**5. Problem Statement**

1. Generate movie name and its duration

**Eclicpse application code:**

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**object** sess5\_assgn2 {

**def** main(args: Array[*String*]) {

**val** conf = **new** SparkConf().setAppName("sess5\_assgn2")

**val** sc = **new** SparkContext(conf)

**val** a = sc.textFile("file:////my-folder/src\_file/sess5\_assgn2\_src01.txt")

/\*

id,name,year,rating,duration

1,The Nightmare Before Christmas,1993,3.9,4568

1,The Nightmare Before Christmas,1993,3.9,4568

1,The Nightmare Before Christmas,1993,3.9,4568

2,The Mummy,1932,3.5,4388

3,Orphans of the Storm,1921,3.2,9062

4,The Object of Beauty,1991,2.8,6150

5,Night Tide,1963,2.8,5126

5,Night Tide,1963,2.8,5126

5,Night Tide,1963,2.8,5126

6,One Magic Christmas,1985,3.8,5333

7,Muriel's Wedding,1994,3.5,6323

8,Mother's Boys,1994,3.4,5733

9,Nosferatu: Original Version,1929,3.5,5651

10,Nick of Time,1995,3.4,5333

9,Nosferatu: Original Version,1929,3.5,5651

\*/

**val** a\_Header = a.first()

**val** a\_withoutHeader = a.filter(record=>record!=a\_Header)

**val** b = a\_withoutHeader.map(\_.split(","))

//val c = b.map(x=>(x(1),x(4))).sortByKey().collect().mkString(",")

**val** c = b.map(x=>(x(1),x(4))).sortByKey().collect().foreach(println)

//println(c)

}

}

**Command Prompt:**

Building Package:

bash-4.1$ sbt package

[warn] Executing in batch mode.

[warn] For better performance, hit [ENTER] to switch to interactive mode, or

[warn] consider launching sbt without any commands, or explicitly passing 'shell'

[info] Loading global plugins from /var/lib/hadoop-hdfs/.sbt/0.13/plugins

[info] Loading project definition from /tmp/hello-world-scala/project

[info] Set current project to sample (in build file:/tmp/hello-world-scala/)

[info] Compiling 1 Scala source to /tmp/hello-world-scala/target/scala-2.10/classes...

[warn] Multiple main classes detected. Run 'show discoveredMainClasses' to see the list

[info] Packaging /tmp/hello-world-scala/target/scala-2.10/sample\_2.10-0.1-SNAPSHOT.jar ...

[info] Done packaging.

[success] Total time: 11 s, completed Jun 11, 2017 9:17:44 AM

Submitting Application:

bash-4.1$ spark-submit --class sess5\_assgn2 --master spark://sudipta.com:7077 --deploy-mode client target/scala-2.10/sample\_2.10-0.1-SNAPSHOT.jar

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: Found binding in [jar:file:/usr/lib/parquet/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/lib/avro/avro-tools-1.7.6-cdh5.11.0.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]

17/06/11 09:26:42 INFO spark.SparkContext: Running Spark version 1.6.0

17/06/11 09:26:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/06/11 09:26:43 INFO spark.SecurityManager: Changing view acls to: hdfs

17/06/11 09:26:43 INFO spark.SecurityManager: Changing modify acls to: hdfs

17/06/11 09:26:43 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hdfs); users with modify permissions: Set(hdfs)

17/06/11 09:26:43 INFO util.Utils: Successfully started service 'sparkDriver' on port 55942.

17/06/11 09:26:44 INFO slf4j.Slf4jLogger: Slf4jLogger started

17/06/11 09:26:44 INFO Remoting: Starting remoting

17/06/11 09:26:44 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.252.128:40837]

17/06/11 09:26:44 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriverActorSystem@192.168.252.128:40837]

17/06/11 09:26:44 INFO util.Utils: Successfully started service 'sparkDriverActorSystem' on port 40837.

