + ',' + values[3]))

In [7]: userreqs = sc \
   ...:  .textFile("/loudacre/weblogs/*6") \
   ...:  .map(lambda line: line.split()) \
   ...:  .map(lambda words: (words[2],1)) \
   ...:  .reduceByKey(lambda v1,v2: v1+v2)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=602822, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=883065, maxMem=280248975
24/05/04 16:31:54 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
24/05/04 16:31:54 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:59196 (size: 20.7 KB, free: 267.2 MB)
24/05/04 16:31:54 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/04 16:31:54 INFO spark.SparkContext: Created broadcast 2 from textFile at NativeMethodAccessorImpl.java:-2
24/05/04 16:31:54 INFO mapred.FileInputFormat: Total input paths to process : 1

In [8]: accounthits = accountsByID.join(userreqs)\
   ...: .values()

In [9]: print(accounts.toDebugString())
(7) /loudacre/accounts/* MapPartitionsRDD[3] at textFile at NativeMethodAccessorImpl.java:-2 []
 |  /loudacre/accounts/* HadoopRDD[2] at textFile at NativeMethodAccessorImpl.java:-2 []

In [10]: accounthits.toDebugString()
Out[10]: '(9) PythonRDD[18] at RDD at PythonRDD.scala:42 []\n |  MapPartitionsRDD[17] at mapPartitions at PythonRDD.scala:338 []\n |  ShuffledRDD[16] at partitionBy at NativeMethodA
ccessorImpl.java:-2 []\n +-(9) PairwiseRDD[15] at join at <ipython-input-8-0bfe05f71275>:1 []\n    |  PythonRDD[14] at join at <ipython-input-8-0bfe05f71275>:1 []\n    |  UnionRDD[1
3] at union at NativeMethodAccessorImpl.java:-2 []\n    |  PythonRDD[11] at RDD at PythonRDD.scala:42 []\n    |  /loudacre/accounts/* MapPartitionsRDD[3] at textFile at NativeMethod
AccessorImpl.java:-2 []\n    |  /loudacre/accounts/* HadoopRDD[2] at textFile at NativeMethodAccessorImpl.java:-2 []\n    |  PythonRDD[12] at RDD at PythonRDD.scala:42 []\n    |  Ma
pPartitionsRDD[10] at mapPartitions at PythonRDD.scala:338 []\n    |  ShuffledRDD[9] at partitionBy at NativeMethodAccessorImpl.java:-2 []\n    +-(2) PairwiseRDD[8] at reduceByKey a
t <ipython-input-7-655863e70dd0>:1 []\n       |  PythonRDD[7] at reduceByKey at <ipython-input-7-655863e70dd0>:1 []\n       |  /loudacre/weblogs/*6 MapPartitionsRDD[5] at textFile a
t NativeMethodAccessorImpl.java:-2 []\n       |  /loudacre/weblogs/*6 HadoopRDD[4] at textFile at NativeMethodAccessorImpl.java:-2 []'

In [11]: sxk230064
```

- There are 3 stages in the job, as noted by indentations.
- Stage 3 is dependent on Stage 2 and Stage 2 is dependent on Stage 1.
- Stage 1 has 32 tasks, stage 2 and 3 have 38 tasks making a total of 107 tasks.

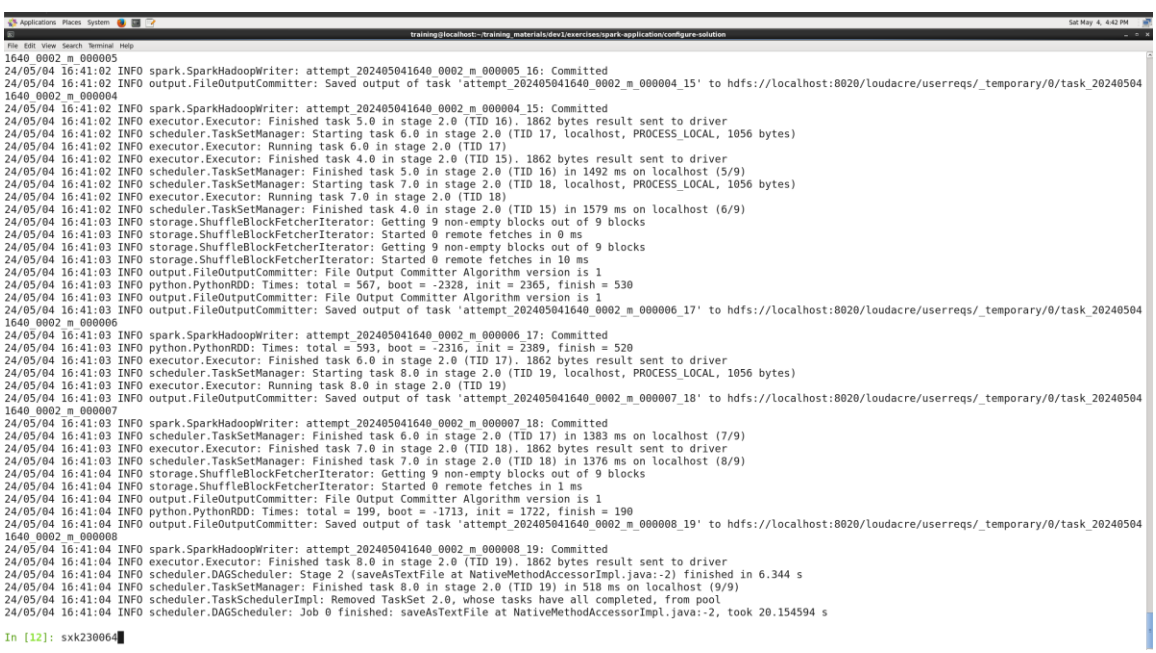
10. Opened the Spark Application UI to check if there is any job running.



11. I ran the following command in the command prompt.
accounthits.\ saveAsTextFile("/loudacre/userreqs")



```
AccessorImpl.java:-2 []\n    |  /loudacre/accounts/* HadoopRDD[2] at textFile at NativeMethodAccessorImpl.java:-2 []\n    |  PythonRDD[12] at RDD at PythonRDD.scala:42 []\n    |  MapPartitionsRDD[10] at mapPartitions at PythonRDD.scala:338 []\n    |  ShuffledRDD[9] at partitionBy at NativeMethodAccessorImpl.java:-2 []\n    +-(2) PairwiseRDD[8] at reduceByKey at <ipython-input-7-655863e70dd0>:1 []\n       |  PythonRDD[7] at reduceByKey at <ipython-input-7-655863e70dd0>:1 []\n       |  /loudacre/weblogs/*6 MapPartitionsRDD[5] at textFile at NativeMethodAccessorImpl.java:-2 []\n       |  /loudacre/weblogs/*6 HadoopRDD[4] at textFile at NativeMethodAccessorImpl.java:-2 []'

In [11]: accounthits.\
   ....:  saveAsTextFile("/loudacre/userreqs")
24/05/04 16:40:43 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/05/04 16:40:43 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/05/04 16:40:43 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/05/04 16:40:43 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/05/04 16:40:43 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/05/04 16:40:44 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/04 16:40:44 INFO spark.SparkContext: Starting job: saveAsTextFile at NativeMethodAccessorImpl.java:-2
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Registering RDD 8 (reduceByKey at <ipython-input-7-655863e70dd0>:1)
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Registering RDD 15 (join at <ipython-input-8-0bfe05f71275>:1)
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) with 9 output partitions (allowLocal=false)
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Final stage: Stage 2(saveAsTextFile at NativeMethodAccessorImpl.java:-2)
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 1)
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Missing parents: List(Stage 1)
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Submitting Stage 0 (PairwiseRDD[8] at reduceByKey at <ipython-input-7-655863e70dd0>:1), which has no missing parents
24/05/04 16:40:44 INFO storage.MemoryStore: ensureFreeSpace(7896) called with curMem=904269, maxMem=280248975
24/05/04 16:40:44 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 7.7 KB, free 266.4 MB)
24/05/04 16:40:44 INFO storage.MemoryStore: ensureFreeSpace(5072) called with curMem=912165, maxMem=280248975
24/05/04 16:40:44 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 5.0 KB, free 266.4 MB)
24/05/04 16:40:44 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:59196 (size: 5.0 KB, free: 267.2 MB)
24/05/04 16:40:44 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
24/05/04 16:40:44 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:839
24/05/04 16:40:44 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from Stage 0 (PairwiseRDD[8] at reduceByKey at <ipython-input-7-655863e70dd0>:1)
24/05/04 16:40:44 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 2 tasks
24/05/04 16:40:45 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1315 bytes)
24/05/04 16:40:45 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, localhost, PROCESS_LOCAL, 1315 bytes)
24/05/04 16:40:45 INFO executor.Executor: Running task 1.0 in stage 0.0 (TID 1)
24/05/04 16:40:45 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/04 16:40:45 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:0+265673
24/05/04 16:40:45 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/weblogs/FlumeData.1423586038966:265673+265673
24/05/04 16:40:46 INFO python.PythonRDD: Times: total = 1168, boot = 700, init = 265, finish = 203
24/05/04 16:40:46 INFO python.PythonRDD: Times: total = 1208, boot = 691, init = 262, finish = 255
24/05/04 16:40:46 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 1995 bytes result sent to driver
```

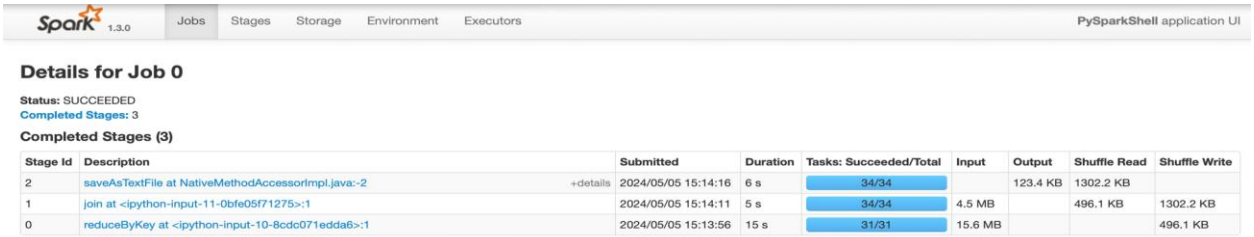


