Name: Arun Kumar N
Student ID : 2017CBDE030
Spark Assignment

Pig Run Logs:

1. **Single Row Lookup**

```
[root@ip-10-0-0-229 ~]# pig firstprob.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2018-05-02 11:34:43,026 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.14.2 (rUnversioned directory) compiled Mar 27 2018, 13:35:40
2018-05-02 11:34:43,027 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1525260883003.log
2018-05-02 11:34:43,844 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2018-05-02 11:34:43,953 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-02 11:34:43,954 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:34:43,954 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://ip-10-0-0-100.ec2.internal:8020
2018-05-02 11:34:44,432 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:34:44,475 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:34:44,513 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:34:44,551 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:34:44,599 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:34:44,633 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

2018-05-02 11:34:44,666 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:34:44,699 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:34:45,418 [main] WARN org.apache.pig.PigServer - Encountered
Warning IMPLICIT_CAST_TO_CHARARRAY 5 time(s).
2018-05-02 11:34:45,437 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig
features used in the script: FILTER
2018-05-02 11:34:45,469 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter,
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}
2018-05-02 11:34:45,490 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation -
mapred.textoutputformat.separator is deprecated. Instead, use
mapreduce.output.textoutputformat.separator
2018-05-02 11:34:45,566 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File
concatenation threshold: 100 optimistic? false
2018-05-02 11:34:45,587 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
r - MR plan size before optimization: 1
2018-05-02 11:34:45,587 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
r - MR plan size after optimization: 1
2018-05-02 11:34:45,705 [main] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.229:8032
2018-05-02 11:34:46,420 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig
script settings are added to the job
2018-05-02 11:34:46,478 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation -
mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use
mapreduce.reduce.markreset.buffer.percent

2018-05-02 11:34:46,478 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-05-02 11:34:46,478 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is
deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2018-05-02 11:34:49,643 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- creating jar file Job1657741534035870248.jar
2018-05-02 11:34:53,484 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- jar file Job1657741534035870248.jar created
2018-05-02 11:34:53,484 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.jar is deprecated. Instead,
use mapreduce.job.jar
2018-05-02 11:34:53,503 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting up single store job
2018-05-02 11:34:53,554 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 1 map-reduce job(s) waiting for submission.
2018-05-02 11:34:53,555 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is
deprecated. Instead, use mapreduce.jobtracker.http.address
2018-05-02 11:34:53,561 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.229:8032
2018-05-02 11:34:53,582 [JobControl] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:34:54,524 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
10
2018-05-02 11:34:54,534 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 115
2018-05-02 11:34:54,626 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(combined) to process : 36
2018-05-02 11:34:55,031 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:36

2018-05-02 11:34:55,684 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job:
job_1525255285129_0002
2018-05-02 11:34:55,976 [JobControl] INFO
org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application
application_1525255285129_0002
2018-05-02 11:34:56,012 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The
url to track the job: http://ip-10-0-0-
229.ec2.internal:8088/proxy/application_1525255285129_0002/
2018-05-02 11:34:56,013 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - HadoopJobId: job_1525255285129_0002
2018-05-02 11:34:56,013 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Processing aliases data,filtered
2018-05-02 11:34:56,013 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - detailed locations: M: data[1,7],filtered[20,11] C: R:
2018-05-02 11:34:56,052 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 0% complete
2018-05-02 11:35:26,266 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 4% complete
2018-05-02 11:35:41,637 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 8% complete
2018-05-02 11:35:57,987 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 12% complete
2018-05-02 11:36:11,147 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 16% complete
2018-05-02 11:36:27,753 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 20% complete
2018-05-02 11:36:43,547 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 25% complete

2018-05-02 11:36:59,614 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 29% complete
2018-05-02 11:37:18,243 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 34% complete
2018-05-02 11:37:35,740 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 38% complete
2018-05-02 11:37:49,935 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 43% complete
2018-05-02 11:38:06,144 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 47% complete
2018-05-02 11:38:21,712 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is
deprecated. Instead, use mapreduce.job.reduces
2018-05-02 11:38:21,750 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 100% complete
2018-05-02 11:38:21,752 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats -
Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.14.2 0.12.0-cdh5.14.2 root 2018-05-02 11:34:46 2018-05-02 11:38:21
FILTER

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTIme AvgMapTime
MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime
MedianReducetime Alias Feature Outputs
job_1525255285129_0002 36 0 19 9 14 14 n/a n/a n/a n/a
data,filtered MAP_ONLY
/user/root/spark_assignment/input_dataset/yellow_tripdata_*/output7.out,

Input(s):
Successfully read 54549040 records (4812997267 bytes) from:
"/user/root/spark_assignment/input_dataset/yellow_tripdata_*"

Output(s):
Successfully stored 1 records (90 bytes) in:
"/user/root/spark_assignment/input_dataset/yellow_tripdata_*/output7.out"

Counters:
Total records written : 1
Total bytes written : 90
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1525255285129_0002

2018-05-02 11:38:21,860 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
[root@ip-10-0-0-229 ~]#

2. Filter

[root@ip-10-0-0-229 ~]# pig secondprob.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2018-05-02 11:42:12,253 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.14.2 (rUnversioned directory) compiled Mar 27 2018, 13:35:40
2018-05-02 11:42:12,254 [main] INFO org.apache.pig.Main - Logging error messages to: /root/pig_1525261332233.log
2018-05-02 11:42:13,101 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2018-05-02 11:42:13,216 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-05-02 11:42:13,216 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

2018-05-02 11:42:13,216 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://ip-10-0-0-229.ec2.internal:8020
2018-05-02 11:42:13,720 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,762 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,794 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,847 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,880 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,913 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,945 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:13,974 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:42:14,632 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2018-05-02 11:42:14,665 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2018-05-02 11:42:14,690 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator

2018-05-02 11:42:14,781 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2018-05-02 11:42:14,804 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2018-05-02 11:42:14,804 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2018-05-02 11:42:14,913 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.100:8032
2018-05-02 11:42:15,126 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2018-05-02 11:42:15,191 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation -
mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2018-05-02 11:42:15,191 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-05-02 11:42:15,191 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2018-05-02 11:42:16,173 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job4431028902542644732.jar
2018-05-02 11:42:19,862 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job4431028902542644732.jar created
2018-05-02 11:42:19,862 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.jar is deprecated. Instead, use mapreduce.job.jar
2018-05-02 11:42:19,883 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2018-05-02 11:42:19,890 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2018-05-02 11:42:19,891 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2018-05-02 11:42:19,891 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []

2018-05-02 11:42:19,938 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 1 map-reduce job(s) waiting for submission.
2018-05-02 11:42:19,939 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is
deprecated. Instead, use mapreduce.jobtracker.http.address
2018-05-02 11:42:19,944 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.100:8032
2018-05-02 11:42:19,983 [JobControl] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:42:20,859 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
11
2018-05-02 11:42:20,880 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 151
2018-05-02 11:42:20,976 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(combined) to process : 36
2018-05-02 11:42:21,609 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:36
2018-05-02 11:42:22,330 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job:
job_1525255285129_0003
2018-05-02 11:42:22,605 [JobControl] INFO
org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application
application_1525255285129_0003
2018-05-02 11:42:22,641 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The
url to track the job: http://ip-10-0-0-
229.ec2.internal:8088/proxy/application_1525255285129_0003/
2018-05-02 11:42:22,641 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - HadoopJobId: job_1525255285129_0003
2018-05-02 11:42:22,641 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Processing aliases data,filtered
2018-05-02 11:42:22,641 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - detailed locations: M: data[1,6],data[-1,-1],filtered[20,11] C: R:

2018-05-02 11:42:22,677 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 0% complete
2018-05-02 11:42:55,417 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 4% complete
2018-05-02 11:43:18,909 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 8% complete
2018-05-02 11:43:40,948 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 12% complete
2018-05-02 11:44:03,313 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 16% complete
2018-05-02 11:44:25,268 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 20% complete
2018-05-02 11:44:48,324 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 25% complete
2018-05-02 11:45:10,166 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 29% complete
2018-05-02 11:45:29,556 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 33% complete
2018-05-02 11:45:53,994 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 37% complete
2018-05-02 11:46:14,637 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 41% complete
2018-05-02 11:46:34,578 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 45% complete
2018-05-02 11:46:55,675 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 50% complete

2018-05-02 11:46:58,544 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2018-05-02 11:46:58,598 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-05-02 11:46:58,600 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.14.2 0.12.0-cdh5.14.2 root 2018-05-02 11:42:15 2018-05-02 11:46:58 FILTER

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTIme AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1525255285129_0003 36 0 23 17 20 20 n/a n/a n/a n/a data,filtered MAP_ONLY /user/root/spark_assignment/input_dataset/yellow_tripdata_*/output5.out,

Input(s):
Successfully read 54549041 records (4812997496 bytes) from: "/user/root/spark_assignment/input_dataset/yellow_tripdata_*"

Output(s):
Successfully stored 31029 records (2917814 bytes) in: "/user/root/spark_assignment/input_dataset/yellow_tripdata_*/output5.out"

Counters:
Total records written : 31029
Total bytes written : 2917814
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1525255285129_0003

2018-05-02 11:46:58,734 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 31120 time(s).
2018-05-02 11:46:58,734 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 496688 time(s).
2018-05-02 11:46:58,734 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - Success!