17/06/11 09:26:44 INFO spark.SparkEnv: Registering MapOutputTracker

17/06/11 09:26:44 INFO spark.SparkEnv: Registering BlockManagerMaster

17/06/11 09:26:44 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-9f444457-2a3e-42a3-8cc0-4193f25c2d80

17/06/11 09:26:44 INFO storage.MemoryStore: MemoryStore started with capacity 534.5 MB

17/06/11 09:26:45 INFO spark.SparkEnv: Registering OutputCommitCoordinator

17/06/11 09:26:45 INFO server.Server: jetty-8.y.z-SNAPSHOT

17/06/11 09:26:45 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040

17/06/11 09:26:45 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.

17/06/11 09:26:45 INFO ui.SparkUI: Started SparkUI at http://192.168.252.128:4040

17/06/11 09:26:45 INFO spark.SparkContext: Added JAR file:/tmp/hello-world-scala/target/scala-2.10/sample\_2.10-0.1-SNAPSHOT.jar at spark://192.168.252.128:55942/jars/sample\_2.10-0.1-SNAPSHOT.jar with timestamp 1497166005761

17/06/11 09:26:45 INFO client.AppClient$ClientEndpoint: Connecting to master spark://sudipta.com:7077...

17/06/11 09:26:46 INFO cluster.SparkDeploySchedulerBackend: Connected to Spark cluster with app ID app-20170611092646-0006

17/06/11 09:26:46 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 50081.

17/06/11 09:26:46 INFO netty.NettyBlockTransferService: Server created on 50081

17/06/11 09:26:46 INFO storage.BlockManagerMaster: Trying to register BlockManager

17/06/11 09:26:46 INFO storage.BlockManagerMasterEndpoint: Registering block manager 192.168.252.128:50081 with 534.5 MB RAM, BlockManagerId(driver, 192.168.252.128, 50081)

17/06/11 09:26:46 INFO storage.BlockManagerMaster: Registered BlockManager

17/06/11 09:26:46 INFO client.AppClient$ClientEndpoint: Executor added: app-20170611092646-0006/0 on worker-20170611083223-192.168.252.128-7078 (192.168.252.128:7078) with 1 cores

17/06/11 09:26:46 INFO cluster.SparkDeploySchedulerBackend: Granted executor ID app-20170611092646-0006/0 on hostPort 192.168.252.128:7078 with 1 cores, 1024.0 MB RAM

17/06/11 09:26:46 INFO client.AppClient$ClientEndpoint: Executor updated: app-20170611092646-0006/0 is now RUNNING

17/06/11 09:26:47 INFO cluster.SparkDeploySchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0

17/06/11 09:26:49 INFO storage.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 149.4 KB, free 534.4 MB)

17/06/11 09:26:49 INFO storage.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 16.3 KB, free 534.4 MB)

17/06/11 09:26:49 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 192.168.252.128:50081 (size: 16.3 KB, free: 534.5 MB)

17/06/11 09:26:49 INFO spark.SparkContext: Created broadcast 0 from textFile at sess5\_assgn2.scala:8

17/06/11 09:26:53 INFO mapred.FileInputFormat: Total input paths to process : 1

17/06/11 09:26:53 INFO spark.SparkContext: Starting job: first at sess5\_assgn2.scala:27

17/06/11 09:26:53 INFO scheduler.DAGScheduler: Got job 0 (first at sess5\_assgn2.scala:27) with 1 output partitions

17/06/11 09:26:53 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (first at sess5\_assgn2.scala:27)

17/06/11 09:26:53 INFO scheduler.DAGScheduler: Parents of final stage: List()

17/06/11 09:26:54 INFO scheduler.DAGScheduler: Missing parents: List()

17/06/11 09:26:54 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (file:////my-folder/src\_file/sess5\_assgn2\_src01.txt MapPartitionsRDD[1] at textFile at sess5\_assgn2.scala:8), which has no missing parents

17/06/11 09:26:54 INFO storage.MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 3.2 KB, free 534.4 MB)

17/06/11 09:26:54 INFO storage.MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 1890.0 B, free 534.4 MB)

17/06/11 09:26:54 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on 192.168.252.128:50081 (size: 1890.0 B, free: 534.5 MB)

