

Note: Replace the shown spark version with the actual version given by the instructor

1. Un-tar Spark

```
notroot@ubuntu:~$ cd lab/software
notroot@ubuntu:~/lab/software$ tar -xvf ../../downloads/spark-1.4.0-bin-hadoop2.6.tgz
```

```
notroot@ubuntu:~/lab/software$ ls -all
total 20
drwxrwxr-x  5 notroot notroot 4096 Oct 18 10:13 .
drwxrwxr-x  5 notroot notroot 4096 Oct 17 19:26 ..
drwxr-xr-x 10 notroot notroot 4096 Oct 17 19:37 hadoop-2.7.0
drwxr-xr-x  8 notroot notroot 4096 Feb 25  2013 scala-2.9.3
drwxr-xr-x 11 notroot notroot 4096 Jun  2 18:30 spark-1.4.0-bin-hadoop2.6
```

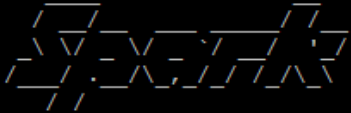
2. Make entries in the .bashrc

```
export SPARK_HOME=/home/notroot/lab/software/spark-1.6.0-bin-hadoop2.6
export PATH=$SPARK_HOME/bin:$PATH
```

3. Create a simple file called `sample` in the `lab/data` directory and add some content to it.

4. Start spark by typing spark-shell

```
notroot@ubuntu:~$ spark-shell
15/10/18 10:25:50 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
15/10/18 10:25:51 INFO spark.SecurityManager: Changing view acls to: notroot
15/10/18 10:25:51 INFO spark.SecurityManager: Changing modify acls to: notroot
15/10/18 10:25:51 INFO spark.SecurityManager: SecurityManager: authentication di
sabled; ui acls disabled; users with view permissions: Set(notroot); users with
modify permissions: Set(notroot)
15/10/18 10:25:51 INFO spark.HttpServer: Starting HTTP Server
15/10/18 10:25:52 INFO server.Server: jetty-8.y.z-SNAPSHOT
15/10/18 10:25:52 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0
:37245
15/10/18 10:25:52 INFO util.Utils: Successfully started service 'HTTP class serv
er' on port 37245.
Welcome to

 version 1.4.0

Using Scala version 2.10.4 (OpenJDK 64-Bit Server VM, Java 1.7.0_79)
Type in expressions to have them evaluated.
Type :help for more information.
```

```

15/10/18 10:26:13 WARN util.Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 192.168.235.207 instead (on interface eth0)
15/10/18 10:26:13 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
15/10/18 10:26:13 INFO spark.SparkContext: Running Spark version 1.4.0
15/10/18 10:26:13 INFO spark.SecurityManager: Changing view acls to: notroot
15/10/18 10:26:13 INFO spark.SecurityManager: Changing modify acls to: notroot
15/10/18 10:26:13 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(notroot); users with modify permissions: Set(notroot)
15/10/18 10:26:15 INFO slf4j.Slf4jLogger: Slf4jLogger started
15/10/18 10:26:16 INFO Remoting: Starting remoting
15/10/18 10:26:16 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.235.207:33362]
15/10/18 10:26:16 INFO util.Utils: Successfully started service 'sparkDriver' on port 33362.
15/10/18 10:26:16 INFO spark.SparkEnv: Registering MapOutputTracker
15/10/18 10:26:16 INFO spark.SparkEnv: Registering BlockManagerMaster
15/10/18 10:26:17 INFO storage.DiskBlockManager: Created local directory at /tmp/spark-dbf45b2-c687-4036-a63d-6d51884c9edc/blockmgr-4f14d743-5057-4e39-a80d-8d12bc36860b
15/10/18 10:26:17 INFO storage.MemoryStore: MemoryStore started with capacity 267.3 MB
15/10/18 10:26:17 INFO spark.HttpFileServer: HTTP File server directory is /tmp/spark-dbf45b2-c687-4036-a63d-6d51884c9edc/httpd-c3290e65-c943-414f-90af-dba90017f3fc
15/10/18 10:26:17 INFO spark.HttpServer: Starting HTTP Server
15/10/18 10:26:17 INFO server.Server: jetty-8.y.z-SNAPSHOT
15/10/18 10:26:17 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:40318
15/10/18 10:26:17 INFO util.Utils: Successfully started service 'HTTP file server' on port 40318.
15/10/18 10:26:17 INFO spark.SparkEnv: Registering OutputCommitCoordinator
15/10/18 10:26:19 INFO server.Server: jetty-8.y.z-SNAPSHOT
15/10/18 10:26:19 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
15/10/18 10:26:19 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
15/10/18 10:26:19 INFO ui.SparkUI: Started SparkUI at http://192.168.235.207:4040
15/10/18 10:26:19 INFO executor.Executor: Starting executor ID driver on host localhost
15/10/18 10:26:20 INFO executor.Executor: Using REPL class URI: http://192.168.235.207:37245
15/10/18 10:26:20 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 59729.
15/10/18 10:26:20 INFO netty.NettyBlockTransferService: Server created on 59729
15/10/18 10:26:20 INFO storage.BlockManagerMaster: Trying to register BlockManager
15/10/18 10:26:20 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:59729 with 267.3 MB RAM, BlockManagerId(driver, localhost, 59729)
15/10/18 10:26:20 INFO storage.BlockManagerMaster: Registered BlockManager
15/10/18 10:26:22 INFO repl.SparkILoop: Created spark context..
Spark context available as sc.
15/10/18 10:26:24 INFO hive.HiveContext: Initializing execution hive, version 0.13.1
15/10/18 10:26:27 INFO metastore.HiveMetaStore: 0: Opening raw store with implementation class:org.apache.hadoop.hive.metastore.ObjectStore
15/10/18 10:26:27 INFO metastore.ObjectStore: ObjectStore, initialize called
15/10/18 10:26:29 INFO DataNucleus.Persistence: Property datanucleus.cache.level2 unknown - will be ignored
15/10/18 10:26:29 INFO DataNucleus.Persistence: Property hive.metastore.integral.jdo.pushdown unknown - will be ignored
15/10/18 10:26:30 WARN DataNucleus.Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
15/10/18 10:26:31 WARN DataNucleus.Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)
15/10/18 10:26:38 INFO metastore.ObjectStore: Setting MetaStore object pin classes with hive.metastore.cache.pinobjtypes="Table,StorageDescriptor,SerdeInfo,Partition,Database,Type,FieldSchema,Order"
15/10/18 10:26:39 INFO metastore.MetaStoreDirectSql: MySQL check failed, assuming we are not on mysql: Lexical error at line 1, column 5. Encountered: "@" (64), after : "".