```
1640_0002_m_000005
24/05/04 16:41:02 INFO spark.SparkHadoopWriter: attempt_202405041640_0002_m_000005_16: Committed
24/05/04 16:41:02 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405041640_0002_m_000004_15' to hdfs://localhost:8020/loudacre/userreqs/_temporary/0/task_20240504
1640_0002_m_000004
24/05/04 16:41:02 INFO spark.SparkHadoopWriter: attempt_202405041640_0002_m_000004_15: Committed
24/05/04 16:41:02 INFO executor.Executor: Finished task 5.0 in stage 2.0 (TID 16). 1862 bytes result sent to driver
24/05/04 16:41:02 INFO scheduler.TaskSetManager: Starting task 6.0 in stage 2.0 (TID 17, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 16:41:02 INFO executor.Executor: Running task 6.0 in stage 2.0 (TID 17)
24/05/04 16:41:02 INFO executor.Executor: Finished task 4.0 in stage 2.0 (TID 15). 1862 bytes result sent to driver
24/05/04 16:41:02 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 2.0 (TID 16) in 1492 ms on localhost (5/9)
24/05/04 16:41:02 INFO scheduler.TaskSetManager: Starting task 7.0 in stage 2.0 (TID 18, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 16:41:02 INFO executor.Executor: Running task 7.0 in stage 2.0 (TID 18)
24/05/04 16:41:02 INFO scheduler.TaskSetManager: Finished task 4.0 in stage 2.0 (TID 15) in 1579 ms on localhost (6/9)
24/05/04 16:41:03 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 16:41:03 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
24/05/04 16:41:03 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 16:41:03 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 10 ms
24/05/04 16:41:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/04 16:41:03 INFO python.PythonRDD: Times: total = 567, boot = -2328, init = 2365, finish = 530
24/05/04 16:41:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/04 16:41:03 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405041640_0002_m_000006_17' to hdfs://localhost:8020/loudacre/userreqs/_temporary/0/task_20240504
1640_0002_m_000006
24/05/04 16:41:03 INFO spark.SparkHadoopWriter: attempt_202405041640_0002_m_000006_17: Committed
24/05/04 16:41:03 INFO python.PythonRDD: Times: total = 593, boot = -2316, init = 2389, finish = 520
24/05/04 16:41:03 INFO executor.Executor: Finished task 6.0 in stage 2.0 (TID 17). 1862 bytes result sent to driver
24/05/04 16:41:03 INFO scheduler.TaskSetManager: Starting task 8.0 in stage 2.0 (TID 19, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 16:41:03 INFO executor.Executor: Running task 8.0 in stage 2.0 (TID 19)
24/05/04 16:41:03 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405041640_0002_m_000007_18' to hdfs://localhost:8020/loudacre/userreqs/_temporary/0/task_20240504
1640_0002_m_000007
24/05/04 16:41:03 INFO spark.SparkHadoopWriter: attempt_202405041640_0002_m_000007_18: Committed
24/05/04 16:41:03 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 2.0 (TID 17) in 1383 ms on localhost (7/9)
24/05/04 16:41:03 INFO executor.Executor: Finished task 7.0 in stage 2.0 (TID 18). 1862 bytes result sent to driver
24/05/04 16:41:03 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 2.0 (TID 18) in 1376 ms on localhost (8/9)
24/05/04 16:41:04 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 16:41:04 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/05/04 16:41:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/04 16:41:04 INFO python.PythonRDD: Times: total = 199, boot = -1713, init = 1722, finish = 190
24/05/04 16:41:04 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405041640_0002_m_000008_19' to hdfs://localhost:8020/loudacre/userreqs/_temporary/0/task_20240504
1640_0002_m_000008
24/05/04 16:41:04 INFO spark.SparkHadoopWriter: attempt_202405041640_0002_m_000008_19: Committed
24/05/04 16:41:04 INFO executor.Executor: Finished task 8.0 in stage 2.0 (TID 19). 1862 bytes result sent to driver
24/05/04 16:41:04 INFO scheduler.DAGScheduler: Stage 2 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 6.344 s
24/05/04 16:41:04 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 2.0 (TID 19) in 518 ms on localhost (9/9)
24/05/04 16:41:04 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
24/05/04 16:41:04 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 20.154594 s

In [12]: sxk230064
```

12. Reloaded the Spark UI Jobs page in the browser.

13. Then I clicked on job description and received the information that was expected from toDebugString output.
14.

- There are 3 stages in the job.
- Stage 3 is dependent on Stage 2 and Stage 2 is dependent on Stage 1. Stage 1 was completed first, followed by 2 and 3
    - Stage 1: 05:24:09
    - Stage 2: 05:24:21
    - Stage 3: 05:24:39
- There are total of 120 tasks in total.
    - Stage 1: 32
    - Stage 2: 39
    - Stage 3: 39
- The amount of data copied between tasks is indicated in the Shuffle Read and Write column.



15. Next, I clicked on stages tab for viewing more details

Spark 1.3.0 | Jobs | Stages | Storage | Environment | Executors | PySparkShell application UI

Details for Stage 0

Total task time across all tasks: 29 s
Input Size / Records: 15.6 MB / 107962
Shuffle write: 496.1 KB / 2883

▸ Show additional metrics

Summary Metrics for 31 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.1 s	0.2 s	0.3 s	0.6 s	7 s
GC Time	0 ms	0 ms	0 ms	0 ms	0.7 s
Input Size / Records	515.4 KB / 3472	515.4 KB / 3480	515.5 KB / 3482	515.5 KB / 3486	515.6 KB / 3495
Shuffle Write Size / Records	15.1 KB / 93	15.8 KB / 93	16.1 KB / 93	16.3 KB / 93	16.6 KB / 93

Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Write Size / Records
<driver>	localhost:33877	30 s	31	0	31	15.6 MB / 107962	496.1 KB / 2883

Tasks

Index	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	GC Time	Input Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	0	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:13:56	2 s		515.4 KB (hadoop) / 3492	25 ms	15.1 KB / 93	
1	1	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:13:56	2 s		515.6 KB (hadoop) / 3492	30 ms	15.7 KB / 93	
2	2	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:13:58	7 s		515.5 KB (hadoop) / 3481	5 ms	16.3 KB / 93	
3	3	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:13:58	7 s		515.5 KB (hadoop) / 3482	6 ms	16.3 KB / 93	
4	4	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:14:05	0.6 s	35 ms	515.4 KB (hadoop) / 3480	2 ms	16.6 KB / 93	
5	5	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:14:05	0.6 s	35 ms	515.5 KB (hadoop) / 3472	7 ms	16.5 KB / 93	
6	6	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:14:06	0.3 s		515.5 KB (hadoop) / 3483	9 ms	16.4 KB / 93	
7	7	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:14:06	0.3 s		515.4 KB (hadoop) / 3488	6 ms	15.5 KB / 93	
8	8	0	SUCCESS	ANY	<driver> / localhost	2024/05/05 15:14:06	0.4 s		515.5 KB (hadoop) / 3479	3 ms	15.9 KB / 93	

Spark 1.3.0 | Jobs | Stages | Storage | Environment | Executors | PySparkShell application UI

Details for Stage 1

Total task time across all tasks: 9 s
Input Size / Records: 4.5 MB / 32440
Shuffle read: 496.1 KB / 2883
Shuffle write: 1302.2 KB / 2817

▸ Show additional metrics

Summary Metrics for 34 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.2 s	0.2 s	0.2 s	0.2 s	0.8 s
GC Time	0 ms	0 ms	0 ms	6 ms	0.3 s
Input Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 0	0.0 B / 0	1536.0 KB / 10837
Shuffle Read Size / Records	0.0 B / 0	15.1 KB / 93	16.0 KB / 93	16.5 KB / 93	17.5 KB / 93
Shuffle Write Size / Records	12.1 KB / 66	12.8 KB / 68	13.0 KB / 68	13.3 KB / 68	301.0 KB / 238

Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Read Size / Records	Shuffle Write Size / Records
<driver>	localhost:33877	9 s	34	0	34	4.5 MB / 32440	496.1 KB / 2883	1302.2 KB / 2817

Tasks

Index	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	GC Time	Input Size / Records	Shuffle Read Size / Records	Write Time	Shuffle Write Size / Records	Errors
3	31	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:11	0.8 s	0.3 s	/	15.7 KB / 93	2 ms	13.1 KB / 68	
4	32	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:11	0.7 s	0.3 s	/	16.1 KB / 93	9 ms	13.1 KB / 68	
5	33	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:12	0.2 s		/	16.6 KB / 93	2 ms	13.0 KB / 68	
6	34	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:12	0.2 s		/	17.3 KB / 93	4 ms	13.3 KB / 68	

Spark 1.3.0 Jobs Stages Storage Environment Executors PySparkShell application UI

Details for Stage 2

Total task time across all tasks: 12 s
Output: 123.4 KB / 5689
Shuffle read: 1302.2 KB / 2817

▸ Show additional metrics

Summary Metrics for 34 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.2 s	0.2 s	0.3 s	0.4 s	0.8 s
GC Time	0 ms	0 ms	0 ms	9 ms	14 ms
Output Size / Records	3.3 KB / 152	3.4 KB / 158	3.6 KB / 167	3.8 KB / 176	4.0 KB / 186
Shuffle Read Size / Records	36.4 KB / 82	37.6 KB / 83	38.4 KB / 83	39.1 KB / 83	40.1 KB / 83

Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Output Size / Records	Shuffle Read Size / Records
<driver>	localhost:33877	12 s	34	0	34	123.4 KB / 5689	1302.2 KB / 2817

Tasks

Index	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	GC Time	Output Size / Records	Shuffle Read Size / Records	Errors
0	65	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:16	0.6 s		3.6 KB / 169	38.9 KB / 83	
1	66	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:16	0.6 s		3.9 KB / 181	39.6 KB / 83	
2	67	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:17	0.6 s	13 ms	3.4 KB / 155	36.8 KB / 83	
3	68	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:17	0.6 s	13 ms	3.4 KB / 159	37.3 KB / 83	
4	69	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:17	0.4 s		4.0 KB / 183	37.7 KB / 83	
5	70	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:17	0.4 s		3.5 KB / 160	37.7 KB / 82	
6	71	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:18	0.4 s	14 ms	3.7 KB / 172	38.2 KB / 83	
7	72	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:18	0.4 s	14 ms	3.8 KB / 176	37.6 KB / 83	
8	73	0	SUCCESS	PROCESS_LOCAL	<driver> / localhost	2024/05/05 15:14:18	0.3 s		3.5 KB / 163	37.9 KB / 83	

- The Summary Metrics area shows you how much time was spend on various steps. This can help you narrow down performance problems
- Each task is listed in the task pane with a location level that indicates whether the process was running on the same node as the partition, and other statistics
- The Real-World Executor ID column would show the worker nodes running the tasks. In this case all tasks run on the same host.

16. When the job is finished, I went back to the Jobs tab to see the final statistics and the time it took to complete the job, as in the below screenshot.

cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 16 Player (Non-commercial use only)

Player

Applications Places System Wed Aug 3, 9:06 PM

PySparkShell - Spark Jobs - Mozilla Firefox

All Applications | Hue - File Browser | PySparkShell - Spark Jobs

localhost:4040/jobs/

Most Visited Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc

Spark 1.3.0 Jobs Stages Storage Environment Executors PySparkShell application UI

Spark Jobs (?)

Total Duration: 4.5 h
Scheduling Mode: FIFO
Completed Jobs: 1

Completed Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	saveAsTextFile at NativeMethodAccessorImpl.java:-2	2022/08/03 17:24:08	35 s	3/3	110/110

LAB CHAPTER 15: PERSIST AN RDD

Explore the Performance effect of caching an RDD

1) I started the Spark shell in local mode with two threads and ran the provided job setup code, as shown below.