[root@ip-10-0-0-229 ~]#

3. **GroupBy**

[root@ip-10-0-0-229 ~]# pig taxi3.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2018-05-02 11:50:04,610 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-
cdh5.14.2 (rUnversioned directory) compiled Mar 27 2018, 13:35:40
2018-05-02 11:50:04,610 [main] INFO org.apache.pig.Main - Logging error messages to:
/root/pig_1525261804587.log
2018-05-02 11:50:05,413 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file
/root/.pigbootup not found
2018-05-02 11:50:05,522 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
Instead, use mapreduce.jobtracker.address
2018-05-02 11:50:05,522 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:05,522 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to
hadoop file system at: hdfs://ip-10-0-0-229.ec2.internal:8020
2018-05-02 11:50:05,993 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:06,039 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:06,077 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS

2018-05-02 11:50:06,115 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:06,188 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:06,226 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:06,278 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:06,313 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:50:07,049 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig
features used in the script: GROUP_BY,ORDER_BY
2018-05-02 11:50:07,082 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter,
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}
2018-05-02 11:50:07,104 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation -
mapred.textoutputformat.separator is deprecated. Instead, use
mapreduce.output.textoutputformat.separator
2018-05-02 11:50:07,190 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File
concatenation threshold: 100 optimistic? false
2018-05-02 11:50:07,226 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer
- Choosing to move algebraic foreach to combiner
2018-05-02 11:50:07,248 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
r - MR plan size before optimization: 3
2018-05-02 11:50:07,249 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimize
r - MR plan size after optimization: 3

2018-05-02 11:50:07,370 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.229:8032
2018-05-02 11:50:07,585 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2018-05-02 11:50:07,663 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2018-05-02 11:50:07,663 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-05-02 11:50:07,664 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2018-05-02 11:50:07,666 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2018-05-02 11:50:07,668 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2018-05-02 11:50:07,715 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=4813731354
2018-05-02 11:50:07,715 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 5
2018-05-02 11:50:07,715 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2018-05-02 11:50:08,717 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job2459744513782301268.jar
2018-05-02 11:50:12,531 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job2459744513782301268.jar created
2018-05-02 11:50:12,531 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.jar is deprecated. Instead, use mapreduce.job.jar

2018-05-02 11:50:12,549 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2018-05-02 11:50:12,558 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2018-05-02 11:50:12,558 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2018-05-02 11:50:12,558 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2018-05-02 11:50:12,646 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2018-05-02 11:50:12,647 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.address
2018-05-02 11:50:12,653 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.229:8032
2018-05-02 11:50:12,673 [JobControl] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-05-02 11:50:13,497 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 12
2018-05-02 11:50:13,519 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 187
2018-05-02 11:50:13,657 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 36
2018-05-02 11:50:14,112 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:36
2018-05-02 11:50:14,740 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1525255285129_0004
2018-05-02 11:50:14,928 [JobControl] INFO
org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1525255285129_0004
2018-05-02 11:50:14,964 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-10-0-0-229.ec2.internal:8088/proxy/application_1525255285129_0004/

2018-05-02 11:50:14,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1525255285129_0004
2018-05-02 11:50:14,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases counted2,counted3,data,grouped
2018-05-02 11:50:14,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: data[1,6],data[-1,-1],counted2[23,11],grouped[21,10] C: counted2[23,11],grouped[21,10] R: counted2[23,11],counted3[24,11]
2018-05-02 11:50:15,019 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2018-05-02 11:51:45,733 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4% complete
2018-05-02 11:53:08,214 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 8% complete
2018-05-02 11:54:29,853 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 12% complete
2018-05-02 11:55:36,774 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 16% complete
2018-05-02 11:56:07,930 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 21% complete
2018-05-02 11:56:11,709 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 26% complete
2018-05-02 11:56:16,033 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 33% complete
2018-05-02 11:56:21,506 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2018-05-02 11:56:21,519 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3

2018-05-02 11:56:21,520 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Reduce phase detected, estimating # of required reducers.
2018-05-02 11:56:21,520 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Using reducer estimator:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEst
imator
2018-05-02 11:56:21,526 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEst
imator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=84
2018-05-02 11:56:21,527 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting Parallelism to 1
2018-05-02 11:56:22,099 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- creating jar file Job9118603956817390515.jar
2018-05-02 11:56:25,732 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- jar file Job9118603956817390515.jar created
2018-05-02 11:56:25,742 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting up single store job
2018-05-02 11:56:25,742 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key
[pig.schematuple] is false, will not generate code.
2018-05-02 11:56:25,743 [main] INFO org.apache.pig.data.SchemaTupleFrontend -
Starting process to move generated code to distributed cache
2018-05-02 11:56:25,743 [main] INFO org.apache.pig.data.SchemaTupleFrontend -
Setting key [pig.schematuple.classes] with classes to deserialize []
2018-05-02 11:56:25,764 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 1 map-reduce job(s) waiting for submission.
2018-05-02 11:56:25,769 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.229:8032
2018-05-02 11:56:25,774 [JobControl] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:56:26,260 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
5

2018-05-02 11:56:26,260 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 5
2018-05-02 11:56:26,260 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2018-05-02 11:56:26,283 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2018-05-02 11:56:26,714 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1525255285129_0005
2018-05-02 11:56:26,943 [JobControl] INFO
org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1525255285129_0005
2018-05-02 11:56:26,944 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-10-0-0-229.ec2.internal:8088/proxy/application_1525255285129_0005/
2018-05-02 11:56:26,945 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1525255285129_0005
2018-05-02 11:56:26,945 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases final
2018-05-02 11:56:26,945 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: final[25,8] C: R:
2018-05-02 11:56:38,761 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2018-05-02 11:56:46,201 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 66% complete
2018-05-02 11:56:52,525 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2018-05-02 11:56:52,534 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-05-02 11:56:52,535 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.

2018-05-02 11:56:52,535 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting Parallelism to 1
2018-05-02 11:56:53,472 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- creating jar file Job952175026799713181.jar
2018-05-02 11:56:57,061 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- jar file Job952175026799713181.jar created
2018-05-02 11:56:57,070 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting up single store job
2018-05-02 11:56:57,070 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key
[pig.schematuple] is false, will not generate code.
2018-05-02 11:56:57,070 [main] INFO org.apache.pig.data.SchemaTupleFrontend -
Starting process to move generated code to distributed cache
2018-05-02 11:56:57,071 [main] INFO org.apache.pig.data.SchemaTupleFrontend -
Setting key [pig.schematuple.classes] with classes to deserialize []
2018-05-02 11:56:57,087 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaunche
r - 1 map-reduce job(s) waiting for submission.
2018-05-02 11:56:57,089 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at ip-10-0-0-229.ec2.internal/10.0.0.229:8032
2018-05-02 11:56:57,094 [JobControl] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
Instead, use fs.defaultFS
2018-05-02 11:56:57,166 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
5
2018-05-02 11:56:57,166 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
process : 5
2018-05-02 11:56:57,166 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(combined) to process : 1
2018-05-02 11:56:57,998 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2018-05-02 11:56:58,014 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job:
job_1525255285129_0006

2018-05-02 11:56:58,048 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1525255285129_0006
2018-05-02 11:56:58,049 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://ip-10-0-0-229.ec2.internal:8088/proxy/application_1525255285129_0006/
2018-05-02 11:56:58,049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1525255285129_0006
2018-05-02 11:56:58,049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases final
2018-05-02 11:56:58,049 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: final[25,8] C: R:
2018-05-02 11:57:10,108 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 83% complete
2018-05-02 11:57:18,592 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-05-02 11:57:18,606 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.14.2 0.12.0-cdh5.14.2 root 2018-05-02 11:50:07 2018-05-02 11:57:18 GROUP_BY,ORDER_BY

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTIme AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1525255285129_0004 36 5 28 19 25 25 69 419 5 counted2,counted3,data,grouped GROUP_BY,COMBINER
job_1525255285129_0005 1 1 3 3 3 3 5 55 5 final SAMPLER
job_1525255285129_0006 1 1 4 4 4 4 3 33 3 final ORDER_BY
/user/root/spark_assignment/input_dataset/yellow_tripdata_*/output6.out,

Input(s):
Successfully read 54580070 records (4815920314 bytes) from:
"/user/root/spark_assignment/input_dataset/yellow_tripdata_*"

Output(s):
Successfully stored 5 records (42 bytes) in:
"/user/root/spark_assignment/input_dataset/yellow_tripdata_*/output6.out"

Counters:
Total records written : 5
Total bytes written : 42
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1525255285129_0004 -> job_1525255285129_0005,
job_1525255285129_0005 -> job_1525255285129_0006,
job_1525255285129_0006

2018-05-02 11:57:19,608 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: ip-10-0-0-229.ec2.internal/10.0.0.229:46177. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2018-05-02 11:57:20,609 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: ip-10-0-0-229.ec2.internal/10.0.0.229:46177. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2018-05-02 11:57:21,609 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: ip-10-0-0-229.ec2.internal/10.0.0.229:46177. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2018-05-02 11:57:21,716 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-02 11:57:23,001 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: ip-10-0-0-229.ec2.internal/10.0.0.229:34278. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)

2018-05-02 11:57:24,002 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: ip-10-0-0-229.ec2.internal/10.0.0.229:34278. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2018-05-02 11:57:25,002 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: ip-10-0-0-229.ec2.internal/10.0.0.229:34278. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2018-05-02 11:57:25,105 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-02 11:57:25,276 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-05-02 11:57:25,331 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 993152 time(s).
2018-05-02 11:57:25,331 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 62149 time(s).
2018-05-02 11:57:25,331 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
[root@ip-10-0-0-229 ~]#

Spark Run Logs

1. **Single Row Lookup**

   [root@ip-10-0-0-229 ~]# spark-submit --class com.spark.Assignment1.SparkTask1 --master yarn --deploy-mode client --name testproject --conf "spark.app.id=testproject" testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar /user/root/spark_assignment/input_dataset/yellow_tripdata_* /user/root/spark_assignment/output/single_row_lookup_SparkRDD
   18/05/03 05:11:57 INFO spark.SparkContext: Running Spark version 1.6.0
   18/05/03 05:11:58 INFO spark.SecurityManager: Changing view acls to: root
   18/05/03 05:11:58 INFO spark.SecurityManager: Changing modify acls to: root
   18/05/03 05:11:58 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); users with modify permissions: Set(root)

18/05/03 05:11:58 INFO util.Utils: Successfully started service 'sparkDriver' on port 43970.
18/05/03 05:11:59 INFO slf4j.Slf4jLogger: Slf4jLogger started
18/05/03 05:11:59 INFO Remoting: Starting remoting
18/05/03 05:11:59 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.0.229:45397]
18/05/03 05:11:59 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriverActorSystem@10.0.0.229:45397]
18/05/03 05:11:59 INFO util.Utils: Successfully started service 'sparkDriverActorSystem' on port 45397.
18/05/03 05:11:59 INFO spark.SparkEnv: Registering MapOutputTracker
18/05/03 05:11:59 INFO spark.SparkEnv: Registering BlockManagerMaster
18/05/03 05:11:59 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-900364c9-1348-4b1e-b226-70cc6627146a
18/05/03 05:11:59 INFO storage.MemoryStore: MemoryStore started with capacity 530.0 MB
18/05/03 05:11:59 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/05/03 05:11:59 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
18/05/03 05:11:59 INFO ui.SparkUI: Started SparkUI at http://10.0.0.229:4040
18/05/03 05:11:59 INFO spark.SparkContext: Added JAR file:/root/testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar at spark://10.0.0.229:43970/jars/testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar with timestamp 1525324319684
18/05/03 05:11:59 INFO executor.Executor: Starting executor ID driver on host localhost
18/05/03 05:11:59 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40313.