```

```

15/10/18 10:26:42 INFO DataNucleus.Datastore: The class "org.apache.hadoop.hive.metastore.model.MF
ieldSchema" is tagged as "embedded-only" so does not have its own datastore table.
15/10/18 10:26:42 INFO DataNucleus.Datastore: The class "org.apache.hadoop.hive.metastore.model.MO
rder" is tagged as "embedded-only" so does not have its own datastore table.
15/10/18 10:26:48 INFO DataNucleus.Datastore: The class "org.apache.hadoop.hive.metastore.model.MF
ieldSchema" is tagged as "embedded-only" so does not have its own datastore table.
15/10/18 10:26:48 INFO DataNucleus.Datastore: The class "org.apache.hadoop.hive.metastore.model.MO
rder" is tagged as "embedded-only" so does not have its own datastore table.
15/10/18 10:26:48 INFO metastore.ObjectStore: Initialized ObjectStore
15/10/18 10:26:49 WARN metastore.ObjectStore: Version information not found in metastore. hive.met
astore.schema.verification is not enabled so recording the schema version 0.13.1aa
15/10/18 10:26:50 INFO metastore.HiveMetaStore: Added admin role in metastore
15/10/18 10:26:50 INFO metastore.HiveMetaStore: Added public role in metastore
15/10/18 10:26:51 INFO metastore.HiveMetaStore: No user is added in admin role, since config is em
pty
15/10/18 10:26:51 INFO session.SessionState: No Tez session required at this point. hive.execution
.engine=mr.
15/10/18 10:26:51 INFO repl.SparkILoop: Created sql context (with Hive support)..
SQL context available as sqlContext.

```

- Let us try the canonical Word Count example in Spark.

```
val textFile = sc.textFile("file:///home/notroot/lab/data/sample")
```

[Note that the source file can be in the Local File System or in HDFS. In the screen shot below, it is in HDFS, whereas we are not trying it from the local File System and hence the difference in the screen shot on the first line]

```
val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
```

Note: There is a space in the split function in between " "
counts.saveAsTextFile("/wordcount")

-

```

scala> val textFile = sc.textFile("/input/sample")
15/10/18 10:35:47 INFO storage.MemoryStore: ensureFreeSpace(234360) called with curMem=362814, max
Mem=280248975
15/10/18 10:35:47 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimate
d size 228.9 KB, free 266.7 MB)
15/10/18 10:35:47 INFO storage.MemoryStore: ensureFreeSpace(19963) called with curMem=597174, maxM
em=280248975
15/10/18 10:35:47 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (es
timated size 19.5 KB, free 266.7 MB)
15/10/18 10:35:47 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:5
9729 (size: 19.5 KB, free: 267.2 MB)
15/10/18 10:35:47 INFO spark.SparkContext: Created broadcast 2 from textFile at <console>:21
textFile: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[11] at textFile at <console>:21

```

```

scala> val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_
+ _)
15/10/18 10:35:52 INFO mapred.FileInputFormat: Total input paths to process : 1
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[14] at reduceByKey at <console>:23