17/06/11 09:26:54 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1006

17/06/11 09:26:54 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (file:////my-folder/src\_file/sess5\_assgn2\_src01.txt MapPartitionsRDD[1] at textFile at sess5\_assgn2.scala:8)

17/06/11 09:26:54 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks

17/06/11 09:26:54 INFO cluster.SparkDeploySchedulerBackend: Registered executor NettyRpcEndpointRef(null) (sudipta.com:54906) with ID 0

17/06/11 09:26:55 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, sudipta.com, executor 0, partition 0, PROCESS\_LOCAL, 2222 bytes)

17/06/11 09:26:55 INFO storage.BlockManagerMasterEndpoint: Registering block manager sudipta.com:49027 with 534.5 MB RAM, BlockManagerId(0, sudipta.com, 49027)

17/06/11 09:26:56 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on sudipta.com:49027 (size: 1890.0 B, free: 534.5 MB)

17/06/11 09:26:56 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on sudipta.com:49027 (size: 16.3 KB, free: 534.5 MB)

17/06/11 09:26:58 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 3752 ms on sudipta.com (executor 0) (1/1)

17/06/11 09:26:58 INFO scheduler.DAGScheduler: ResultStage 0 (first at sess5\_assgn2.scala:27) finished in 4.499 s

17/06/11 09:26:58 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Job 0 finished: first at sess5\_assgn2.scala:27, took 4.944395 s

17/06/11 09:26:58 INFO spark.SparkContext: Starting job: sortByKey at sess5\_assgn2.scala:32

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Got job 1 (sortByKey at sess5\_assgn2.scala:32) with 2 output partitions

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (sortByKey at sess5\_assgn2.scala:32)

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Parents of final stage: List()

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Missing parents: List()

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[6] at sortByKey at sess5\_assgn2.scala:32), which has no missing parents

17/06/11 09:26:58 INFO storage.MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 4.1 KB, free 534.4 MB)

17/06/11 09:26:58 INFO storage.MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 2.2 KB, free 534.4 MB)

17/06/11 09:26:58 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on 192.168.252.128:50081 (size: 2.2 KB, free: 534.5 MB)

17/06/11 09:26:58 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006

17/06/11 09:26:58 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (MapPartitionsRDD[6] at sortByKey at sess5\_assgn2.scala:32)

17/06/11 09:26:58 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 2 tasks

17/06/11 09:26:58 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, sudipta.com, executor 0, partition 0, PROCESS\_LOCAL, 2222 bytes)

17/06/11 09:26:59 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on sudipta.com:49027 (size: 2.2 KB, free: 534.5 MB)

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 1.0 (TID 2, sudipta.com, executor 0, partition 1, PROCESS\_LOCAL, 2222 bytes)

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 204 ms on sudipta.com (executor 0) (1/2)

17/06/11 09:26:59 INFO scheduler.DAGScheduler: ResultStage 1 (sortByKey at sess5\_assgn2.scala:32) finished in 0.243 s

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Job 1 finished: sortByKey at sess5\_assgn2.scala:32, took 0.297591 s

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 1.0 (TID 2) in 66 ms on sudipta.com (executor 0) (2/2)

17/06/11 09:26:59 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool

17/06/11 09:26:59 INFO spark.SparkContext: Starting job: collect at sess5\_assgn2.scala:32

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Registering RDD 4 (map at sess5\_assgn2.scala:32)

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Got job 2 (collect at sess5\_assgn2.scala:32) with 2 output partitions

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Final stage: ResultStage 3 (collect at sess5\_assgn2.scala:32)

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 2)

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 2)

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 2 (MapPartitionsRDD[4] at map at sess5\_assgn2.scala:32), which has no missing parents

17/06/11 09:26:59 INFO storage.MemoryStore: Block broadcast\_3 stored as values in memory (estimated size 4.6 KB, free 534.4 MB)

17/06/11 09:26:59 INFO storage.MemoryStore: Block broadcast\_3\_piece0 stored as bytes in memory (estimated size 2.6 KB, free 534.4 MB)

17/06/11 09:26:59 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on 192.168.252.128:50081 (size: 2.6 KB, free: 534.5 MB)