18/05/03 05:11:59 INFO netty.NettyBlockTransferService: Server created on 40313
18/05/03 05:11:59 INFO storage.BlockManager: external shuffle service port = 7337
18/05/03 05:11:59 INFO storage.BlockManagerMaster: Trying to register BlockManager
18/05/03 05:11:59 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:40313 with 530.0 MB RAM, BlockManagerId(driver, localhost, 40313)
18/05/03 05:11:59 INFO storage.BlockManagerMaster: Registered BlockManager
18/05/03 05:12:00 INFO scheduler.EventLoggingListener: Logging events to hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/spark/applicationHistory/local-1525324319746

18/05/03 05:12:01 INFO spark.SparkContext: Registered listener com.cloudera.spar
k.lineage.ClouderaNavigatorListener
18/05/03 05:12:01 INFO storage.MemoryStore: Block broadcast_0 stored as values i
n memory (estimated size 303.4 KB, free 529.7 MB)
18/05/03 05:12:01 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as
b
ytes in memory (estimated size 26.1 KB, free 529.7 MB)
18/05/03 05:12:01 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in
mem
ory on localhost:40313 (size: 26.1 KB, free: 530.0 MB)
18/05/03 05:12:01 INFO spark.SparkContext: Created broadcast 0 from textFile at
SparkTask1.java:21
18/05/03 05:12:01 INFO mapred.FileInputFormat: Total input paths to process : 7
18/05/03 05:12:01 INFO spark.SparkContext: Starting job: foreach at SparkTask1.j
ava:24
18/05/03 05:12:01 INFO scheduler.DAGScheduler: Got job 0 (foreach at SparkTask1.
java:24) with 46 output partitions
18/05/03 05:12:01 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (forea
ch at SparkTask1.java:24)
18/05/03 05:12:01 INFO scheduler.DAGScheduler: Parents of final stage: List()
18/05/03 05:12:01 INFO scheduler.DAGScheduler: Missing parents: List()
18/05/03 05:12:01 INFO scheduler.DAGScheduler: Submitting ResultStage 0
(MapPart
itionsRDD[4] at map at SparkTask1.java:24), which has no missing parents
18/05/03 05:12:01 INFO storage.MemoryStore: Block broadcast_1 stored as values i
n memory (estimated size 5.0 KB, free 529.7 MB)
18/05/03 05:12:01 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as
b
ytes in memory (estimated size 2.7 KB, free 529.7 MB)
18/05/03 05:12:01 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in
mem
ory on localhost:40313 (size: 2.7 KB, free: 530.0 MB)
18/05/03 05:12:01 INFO spark.SparkContext: Created broadcast 1 from broadcast at
DAGScheduler.scala:1004
18/05/03 05:12:01 INFO scheduler.DAGScheduler: Submitting 46 missing tasks from
ResultStage 0 (MapPartitionsRDD[4] at map at SparkTask1.java:24) (first 15 tasks
are for partitions Vector(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14))
18/05/03 05:12:01 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 46
tasks
18/05/03 05:12:01 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0
(TID 0, localhost, executor driver, partition 0, ANY, 2316 bytes)

18/05/03 05:12:01 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, localhost, executor driver, partition 1, ANY, 2316 bytes)
18/05/03 05:12:01 INFO scheduler.TaskSetManager: Starting task 2.0 in stage 0.0 (TID 2, localhost, executor driver, partition 2, ANY, 2316 bytes)
18/05/03 05:12:01 INFO scheduler.TaskSetManager: Starting task 3.0 in stage 0.0 (TID 3, localhost, executor driver, partition 3, ANY, 2316 bytes)
18/05/03 05:12:01 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
18/05/03 05:12:01 INFO spark.ExecutorAllocationManager: New executor driver has registered (new total is 1)
18/05/03 05:12:01 INFO executor.Executor: Running task 2.0 in stage 0.0 (TID 2)
18/05/03 05:12:01 INFO executor.Executor: Running task 3.0 in stage 0.0 (TID 3)
18/05/03 05:12:01 INFO executor.Executor: Running task 1.0 in stage 0.0 (TID 1)
18/05/03 05:12:01 INFO executor.Executor: Fetching spark://10.0.0.229:43970/jars/testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar with timestamp 15253243196
84
18/05/03 05:12:02 INFO util.Utils: Fetching spark://10.0.0.229:43970/jars/testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar to /tmp/spark-a3c1fcac-b49c-4350-
ac5a-e4981818e389/userFiles-d4ea7c8c-a7bb-41b7-bb3b-5aadbf2038f9/fetchFileTemp12
16204959957328190.tmp
18/05/03 05:12:02 INFO executor.Executor: Adding file:/tmp/spark-a3c1fcac-b49c-4350-ac5a-e4981818e389/userFiles-d4ea7c8c-a7bb-41b7-bb3b-5aadbf2038f9/testproject
-0.0.1-SNAPSHOT-jar-with-dependencies.jar to class loader
18/05/03 05:12:02 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:0+134217728
18/05/03 05:12:02 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:402653184+134217728
18/05/03 05:12:02 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:134217728+134217728
18/05/03 05:12:02 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-

1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:268435456+134217728
18/05/03 05:12:02 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
18/05/03 05:12:02 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
18/05/03 05:12:02 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
18/05/03 05:12:02 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
18/05/03 05:12:02 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
18/05/03 05:12:02 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:12:02 WARN spark.ExecutorAllocationManager: Unable to reach the cluster manager to request 2 total executors!
18/05/03 05:12:03 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:12:03 WARN spark.ExecutorAllocationManager: Unable to reach the cluster manager to request 3 total executors!
18/05/03 05:12:04 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:12:04 WARN spark.ExecutorAllocationManager: Unable to reach the cluster manager to request 4 total executors!
18/05/03 05:12:05 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:12:05 WARN spark.ExecutorAllocationManager: Unable to reach the cluster manager to request 5 total executors!
18/05/03 05:12:06 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode

18/05/03 05:12:06 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 6 total executors!
18/05/03 05:12:07 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:07 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 7 total executors!
18/05/03 05:12:08 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:08 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 8 total executors!
18/05/03 05:12:09 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:09 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 9 total executors!
18/05/03 05:12:10 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:10 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 10 total executors!
18/05/03 05:12:11 INFO executor.Executor: Finished task 1.0 in stage 0.0 (TID 1)
. 2002 bytes result sent to driver
18/05/03 05:12:11 INFO executor.Executor: Finished task 2.0 in stage 0.0 (TID 2)
. 2002 bytes result sent to driver
18/05/03 05:12:11 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:11 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 11 total executors!
18/05/03 05:12:11 INFO scheduler.TaskSetManager: Starting task 4.0 in stage 0.0
(TID 4, localhost, executor driver, partition 4, ANY, 2316 bytes)
18/05/03 05:12:11 INFO executor.Executor: Running task 4.0 in stage 0.0 (TID 4)

18/05/03 05:12:11 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:536870912+134217728
18/05/03 05:12:11 INFO scheduler.TaskSetManager: Starting task 5.0 in stage 0.0
(TID 5, localhost, executor driver, partition 5, ANY, 2316 bytes)
18/05/03 05:12:11 INFO executor.Executor: Running task 5.0 in stage 0.0 (TID 5)
18/05/03 05:12:11 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:671088640+86047889
18/05/03 05:12:11 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 0.0
(TID 1) in 9987 ms on localhost (executor driver) (1/46)
18/05/03 05:12:11 INFO scheduler.TaskSetManager: Finished task 2.0 in stage 0.0
(TID 2) in 9990 ms on localhost (executor driver) (2/46)
18/05/03 05:12:11 INFO executor.Executor: Finished task 3.0 in stage 0.0 (TID 3)
. 2002 bytes result sent to driver
18/05/03 05:12:11 INFO scheduler.TaskSetManager: Starting task 6.0 in stage 0.0
(TID 6, localhost, executor driver, partition 6, ANY, 2316 bytes)
18/05/03 05:12:11 INFO executor.Executor: Running task 6.0 in stage 0.0 (TID 6)
18/05/03 05:12:11 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-08.csv:0+134217728
18/05/03 05:12:11 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 0.0
(TID 3) in 10034 ms on localhost (executor driver) (3/46)
18/05/03 05:12:12 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0)
. 2057 bytes result sent to driver
18/05/03 05:12:12 INFO scheduler.TaskSetManager: Starting task 7.0 in stage 0.0
(TID 7, localhost, executor driver, partition 7, ANY, 2316 bytes)
18/05/03 05:12:12 INFO executor.Executor: Running task 7.0 in stage 0.0 (TID 7)
18/05/03 05:12:12 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0
(TID 0) in 10225 ms on localhost (executor driver) (4/46)
18/05/03 05:12:12 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-08.csv:134217728+134217728
18/05/03 05:12:12 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode

18/05/03 05:12:12 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 12 total executors!
18/05/03 05:12:13 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:13 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 13 total executors!
18/05/03 05:12:14 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:14 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 14 total executors!
18/05/03 05:12:15 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:15 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 15 total executors!
18/05/03 05:12:16 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:16 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 16 total executors!
18/05/03 05:12:16 INFO executor.Executor: Finished task 5.0 in stage 0.0 (TID 5)
. 2057 bytes result sent to driver
18/05/03 05:12:16 INFO scheduler.TaskSetManager: Starting task 8.0 in stage 0.0
(TID 8, localhost, executor driver, partition 8, ANY, 2316 bytes)
18/05/03 05:12:16 INFO executor.Executor: Running task 8.0 in stage 0.0 (TID 8)
18/05/03 05:12:16 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-08.csv:268435456+134217728
18/05/03 05:12:16 INFO scheduler.TaskSetManager: Finished task 5.0 in stage 0.0
(TID 5) in 5081 ms on localhost (executor driver) (5/46)
18/05/03 05:12:17 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode

18/05/03 05:12:17 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 17 total executors!
18/05/03 05:12:18 INFO executor.Executor: Finished task 4.0 in stage 0.0 (TID 4)
. 2057 bytes result sent to driver
18/05/03 05:12:18 INFO scheduler.TaskSetManager: Starting task 9.0 in stage 0.0
(TID 9, localhost, executor driver, partition 9, ANY, 2316 bytes)
18/05/03 05:12:18 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:18 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 18 total executors!
18/05/03 05:12:18 INFO executor.Executor: Running task 9.0 in stage 0.0 (TID 9)
18/05/03 05:12:18 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-08.csv:402653184+134217728
18/05/03 05:12:18 INFO scheduler.TaskSetManager: Finished task 4.0 in stage 0.0
(TID 4) in 7040 ms on localhost (executor driver) (6/46)
18/05/03 05:12:19 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:19 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 19 total executors!
18/05/03 05:12:19 INFO executor.Executor: Finished task 7.0 in stage 0.0 (TID 7)
. 2057 bytes result sent to driver
18/05/03 05:12:19 INFO scheduler.TaskSetManager: Starting task 10.0 in stage 0.0
(TID 10, localhost, executor driver, partition 10, ANY, 2316 bytes)
18/05/03 05:12:19 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 0.0
(TID 7) in 7863 ms on localhost (executor driver) (7/46)
18/05/03 05:12:19 INFO executor.Executor: Running task 10.0 in stage 0.0 (TID 10
)
18/05/03 05:12:19 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-08.csv:536870912+134217728
18/05/03 05:12:20 INFO executor.Executor: Finished task 6.0 in stage 0.0 (TID 6)
. 2057 bytes result sent to driver

18/05/03 05:12:20 INFO scheduler.TaskSetManager: Starting task 11.0 in stage 0.0 (TID 11, localhost, executor driver, partition 11, ANY, 2316 bytes)
18/05/03 05:12:20 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 0.0 (TID 6) in 8150 ms on localhost (executor driver) (8/46)
18/05/03 05:12:20 INFO executor.Executor: Running task 11.0 in stage 0.0 (TID 11
)
18/05/03 05:12:20 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-08.csv:671088640+71329760
18/05/03 05:12:20 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:20 WARN spark.ExecutorAllocationManager: Unable to reach the
clus
ter manager to request 20 total executors!