```
[training@localhost ~]$ pyspark --master local[2]
Python 2.7.8 (default, Aug 27 2015, 05:23:36)
Type "copyright", "credits" or "license" for more information.

IPython 3.2.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/05/04 17:01:00 INFO spark.SparkContext: Running Spark version 1.3.0
24/05/04 17:01:00 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.188.133 instead (on
interface eth0)
24/05/04 17:01:00 WARN util.Utils: Set SPARK LOCAL IP if you need to bind to another address
24/05/04 17:01:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/05/04 17:01:02 INFO spark.SecurityManager: Changing view acls to: training
24/05/04 17:01:02 INFO spark.SecurityManager: Changing modify acls to: training
24/05/04 17:01:02 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training);
 users with modify permissions: Set(training)
24/05/04 17:01:03 INFO slf4j.Slf4jLogger: Slf4jLogger started
24/05/04 17:01:04 INFO Remoting: Starting remoting
24/05/04 17:01:04 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.188.133:54777]
24/05/04 17:01:04 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriver@192.168.188.133:54777]
24/05/04 17:01:04 INFO util.Utils: Successfully started service 'sparkDriver' on port 54777.
24/05/04 17:01:04 INFO spark.SparkEnv: Registering MapOutputTracker
24/05/04 17:01:04 INFO spark.SparkEnv: Registering BlockManagerMaster
24/05/04 17:01:04 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-8d0a27ca-9aa8-4956-8de5-0b887c95e65b/blockmgr-3494d3c9-8f8f-4e
2e-a225-46e62e45254b
24/05/04 17:01:04 INFO storage.MemoryStore: MemoryStore started with capacity 267.3 MB
24/05/04 17:01:04 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-499f99d4-189b-4574-8a43-ab05d6edc2b5/httpd-fbec09de-e6d3-4e32-8
439-e3da684b5994
24/05/04 17:01:04 INFO spark.HttpServer: Starting HTTP Server
24/05/04 17:01:05 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/04 17:01:05 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:53546
24/05/04 17:01:05 INFO util.Utils: Successfully started service 'HTTP file server' on port 53546.
24/05/04 17:01:05 INFO spark.SparkEnv: Registering OutputCommitCoordinator
24/05/04 17:01:05 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/04 17:01:05 WARN component.AbstractLifeCycle: FAILED SelectChannelConnector@0.0.0.0:4040: java.net.BindException: Address already in use
java.net.BindException: Address already in use
        at sun.nio.ch.Net.bind0(Native Method)
        at sun.nio.ch.Net.bind(Net.java:444)
        at sun.nio.ch.Net.bind(Net.java:436)
        at sun.nio.ch.ServerSocketChannelImpl.bind(ServerSocketChannelImpl.java:214)
```

```
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/kill,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/static,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/threadDump,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/executors,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/environment,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 17:14:46 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 17:14:46 WARN util.Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
24/05/04 17:14:46 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/04 17:14:46 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4045
24/05/04 17:14:46 INFO util.Utils: Successfully started service 'SparkUI' on port 4045.
24/05/04 17:14:46 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4045
24/05/04 17:14:46 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/04 17:14:47 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:37116/user/HeartbeatReceiver
24/05/04 17:14:47 INFO netty.NettyBlockTransferService: Server created on 60979
24/05/04 17:14:47 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/04 17:14:47 INFO storage.BlockManagerMasterActor: Registering block manager localhost:60979 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 60979)
24/05/04 17:14:47 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/04 17:14:49 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/04 17:14:50 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714868086872
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: sxk230064
```

2) Now in this step we count the number of user accounts with a total hit count greater than five as in below.
accounthits.filter(lambda (firstlast,hitcount): hitcount > 5)\ .count()

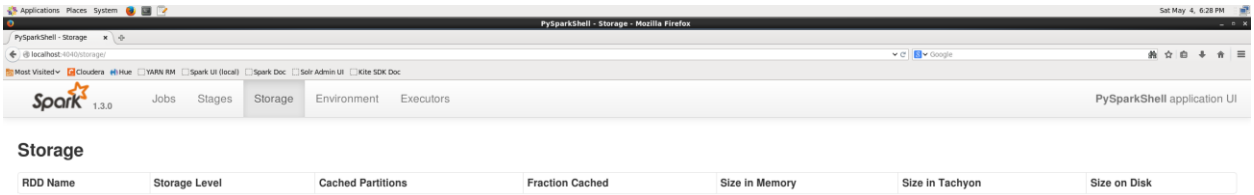
```
In [18]: accounthits=accountsByID.join(userreqs).values()

In [19]: accounthits\
   ....: .filter(lambda (firstlast,hitcount): hitcount>5)\
   ....: .count()
24/05/04 18:23:26 INFO spark.SparkContext: Starting job: count at <ipython-input-19-8b22495f6661>:1
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Registering RDD 10 (reduceByKey at <ipython-input-9-655863e70dd0>:1)
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Registering RDD 42 (join at <ipython-input-18-bc06df63f018>:1)
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Got job 0 (count at <ipython-input-19-8b22495f6661>:1) with 9 output partitions (allowLocal=false)
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Final stage: Stage 2(count at <ipython-input-19-8b22495f6661>:1)
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 1)
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Missing parents: List(Stage 1)
24/05/04 18:23:26 INFO scheduler.DAGScheduler: Submitting Stage 0 (PairwiseRDD[10] at reduceByKey at <ipython-input-9-655863e70dd0>:1), which has no
missing parents
24/05/04 18:23:27 INFO storage.BlockManager: Removing broadcast 1
24/05/04 18:23:27 INFO storage.BlockManager: Removing block broadcast_1_piece0
24/05/04 18:23:27 INFO storage.MemoryStore: Block broadcast_1_piece0 of size 21204 dropped from memory (free 279064463)
24/05/04 18:23:27 INFO storage.BlockManagerInfo: Removed broadcast_1_piece0 on localhost:33121 in memory (size: 20.7 KB, free: 267.2 MB)
24/05/04 18:23:27 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 18:23:27 INFO storage.BlockManager: Removing block broadcast_1
24/05/04 18:23:27 INFO storage.MemoryStore: Block broadcast_1 of size 280243 dropped from memory (free 279344706)
24/05/04 18:23:27 INFO storage.MemoryStore: ensureFreeSpace(7896) called with curMem=904269, maxMem=280248975
24/05/04 18:23:27 INFO storage.MemoryStore: Block broadcast_4 stored as values in memory (estimated size 7.7 KB, free 266.4 MB)
24/05/04 18:23:27 INFO spark.ContextCleaner: Cleaned broadcast 1
24/05/04 18:23:27 INFO storage.MemoryStore: ensureFreeSpace(5078) called with curMem=912165, maxMem=280248975
24/05/04 18:23:27 INFO storage.MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 5.0 KB, free 266.4 MB)
24/05/04 18:23:27 INFO storage.BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:33121 (size: 5.0 KB, free: 267.2 MB)
```

```
24/05/04 18:23:40 INFO python.PythonRDD: Times: total = 441, boot = -839, init = 896, finish = 384
24/05/04 18:23:40 INFO executor.Executor: Finished task 3.0 in stage 2.0 (TID 14). 922 bytes result sent to driver
24/05/04 18:23:40 INFO scheduler.TaskSetManager: Starting task 5.0 in stage 2.0 (TID 16, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 18:23:40 INFO executor.Executor: Running task 5.0 in stage 2.0 (TID 16)
24/05/04 18:23:40 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 2.0 (TID 14) in 490 ms on localhost (4/9)
24/05/04 18:23:40 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 18:23:40 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 6 ms
24/05/04 18:23:41 INFO python.PythonRDD: Times: total = 438, boot = -447, init = 473, finish = 412
24/05/04 18:23:41 INFO executor.Executor: Finished task 4.0 in stage 2.0 (TID 15). 922 bytes result sent to driver
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Starting task 6.0 in stage 2.0 (TID 17, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 18:23:41 INFO executor.Executor: Running task 6.0 in stage 2.0 (TID 17)
24/05/04 18:23:41 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 18:23:41 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Finished task 4.0 in stage 2.0 (TID 15) in 503 ms on localhost (5/9)
24/05/04 18:23:41 INFO python.PythonRDD: Times: total = 434, boot = -498, init = 541, finish = 391
24/05/04 18:23:41 INFO executor.Executor: Finished task 5.0 in stage 2.0 (TID 16). 922 bytes result sent to driver
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Starting task 7.0 in stage 2.0 (TID 18, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 18:23:41 INFO executor.Executor: Running task 7.0 in stage 2.0 (TID 18)
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 2.0 (TID 16) in 494 ms on localhost (6/9)
24/05/04 18:23:41 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 18:23:41 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/05/04 18:23:41 INFO python.PythonRDD: Times: total = 359, boot = -508, init = 542, finish = 325
24/05/04 18:23:41 INFO executor.Executor: Finished task 6.0 in stage 2.0 (TID 17). 922 bytes result sent to driver
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Starting task 8.0 in stage 2.0 (TID 19, localhost, PROCESS_LOCAL, 1056 bytes)
24/05/04 18:23:41 INFO executor.Executor: Running task 8.0 in stage 2.0 (TID 19)
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 2.0 (TID 17) in 416 ms on localhost (7/9)
24/05/04 18:23:41 INFO storage.ShuffleBlockFetcherIterator: Getting 9 non-empty blocks out of 9 blocks
24/05/04 18:23:41 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/05/04 18:23:41 INFO python.PythonRDD: Times: total = 344, boot = -419, init = 444, finish = 319
24/05/04 18:23:41 INFO executor.Executor: Finished task 7.0 in stage 2.0 (TID 18). 922 bytes result sent to driver
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 2.0 (TID 18) in 388 ms on localhost (8/9)
24/05/04 18:23:41 INFO python.PythonRDD: Times: total = 243, boot = -390, init = 409, finish = 224
24/05/04 18:23:41 INFO executor.Executor: Finished task 8.0 in stage 2.0 (TID 19). 922 bytes result sent to driver
24/05/04 18:23:41 INFO scheduler.DAGScheduler: Stage 2 (count at <ipython-input-19-8b22495f6661>:1) finished in 2.110 s
24/05/04 18:23:41 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 2.0 (TID 19) in 286 ms on localhost (9/9)
24/05/04 18:23:41 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
24/05/04 18:23:41 INFO scheduler.DAGScheduler: Job 0 finished: count at <ipython-input-19-8b22495f6661>:1, took 15.005109 s
Out[19]: 105