```

```
scala> counts.saveAsTextFile("/wordcount")
15/10/18 10:36:16 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
15/10/18 10:36:16 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
15/10/18 10:36:16 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
15/10/18 10:36:16 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
15/10/18 10:36:16 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
15/10/18 10:36:18 INFO spark.SparkContext: Starting job: saveAsTextFile at <console>:26
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Registering RDD 13 (map at <console>:23)
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Got job 0 (saveAsTextFile at <console>:26) with 1 output partitions (allowLocal=false)
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Final stage: ResultStage 1(saveAsTextFile at <console>:26)
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[13] at map at <console>:23), which has no missing parents
15/10/18 10:36:18 INFO storage.MemoryStore: ensureFreeSpace(4096) called with curMem=617137, maxMem=280248975
15/10/18 10:36:18 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 4.0 KB, free 266.7 MB)
15/10/18 10:36:18 INFO storage.MemoryStore: ensureFreeSpace(2287) called with curMem=621233, maxMem=280248975
15/10/18 10:36:18 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 2.2 KB, free 266.7 MB)
15/10/18 10:36:18 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:59729 (size: 2.2 KB, free: 267.2 MB)
15/10/18 10:36:18 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:874
```

```
15/10/18 10:36:18 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[13] at map at <console>:23)
15/10/18 10:36:18 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
15/10/18 10:36:18 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, ANY, 1396 bytes)
15/10/18 10:36:19 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
15/10/18 10:36:19 INFO rdd.HadoopRDD: Input split: hdfs://localhost:9000/input/sample:0+24
15/10/18 10:36:19 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2001 bytes result sent to driver
15/10/18 10:36:19 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1353 ms on localhost (1/1)
15/10/18 10:36:19 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
15/10/18 10:36:19 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (map at <console>:23) finished in 1.416 s
15/10/18 10:36:19 INFO scheduler.DAGScheduler: looking for newly runnable stages
15/10/18 10:36:19 INFO scheduler.DAGScheduler: running: Set()
15/10/18 10:36:19 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
15/10/18 10:36:19 INFO scheduler.DAGScheduler: failed: Set()
15/10/18 10:36:19 INFO scheduler.DAGScheduler: Missing parents for ResultStage 1: List()
15/10/18 10:36:19 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[15] at saveAsTextFile at <console>:26), which is now runnable
15/10/18 10:36:20 INFO storage.MemoryStore: ensureFreeSpace(128016) called with curMem=623520, maxMem=280248975
15/10/18 10:36:20 INFO storage.MemoryStore: Block broadcast_4 stored as values in memory (estimated size 125.0 KB, free 266.5 MB)
15/10/18 10:36:20 INFO storage.MemoryStore: ensureFreeSpace(42968) called with curMem=751536, maxMem=280248975
15/10/18 10:36:20 INFO storage.MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 42.0 KB, free 266.5 MB)
15/10/18 10:36:20 INFO storage.BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:59729 (size: 42.0 KB, free: 267.2 MB)
15/10/18 10:36:20 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:874
15/10/18 10:36:20 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[15] at saveAsTextFile at <console>:26)
15/10/18 10:36:20 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
15/10/18 10:36:20 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, PROCESS_LOCAL, 1165 bytes)
```



```

15/10/18 10:36:20 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
15/10/18 10:36:20 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
15/10/18 10:36:20 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 18 ms
15/10/18 10:36:20 INFO output.FileOutputCommitter: Saved output of task 'attempt_201510181036_0001_m_000000_1' to hdfs://localhost:9000/wordcount/_temporary/0/task_201510181036_0001_m_000000
15/10/18 10:36:20 INFO mapred.SparkHadoopMapRedUtil: attempt_201510181036_0001_m_000000_1: Committed
15/10/18 10:36:20 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 1828 bytes result sent to driver
15/10/18 10:36:21 INFO scheduler.DAGScheduler: ResultStage 1 (saveAsTextFile at <console>:26) finished in 0.877 s
15/10/18 10:36:21 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 878 ms on localhost (1/1)
15/10/18 10:36:21 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
15/10/18 10:36:21 INFO scheduler.DAGScheduler: Job 0 finished: saveAsTextFile at <console>:26, took 3.004800 s

```

7. Checking the contents.

```

notroot@ubuntu:~$ hadoop fs -ls /
Found 4 items
drwxr-xr-x  - notroot supergroup          0 2015-10-17 20:26 /benchmark_gen2
drwxr-xr-x  - notroot supergroup          0 2015-10-18 10:23 /input
drwxrwx---  - notroot supergroup          0 2015-10-17 19:49 /tmp
drwxr-xr-x  - notroot supergroup          0 2015-10-18 10:36 /wordcount
notroot@ubuntu:~$ hadoop fs -ls /wordcount
Found 2 items
-rw-r--r--  1 notroot supergroup          0 2015-10-18 10:36 /wordcount/_SUCCESS
-rw-r--r--  1 notroot supergroup        24 2015-10-18 10:36 /wordcount/part-00000
notroot@ubuntu:~$ hadoop fs -cat /wordcount/part-00000
(are,2)
(you,2)
(How,2)
notroot@ubuntu:~$ █

```