18/05/03 05:12:21 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:21 WARN spark.ExecutorAllocationManager: Unable to reach the
clus
ter manager to request 21 total executors!
18/05/03 05:12:22 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:22 WARN spark.ExecutorAllocationManager: Unable to reach the
clus
ter manager to request 22 total executors!
18/05/03 05:12:23 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:23 WARN spark.ExecutorAllocationManager: Unable to reach the
clus
ter manager to request 23 total executors!
18/05/03 05:12:24 INFO executor.Executor: Finished task 8.0 in stage 0.0 (TID 8)
. 2057 bytes result sent to driver
18/05/03 05:12:24 INFO scheduler.TaskSetManager: Starting task 12.0 in stage 0.0 (TID 12, localhost, executor driver, partition 12, ANY, 2316 bytes)
18/05/03 05:12:24 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 0.0 (TID 8) in 7300 ms on localhost (executor driver) (9/46)

18/05/03 05:12:24 INFO executor.Executor: Running task 12.0 in stage 0.0 (TID 12
)
18/05/03 05:12:24 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-09.csv:0+134217728
18/05/03 05:12:24 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:24 WARN spark.ExecutorAllocationManager: Unable to reach the
clus
ter manager to request 24 total executors!
18/05/03 05:12:25 INFO executor.Executor: Finished task 11.0 in stage 0.0 (TID 1
1). 2057 bytes result sent to driver
18/05/03 05:12:25 INFO scheduler.TaskSetManager: Starting task 13.0 in stage 0.0
(TID 13, localhost, executor driver, partition 13, ANY, 2316 bytes)
18/05/03 05:12:25 INFO scheduler.TaskSetManager: Finished task 11.0 in stage 0.0
(TID 11) in 5707 ms on localhost (executor driver) (10/46)
18/05/03 05:12:25 INFO executor.Executor: Running task 13.0 in stage 0.0 (TID 13
)
18/05/03 05:12:25 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-09.csv:134217728+134217728
18/05/03 05:12:25 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:25 WARN spark.ExecutorAllocationManager: Unable to reach the
clus
ter manager to request 25 total executors!
18/05/03 05:12:26 INFO executor.Executor: Finished task 9.0 in stage 0.0 (TID 9)
. 2057 bytes result sent to driver
18/05/03 05:12:26 INFO scheduler.TaskSetManager: Starting task 14.0 in stage 0.0
(TID 14, localhost, executor driver, partition 14, ANY, 2316 bytes)
18/05/03 05:12:26 INFO executor.Executor: Running task 14.0 in stage 0.0 (TID 14
)
18/05/03 05:12:26 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-09.csv:268435456+134217728

18/05/03 05:12:26 INFO scheduler.TaskSetManager: Finished task 9.0 in stage 0.0 (TID 9) in 7770 ms on localhost (executor driver) (11/46)
18/05/03 05:12:26 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:26 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 26 total executors!
18/05/03 05:12:27 INFO executor.Executor: Finished task 10.0 in stage 0.0 (TID 1
0). 2057 bytes result sent to driver
18/05/03 05:12:27 INFO scheduler.TaskSetManager: Starting task 15.0 in stage 0.0 (TID 15, localhost, executor driver, partition 15, ANY, 2316 bytes)
18/05/03 05:12:27 INFO scheduler.TaskSetManager: Finished task 10.0 in stage 0.0 (TID 10) in 7193 ms on localhost (executor driver) (12/46)
18/05/03 05:12:27 INFO executor.Executor: Running task 15.0 in stage 0.0 (TID 15
)
18/05/03 05:12:27 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-09.csv:402653184+134217728
18/05/03 05:12:27 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:27 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 27 total executors!
18/05/03 05:12:28 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:28 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 28 total executors!
18/05/03 05:12:29 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:29 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 29 total executors!
18/05/03 05:12:30 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode

18/05/03 05:12:30 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 30 total executors!
18/05/03 05:12:31 INFO executor.Executor: Finished task 12.0 in stage 0.0 (TID 1
2). 2057 bytes result sent to driver
18/05/03 05:12:31 INFO scheduler.TaskSetManager: Starting task 16.0 in stage 0.0
(TID 16, localhost, executor driver, partition 16, ANY, 2316 bytes)
18/05/03 05:12:31 INFO scheduler.TaskSetManager: Finished task 12.0 in stage 0.0
(TID 12) in 7264 ms on localhost (executor driver) (13/46)
18/05/03 05:12:31 INFO executor.Executor: Running task 16.0 in stage 0.0 (TID 16
)
18/05/03 05:12:31 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-09.csv:536870912+134217728
18/05/03 05:12:31 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:31 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 31 total executors!
18/05/03 05:12:32 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:32 WARN spark.ExecutorAllocationManager: Unable to reach the clus
ter manager to request 32 total executors!
18/05/03 05:12:33 INFO executor.Executor: Finished task 13.0 in stage 0.0 (TID 1
3). 2057 bytes result sent to driver
18/05/03 05:12:33 INFO scheduler.TaskSetManager: Starting task 17.0 in stage 0.0
(TID 17, localhost, executor driver, partition 17, ANY, 2316 bytes)
18/05/03 05:12:33 INFO executor.Executor: Running task 17.0 in stage 0.0 (TID 17
)
18/05/03 05:12:33 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-09.csv:671088640+117983715
18/05/03 05:12:33 INFO scheduler.TaskSetManager: Finished task 13.0 in stage 0.0
(TID 13) in 7853 ms on localhost (executor driver) (14/46)
18/05/03 05:12:34 INFO executor.Executor: Finished task 14.0 in stage 0.0 (TID 1
4). 2057 bytes result sent to driver

18/05/03 05:12:34 INFO scheduler.TaskSetManager: Starting task 18.0 in stage 0.0 (TID 18, localhost, executor driver, partition 18, ANY, 2316 bytes)
18/05/03 05:12:34 INFO scheduler.TaskSetManager: Finished task 14.0 in stage 0.0 (TID 14) in 7857 ms on localhost (executor driver) (15/46)
18/05/03 05:12:34 INFO executor.Executor: Running task 18.0 in stage 0.0 (TID 18
)
18/05/03 05:12:34 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-10.csv:0+134217728
18/05/03 05:12:34 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:34 INFO executor.Executor: Finished task 15.0 in stage 0.0 (TID 1
5). 2057 bytes result sent to driver
18/05/03 05:12:34 INFO scheduler.TaskSetManager: Starting task 19.0 in stage 0.0 (TID 19, localhost, executor driver, partition 19, ANY, 2316 bytes)
18/05/03 05:12:34 INFO executor.Executor: Running task 19.0 in stage 0.0 (TID 19
)
18/05/03 05:12:34 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-10.csv:134217728+134217728
18/05/03 05:12:34 INFO scheduler.TaskSetManager: Finished task 15.0 in stage 0.0 (TID 15) in 7713 ms on localhost (executor driver) (16/46)
18/05/03 05:12:34 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
[2, 2017-10-01 00:15:30, 2017-10-01 00:25:11, 1, 2.17, 1, N, 141, 142, 1, 9, 0.5
, 0.5, 2.06, 0, 0.3, 12.36]
18/05/03 05:12:41 INFO executor.Executor: Finished task 16.0 in stage 0.0 (TID 1
6). 2057 bytes result sent to driver
18/05/03 05:12:41 INFO scheduler.TaskSetManager: Starting task 20.0 in stage 0.0 (TID 20, localhost, executor driver, partition 20, ANY, 2316 bytes)
18/05/03 05:12:41 INFO scheduler.TaskSetManager: Finished task 16.0 in stage 0.0 (TID 16) in 9942 ms on localhost (executor driver) (17/46)
18/05/03 05:12:41 INFO executor.Executor: Running task 20.0 in stage 0.0 (TID 20
)
18/05/03 05:12:41 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-

1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-10.csv:268435456+134217728
18/05/03 05:12:41 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:42 INFO executor.Executor: Finished task 17.0 in stage 0.0 (TID 17). 2057 bytes result sent to driver
18/05/03 05:12:42 INFO scheduler.TaskSetManager: Starting task 21.0 in stage 0.0 (TID 21, localhost, executor driver, partition 21, ANY, 2316 bytes)
18/05/03 05:12:42 INFO scheduler.TaskSetManager: Finished task 17.0 in stage 0.0 (TID 17) in 9319 ms on localhost (executor driver) (18/46)
18/05/03 05:12:42 INFO executor.Executor: Running task 21.0 in stage 0.0 (TID 21)
18/05/03 05:12:42 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-10.csv:402653184+134217728
18/05/03 05:12:42 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:44 INFO executor.Executor: Finished task 19.0 in stage 0.0 (TID 19). 2057 bytes result sent to driver
18/05/03 05:12:44 INFO scheduler.TaskSetManager: Starting task 22.0 in stage 0.0 (TID 22, localhost, executor driver, partition 22, ANY, 2316 bytes)
18/05/03 05:12:44 INFO scheduler.TaskSetManager: Finished task 19.0 in stage 0.0 (TID 19) in 9415 ms on localhost (executor driver) (19/46)
18/05/03 05:12:44 INFO executor.Executor: Running task 22.0 in stage 0.0 (TID 22)
18/05/03 05:12:44 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-10.csv:536870912+134217728
18/05/03 05:12:44 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:12:44 INFO executor.Executor: Finished task 18.0 in stage 0.0 (TID 18). 2057 bytes result sent to driver
18/05/03 05:12:44 INFO scheduler.TaskSetManager: Starting task 23.0 in stage 0.0 (TID 23, localhost, executor driver, partition 23, ANY, 2316 bytes)
18/05/03 05:12:44 INFO scheduler.TaskSetManager: Finished task 18.0 in stage 0.0 (TID 18) in 9989 ms on localhost (executor driver) (20/46)

18/05/03 05:12:44 INFO executor.Executor: Running task 23.0 in stage 0.0 (TID 23
)
18/05/03 05:12:44 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-10.csv:671088640+134217728
18/05/03 05:12:44 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:50 INFO executor.Executor: Finished task 20.0 in stage 0.0 (TID 2
0). 2057 bytes result sent to driver
18/05/03 05:12:50 INFO scheduler.TaskSetManager: Starting task 24.0 in stage 0.0