In [20]: sxk230064
```

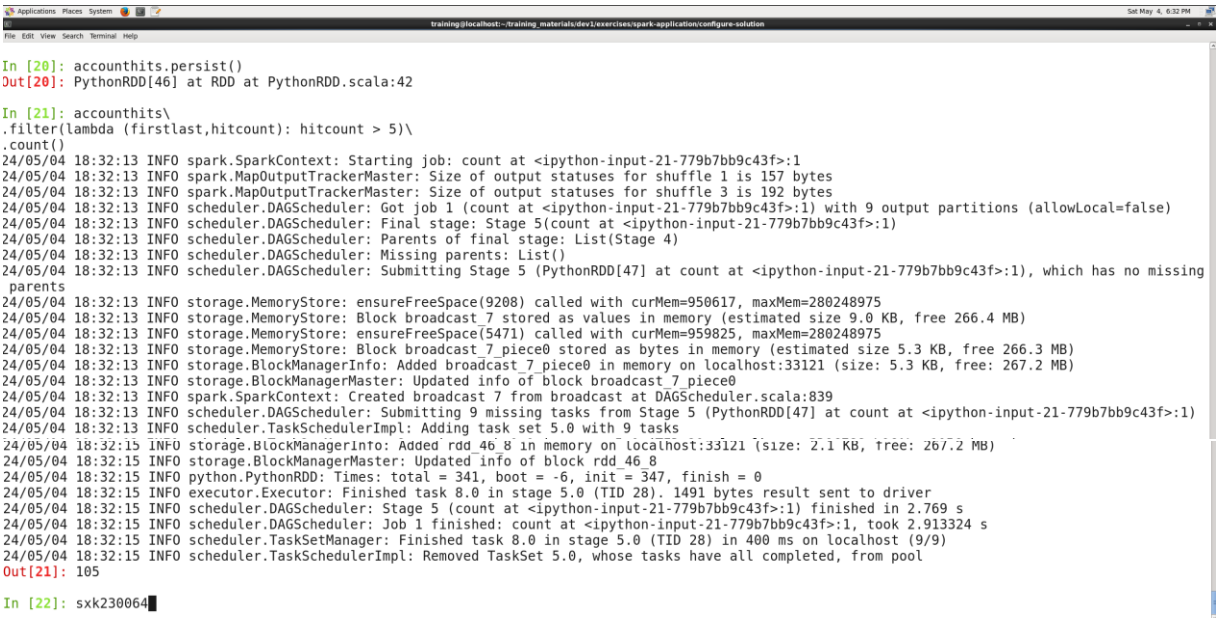
3) The RDD is then cached or persisted to memory by calling accounthits.persist(), as shown below.

```
In [20]: accounthits.persist()
Out[20]: PythonRDD[46] at RDD at PythonRDD.scala:42
```

4) As I didn't perform an action, the Storage tab in Spark UI is blank.



5) Next, in the Spark shell, I execute the count again as follows,
accounthits.filter(lambda (firstlast,hitcount): hitcount > 5)\ .count()



```
In [20]: accounthits.persist()
Out[20]: PythonRDD[46] at RDD at PythonRDD.scala:42

In [21]: accounthits\
.filter(lambda (firstlast,hitcount): hitcount > 5)\
.count()
24/05/04 18:32:13 INFO spark.SparkContext: Starting job: count at <ipython-input-21-779b7bb9c43f>:1
24/05/04 18:32:13 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 1 is 157 bytes
24/05/04 18:32:13 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 3 is 192 bytes
24/05/04 18:32:13 INFO scheduler.DAGScheduler: Got job 1 (count at <ipython-input-21-779b7bb9c43f>:1) with 9 output partitions (allowLocal=false)
24/05/04 18:32:13 INFO scheduler.DAGScheduler: Final stage: Stage 5(count at <ipython-input-21-779b7bb9c43f>:1)
24/05/04 18:32:13 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 4)
24/05/04 18:32:13 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/04 18:32:13 INFO scheduler.DAGScheduler: Submitting Stage 5 (PythonRDD[47] at count at <ipython-input-21-779b7bb9c43f>:1), which has no missing
 parents
24/05/04 18:32:13 INFO storage.MemoryStore: ensureFreeSpace(9208) called with curMem=950617, maxMem=280248975
24/05/04 18:32:13 INFO storage.MemoryStore: Block broadcast_7 stored as values in memory (estimated size 9.0 KB, free 266.4 MB)
24/05/04 18:32:13 INFO storage.MemoryStore: ensureFreeSpace(5471) called with curMem=959825, maxMem=280248975
24/05/04 18:32:13 INFO storage.MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 5.3 KB, free 266.3 MB)
24/05/04 18:32:13 INFO storage.BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:33121 (size: 5.3 KB, free: 267.2 MB)
24/05/04 18:32:13 INFO storage.BlockManagerMaster: Updated info of block broadcast_7_piece0
24/05/04 18:32:13 INFO spark.SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:839
24/05/04 18:32:13 INFO scheduler.DAGScheduler: Submitting 9 missing tasks from Stage 5 (PythonRDD[47] at count at <ipython-input-21-779b7bb9c43f>:1)
24/05/04 18:32:13 INFO scheduler.TaskSchedulerImpl: Adding task set 5.0 with 9 tasks
24/05/04 18:32:15 INFO storage.BlockManagerInfo: Added rdd_46_8 in memory on localhost:33121 (size: 2.1 KB, free: 267.2 MB)
24/05/04 18:32:15 INFO storage.BlockManagerMaster: Updated info of block rdd_46_8
24/05/04 18:32:15 INFO python.PythonRDD: Times: total = 341, boot = -6, init = 347, finish = 0
24/05/04 18:32:15 INFO executor.Executor: Finished task 8.0 in stage 5.0 (TID 28). 1491 bytes result sent to driver
24/05/04 18:32:15 INFO scheduler.DAGScheduler: Stage 5 (count at <ipython-input-21-779b7bb9c43f>:1) finished in 2.769 s
24/05/04 18:32:15 INFO scheduler.DAGScheduler: Job 1 finished: count at <ipython-input-21-779b7bb9c43f>:1, took 2.913324 s
24/05/04 18:32:15 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 5.0 (TID 28) in 400 ms on localhost (9/9)
24/05/04 18:32:15 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
Out[21]: 105

In [22]: sxk230064
```

6) Now, I view the RDD's 'toDebugString'. Here, the output shows the persistence level selected.

```
24/05/04 18:32:15 INFO scheduler.DAGScheduler: Stage 5 (count at <ipython-input-21-779b7bb9c43f>:1) finished in 2.769 s
24/05/04 18:32:15 INFO scheduler.DAGScheduler: Job 1 finished: count at <ipython-input-21-779b7bb9c43f>:1, took 2.913324 s
24/05/04 18:32:15 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 5.0 (TID 28) in 400 ms on localhost (9/9)
24/05/04 18:32:15 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
Out[21]: 105

In [22]: print accounts.toDebugString()
(7) /loudacre/accounts/* MapPartitionsRDD[5] at textFile at NativeMethodAccessorImpl.java:-2 []
 |  /loudacre/accounts/* HadoopRDD[4] at textFile at NativeMethodAccessorImpl.java:-2 []

In [23]: sxk230064
```

7) Reload the Storage tab in your browser, and this time note that the RDD you persisted is shown. Click on the RDD ID to see details about partitions and persistence

Spark 1.3.0 Jobs Stages Storage Environment Executors PySparkShell application U

RDD Storage Info for 17

Storage Level: Memory Serialized 1x Replicated
Cached Partitions: 33
Total Partitions: 33
Memory Size: 108.2 KB
Disk Size: 0.0 B

Data Distribution on 1 Executors

Host	Memory Usage	Disk Usage
localhost:38238	108.2 KB (267.2 MB Remaining)	0.0 B

33 Partitions

Block Name	Storage Level	Size in Memory	Size on Disk	Executors
rdd_17_0	Memory Serialized 1x Replicated	3.1 KB	0.0 B	localhost:38238
rdd_17_1	Memory Serialized 1x Replicated	3.1 KB	0.0 B	localhost:38238
rdd_17_10	Memory Serialized 1x Replicated	3.3 KB	0.0 B	localhost:38238
rdd_17_11	Memory Serialized 1x Replicated	3.2 KB	0.0 B	localhost:38238
rdd_17_12	Memory Serialized 1x Replicated	3.1 KB	0.0 B	localhost:38238
rdd_17_13	Memory Serialized 1x Replicated	3.2 KB	0.0 B	localhost:38238
rdd_17_14	Memory Serialized 1x Replicated	3.5 KB	0.0 B	localhost:38238
rdd_17_15	Memory Serialized 1x Replicated	3.3 KB	0.0 B	localhost:38238
rdd_17_16	Memory Serialized 1x Replicated	3.6 KB	0.0 B	localhost:38238
rdd_17_17	Memory Serialized 1x Replicated	3.4 KB	0.0 B	localhost:38238
rdd_17_18	Memory Serialized 1x Replicated	3.2 KB	0.0 B	localhost:38238
rdd_17_19	Memory Serialized 1x Replicated	3.5 KB	0.0 B	localhost:38238
rdd_17_2	Memory Serialized 1x Replicated	3.2 KB	0.0 B	localhost:38238
rdd_17_20	Memory Serialized 1x Replicated	3.3 KB	0.0 B	localhost:38238

8) After that, I went to the Executors tab. In our example, not the entire dataset is cached. Here, more nodes can have more memory and cache more data

LAB CHAPTER 16: IMPLEMENT AN ITERATIVE ALGORITHM WITH SPARK

Review the Data

In this Exercise, I will practice implementing iterative algorithms in Spark by calculating k-means for a set of points.

1) First, I copied $DEV1DATA/devicestatus.txt to HDFS directory /loudacre/.
hdfs dfs -put $DEV1DATA/devicestatus.txt /loudacre