17/06/11 09:26:59 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1006

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage 2 (MapPartitionsRDD[4] at map at sess5\_assgn2.scala:32)

17/06/11 09:26:59 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 2 tasks

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 3, sudipta.com, executor 0, partition 0, PROCESS\_LOCAL, 2211 bytes)

17/06/11 09:26:59 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on sudipta.com:49027 (size: 2.6 KB, free: 534.5 MB)

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 2.0 (TID 4, sudipta.com, executor 0, partition 1, PROCESS\_LOCAL, 2211 bytes)

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 3) in 317 ms on sudipta.com (executor 0) (1/2)

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 2.0 (TID 4) in 67 ms on sudipta.com (executor 0) (2/2)

17/06/11 09:26:59 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool

17/06/11 09:26:59 INFO scheduler.DAGScheduler: ShuffleMapStage 2 (map at sess5\_assgn2.scala:32) finished in 0.293 s

17/06/11 09:26:59 INFO scheduler.DAGScheduler: looking for newly runnable stages

17/06/11 09:26:59 INFO scheduler.DAGScheduler: running: Set()

17/06/11 09:26:59 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 3)

17/06/11 09:26:59 INFO scheduler.DAGScheduler: failed: Set()

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Submitting ResultStage 3 (ShuffledRDD[7] at sortByKey at sess5\_assgn2.scala:32), which has no missing parents

17/06/11 09:26:59 INFO storage.MemoryStore: Block broadcast\_4 stored as values in memory (estimated size 2.8 KB, free 534.3 MB)

17/06/11 09:26:59 INFO storage.MemoryStore: Block broadcast\_4\_piece0 stored as bytes in memory (estimated size 1671.0 B, free 534.3 MB)

17/06/11 09:26:59 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on 192.168.252.128:50081 (size: 1671.0 B, free: 534.5 MB)

17/06/11 09:26:59 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1006

17/06/11 09:26:59 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ResultStage 3 (ShuffledRDD[7] at sortByKey at sess5\_assgn2.scala:32)

17/06/11 09:26:59 INFO scheduler.TaskSchedulerImpl: Adding task set 3.0 with 2 tasks

17/06/11 09:26:59 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 5, sudipta.com, executor 0, partition 0, NODE\_LOCAL, 1967 bytes)

17/06/11 09:26:59 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on sudipta.com:49027 (size: 1671.0 B, free: 534.5 MB)

17/06/11 09:26:59 INFO spark.MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to sudipta.com:54906

17/06/11 09:26:59 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 153 bytes

17/06/11 09:27:00 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 3.0 (TID 6, sudipta.com, executor 0, partition 1, NODE\_LOCAL, 1967 bytes)

17/06/11 09:27:00 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 5) in 315 ms on sudipta.com (executor 0) (1/2)

17/06/11 09:27:00 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 3.0 (TID 6) in 64 ms on sudipta.com (executor 0) (2/2)

17/06/11 09:27:00 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool

17/06/11 09:27:00 INFO scheduler.DAGScheduler: ResultStage 3 (collect at sess5\_assgn2.scala:32) finished in 0.367 s

17/06/11 09:27:00 INFO scheduler.DAGScheduler: Job 2 finished: collect at sess5\_assgn2.scala:32, took 0.821329 s

(Mother's Boys,5733)

(Muriel's Wedding,6323)

(Nick of Time,5333)

(Night Tide,5126)

(Night Tide,5126)

(Night Tide,5126)

(Nosferatu: Original Version,5651)

(Nosferatu: Original Version,5651)

(One Magic Christmas,5333)

(Orphans of the Storm,9062)

(The Mummy,4388)

(The Nightmare Before Christmas,4568)

(The Nightmare Before Christmas,4568)

(The Nightmare Before Christmas,4568)

(The Object of Beauty,6150)

17/06/11 09:27:00 INFO spark.SparkContext: Invoking stop() from shutdown hook

2. How many distinct movies are there?

**Eclipse Application code:**

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**object** sess5\_assgn2\_ques2 {