(TID 24, localhost, executor driver, partition 24, ANY, 2316 bytes)
18/05/03 05:12:50 INFO executor.Executor: Running task 24.0 in stage 0.0 (TID 24
)
18/05/03 05:12:50 INFO scheduler.TaskSetManager: Finished task 20.0 in stage 0.0
(TID 20) in 8788 ms on localhost (executor driver) (21/46)
18/05/03 05:12:50 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-10.csv:805306368+56688482
18/05/03 05:12:50 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:51 INFO executor.Executor: Finished task 21.0 in stage 0.0 (TID 2
1). 2057 bytes result sent to driver
18/05/03 05:12:51 INFO scheduler.TaskSetManager: Starting task 25.0 in stage 0.0
(TID 25, localhost, executor driver, partition 25, ANY, 2316 bytes)
18/05/03 05:12:51 INFO scheduler.TaskSetManager: Finished task 21.0 in stage 0.0
(TID 21) in 8642 ms on localhost (executor driver) (22/46)
18/05/03 05:12:51 INFO executor.Executor: Running task 25.0 in stage 0.0 (TID 25
)
18/05/03 05:12:51 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-11.csv:0+134217728
18/05/03 05:12:51 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:12:52 INFO executor.Executor: Finished task 22.0 in stage 0.0 (TID 2
2). 2057 bytes result sent to driver

18/05/03 05:12:52 INFO scheduler.TaskSetManager: Starting task 26.0 in stage 0.0 (TID 26, localhost, executor driver, partition 26, ANY, 2316 bytes)
18/05/03 05:12:52 INFO executor.Executor: Running task 26.0 in stage 0.0 (TID 26)
18/05/03 05:12:52 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv:134217728+134217728
18/05/03 05:12:52 INFO scheduler.TaskSetManager: Finished task 22.0 in stage 0.0 (TID 22) in 8081 ms on localhost (executor driver) (23/46)
18/05/03 05:12:52 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:12:52 INFO executor.Executor: Finished task 23.0 in stage 0.0 (TID 23). 2057 bytes result sent to driver
18/05/03 05:12:52 INFO scheduler.TaskSetManager: Starting task 27.0 in stage 0.0 (TID 27, localhost, executor driver, partition 27, ANY, 2316 bytes)
18/05/03 05:12:52 INFO scheduler.TaskSetManager: Finished task 23.0 in stage 0.0 (TID 23) in 8038 ms on localhost (executor driver) (24/46)
18/05/03 05:12:52 INFO executor.Executor: Running task 27.0 in stage 0.0 (TID 27)
18/05/03 05:12:52 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv:268435456+134217728
18/05/03 05:12:52 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:12:54 INFO executor.Executor: Finished task 24.0 in stage 0.0 (TID 24). 2057 bytes result sent to driver
18/05/03 05:12:54 INFO scheduler.TaskSetManager: Starting task 28.0 in stage 0.0 (TID 28, localhost, executor driver, partition 28, ANY, 2316 bytes)
18/05/03 05:12:54 INFO executor.Executor: Running task 28.0 in stage 0.0 (TID 28)
18/05/03 05:12:54 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv:402653184+134217728
18/05/03 05:12:54 INFO scheduler.TaskSetManager: Finished task 24.0 in stage 0.0 (TID 24) in 4315 ms on localhost (executor driver) (25/46)

18/05/03 05:12:54 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:00 INFO executor.Executor: Finished task 25.0 in stage 0.0 (TID 2
5). 2057 bytes result sent to driver
18/05/03 05:13:00 INFO scheduler.TaskSetManager: Starting task 29.0 in stage 0.0
(TID 29, localhost, executor driver, partition 29, ANY, 2316 bytes)
18/05/03 05:13:00 INFO scheduler.TaskSetManager: Finished task 25.0 in stage 0.0
(TID 25) in 8497 ms on localhost (executor driver) (26/46)
18/05/03 05:13:00 INFO executor.Executor: Running task 29.0 in stage 0.0 (TID 29
)
18/05/03 05:13:00 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-11.csv:536870912+134217728
18/05/03 05:13:00 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:00 INFO executor.Executor: Finished task 26.0 in stage 0.0 (TID 2
6). 2057 bytes result sent to driver
18/05/03 05:13:00 INFO scheduler.TaskSetManager: Starting task 30.0 in stage 0.0
(TID 30, localhost, executor driver, partition 30, ANY, 2316 bytes)
18/05/03 05:13:00 INFO scheduler.TaskSetManager: Finished task 26.0 in stage 0.0
(TID 26) in 8643 ms on localhost (executor driver) (27/46)
18/05/03 05:13:00 INFO executor.Executor: Running task 30.0 in stage 0.0 (TID 30
)
18/05/03 05:13:00 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-11.csv:671088640+134217728
18/05/03 05:13:01 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:01 INFO executor.Executor: Finished task 27.0 in stage 0.0 (TID 2
7). 2057 bytes result sent to driver
18/05/03 05:13:01 INFO scheduler.TaskSetManager: Starting task 31.0 in stage 0.0
(TID 31, localhost, executor driver, partition 31, ANY, 2316 bytes)
18/05/03 05:13:01 INFO executor.Executor: Running task 31.0 in stage 0.0 (TID 31
)
18/05/03 05:13:01 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-

1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv:805306368+13877504
18/05/03 05:13:01 INFO scheduler.TaskSetManager: Finished task 27.0 in stage 0.0 (TID 27) in 8612 ms on localhost (executor driver) (28/46)
18/05/03 05:13:01 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:03 INFO executor.Executor: Finished task 31.0 in stage 0.0 (TID 3
1). 2057 bytes result sent to driver
18/05/03 05:13:03 INFO scheduler.TaskSetManager: Starting task 32.0 in stage 0.0 (TID 32, localhost, executor driver, partition 32, ANY, 2318 bytes)
18/05/03 05:13:03 INFO executor.Executor: Running task 32.0 in stage 0.0 (TID 32
)
18/05/03 05:13:03 INFO scheduler.TaskSetManager: Finished task 31.0 in stage 0.0 (TID 31) in 2050 ms on localhost (executor driver) (29/46)
18/05/03 05:13:03 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv.1:0+134217728
18/05/03 05:13:03 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:03 INFO executor.Executor: Finished task 28.0 in stage 0.0 (TID 2
8). 2057 bytes result sent to driver
18/05/03 05:13:03 INFO scheduler.TaskSetManager: Starting task 33.0 in stage 0.0 (TID 33, localhost, executor driver, partition 33, ANY, 2318 bytes)
18/05/03 05:13:03 INFO executor.Executor: Running task 33.0 in stage 0.0 (TID 33
)
18/05/03 05:13:03 INFO scheduler.TaskSetManager: Finished task 28.0 in stage 0.0 (TID 28) in 9020 ms on localhost (executor driver) (30/46)
18/05/03 05:13:03 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv.1:134217728+134217728
18/05/03 05:13:03 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:08 INFO executor.Executor: Finished task 29.0 in stage 0.0 (TID 2
9). 2057 bytes result sent to driver
18/05/03 05:13:08 INFO scheduler.TaskSetManager: Starting task 34.0 in stage 0.0 (TID 34, localhost, executor driver, partition 34, ANY, 2318 bytes)

18/05/03 05:13:08 INFO scheduler.TaskSetManager: Finished task 29.0 in stage 0.0 (TID 29) in 8325 ms on localhost (executor driver) (31/46)
18/05/03 05:13:08 INFO executor.Executor: Running task 34.0 in stage 0.0 (TID 34)
18/05/03 05:13:08 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv.1:268435456+134217728
18/05/03 05:13:08 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:13:08 INFO executor.Executor: Finished task 30.0 in stage 0.0 (TID 30). 2057 bytes result sent to driver
18/05/03 05:13:08 INFO scheduler.TaskSetManager: Starting task 35.0 in stage 0.0 (TID 35, localhost, executor driver, partition 35, ANY, 2318 bytes)
18/05/03 05:13:08 INFO scheduler.TaskSetManager: Finished task 30.0 in stage 0.0 (TID 30) in 7768 ms on localhost (executor driver) (32/46)
18/05/03 05:13:08 INFO executor.Executor: Running task 35.0 in stage 0.0 (TID 35)
18/05/03 05:13:08 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv.1:402653184+134217728
18/05/03 05:13:08 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:13:11 INFO executor.Executor: Finished task 32.0 in stage 0.0 (TID 32). 2057 bytes result sent to driver
18/05/03 05:13:11 INFO scheduler.TaskSetManager: Starting task 36.0 in stage 0.0 (TID 36, localhost, executor driver, partition 36, ANY, 2318 bytes)
18/05/03 05:13:11 INFO executor.Executor: Running task 36.0 in stage 0.0 (TID 36)
18/05/03 05:13:11 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-11.csv.1:536870912+134217728
18/05/03 05:13:11 INFO scheduler.TaskSetManager: Finished task 32.0 in stage 0.0 (TID 32) in 8379 ms on localhost (executor driver) (33/46)
18/05/03 05:13:11 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode

18/05/03 05:13:11 INFO executor.Executor: Finished task 33.0 in stage 0.0 (TID 3
3). 2057 bytes result sent to driver
18/05/03 05:13:11 INFO scheduler.TaskSetManager: Starting task 37.0 in stage 0.0
(TID 37, localhost, executor driver, partition 37, ANY, 2318 bytes)
18/05/03 05:13:11 INFO executor.Executor: Running task 37.0 in stage 0.0 (TID 37
)
18/05/03 05:13:11 INFO scheduler.TaskSetManager: Finished task 33.0 in stage 0.0
(TID 33) in 8295 ms on localhost (executor driver) (34/46)
18/05/03 05:13:11 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-11.csv.1:671088640+134217728
18/05/03 05:13:11 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:13:16 INFO executor.Executor: Finished task 34.0 in stage 0.0 (TID 3
4). 2057 bytes result sent to driver
18/05/03 05:13:16 INFO scheduler.TaskSetManager: Starting task 38.0 in stage 0.0
(TID 38, localhost, executor driver, partition 38, ANY, 2318 bytes)
18/05/03 05:13:16 INFO executor.Executor: Running task 38.0 in stage 0.0 (TID 38
)
18/05/03 05:13:16 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-11.csv.1:805306368+13877504
18/05/03 05:13:16 INFO scheduler.TaskSetManager: Finished task 34.0 in stage 0.0
(TID 34) in 8482 ms on localhost (executor driver) (35/46)
18/05/03 05:13:16 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:13:17 INFO executor.Executor: Finished task 35.0 in stage 0.0 (TID 3
5). 2057 bytes result sent to driver