```
[training@localhost configure-solution]$ cd ~
[training@localhost ~]$ hdfs dfs -put $DEV1DATA/devicestatus.txt /loudacre
[training@localhost ~]$ sxk230064
```

2) Then, as shown below, I run the Spark script $DEV1/exercises/spark-etl/bonus/DeviceStatusETL.
devstatus = sc.textFile("/loudacre/devicestatus.txt")

```
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/rdd,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/storage,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/pool,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/stage,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/stages,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/job,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs/json,null}
24/05/04 18:45:33 INFO handler.ContextHandler: stopped o.s.j.s.ServletContextHandler{/jobs,null}
24/05/04 18:45:33 WARN util.Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
24/05/04 18:45:33 INFO server.Server: jetty-8.y.z-SNAPSHOT
24/05/04 18:45:33 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4046
24/05/04 18:45:33 INFO util.Utils: Successfully started service 'SparkUI' on port 4046.
24/05/04 18:45:33 INFO ui.SparkUI: Started SparkUI at http://192.168.188.133:4046
24/05/04 18:45:33 INFO executor.Executor: Starting executor ID <driver> on host localhost
24/05/04 18:45:33 INFO util.AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sparkDriver@192.168.188.133:49852/user/HeartbeatReceiver
24/05/04 18:45:34 INFO netty.NettyBlockTransferService: Server created on 43851
24/05/04 18:45:34 INFO storage.BlockManagerMaster: Trying to register BlockManager
24/05/04 18:45:34 INFO storage.BlockManagerMasterActor: Registering block manager localhost:43851 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 43851)
24/05/04 18:45:34 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/04 18:45:36 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/04 18:45:37 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714873533636
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__/ .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: devstatus=sc.textFile("/loudacre/devicestatus.txt")
24/05/04 18:46:49 INFO storage.MemoryStore: ensureFreeSpace(280171) called with curMem=0, maxMem=280248975
24/05/04 18:46:49 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 273.6 KB, free 267.0 MB)
24/05/04 18:46:50 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=280171, maxMem=280248975
24/05/04 18:46:50 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 20.7 KB, free 267.0 MB)
24/05/04 18:46:50 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:43851 (size: 20.7 KB, free: 267.2 MB)
24/05/04 18:46:50 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/04 18:46:50 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:-2

In [2]: sxk230064
```

3) We loaded the data file, filtered out lines with less than 10 characters, used the 20th character as delimiter, analyzed the lines, and filtered out bad lines. I also created a new RDD containing the date, manufacturer, device ID, latitude, and longitude.

**cleanstatus = devstatus. **
**filter(lambda line: len(line) > 20). **
**map(lambda line: line.split(line[19:20])). **
filter(lambda values: len(values) == 14)

```
In [2]: cleanstatus=devstatus.\
   ...: filter(lambda line:len(line)>20).\
   ...: map(lambda line:line.split(line[19:20])).\
   ...: filter(lambda values:len(values)==14)
```

4) Making a new RDD with the following information: date, manufacturer, device ID, latitude, and longitude.

**devicedata = cleanstatus. **
map(lambda values: (values[0], values[1].split(' ')[0], values[2], values[12], values[13]))

```
In [2]: cleanstatus=devstatus.\
   ...: filter(lambda line:len(line)>20).\
   ...: map(lambda line:line.split(line[19:20])).\
   ...: filter(lambda values:len(values)==14)

In [3]: devicedata = cleanstatus.\
   ...: map(lambda values: (values[0],values[1].split(' ')[0],values[2],values[12],values[13]))

In [4]: sxk230064
```

5) I saved the CSV file as a comma delimited string by using below command,

**devicedata. **
**map(lambda values: ','.join(values)). **
saveAsTextFile("/loudacre/devicestatus_etl")

```
In [4]: devicedata. \
   ...: map(lambda values: ','.join(values)). \
   ...: saveAsTextFile("/loudacre/devicestatus_etl")
24/05/04 18:56:46 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
24/05/04 18:56:46 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
24/05/04 18:56:46 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
24/05/04 18:56:46 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
24/05/04 18:56:46 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
24/05/04 18:56:46 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/04 18:56:46 INFO mapred.FileInputFormat: Total input paths to process : 1
24/05/04 18:56:46 INFO spark.SparkContext: Starting job: saveAsTextFile at NativeMethodAccessorImpl.java:-2
24/05/04 18:56:46 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) with 1 output partitions (allowLocal=false)
24/05/04 18:56:46 INFO scheduler.DAGScheduler: Final stage: Stage 0(saveAsTextFile at NativeMethodAccessorImpl.java:-2)
24/05/04 18:56:46 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/04 18:56:46 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/04 18:56:46 INFO scheduler.DAGScheduler: Submitting Stage 0 (MapPartitionsRDD[4] at saveAsTextFile at NativeMethodAccessorImpl.java:-2), which has no missing parents
24/05/04 18:56:47 INFO storage.MemoryStore: ensureFreeSpace(138696) called with curMem=301375, maxMem=280248975
24/05/04 18:56:47 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 135.4 KB, free 266.8 MB)
24/05/04 18:56:47 INFO storage.MemoryStore: ensureFreeSpace(48970) called with curMem=440071, maxMem=280248975
24/05/04 18:56:47 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 47.8 KB, free 266.8 MB)
24/05/04 18:56:47 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:43851 (size: 47.8 KB, free: 267.2 MB)
24/05/04 18:56:47 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 18:56:47 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:839
24/05/04 18:56:47 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 0 (MapPartitionsRDD[4] at saveAsTextFile at NativeMethodAccessorImpl.java:-2)
24/05/04 18:56:47 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
24/05/04 18:56:47 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1311 bytes)
24/05/04 18:56:47 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/04 18:56:48 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/devicestatus.txt:0+64124001
24/05/04 18:56:48 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/04 18:57:03 INFO python.PythonRDD: Times: total = 15415, boot = 751, init = 205, finish = 14459
24/05/04 18:57:03 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405041856_0000_m_000000_0' to hdfs://localhost:8020/loudacre/devicestatus_etl/_temporary/0/task_2
0240504185 6_0000_m_000000
24/05/04 18:57:03 INFO spark.SparkHadoopWriter: attempt_202405041856_0000_m_000000_0: Committed
24/05/04 18:57:03 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2105 bytes result sent to driver
24/05/04 18:57:03 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 16239 ms on localhost (1/1)
24/05/04 18:57:03 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
24/05/04 18:57:03 INFO scheduler.DAGScheduler: Stage 0 (saveAsTextFile at NativeMethodAccessorImpl.java:-2) finished in 16.351 s
24/05/04 18:57:03 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at NativeMethodAccessorImpl.java:-2, took 17.104238 s

In [5]: sxk230064
```

Calculate k-means for device location -

1) I have executed the code provided by the professor as below, In this step for a point p and an array of points, return the index in the array of the point closest to p.

2) I defined a function that returns squared distances between two points.

```
Applications  Places  System                                                    Sat May 4, 7:15 PM
training@localhost:~
File  Edit  View  Search  Terminal  Help

In [19]: def closestPoint(p,points):
   ....:     bestIndex=0
   ....:     closest=float("+inf")
   ....:     for i in range(len(points)):
   ....:         dist = distanceSquared(p,points[i])
   ....:         if dist < closest:
   ....:             closest = dist
   ....:             bestIndex = i
   ....:         return bestIndex
```

3) Then, I defined a function that returns the sum of two points.

```
   .....

In [20]: def distanceSquared(p1,p2):
   ....:     return(p1[0]-p2[0])**2+(p1[1]-p2[1])**2
   ....:

In [21]: def addPoints(p1,p2):
   ....:     return[p1[0]+p2[0],p1[1]+p2[1]]
   ....:
```

4) Now I create a file variable and set the number of means (K) to 5 as follows

5) Now I set the variable convergeDist=0.1 for iteration.

```
In [19]: def closestPoint(p,points):
   ....:     bestIndex=0
   ....:     closest=float("+inf")
   ....:     for i in range(len(points)):
   ....:         dist = distanceSquared(p,points[i])
   ....:         if dist < closest:
   ....:             closest = dist
   ....:             bestIndex = i
   ....:         return bestIndex
   ....:

In [20]: def distanceSquared(p1,p2):
   ....:     return(p1[0]-p2[0])**2+(p1[1]-p2[1])**2
   ....:

In [21]: def addPoints(p1,p2):
   ....:     return[p1[0]+p2[0],p1[1]+p2[1]]
   ....:

In [22]: filename =
  File "<ipython-input-22-1eebd11c9e56>", line 1
    filename =

SyntaxError: invalid syntax


In [23]: filename = "/loudacre/devicestatus_etl/*"

In [24]: K=5

In [25]: convergDist=.1

In [26]: sxk230064
```

6) I parse the latitude and longitude into two-element arrays and filter out records where the latitude/longitude is unavailable, caching it for later use as follows:
points = sc.textFile(filename).map(lambda line: line.split(",")).map(lambda fields: [float(fields[3]), float(fields[4])]).filter(lambda point: sum(point) != 0).persist()

```
In [26]: points = sc.textFile(filename).map(lambda line: line.split(",")).map(lambda fields: [float(fields[3]), float(fields[4])]).filter(lambda point: sum(point) != 0).persist()
24/05/04 19:17:14 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=489041, maxMem=280248975
24/05/04 19:17:14 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 273.7 KB, free 266.5 MB)
24/05/04 19:17:14 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=769284, maxMem=280248975
24/05/04 19:17:14 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.5 MB)
24/05/04 19:17:14 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:43851 (size: 20.7 KB, free: 267.2 MB)
24/05/04 19:17:14 INFO storage.BlockManagerMaster: Updated info of block broadcast_2_piece0
24/05/04 19:17:14 INFO spark.SparkContext: Created broadcast 2 from textFile at NativeMethodAccessorImpl.java:-2