**def** main(args: Array[*String*]) {

**val** conf = **new** SparkConf().setAppName("sess5\_assgn2\_ques2")

**val** sc = **new** SparkContext(conf)

**val** a = sc.textFile("file:////my-folder/src\_file/sess5\_assgn2\_src01.txt")

/\*

id,name,year,rating,duration

1,The Nightmare Before Christmas,1993,3.9,4568

1,The Nightmare Before Christmas,1993,3.9,4568

1,The Nightmare Before Christmas,1993,3.9,4568

2,The Mummy,1932,3.5,4388

3,Orphans of the Storm,1921,3.2,9062

4,The Object of Beauty,1991,2.8,6150

5,Night Tide,1963,2.8,5126

5,Night Tide,1963,2.8,5126

5,Night Tide,1963,2.8,5126

6,One Magic Christmas,1985,3.8,5333

7,Muriel's Wedding,1994,3.5,6323

8,Mother's Boys,1994,3.4,5733

9,Nosferatu: Original Version,1929,3.5,5651

10,Nick of Time,1995,3.4,5333

9,Nosferatu: Original Version,1929,3.5,5651

\*/

**val** a\_Header = a.first()

**val** a\_withoutHeader = a.filter(record=>record!=a\_Header)

**val** b = a\_withoutHeader.map(\_.split(","))

//val c = b.map(x=>(x(1))).sortByKey().collect().mkString(",")

**val** c = b.map(x=>(x(1))).sortBy({x=>x},**false**).collect().foreach(println)

//println(c)

}

}

**Command Prompt:**

bash-4.1$ sbt package

[warn] Executing in batch mode.

[warn] For better performance, hit [ENTER] to switch to interactive mode, or

[warn] consider launching sbt without any commands, or explicitly passing 'shell'

[info] Loading global plugins from /var/lib/hadoop-hdfs/.sbt/0.13/plugins

[info] Loading project definition from /tmp/hello-world-scala/project

[info] Set current project to sample (in build file:/tmp/hello-world-scala/)

[info] Compiling 1 Scala source to /tmp/hello-world-scala/target/scala-2.10/classes...

[warn] Multiple main classes detected. Run 'show discoveredMainClasses' to see the list

[info] Packaging /tmp/hello-world-scala/target/scala-2.10/sample\_2.10-0.1-SNAPSHOT.jar ...

[info] Done packaging.

[success] Total time: 34 s, completed Jun 11, 2017 9:39:42 AM

bash-4.1$ spark-submit --class sess5\_assgn2\_ques2 --master spark://sudipta.com:7077 --deploy-mode client target/scala-2.10/sample\_2.10-0.1-SNAPSHOT.jar

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: Found binding in [jar:file:/usr/lib/parquet/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/lib/avro/avro-tools-1.7.6-cdh5.11.0.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]

17/06/11 09:40:16 INFO spark.SparkContext: Running Spark version 1.6.0

17/06/11 09:40:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/06/11 09:40:17 INFO spark.SecurityManager: Changing view acls to: hdfs

17/06/11 09:40:17 INFO spark.SecurityManager: Changing modify acls to: hdfs

17/06/11 09:40:17 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hdfs); users with modify permissions: Set(hdfs)

17/06/11 09:40:18 INFO util.Utils: Successfully started service 'sparkDriver' on port 48277.

17/06/11 09:40:18 INFO slf4j.Slf4jLogger: Slf4jLogger started

17/06/11 09:40:19 INFO Remoting: Starting remoting

17/06/11 09:40:19 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.252.128:37252]

17/06/11 09:40:19 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriverActorSystem@192.168.252.128:37252]

17/06/11 09:40:19 INFO util.Utils: Successfully started service 'sparkDriverActorSystem' on port 37252.

17/06/11 09:40:19 INFO spark.SparkEnv: Registering MapOutputTracker

17/06/11 09:40:19 INFO spark.SparkEnv: Registering BlockManagerMaster

17/06/11 09:40:19 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-c74eab39-1e7f-4b96-976c-08eb6bd37b05

17/06/11 09:40:19 INFO storage.MemoryStore: MemoryStore started with capacity 534.5 MB

17/06/11 09:40:19 INFO spark.SparkEnv: Registering OutputCommitCoordinator

17/06/11 09:40:20 INFO server.Server: jetty-8.y.z-SNAPSHOT

17/06/11 09:40:20 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040

17/06/11 09:40:20 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.