18/05/03 05:13:17 INFO scheduler.TaskSetManager: Starting task 39.0 in stage 0.0
(TID 39, localhost, executor driver, partition 39, ANY, 2316 bytes)
18/05/03 05:13:17 INFO executor.Executor: Running task 39.0 in stage 0.0 (TID 39
)
18/05/03 05:13:17 INFO scheduler.TaskSetManager: Finished task 35.0 in stage 0.0
(TID 35) in 8481 ms on localhost (executor driver) (36/46)
18/05/03 05:13:17 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-

1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-12.csv:0+134217728
18/05/03 05:13:17 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:18 INFO executor.Executor: Finished task 38.0 in stage 0.0 (TID 38). 2057 bytes result sent to driver
18/05/03 05:13:18 INFO scheduler.TaskSetManager: Starting task 40.0 in stage 0.0 (TID 40, localhost, executor driver, partition 40, ANY, 2316 bytes)
18/05/03 05:13:18 INFO executor.Executor: Running task 40.0 in stage 0.0 (TID 40)
18/05/03 05:13:18 INFO scheduler.TaskSetManager: Finished task 38.0 in stage 0.0 (TID 38) in 1438 ms on localhost (executor driver) (37/46)
18/05/03 05:13:18 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-12.csv:134217728+134217728
18/05/03 05:13:18 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:19 INFO executor.Executor: Finished task 36.0 in stage 0.0 (TID 36). 2057 bytes result sent to driver
18/05/03 05:13:19 INFO scheduler.TaskSetManager: Starting task 41.0 in stage 0.0 (TID 41, localhost, executor driver, partition 41, ANY, 2316 bytes)
18/05/03 05:13:19 INFO scheduler.TaskSetManager: Finished task 36.0 in stage 0.0 (TID 36) in 7754 ms on localhost (executor driver) (38/46)
18/05/03 05:13:19 INFO executor.Executor: Running task 41.0 in stage 0.0 (TID 41)
18/05/03 05:13:19 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-12.csv:268435456+134217728
18/05/03 05:13:19 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:19 INFO executor.Executor: Finished task 37.0 in stage 0.0 (TID 37). 2057 bytes result sent to driver
18/05/03 05:13:19 INFO scheduler.TaskSetManager: Starting task 42.0 in stage 0.0 (TID 42, localhost, executor driver, partition 42, ANY, 2316 bytes)
18/05/03 05:13:19 INFO executor.Executor: Running task 42.0 in stage 0.0 (TID 42)

18/05/03 05:13:19 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-12.csv:402653184+134217728
18/05/03 05:13:19 INFO scheduler.TaskSetManager: Finished task 37.0 in stage 0.0
(TID 37) in 7757 ms on localhost (executor driver) (39/46)
18/05/03 05:13:19 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:13:26 INFO executor.Executor: Finished task 39.0 in stage 0.0 (TID 3
9). 2057 bytes result sent to driver
18/05/03 05:13:26 INFO scheduler.TaskSetManager: Starting task 43.0 in stage 0.0
(TID 43, localhost, executor driver, partition 43, ANY, 2316 bytes)
18/05/03 05:13:26 INFO executor.Executor: Running task 43.0 in stage 0.0 (TID 43
)
18/05/03 05:13:26 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-12.csv:536870912+134217728
18/05/03 05:13:26 INFO scheduler.TaskSetManager: Finished task 39.0 in stage 0.0
(TID 39) in 8855 ms on localhost (executor driver) (40/46)
18/05/03 05:13:26 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:13:26 INFO executor.Executor: Finished task 40.0 in stage 0.0 (TID 4
0). 2057 bytes result sent to driver
18/05/03 05:13:26 INFO scheduler.TaskSetManager: Starting task 44.0 in stage 0.0
(TID 44, localhost, executor driver, partition 44, ANY, 2316 bytes)
18/05/03 05:13:26 INFO scheduler.TaskSetManager: Finished task 40.0 in stage 0.0
(TID 40) in 8315 ms on localhost (executor driver) (41/46)
18/05/03 05:13:26 INFO executor.Executor: Running task 44.0 in stage 0.0 (TID 44
)
18/05/03 05:13:26 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-12.csv:671088640+134217728
18/05/03 05:13:26 WARN spark.SparkContext: Requesting executors is only
supporte
d in coarse-grained mode
18/05/03 05:13:27 INFO executor.Executor: Finished task 41.0 in stage 0.0 (TID 4
1). 2057 bytes result sent to driver

18/05/03 05:13:27 INFO scheduler.TaskSetManager: Starting task 45.0 in stage 0.0 (TID 45, localhost, executor driver, partition 45, ANY, 2316 bytes)
18/05/03 05:13:27 INFO executor.Executor: Running task 45.0 in stage 0.0 (TID 45)
18/05/03 05:13:27 INFO scheduler.TaskSetManager: Finished task 41.0 in stage 0.0 (TID 41) in 8637 ms on localhost (executor driver) (42/46)
18/05/03 05:13:27 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-12.csv:805306368+32783040
18/05/03 05:13:27 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:13:28 INFO executor.Executor: Finished task 42.0 in stage 0.0 (TID 42). 2057 bytes result sent to driver
18/05/03 05:13:28 INFO scheduler.TaskSetManager: Finished task 42.0 in stage 0.0 (TID 42) in 8844 ms on localhost (executor driver) (43/46)
18/05/03 05:13:28 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:13:30 INFO executor.Executor: Finished task 45.0 in stage 0.0 (TID 45). 2057 bytes result sent to driver
18/05/03 05:13:30 INFO scheduler.TaskSetManager: Finished task 45.0 in stage 0.0 (TID 45) in 2058 ms on localhost (executor driver) (44/46)
18/05/03 05:13:30 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:13:31 INFO executor.Executor: Finished task 43.0 in stage 0.0 (TID 43). 2057 bytes result sent to driver
18/05/03 05:13:31 INFO scheduler.TaskSetManager: Finished task 43.0 in stage 0.0 (TID 43) in 5873 ms on localhost (executor driver) (45/46)
18/05/03 05:13:31 WARN spark.SparkContext: Requesting executors is only supported in coarse-grained mode
18/05/03 05:13:32 INFO executor.Executor: Finished task 44.0 in stage 0.0 (TID 44). 2057 bytes result sent to driver
18/05/03 05:13:32 INFO scheduler.TaskSetManager: Finished task 44.0 in stage 0.0 (TID 44) in 5507 ms on localhost (executor driver) (46/46)
18/05/03 05:13:32 INFO scheduler.DAGScheduler: ResultStage 0 (foreach at SparkTask1.java:24) finished in 90.260 s

18/05/03 05:13:32 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose t
asks have all completed, from pool
18/05/03 05:13:32 INFO scheduler.DAGScheduler: Job 0 finished: foreach at SparkT
ask1.java:24, took 90.396312 s
18/05/03 05:13:32 WARN spark.SparkContext: Requesting executors is only supporte
d in coarse-grained mode
18/05/03 05:13:32 INFO ui.SparkUI: Stopped Spark web UI at http://10.0.0.229:404
0
18/05/03 05:13:32 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMas
terEndpoint stopped!
18/05/03 05:13:32 INFO storage.MemoryStore: MemoryStore cleared
18/05/03 05:13:32 INFO storage.BlockManager: BlockManager stopped
18/05/03 05:13:32 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
18/05/03 05:13:32 INFO
scheduler.OutputCommitCoordinator$OutputCommitCoordinator
Endpoint: OutputCommitCoordinator stopped!
18/05/03 05:13:32 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shuttin
g down remote daemon.
18/05/03 05:13:32 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote
daemon shut down; proceeding with flushing remote transports.
18/05/03 05:13:32 INFO spark.SparkContext: Successfully stopped SparkContext
1525324317199----------1525324412250
18/05/03 05:13:32 INFO util.ShutdownHookManager: Shutdown hook called
18/05/03 05:13:32 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a
3c1fcac-b49c-4350-ac5a-e4981818e389
18/05/03 05:13:32 INFO Remoting: Remoting shut down
18/05/03 05:13:32 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remotin
g shut down.

2. **Filter**

[root@ip-10-0-0-229 ~]# spark-submit --class com.spark.Assignment2.SparkTask2 --master yarn --deploy-mode client --name testproject --conf "spark.app.id=testpro
ject" testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar /user/root/spark_assi

```
gnment/input_dataset/yellow_tripdata_*
/user/root/spark_assignment/output/Filter
/
18/05/03 05:27:38 INFO spark.SparkContext: Running Spark version 1.6.0
18/05/03 05:27:39 INFO spark.SecurityManager: Changing view acls to: root
18/05/03 05:27:39 INFO spark.SecurityManager: Changing modify acls to: root
18/05/03 05:27:39 INFO spark.SecurityManager: SecurityManager: authentication di
sabled; ui acls disabled; users with view permissions: Set(root); users with mod
ify permissions: Set(root)
18/05/03 05:27:39 INFO util.Utils: Successfully started service 'sparkDriver' on
port 37523.
18/05/03 05:27:39 INFO slf4j.Slf4jLogger: Slf4jLogger started
18/05/03 05:27:39 INFO Remoting: Starting remoting
18/05/03 05:27:39 INFO Remoting: Remoting started; listening on addresses :[akka
.tcp://sparkDriverActorSystem@10.0.0.229:41278]
18/05/03 05:27:39 INFO Remoting: Remoting now listens on addresses: [akka.tcp://
sparkDriverActorSystem@10.0.0.229:41278]
18/05/03 05:27:39 INFO util.Utils: Successfully started service 'sparkDriverActo
rSystem' on port 41278.
18/05/03 05:27:39 INFO spark.SparkEnv: Registering MapOutputTracker
18/05/03 05:27:40 INFO spark.SparkEnv: Registering BlockManagerMaster
18/05/03 05:27:40 INFO storage.DiskBlockManager: Created local directory at /tmp
/blockmgr-7123f2fd-a005-4ec6-b974-c2571082522e
18/05/03 05:27:40 INFO storage.MemoryStore: MemoryStore started with capacity
53
0.0 MB
18/05/03 05:27:40 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/05/03 05:27:40 INFO util.Utils: Successfully started service 'SparkUI' on por
t 4040.
18/05/03 05:27:40 INFO ui.SparkUI: Started SparkUI at http://10.0.0.229:4040
18/05/03 05:27:40 INFO spark.SparkContext: Added JAR file:/root/testproject-0.0.
1-SNAPSHOT-jar-with-dependencies.jar at spark://10.0.0.229:37523/jars/testprojec
t-0.0.1-SNAPSHOT-jar-with-dependencies.jar with timestamp 1525325260379
18/05/03 05:27:40 INFO executor.Executor: Starting executor ID driver on host lo
calhost
18/05/03 05:27:40 INFO util.Utils: Successfully started service 'org.apache.spar
k.network.netty.NettyBlockTransferService' on port 36563.