In [27]: kPoints = points.takeSample(False, K, 42)
24/05/04 19:17:43 INFO mapred.FileInputFormat: Total input paths to process : 1
24/05/04 19:17:43 INFO spark.SparkContext: Starting job: takeSample at <ipython-input-27-7bdb5298460a>:1
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Got job 1 (takeSample at <ipython-input-27-7bdb5298460a>:1) with 1 output partitions (allowLocal=false)
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Final stage: Stage 1(takeSample at <ipython-input-27-7bdb5298460a>:1)
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Submitting Stage 1 (PythonRDD[8] at takeSample at <ipython-input-27-7bdb5298460a>:1), which has no missing parents
24/05/04 19:17:43 INFO storage.MemoryStore: ensureFreeSpace(7176) called with curMem=790488, maxMem=280248975
24/05/04 19:17:43 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 7.0 KB, free 266.5 MB)
24/05/04 19:17:43 INFO storage.MemoryStore: ensureFreeSpace(4195) called with curMem=797664, maxMem=280248975
24/05/04 19:17:43 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 4.1 KB, free 266.5 MB)
24/05/04 19:17:43 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:43851 (size: 4.1 KB, free: 267.2 MB)
24/05/04 19:17:43 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
24/05/04 19:17:43 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:839
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (PythonRDD[8] at takeSample at <ipython-input-27-7bdb5298460a>:1)
24/05/04 19:17:43 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
24/05/04 19:17:44 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS LOCAL, 1322 bytes)
```

7) Now, I created a K-length array by taking random sample of K location points from the RDD as starting center points
kPoints = points.takeSample(False, K, 42)

```
In [27]: kPoints = points.takeSample(False, K, 42)
24/05/04 19:17:43 INFO mapred.FileInputFormat: Total input paths to process : 1
24/05/04 19:17:43 INFO spark.SparkContext: Starting job: takeSample at <ipython-input-27-7bdb5298460a>:1
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Got job 1 (takeSample at <ipython-input-27-7bdb5298460a>:1) with 1 output partitions (allowLocal=false)
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Final stage: Stage 1(takeSample at <ipython-input-27-7bdb5298460a>:1)
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Submitting Stage 1 (PythonRDD[8] at takeSample at <ipython-input-27-7bdb5298460a>:1), which has no missing parents
24/05/04 19:17:43 INFO storage.MemoryStore: ensureFreeSpace(7176) called with curMem=790488, maxMem=280248975
24/05/04 19:17:43 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 7.0 KB, free 266.5 MB)
24/05/04 19:17:43 INFO storage.MemoryStore: ensureFreeSpace(4195) called with curMem=797664, maxMem=280248975
24/05/04 19:17:43 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 4.1 KB, free 266.5 MB)
24/05/04 19:17:43 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:43851 (size: 4.1 KB, free: 267.2 MB)
24/05/04 19:17:43 INFO storage.BlockManagerMaster: Updated info of block broadcast_3_piece0
24/05/04 19:17:43 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:839
24/05/04 19:17:43 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (PythonRDD[8] at takeSample at <ipython-input-27-7bdb5298460a>:1)
24/05/04 19:17:43 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
24/05/04 19:17:44 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS LOCAL, 1322 bytes)
```

```
24/05/04 19:17:52 INFO storage.MemoryStore: Block broadcast_4 stored as values in memory (estimated size 6.4 KB, free 257.3 MB)
24/05/04 19:17:52 INFO storage.MemoryStore: ensureFreeSpace(4087) called with curMem=10425028, maxMem=280248975
24/05/04 19:17:52 INFO storage.MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 4.0 KB, free 257.3 MB)
24/05/04 19:17:52 INFO storage.BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:43851 (size: 4.0 KB, free: 258.0 MB)
24/05/04 19:17:52 INFO storage.BlockManagerMaster: Updated info of block broadcast_4_piece0
24/05/04 19:17:52 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:839
24/05/04 19:17:52 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 2 (PythonRDD[9] at takeSample at <ipython-input-27-7bdb5298460a>:1)
24/05/04 19:17:52 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
24/05/04 19:17:52 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, PROCESS_LOCAL, 1322 bytes)
24/05/04 19:17:52 INFO executor.Executor: Running task 0.0 in stage 2.0 (TID 2)
24/05/04 19:17:53 INFO storage.BlockManager: Found block rdd_7_0 locally
24/05/04 19:17:53 INFO python.PythonRDD: Times: total = 891, boot = -1933, init = 2156, finish = 668
24/05/04 19:17:53 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 2570 bytes result sent to driver
24/05/04 19:17:53 INFO scheduler.DAGScheduler: Stage 2 (takeSample at <ipython-input-27-7bdb5298460a>:1) finished in 1.002 s
24/05/04 19:17:53 INFO scheduler.DAGScheduler: Job 2 finished: takeSample at <ipython-input-27-7bdb5298460a>:1, took 1.217057 s
24/05/04 19:17:53 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 978 ms on localhost (1/1)
24/05/04 19:17:53 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool

In [28]: sxk230064
```

print "Starting Points: " , kPoints

```
In [29]: print"Starting Points:",kPoints
Starting Points: [[38.5789394051, -122.617581626], [38.3292500834, -121.871978282], [34.3235180239, -119.699243774], [32.9779195839, -116.878209392], [42.8031393876, -122.55759398]]

In [30]: sxk230064
```

8) Then I loop until the total distance between the points of one iteration and the next is less than the convergence distance specified, which is as follows and printed final center points.

```
tempDist = float("+inf")

while tempDist > convergeDist:
    closest = points.map(lambda p: (closestPoint(p, kPoints), (p,1)))
    pointstats = closest.reduceByKey(lambda (point1, n1), (point2, n2): (addPoints(point1,
point2), n1 + n2))
    newPoints = pointstats.map(lambda (i,(point,n)): (i,[point[0]/n, point[1]/n])).collect()
    tempDist = 0
    for (i, point) in newPoints: tempDist += distanceSquared(kPoints[i],point)
    print "Distance between iterations:" ,tempDist
    for (i, point) in newPoints: kPoints[i] = point
    print "Final Center Points:" + str(kPoints)
```

Applications Places System Sat May 4, 7:36 PM

training@localhost:~

File Edit View Search Terminal Help

```
In [17]: while tempDist > convergeDist:
   ....:     closest = points.map(lambda p: (closestPoint(p, kPoints), (p,1)))
   ....:     pointstats = closest.reduceByKey(lambda (point1, n1), (point2, n2): (addPoints(point1, point2), n1 + n2)
)
   ....:     newPoints = pointstats.map(lambda (i,(point,n)): (i,[point[0]/n, point[1]/n])).collect()
   ....:     tempDist=0
   ....:     for (i, point) in newPoints: tempDist += distanceSquared(kPoints[i],point)
   ....:     print "Distance between iterations:" ,tempDist
   ....:     for (i, point) in newPoints: kPoints[i] = point
   ....:     print "Final Center Points:" + str(kPoints)
   ....:
24/05/04 19:36:30 INFO spark.SparkContext: Starting job: collect at <ipython-input-17-8312b4879b3a>:4
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Registering RDD 17 (reduceByKey at <ipython-input-17-8312b4879b3a>:3)
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Got job 5 (collect at <ipython-input-17-8312b4879b3a>:4) with 1 output
 partitions (allowLocal=false)
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Final stage: Stage 6(collect at <ipython-input-17-8312b4879b3a>:4)
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 5)
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Missing parents: List(Stage 5)
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Submitting Stage 5 (PairwiseRDD[17] at reduceByKey at <ipython-input-1
7-8312b4879b3a>:3), which has no missing parents
24/05/04 19:36:30 INFO storage.MemoryStore: ensureFreeSpace(10328) called with curMem=20358533, maxMem=280248975
24/05/04 19:36:30 INFO storage.MemoryStore: Block broadcast_8 stored as values in memory (estimated size 10.1 KB, fre
e 247.8 MB)
24/05/04 19:36:30 INFO storage.MemoryStore: ensureFreeSpace(6444) called with curMem=20368861, maxMem=280248975
24/05/04 19:36:30 INFO storage.MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 6.3 KB
, free 247.8 MB)
24/05/04 19:36:30 INFO storage.BlockManagerInfo: Added broadcast_8_piece0 in memory on localhost:52681 (size: 6.3 KB,
 free: 248.8 MB)
24/05/04 19:36:30 INFO storage.BlockManagerMaster: Updated info of block broadcast_8_piece0
24/05/04 19:36:30 INFO spark.SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:839
24/05/04 19:36:30 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 5 (PairwiseRDD[17] at reduceByKe
y at <ipython-input-17-8312b4879b3a>:3)
24/05/04 19:36:30 INFO scheduler.TaskSchedulerImpl: Adding task set 5.0 with 1 tasks
24/05/04 19:36:31 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 5, localhost, PROCESS_LOCAL, 131
1 bytes)
24/05/04 19:36:31 INFO executor.Executor: Running task 0.0 in stage 5.0 (TID 5)
24/05/04 19:36:31 INFO storage.BlockManager: Found block rdd_12_0 locally
24/05/04 19:36:33 INFO python.PythonRDD: Times: total = 2579, boot = 13, init = 29, finish = 2537
24/05/04 19:36:33 INFO executor.Executor: Finished task 0.0 in stage 5.0 (TID 5). 2049 bytes result sent to driver
24/05/04 19:36:33 INFO scheduler.DAGScheduler: Stage 5 (reduceByKey at <ipython-input-17-8312b4879b3a>:3) finished in
 2.824 s
24/05/04 19:36:33 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 2786 ms on localhost (1/1)
24/05/04 19:36:33 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
24/05/04 19:36:33 INFO scheduler.DAGScheduler: looking for newly runnable stages
```

training@local... training@local... PySparkShell -... training@local... PySparkShell -... training@local...

Applications Places System Sat May 4, 7:40 PM

training@localhost:~

File Edit View Search Terminal Help