17/06/11 09:40:20 INFO ui.SparkUI: Started SparkUI at http://192.168.252.128:4040

17/06/11 09:40:20 INFO spark.SparkContext: Added JAR file:/tmp/hello-world-scala/target/scala-2.10/sample\_2.10-0.1-SNAPSHOT.jar at spark://192.168.252.128:48277/jars/sample\_2.10-0.1-SNAPSHOT.jar with timestamp 1497166820576

17/06/11 09:40:20 INFO client.AppClient$ClientEndpoint: Connecting to master spark://sudipta.com:7077...

17/06/11 09:40:21 INFO cluster.SparkDeploySchedulerBackend: Connected to Spark cluster with app ID app-20170611094021-0008

17/06/11 09:40:21 INFO client.AppClient$ClientEndpoint: Executor added: app-20170611094021-0008/0 on worker-20170611083223-192.168.252.128-7078 (192.168.252.128:7078) with 1 cores

17/06/11 09:40:21 INFO cluster.SparkDeploySchedulerBackend: Granted executor ID app-20170611094021-0008/0 on hostPort 192.168.252.128:7078 with 1 cores, 1024.0 MB RAM

17/06/11 09:40:21 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 49287.

17/06/11 09:40:21 INFO netty.NettyBlockTransferService: Server created on 49287

17/06/11 09:40:21 INFO storage.BlockManagerMaster: Trying to register BlockManager

17/06/11 09:40:21 INFO storage.BlockManagerMasterEndpoint: Registering block manager 192.168.252.128:49287 with 534.5 MB RAM, BlockManagerId(driver, 192.168.252.128, 49287)

17/06/11 09:40:21 INFO storage.BlockManagerMaster: Registered BlockManager

17/06/11 09:40:21 INFO client.AppClient$ClientEndpoint: Executor updated: app-20170611094021-0008/0 is now RUNNING

17/06/11 09:40:22 INFO cluster.SparkDeploySchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0

17/06/11 09:40:23 INFO storage.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 149.4 KB, free 534.4 MB)

17/06/11 09:40:24 INFO storage.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 16.3 KB, free 534.4 MB)

17/06/11 09:40:24 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 192.168.252.128:49287 (size: 16.3 KB, free: 534.5 MB)

17/06/11 09:40:24 INFO spark.SparkContext: Created broadcast 0 from textFile at sess5\_assgn2\_ques2.scala:8

17/06/11 09:40:28 INFO mapred.FileInputFormat: Total input paths to process : 1

17/06/11 09:40:30 INFO spark.SparkContext: Starting job: first at sess5\_assgn2\_ques2.scala:27

17/06/11 09:40:30 INFO scheduler.DAGScheduler: Got job 0 (first at sess5\_assgn2\_ques2.scala:27) with 1 output partitions

17/06/11 09:40:30 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (first at sess5\_assgn2\_ques2.scala:27)

17/06/11 09:40:30 INFO scheduler.DAGScheduler: Parents of final stage: List()

17/06/11 09:40:30 INFO scheduler.DAGScheduler: Missing parents: List()

17/06/11 09:40:31 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (file:////my-folder/src\_file/sess5\_assgn2\_src01.txt MapPartitionsRDD[1] at textFile at sess5\_assgn2\_ques2.scala:8), which has no missing parents

17/06/11 09:40:32 INFO cluster.SparkDeploySchedulerBackend: Registered executor NettyRpcEndpointRef(null) (sudipta.com:54233) with ID 0

17/06/11 09:40:32 INFO storage.MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 3.2 KB, free 534.4 MB)

17/06/11 09:40:32 INFO storage.MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 1890.0 B, free 534.4 MB)

17/06/11 09:40:32 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on 192.168.252.128:49287 (size: 1890.0 B, free: 534.5 MB)