18/05/03 05:27:40 INFO netty.NettyBlockTransferService: Server created on 36563
18/05/03 05:27:40 INFO storage.BlockManager: external shuffle service port = 733
7
```

18/05/03 05:27:40 INFO storage.BlockManagerMaster: Trying to register BlockManag
er
18/05/03 05:27:40 INFO storage.BlockManagerMasterEndpoint: Registering block man
ager localhost:36563 with 530.0 MB RAM, BlockManagerId(driver, localhost, 36563)
18/05/03 05:27:40 INFO storage.BlockManagerMaster: Registered BlockManager
18/05/03 05:27:41 INFO scheduler.EventLoggingListener: Logging events to hdfs://
ip-10-0-0-229.ap-south-1.compute.internal:8020/user/spark/applicationHistory/loc
al-1525325260426
18/05/03 05:27:41 INFO spark.SparkContext: Registered listener com.cloudera.spar
k.lineage.ClouderaNavigatorListener
18/05/03 05:27:42 INFO storage.MemoryStore: Block broadcast_0 stored as values i
n memory (estimated size 303.4 KB, free 529.7 MB)
18/05/03 05:27:42 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as b
ytes in memory (estimated size 26.1 KB, free 529.7 MB)
18/05/03 05:27:42 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in mem
ory on localhost:36563 (size: 26.1 KB, free: 530.0 MB)
18/05/03 05:27:42 INFO spark.SparkContext: Created broadcast 0 from textFile at
SparkTask2.java:23
18/05/03 05:27:42 INFO mapred.FileInputFormat: Total input paths to process : 7
18/05/03 05:27:42 INFO spark.SparkContext: Starting job: foreach at SparkTask2.j
ava:26
18/05/03 05:27:42 INFO scheduler.DAGScheduler: Got job 0 (foreach at SparkTask2.
java:26) with 46 output partitions
18/05/03 05:27:42 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (forea
ch at SparkTask2.java:26)
18/05/03 05:27:42 INFO scheduler.DAGScheduler: Parents of final stage: List()
18/05/03 05:27:42 INFO scheduler.DAGScheduler: Missing parents: List()
18/05/03 05:27:42 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (MapPart
itionsRDD[4] at map at SparkTask2.java:26), which has no missing parents
18/05/03 05:27:42 INFO storage.MemoryStore: Block broadcast_1 stored as values i
n memory (estimated size 5.0 KB, free 529.7 MB)
18/05/03 05:27:42 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as b
ytes in memory (estimated size 2.7 KB, free 529.7 MB)

18/05/03 05:27:42 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in mem
ory on localhost:36563 (size: 2.7 KB, free: 530.0 MB)
18/05/03 05:27:42 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1004
18/05/03 05:27:42 INFO scheduler.DAGScheduler: Submitting 46 missing tasks from ResultStage 0 (MapPartitionsRDD[4] at map at SparkTask2.java:26) (first 15 tasks are for partitions Vector(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14))
18/05/03 05:27:42 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 46 tasks
18/05/03 05:27:42 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, ANY, 2316 bytes)
18/05/03 05:27:42 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, localhost, executor driver, partition 1, ANY, 2316 bytes)
18/05/03 05:27:42 INFO scheduler.TaskSetManager: Starting task 2.0 in stage 0.0 (TID 2, localhost, executor driver, partition 2, ANY, 2316 bytes)
18/05/03 05:27:42 INFO scheduler.TaskSetManager: Starting task 3.0 in stage 0.0 (TID 3, localhost, executor driver, partition 3, ANY, 2316 bytes)
18/05/03 05:27:42 INFO executor.Executor: Running task 1.0 in stage 0.0 (TID 1)
18/05/03 05:27:42 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
18/05/03 05:27:42 INFO executor.Executor: Running task 2.0 in stage 0.0 (TID 2)
18/05/03 05:27:42 INFO executor.Executor: Running task 3.0 in stage 0.0 (TID 3)
18/05/03 05:27:42 INFO executor.Executor: Fetching spark://10.0.0.229:37523/jars /testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar with timestamp 15253252603
79
18/05/03 05:27:42 INFO spark.ExecutorAllocationManager: New executor driver has registered (new total is 1)
18/05/03 05:27:42 INFO util.Utils: Fetching spark://10.0.0.229:37523/jars/testpr oject-0.0.1-SNAPSHOT-jar-with-dependencies.jar to /tmp/spark-f9fb3046-c29a-4e57-
9839-05a13e43cfdf/userFiles-e76885e9-6540-4a81-9849-55b46e5b4b8f/fetchFileTemp40
97302059845584229.tmp
18/05/03 05:27:43 INFO executor.Executor: Adding file:/tmp/spark-f9fb3046-c29a-4 e57-9839-05a13e43cfdf/userFiles-e76885e9-6540-4a81-9849-55b46e5b4b8f/testproject
-0.0.1-SNAPSHOT-jar-with-dependencies.jar to class loader
18/05/03 05:27:43 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-

1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:268435456+134217728
18/05/03 05:27:43 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:134217728+134217728
18/05/03 05:27:43 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:0+134217728
18/05/03 05:27:43 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south-1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:402653184+134217728
18/05/03 05:27:43 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
18/05/03 05:27:43 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
18/05/03 05:27:43 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
18/05/03 05:27:43 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
18/05/03 05:27:43 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
18/05/03 05:27:43 ERROR executor.Executor: Exception in task 0.0 in stage 0.0 (TID 0)
java.lang.ArrayIndexOutOfBoundsException: 5
	at com.spark.Assignment2.SparkTask2.lambda$main$19e3a193$2(SparkTask2.java:26)
	at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.scala:78)
	at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.scala:78)
	at scala.collection.Iterator$$anon$14.hasNext(Iterator.scala:390)
	at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:327)
	at scala.collection.Iterator$class.foreach(Iterator.scala:727)
	at scala.collection.AbstractIterator.foreach(Iterator.scala:1157)
	at org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(RDD.scala:912)

```
	at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
	at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.sc
ala:1888)
	at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.sc
ala:1888)
	at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
	at org.apache.spark.scheduler.Task.run(Task.scala:89)
	at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
	at java.lang.Thread.run(Thread.java:748)
[1, 2017-07-07 16:39:57, 2017-07-07 17:02:15, 1, 7.30, 4, N, 132, 265, 2, 25.5,
1, 0.5, 0, 0, 0.3, 27.3]
[1, 2017-07-13 00:52:25, 2017-07-13 01:09:29, 1, 4.90, 4, N, 142, 116, 1, 16.5,
0.5, 0.5, 3.55, 0, 0.3, 21.35]
[2, 2017-07-13 00:07:35, 2017-07-13 00:46:53, 1, 21.15, 4, N, 162, 265, 2, 60, 0
.5, 0.5, 0, 5.76, 0.3, 67.06]
[2, 2017-07-07 16:06:05, 2017-07-07 16:56:07, 1, 17.58, 4, N, 132, 265, 2, 65.5,
1, 0.5, 0, 0, 0.3, 67.3]
[2, 2017-07-18 10:44:19, 2017-07-18 12:05:53, 1, 47.12, 4, N, 138, 265, 2, 212.5
, 0, 0.5, 0, 5.76, 0.3, 219.06]
[1, 2017-07-13 00:34:13, 2017-07-13 01:31:45, 1, 22.60, 4, N, 90, 265, 2, 86.5,
0.5, 0.5, 0, 2.64, 0.3, 90.44]
[2, 2017-07-13 00:24:19, 2017-07-13 00:35:46, 5, 7.15, 4, N, 132, 265, 2, 21.5,
0.5, 0.5, 0, 0, 0.3, 22.8]
[2, 2017-07-07 16:44:11, 2017-07-07 17:15:43, 6, 8.59, 4, N, 132, 265, 1, 36, 1,
0.5, 0, 0, 0.3, 37.8]
[1, 2017-07-07 17:15:48, 2017-07-07 17:52:21, 1, 11.60, 4, N, 132, 265, 2, 54.5,
1, 0.5, 0, 0, 0.3, 56.3]
18/05/03 05:27:43 INFO scheduler.TaskSetManager: Starting task 4.0 in stage 0.0
(TID 4, localhost, executor driver, partition 4, ANY, 2316 bytes)
18/05/03 05:27:43 INFO executor.Executor: Running task 4.0 in stage 0.0 (TID 4)
[2, 2017-07-07 17:40:08, 2017-07-07 19:16:33, 1, 22.37, 4, N, 261, 265, 1, 82, 1
, 0.5, 16.76, 0, 0.3, 100.56]
18/05/03 05:27:43 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
```

1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdata_2017-07.csv:536870912+134217728
18/05/03 05:27:43 WARN scheduler.TaskSetManager: Lost task 0.0 in stage 0.0 (TID 0, localhost, executor driver): java.lang.ArrayIndexOutOfBoundsException: 5
at com.spark.Assignment2.SparkTask2.lambda$main$19e3a193$2(SparkTask2.java:26)
at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.scala:78)
at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.scala:78)
at scala.collection.Iterator$$anon$14.hasNext(Iterator.scala:390)
at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:327)
at scala.collection.Iterator$class.foreach(Iterator.scala:727)
at scala.collection.AbstractIterator.foreach(Iterator.scala:1157)
at org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(RDD.scala:912)
at org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(RDD.scala:912)
at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:1888)
at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:1888)
at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
at org.apache.spark.scheduler.Task.run(Task.scala:89)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:748)

18/05/03 05:27:43 ERROR scheduler.TaskSetManager: Task 0 in stage 0.0 failed 1 times; aborting job
[1, 2017-07-07 17:26:25, 2017-07-07 18:09:09, 1, 16.40, 4, N, 132, 265, 1, 73.5, 1, 0.5, 18.8, 0, 0.3, 94.1]
[2, 2017-07-07 17:01:38, 2017-07-07 17:37:21, 1, 8.48, 4, N, 132, 265, 2, 37, 1, 0.5, 0, 0, 0.3, 38.8]
18/05/03 05:27:43 INFO scheduler.TaskSchedulerImpl: Cancelling stage 0

[2, 2017-07-07 17:38:31, 2017-07-07 18:31:21, 2, 17.74, 4, N, 132, 265, 2, 61, 1
, 0.5, 0, 0, 0.3, 62.8]
18/05/03 05:27:43 INFO scheduler.TaskSchedulerImpl: Stage 0 was cancelled
18/05/03 05:27:43 INFO scheduler.DAGScheduler: ResultStage 0 (foreach at SparkTa
sk2.java:26) failed in 0.730 s due to Job aborted due to stage failure: Task 0 i
n stage 0.0 failed 1 times, most recent failure: Lost task 0.0 in stage 0.0 (TID
0, localhost, executor driver): java.lang.ArrayIndexOutOfBoundsException: 5
	at
com.spark.Assignment2.SparkTask2.lambda$main$19e3a193$2(SparkTask2.ja
va:26)
	at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.sca
la:78)
	at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.sca
la:78)
	at scala.collection.Iterator$$anon$14.hasNext(Iterator.scala:390)
	at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:327)
	at scala.collection.Iterator$class.foreach(Iterator.scala:727)
	at scala.collection.AbstractIterator.foreach(Iterator.scala:1157)
	at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
	at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
	at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.sc
ala:1888)
	at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.sc
ala:1888)
	at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
	at org.apache.spark.scheduler.Task.run(Task.scala:89)
	at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