```
24/05/04 19:36:34 INFO storage.BlockManager: Found block rdd_12_0 locally
24/05/04 19:36:36 INFO python.PythonRDD: Times: total = 2513, boot = -243, init = 265, finish = 2491
24/05/04 19:36:36 INFO executor.Executor: Finished task 0.0 in stage 7.0 (TID 7). 2049 bytes result sent to driver
24/05/04 19:36:37 INFO scheduler.DAGScheduler: Stage 7 (reduceByKey at <ipython-input-17-8312b4879b3a>:3) finished in
 2.641 s
24/05/04 19:36:37 INFO scheduler.DAGScheduler: looking for newly runnable stages
24/05/04 19:36:37 INFO scheduler.DAGScheduler: running: Set()
24/05/04 19:36:37 INFO scheduler.DAGScheduler: waiting: Set(Stage 8)
24/05/04 19:36:37 INFO scheduler.DAGScheduler: failed: Set()
24/05/04 19:36:37 INFO scheduler.DAGScheduler: Missing parents for Stage 8: List()
24/05/04 19:36:37 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 7.0 (TID 7) in 2617 ms on localhost (1/1)
24/05/04 19:36:37 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
24/05/04 19:36:37 INFO scheduler.DAGScheduler: Submitting Stage 8 (PythonRDD[26] at collect at <ipython-input-17-8312
b4879b3a>:4), which is now runnable
24/05/04 19:36:37 INFO storage.MemoryStore: ensureFreeSpace(5592) called with curMem=20401205, maxMem=280248975
24/05/04 19:36:37 INFO storage.MemoryStore: Block broadcast_11 stored as values in memory (estimated size 5.5 KB, fre
e 247.8 MB)
24/05/04 19:36:37 INFO storage.MemoryStore: ensureFreeSpace(3541) called with curMem=20406797, maxMem=280248975
24/05/04 19:36:37 INFO storage.MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.5 K
B, free 247.8 MB)
24/05/04 19:36:37 INFO storage.BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:52681 (size: 3.5 KB
, free: 248.8 MB)
24/05/04 19:36:37 INFO storage.BlockManagerMaster: Updated info of block broadcast_11_piece0
24/05/04 19:36:37 INFO spark.SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:839
24/05/04 19:36:37 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 8 (PythonRDD[26] at collect at <
ipython-input-17-8312b4879b3a>:4)
24/05/04 19:36:37 INFO scheduler.TaskSchedulerImpl: Adding task set 8.0 with 1 tasks
24/05/04 19:36:37 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 8.0 (TID 8, localhost, PROCESS_LOCAL, 105
5 bytes)
24/05/04 19:36:37 INFO executor.Executor: Running task 0.0 in stage 8.0 (TID 8)
24/05/04 19:36:37 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
24/05/04 19:36:37 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
24/05/04 19:36:37 INFO python.PythonRDD: Times: total = 71, boot = -263, init = 334, finish = 0
24/05/04 19:36:37 INFO executor.Executor: Finished task 0.0 in stage 8.0 (TID 8). 948 bytes result sent to driver
24/05/04 19:36:37 INFO scheduler.DAGScheduler: Stage 8 (collect at <ipython-input-17-8312b4879b3a>:4) finished in 0.1
57 s
24/05/04 19:36:37 INFO scheduler.DAGScheduler: Job 6 finished: collect at <ipython-input-17-8312b4879b3a>:4, took 3.0
26764 s
Distance between iterations: 0.0
Final Center Points:[[36.84127339936045, -118.59276074471572], [38.3292500834, -121.871978282], [34.3235180239, -119.
599243774], [32.9779195839, -116.878209392], [42.8031393876, -122.55759398]]
24/05/04 19:36:37 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 8.0 (TID 8) in 143 ms on localhost (1/1)
24/05/04 19:36:37 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool

In [18]: sxk230064
```

LAB CHAPTER 17: USE SPARK SQL FOR ETL

Review the Data in MySQL

In this exercise have used Spark SQL to load data from MySQL, process it, and store it to HDFS.

1) Listed all the columns and types in the table
   mysql -utraining -ptraining loudacre \ -e"describe webpage"

2) I have viewed few lines by using below command.
   mysql -utraining -ptraining loudacre \ -e"select * from webpage limit 5"

```
[training@localhost ~]$ mysql -utraining -ptraining loudacre \
> -e"describe webpage"
+--------------------+--------------+------+-----+---------+--------------
--+
| Field              | Type         | Null | Key | Default | Extra
  |
+--------------------+--------------+------+-----+---------+--------------
--+
| web_page_num       | int(11)      | NO   | PRI | NULL    | auto_incremen
t |
| web_page_file_name | varchar(255) | NO   |     | NULL    |
  |
| associated_files   | varchar(255) | NO   |     | NULL    |
  |
+--------------------+--------------+------+-----+---------+--------------
--+
[training@localhost ~]$ mysql -utraining -ptraining loudacre \
> -e"select * from webpage limit 5"
+--------------+-------------------------+------------------------------
------+
| web_page_num | web_page_file_name      | associated_files
      |
+--------------+-------------------------+------------------------------
------+
|            1 | sorrento_f00l_sales.html | theme.css,code.js,sorrento_f00
l.jpg |
|            2 | titanic_2100_sales.html  | theme.css,code.js,titanic_2100
.jpg  |
|            3 | meetoo_3.0_sales.html    | theme.css,code.js,meetoo_3.0.j
pg    |
|            4 | meetoo_3.1_sales.html    | theme.css,code.js,meetoo_3.1.j
pg    |
|            5 | ifruit_1_sales.html      | theme.css,code.js,ifruit_1.jpg
      |
+--------------+-------------------------+------------------------------
------+
[training@localhost ~]$ sxk230064
```

The data in the associated files column is comma-delimited string. Using Spark SQL, I will extract the data in the column, split the string and create a new dataset in HDFS that will be having new number of webpages.

Load the Data from MySQL

3) Imported the SQLContext class definition, and also defined a SQL context.
   from pyspark.sql import SQLContext
   sqlCtx = SQLContext(sc)

4) Then I created a new DataFrame based on the webpage table from the database:
   url="jdbc:mysql://localhost/loudacre?user=training&password=training"
   webpages=sqlCtx.load(source="jdbc", url=url, dbtable="webpage")

```
24/05/04 19:49:41 INFO storage.BlockManagerMasterActor: Registering block manager localhost:49825 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 49825)
24/05/04 19:49:41 INFO storage.BlockManagerMaster: Registered BlockManager
24/05/04 19:49:55 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/05/04 19:50:00 INFO scheduler.EventLoggingListener: Logging events to hdfs:///user/spark/applicationHistory/local-1714877379063
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: from pyspark.sql import SQLContext

In [2]: sqlCtx = SQLContext(sc)

In [3]: sxk230064
```

5) In this step I have called webpages for examining the schema of the new DataFrame.

```
      /_/

Using Python version 2.7.8 (default, Aug 27 2015 05:23:36)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]: from pyspark.sql import SQLContext

In [2]: sqlCtx = SQLContext(sc)

In [3]: url="jdbc:mysql://localhost/loudacre?user=training&password=training"

In [4]: webpages=sqlCtx.load(source="jdbc", url=url, dbtable="webpage")

In [5]: webpages:printSchema()
  File "<ipython-input-5-bb00636bd07f>", line 1
    webpages:printSchema()

SyntaxError: invalid syntax


In [6]: webpages.printSchema()
root
 |-- web_page_num: integer (nullable = false)
 |-- web_page_file_name: string (nullable = false)
 |-- associated_files: string (nullable = false)


In [7]: assocfiles = webpages.select(webpages.web_page_num,webpages.associated_files)

In [8]: sxk230064
```

6) Now I have created web_page_num and associated_files columns from the existing dataframe.
assocfiles = webpages.select(webpages.web_page_num,webpages.associated_files)

```
In [7]: assocfiles = webpages.select(webpages.web_page_num,webpages.associated_files)

In [8]: sxk230064
```

7) Now, I have used map method to convert this dataframe into a PairRDD . This is the web_page value, and the value is associated files_string.
afilesrdd = assocfiles.map(lambda row: (row.web_page_num,row.associated_files))

8) Then, I used flatMapValues to split by ‘,’ and extracted the required data as in the below screenshot.
afilesrdd2 = afilesrdd.flatMapValues(lambda filestring:filestring.split(','))

```
In [7]: assocfiles = webpages.select(webpages.web_page_num,webpages.associated_files)

In [8]: afilesrdd = assocfiles.map(lambda row: (row.web_page_num,row.associated_files))

In [9]: afilesrdd2 = afilesrdd.flatMapValues(lambda filestring:filestring.split(','))

In [10]: sxk230064
```

9) Next, from ‘afilesrdd2’, I created a new DataFrame as follows,
afiledf = sqlCtx.createDataFrame(afilesrdd2)