17/06/11 09:40:32 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1006

17/06/11 09:40:33 INFO storage.BlockManagerMasterEndpoint: Registering block manager sudipta.com:58488 with 534.5 MB RAM, BlockManagerId(0, sudipta.com, 58488)

17/06/11 09:40:33 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (file:////my-folder/src\_file/sess5\_assgn2\_src01.txt MapPartitionsRDD[1] at textFile at sess5\_assgn2\_ques2.scala:8)

17/06/11 09:40:33 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks

17/06/11 09:40:33 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, sudipta.com, executor 0, partition 0, PROCESS\_LOCAL, 2222 bytes)

17/06/11 09:40:43 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on sudipta.com:58488 (size: 1890.0 B, free: 534.5 MB)

17/06/11 09:40:43 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on sudipta.com:58488 (size: 16.3 KB, free: 534.5 MB)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: ResultStage 0 (first at sess5\_assgn2\_ques2.scala:27) finished in 27.601 s

17/06/11 09:41:01 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 27412 ms on sudipta.com (executor 0) (1/1)

17/06/11 09:41:01 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Job 0 finished: first at sess5\_assgn2\_ques2.scala:27, took 30.470054 s

17/06/11 09:41:01 INFO spark.SparkContext: Starting job: sortBy at sess5\_assgn2\_ques2.scala:32

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Got job 1 (sortBy at sess5\_assgn2\_ques2.scala:32) with 2 output partitions

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (sortBy at sess5\_assgn2\_ques2.scala:32)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Parents of final stage: List()

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Missing parents: List()

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[7] at sortBy at sess5\_assgn2\_ques2.scala:32), which has no missing parents

17/06/11 09:41:01 INFO storage.MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 4.4 KB, free 534.4 MB)

17/06/11 09:41:01 INFO storage.MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 2.3 KB, free 534.4 MB)

17/06/11 09:41:01 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on 192.168.252.128:49287 (size: 2.3 KB, free: 534.5 MB)

17/06/11 09:41:01 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (MapPartitionsRDD[7] at sortBy at sess5\_assgn2\_ques2.scala:32)

17/06/11 09:41:01 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 2 tasks

17/06/11 09:41:01 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, sudipta.com, executor 0, partition 0, PROCESS\_LOCAL, 2222 bytes)

17/06/11 09:41:01 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on sudipta.com:58488 (size: 2.3 KB, free: 534.5 MB)

17/06/11 09:41:01 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 1.0 (TID 2, sudipta.com, executor 0, partition 1, PROCESS\_LOCAL, 2222 bytes)

17/06/11 09:41:01 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 183 ms on sudipta.com (executor 0) (1/2)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: ResultStage 1 (sortBy at sess5\_assgn2\_ques2.scala:32) finished in 0.199 s

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Job 1 finished: sortBy at sess5\_assgn2\_ques2.scala:32, took 0.245604 s

17/06/11 09:41:01 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 1.0 (TID 2) in 59 ms on sudipta.com (executor 0) (2/2)

17/06/11 09:41:01 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool

17/06/11 09:41:01 INFO spark.SparkContext: Starting job: collect at sess5\_assgn2\_ques2.scala:32

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Registering RDD 5 (sortBy at sess5\_assgn2\_ques2.scala:32)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Got job 2 (collect at sess5\_assgn2\_ques2.scala:32) with 2 output partitions

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Final stage: ResultStage 3 (collect at sess5\_assgn2\_ques2.scala:32)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 2)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 2)

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 2 (MapPartitionsRDD[5] at sortBy at sess5\_assgn2\_ques2.scala:32), which has no missing parents

17/06/11 09:41:01 INFO storage.MemoryStore: Block broadcast\_3 stored as values in memory (estimated size 4.9 KB, free 534.4 MB)

17/06/11 09:41:01 INFO storage.MemoryStore: Block broadcast\_3\_piece0 stored as bytes in memory (estimated size 2.7 KB, free 534.4 MB)

17/06/11 09:41:01 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on 192.168.252.128:49287 (size: 2.7 KB, free: 534.5 MB)