```
In [10]: afiledf = sqlCtx.createDataFrame(afilesrdd2)
24/05/04 20:02:00 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/04 20:02:00 INFO scheduler.DAGScheduler: Got job 0 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/04 20:02:00 INFO scheduler.DAGScheduler: Final stage: Stage 0(runJob at PythonRDD.scala:356)
24/05/04 20:02:00 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/04 20:02:00 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/04 20:02:00 INFO scheduler.DAGScheduler: Submitting Stage 0 (PythonRDD[5] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/04 20:02:01 INFO storage.MemoryStore: ensureFreeSpace(8608) called with curMem=0, maxMem=280248975
24/05/04 20:02:01 INFO storage.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 8.4 KB, free 267.3 MB)
24/05/04 20:02:01 INFO storage.MemoryStore: ensureFreeSpace(4890) called with curMem=8608, maxMem=280248975
24/05/04 20:02:01 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 4.8 KB, free 267.3 MB)
24/05/04 20:02:01 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:49825 (size: 4.8 KB, free: 267.3 MB)
24/05/04 20:02:01 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/04 20:02:01 INFO spark.SparkContext: Created broadcast 0 from broadcast at DAGScheduler.scala:839
24/05/04 20:02:01 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 0 (PythonRDD[5] at RDD at PythonRDD.scala:42)
24/05/04 20:02:01 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
24/05/04 20:02:01 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, PROCESS_LOCAL, 1062 bytes)
24/05/04 20:02:02 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
24/05/04 20:02:03 INFO jdbc.JDBCRDD: closed connection
24/05/04 20:02:03 INFO python.PythonRDD: Times: total = 1760, boot = 1477, init = 262, finish = 21
24/05/04 20:02:03 INFO jdbc.JDBCRDD: closed connection
24/05/04 20:02:04 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 675 bytes result sent to driver
24/05/04 20:02:04 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 2075 ms on localhost (1/1)
24/05/04 20:02:04 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Stage 0 (runJob at PythonRDD.scala:356) finished in 2.179 s
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:356, took 3.724561 s
24/05/04 20:02:04 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Got job 1 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Final stage: Stage 1(runJob at PythonRDD.scala:356)
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Parents of final stage: List()
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Missing parents: List()
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Submitting Stage 1 (PythonRDD[7] at RDD at PythonRDD.scala:42), which has no missing parents
24/05/04 20:02:04 INFO storage.MemoryStore: ensureFreeSpace(9896) called with curMem=13498, maxMem=280248975
24/05/04 20:02:04 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 9.7 KB, free 267.2 MB)
24/05/04 20:02:04 INFO storage.MemoryStore: ensureFreeSpace(5775) called with curMem=23394, maxMem=280248975
24/05/04 20:02:04 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.6 KB, free 267.2 MB)
24/05/04 20:02:04 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:49825 (size: 5.6 KB, free: 267.3 MB)
24/05/04 20:02:04 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 20:02:04 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:839
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from Stage 1 (PythonRDD[7] at RDD at PythonRDD.scala:42)
24/05/04 20:02:04 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
24/05/04 20:02:04 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS_LOCAL, 1062 bytes)
24/05/04 20:02:04 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
24/05/04 20:02:04 INFO jdbc.JDBCRDD: closed connection
24/05/04 20:02:04 INFO python.PythonRDD: Times: total = 137, boot = 21, init = 109, finish = 7
24/05/04 20:02:04 INFO jdbc.JDBCRDD: closed connection
24/05/04 20:02:04 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 899 bytes result sent to driver
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Stage 1 (runJob at PythonRDD.scala:356) finished in 0.260 s
24/05/04 20:02:04 INFO scheduler.DAGScheduler: Job 1 finished: runJob at PythonRDD.scala:356, took 0.417538 s
24/05/04 20:02:04 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 231 ms on localhost (1/1)
24/05/04 20:02:04 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool

In [11]: sxk230064
```

10) Now, I have called ‘printSchema’ on the new DataFrame as below,

11) After that, I have created a new DataFrame and used 'printSchema' to check that the new DataFrame has the correct column names or not.
finaldf = afiledf. \ withColumnRenamed('_1','web_page_num'). \ withColumnRenamed('_2','associated_file')

```
In [11]: afiledf.printSchema()
root
 |-- _1: long (nullable = true)
 |-- _2: string (nullable = true)


In [12]: finaldf = afiledf. \
  File "<ipython-input-12-84319e82dbdb>", line 1
    finaldf = afiledf. \

SyntaxError: unexpected character after line continuation character


In [13]: finaldf = afiledf. \
   ....: withColumnRenamed('_1','web_page_num'). \
   ....: withColumnRenamed('_2','associated_file')

In [14]: sxk230064
```

```
In [14]: finaldf.printSchema()
root
 |-- web_page_num: long (nullable = true)
 |-- associated_file: string (nullable = true)


In [15]: sxk230064
```

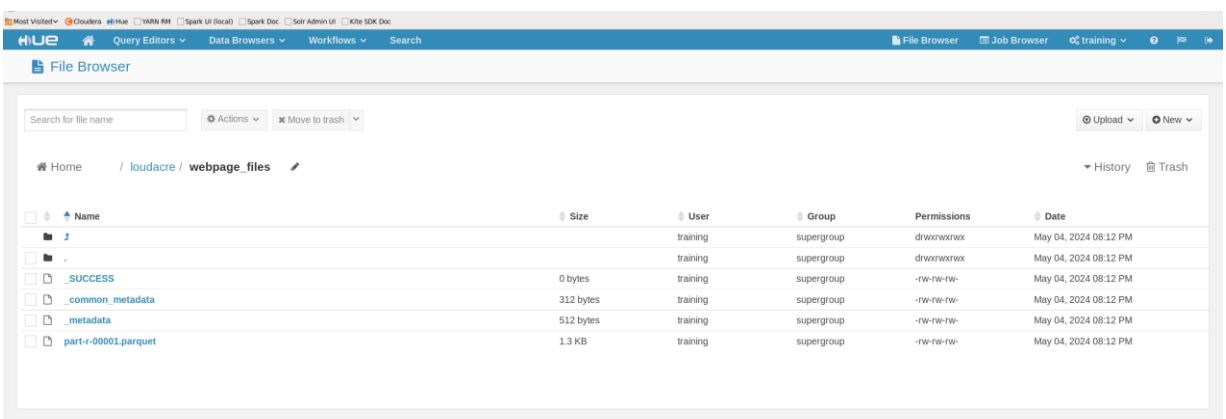
12) Now in this step I have saved the final DataFrame in parquet format to /loudacre/webpage_files
finaldf.save("/loudacre/webpage_files")

```
In [15]: finaldf.save("/loudacre/webpage_files")
24/05/04 20:12:49 INFO storage.BlockManager: Removing broadcast 1
24/05/04 20:12:49 INFO storage.BlockManager: Removing block broadcast_1_piece0
24/05/04 20:12:49 INFO storage.MemoryStore: Block broadcast_1_piece0 of size 5775 dropped from memory (free 280
225581)
24/05/04 20:12:49 INFO storage.BlockManagerInfo: Removed broadcast_1_piece0 on localhost:49825 in memory (size:
 5.6 KB, free: 267.3 MB)
24/05/04 20:12:49 INFO storage.BlockManagerMaster: Updated info of block broadcast_1_piece0
24/05/04 20:12:49 INFO storage.BlockManager: Removing block broadcast_1
24/05/04 20:12:49 INFO storage.MemoryStore: Block broadcast_1 of size 9896 dropped from memory (free 280235477)
24/05/04 20:12:49 INFO spark.ContextCleaner: Cleaned broadcast 1
24/05/04 20:12:49 INFO storage.BlockManager: Removing broadcast 0
24/05/04 20:12:49 INFO storage.BlockManager: Removing block broadcast_0
24/05/04 20:12:49 INFO storage.MemoryStore: Block broadcast_0 of size 8608 dropped from memory (free 280244085)
24/05/04 20:12:49 INFO storage.BlockManager: Removing block broadcast_0_piece0
24/05/04 20:12:49 INFO storage.MemoryStore: Block broadcast_0_piece0 of size 4890 dropped from memory (free 280
248975)
24/05/04 20:12:49 INFO storage.BlockManagerInfo: Removed broadcast_0_piece0 on localhost:49825 in memory (size:
 4.8 KB, free: 267.3 MB)
24/05/04 20:12:49 INFO storage.BlockManagerMaster: Updated info of block broadcast_0_piece0
24/05/04 20:12:49 INFO spark.ContextCleaner: Cleaned broadcast 0
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
24/05/04 20:12:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
```

```
24/05/04 20:12:51 INFO hadoop.ColumnChunkPageWriteStore: written 196B for [web_page_num] INT64: 158 values, 129
B raw, 152B comp, 1 pages, encodings: [PLAIN_DICTIONARY, RLE, BIT_PACKED], dic { 53 entries, 424B raw, 53B comp
}
24/05/04 20:12:51 INFO hadoop.ColumnChunkPageWriteStore: written 198B for [associated_file] BINARY: 158 values,
 129B raw, 141B comp, 1 pages, encodings: [PLAIN_DICTIONARY, RLE, BIT_PACKED], dic { 54 entries, 1,040B raw, 54
B comp}
24/05/04 20:12:51 INFO output.FileOutputCommitter: Saved output of task 'attempt_202405042012_0014_r_000000_0'
to hdfs://localhost:8020/loudacre/webpage_files/_temporary/0/task_202405042012_0014_r_000000
24/05/04 20:12:51 INFO jdbc.JDBCRDD: closed connection
24/05/04 20:12:51 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 640 bytes result sent to driv
er
24/05/04 20:12:51 INFO scheduler.DAGScheduler: Stage 2 (runJob at newParquet.scala:646) finished in 1.176 s
24/05/04 20:12:51 INFO scheduler.DAGScheduler: Job 2 finished: runJob at newParquet.scala:646, took 1.294164 s
24/05/04 20:12:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 1159 ms on localhost
 (1/1)
24/05/04 20:12:51 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from p
ool
24/05/04 20:12:52 INFO hadoop.ParquetFileReader: Initiating action with parallelism: 5

In [16]: sxd230064
```

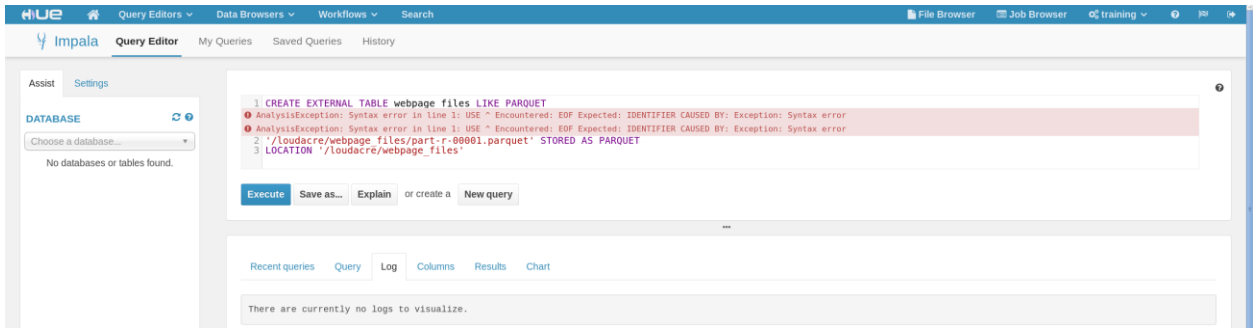
View the Output

13) By using Hue browser, I have listed the files that were saved by Spark SQL



14) Then, in Impala, I ran the following DDL command to create a table to access the new Parquet dataset:

CREATE EXTERNAL TABLE webpage_files LIKE PARQUET '/loudacre/webpage_files/part-r-00001.parquet' STORED AS PARQUET LOCATION '/loudacre/webpage_files'



I am encountering an error due to some technical issues. The error message is: # AnalysisException: Syntax error in line 1: USE ^ Encountered: EOF Expected: IDENTIFIER CAUSED BY: Exception: Syntax error while executing this command.