17/06/11 09:41:01 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1006

17/06/11 09:41:01 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage 2 (MapPartitionsRDD[5] at sortBy at sess5\_assgn2\_ques2.scala:32)

17/06/11 09:41:01 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 2 tasks

17/06/11 09:41:01 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 3, sudipta.com, executor 0, partition 0, PROCESS\_LOCAL, 2211 bytes)

17/06/11 09:41:01 INFO storage.BlockManagerInfo: Added broadcast\_3\_piece0 in memory on sudipta.com:58488 (size: 2.7 KB, free: 534.5 MB)

17/06/11 09:41:03 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 2.0 (TID 4, sudipta.com, executor 0, partition 1, PROCESS\_LOCAL, 2211 bytes)

17/06/11 09:41:03 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 3) in 1823 ms on sudipta.com (executor 0) (1/2)

17/06/11 09:41:03 INFO scheduler.DAGScheduler: ShuffleMapStage 2 (sortBy at sess5\_assgn2\_ques2.scala:32) finished in 1.882 s

17/06/11 09:41:03 INFO scheduler.DAGScheduler: looking for newly runnable stages

17/06/11 09:41:03 INFO scheduler.DAGScheduler: running: Set()

17/06/11 09:41:03 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 3)

17/06/11 09:41:03 INFO scheduler.DAGScheduler: failed: Set()

17/06/11 09:41:03 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 2.0 (TID 4) in 92 ms on sudipta.com (executor 0) (2/2)

17/06/11 09:41:03 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool

17/06/11 09:41:03 INFO scheduler.DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[9] at sortBy at sess5\_assgn2\_ques2.scala:32), which has no missing parents

17/06/11 09:41:03 INFO storage.MemoryStore: Block broadcast\_4 stored as values in memory (estimated size 3.3 KB, free 534.3 MB)

17/06/11 09:41:03 INFO storage.MemoryStore: Block broadcast\_4\_piece0 stored as bytes in memory (estimated size 1932.0 B, free 534.3 MB)

17/06/11 09:41:03 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on 192.168.252.128:49287 (size: 1932.0 B, free: 534.5 MB)

17/06/11 09:41:03 INFO spark.SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1006

17/06/11 09:41:03 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ResultStage 3 (MapPartitionsRDD[9] at sortBy at sess5\_assgn2\_ques2.scala:32)

17/06/11 09:41:03 INFO scheduler.TaskSchedulerImpl: Adding task set 3.0 with 2 tasks

17/06/11 09:41:04 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 5, sudipta.com, executor 0, partition 0, NODE\_LOCAL, 1967 bytes)

17/06/11 09:41:04 INFO storage.BlockManagerInfo: Added broadcast\_4\_piece0 in memory on sudipta.com:58488 (size: 1932.0 B, free: 534.5 MB)

17/06/11 09:41:04 INFO spark.MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to sudipta.com:54233

17/06/11 09:41:04 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 153 bytes

17/06/11 09:41:07 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 3.0 (TID 6, sudipta.com, executor 0, partition 1, NODE\_LOCAL, 1967 bytes)

17/06/11 09:41:07 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 5) in 3777 ms on sudipta.com (executor 0) (1/2)

17/06/11 09:41:08 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 3.0 (TID 6) in 169 ms on sudipta.com (executor 0) (2/2)

17/06/11 09:41:08 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool

17/06/11 09:41:08 INFO scheduler.DAGScheduler: ResultStage 3 (collect at sess5\_assgn2\_ques2.scala:32) finished in 3.934 s

17/06/11 09:41:08 INFO scheduler.DAGScheduler: Job 2 finished: collect at sess5\_assgn2\_ques2.scala:32, took 6.223290 s

The Object of Beauty

The Nightmare Before Christmas

The Nightmare Before Christmas

The Nightmare Before Christmas

The Mummy

Orphans of the Storm

One Magic Christmas

Nosferatu: Original Version

Nosferatu: Original Version

Night Tide

Night Tide

Night Tide

Nick of Time

Muriel's Wedding

Mother's Boys