	at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
[2, 2017-07-07 17:56:53, 2017-07-07 18:57:47, 2, 23.37, 4, N, 138, 265, 1, 86, 1
, 0.5, 17.56, 0, 0.3, 105.36]

```
18/05/03 05:27:43 INFO executor.Executor: Executor is trying to kill task 1.0 in
stage 0.0 (TID 1)
18/05/03 05:27:43 INFO executor.Executor: Executor is trying to kill task 2.0 in
stage 0.0 (TID 2)
18/05/03 05:27:43 INFO executor.Executor: Executor is trying to kill task 3.0 in
stage 0.0 (TID 3)
18/05/03 05:27:43 INFO executor.Executor: Executor is trying to kill task 4.0 in
stage 0.0 (TID 4)
[2, 2017-07-13 00:23:32, 2017-07-13 00:51:37, 1, 17.72, 4, N, 138, 265, 2, 61, 0
.5, 0.5, 0, 0, 0.3, 62.3]
18/05/03 05:27:43 INFO executor.Executor: Executor killed task 1.0 in stage 0.0
(TID 1)
18/05/03 05:27:43 INFO scheduler.DAGScheduler: Job 0 failed: foreach at SparkTas
k2.java:26, took 0.862122 s
18/05/03 05:27:43 INFO executor.Executor: Executor killed task 3.0 in stage 0.0
(TID 3)
18/05/03 05:27:43 INFO executor.Executor: Executor killed task 2.0 in stage 0.0
(TID 2)
18/05/03 05:27:43 INFO executor.Executor: Executor killed task 4.0 in stage 0.0
(TID 4)
Exception in thread "main" org.apache.spark.SparkException: Job aborted due to s
tage failure: Task 0 in stage 0.0 failed 1 times, most recent failure: Lost task
0.0 in stage 0.0 (TID 0, localhost, executor driver): java.lang.ArrayIndexOutOf
BoundsException: 5
        at
com.spark.Assignment2.SparkTask2.lambda$main$19e3a193$2(SparkTask2.ja
va:26)
        at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.sca
la:78)
        at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.sca
la:78)
        at scala.collection.Iterator$$anon$14.hasNext(Iterator.scala:390)
        at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:327)
        at scala.collection.Iterator$class.foreach(Iterator.scala:727)
        at scala.collection.AbstractIterator.foreach(Iterator.scala:1157)
        at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
        at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
```

at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:1888)
at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:1888)
at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
at org.apache.spark.scheduler.Task.run(Task.scala:89)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages(DAGScheduler.scala:1457)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1445)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1444)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1444)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:799)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:799)
at scala.Option.foreach(Option.scala:236)
at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGScheduler.scala:799)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.doOnReceive(DAGScheduler.scala:1668)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1627)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1616)
at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:48)
at org.apache.spark.scheduler.DAGScheduler.runJob(DAGScheduler.scala:620)

at org.apache.spark.SparkContext.runJob(SparkContext.scala:1862)
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1875)
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1888)
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1959)
at org.apache.spark.rdd.RDD$$anonfun$foreach$1.apply(RDD.scala:912)
at org.apache.spark.rdd.RDD$$anonfun$foreach$1.apply(RDD.scala:910)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.s
cala:150)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.s
cala:111)
at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
at org.apache.spark.rdd.RDD.foreach(RDD.scala:910)
at org.apache.spark.api.java.JavaRDDLike$class.foreach(JavaRDDLike.scala
:332)
at org.apache.spark.api.java.AbstractJavaRDDLike.foreach(JavaRDDLike.sca
la:46)
at com.spark.Assignment2.SparkTask2.main(SparkTask2.java:26)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSub
mit$$runMain(SparkSubmit.scala:730)
at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:18
1)
at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)
at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)
at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.lang.ArrayIndexOutOfBoundsException: 5
at
com.spark.Assignment2.SparkTask2.lambda$main$19e3a193$2(SparkTask2.ja
va:26)
at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.sca
la:78)
at org.apache.spark.api.java.JavaRDD$$anonfun$filter$1.apply(JavaRDD.sca
la:78)
at scala.collection.Iterator$$anon$14.hasNext(Iterator.scala:390)
at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:327)
at scala.collection.Iterator$class.foreach(Iterator.scala:727)

```
	at scala.collection.AbstractIterator.foreach(Iterator.scala:1157)
	at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
	at
org.apache.spark.rdd.RDD$$anonfun$foreach$1$$anonfun$apply$32.apply(R
DD.scala:912)
	at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.sc
ala:1888)
	at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.sc
ala:1888)
	at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
	at org.apache.spark.scheduler.Task.run(Task.scala:89)
	at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
	at java.lang.Thread.run(Thread.java:748)
18/05/03 05:27:43 WARN scheduler.TaskSetManager: Lost task 2.0 in stage 0.0 (TID
2, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:27:43 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose
t
asks have all completed, from pool
18/05/03 05:27:43 WARN scheduler.TaskSetManager: Lost task 1.0 in stage 0.0 (TID
1, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:27:43 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose
t
asks have all completed, from pool
18/05/03 05:27:43 WARN scheduler.TaskSetManager: Lost task 4.0 in stage 0.0 (TID
4, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:27:43 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose
t
asks have all completed, from pool
18/05/03 05:27:43 WARN scheduler.TaskSetManager: Lost task 3.0 in stage 0.0 (TID
3, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:27:43 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose
t
asks have all completed, from pool
18/05/03 05:27:43 INFO spark.SparkContext: Invoking stop() from shutdown hook
```

18/05/03 05:27:43 WARN spark.ExecutorAllocationManager: No stages are running, b
ut numRunningTasks != 0
18/05/03 05:27:43 INFO ui.SparkUI: Stopped Spark web UI at http://10.0.0.229:404
0
18/05/03 05:27:43 INFO spark.MapOutputTrackerMasterEndpoint:
MapOutputTrackerMas
terEndpoint stopped!
18/05/03 05:27:43 INFO storage.MemoryStore: MemoryStore cleared
18/05/03 05:27:43 INFO storage.BlockManager: BlockManager stopped
18/05/03 05:27:43 INFO storage.BlockManagerMaster: BlockManagerMaster
stopped
18/05/03 05:27:43 INFO
scheduler.OutputCommitCoordinator$OutputCommitCoordinator
Endpoint: OutputCommitCoordinator stopped!
18/05/03 05:27:43 INFO remote.RemoteActorRefProvider$RemotingTerminator:
Shuttin
g down remote daemon.
18/05/03 05:27:43 INFO remote.RemoteActorRefProvider$RemotingTerminator:
Remote
daemon shut down; proceeding with flushing remote transports.
18/05/03 05:27:43 INFO spark.SparkContext: Successfully stopped SparkContext
18/05/03 05:27:43 INFO util.ShutdownHookManager: Shutdown hook called
18/05/03 05:27:43 INFO util.ShutdownHookManager: Deleting directory
/tmp/spark-f
9fb3046-c29a-4e57-9839-05a13e43cfdf
18/05/03 05:27:43 INFO Remoting: Remoting shut down
18/05/03 05:27:43 INFO remote.RemoteActorRefProvider$RemotingTerminator:
Remotin
g shut down.
[root@ip-10-0-0-229 ~]#

3. GroupBy
   [root@ip-10-0-0-229 ~]# spark-submit --class com.spark.Assignment3.SparkTask3 --
   master yarn --deploy-mode client --name testtproject --conf "spark.app.id=test3p
   roject" testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar /user/root/spark_as
   signment/input_dataset/yellow_tripdata_*
   /user/root/spark_assignment/output/Grou
   pBy_output/
   18/05/03 05:38:36 INFO spark.SparkContext: Running Spark version 1.6.0
   18/05/03 05:38:36 INFO spark.SecurityManager: Changing view acls to: root

18/05/03 05:38:36 INFO spark.SecurityManager: Changing modify acls to: root
18/05/03 05:38:36 INFO spark.SecurityManager: SecurityManager: authentication di
sabled; ui acls disabled; users with view permissions: Set(root); users with mod
ify permissions: Set(root)
18/05/03 05:38:36 INFO util.Utils: Successfully started service 'sparkDriver' on
port 37893.
18/05/03 05:38:37 INFO slf4j.Slf4jLogger: Slf4jLogger started
18/05/03 05:38:37 INFO Remoting: Starting remoting
18/05/03 05:38:37 INFO Remoting: Remoting started; listening on addresses :[akka
.tcp://sparkDriverActorSystem@10.0.0.229:41833]
18/05/03 05:38:37 INFO Remoting: Remoting now listens on addresses: [akka.tcp://
sparkDriverActorSystem@10.0.0.229:41833]
18/05/03 05:38:37 INFO util.Utils: Successfully started service 'sparkDriverActo
rSystem' on port 41833.
18/05/03 05:38:37 INFO spark.SparkEnv: Registering MapOutputTracker
18/05/03 05:38:37 INFO spark.SparkEnv: Registering BlockManagerMaster
18/05/03 05:38:37 INFO storage.DiskBlockManager: Created local directory at /tmp
/blockmgr-1e278638-4b7f-4b55-b439-b3a33ac28afc
18/05/03 05:38:37 INFO storage.MemoryStore: MemoryStore started with capacity
53
0.0 MB
18/05/03 05:38:37 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/05/03 05:38:37 INFO util.Utils: Successfully started service 'SparkUI' on por
t 4040.
18/05/03 05:38:37 INFO ui.SparkUI: Started SparkUI at http://10.0.0.229:4040
18/05/03 05:38:37 INFO spark.SparkContext: Added JAR file:/root/testproject-0.0.
1-SNAPSHOT-jar-with-dependencies.jar at spark://10.0.0.229:37893/jars/testprojec
t-0.0.1-SNAPSHOT-jar-with-dependencies.jar with timestamp 1525325917773
18/05/03 05:38:37 INFO executor.Executor: Starting executor ID driver on host lo
calhost
18/05/03 05:38:37 INFO util.Utils: Successfully started service 'org.apache.spar
k.network.netty.NettyBlockTransferService' on port 34836.
18/05/03 05:38:37 INFO netty.NettyBlockTransferService: Server created on 34836
18/05/03 05:38:37 INFO storage.BlockManager: external shuffle service port = 733
7
18/05/03 05:38:37 INFO storage.BlockManagerMaster: Trying to register
BlockManag
er
18/05/03 05:38:37 INFO storage.BlockManagerMasterEndpoint: Registering block
man
ager localhost:34836 with 530.0 MB RAM, BlockManagerId(driver, localhost, 34836)

18/05/03 05:38:37 INFO storage.BlockManagerMaster: Registered BlockManager
18/05/03 05:38:39 INFO scheduler.EventLoggingListener: Logging events to hdfs://
ip-10-0-0-229.ap-south-1.compute.internal:8020/user/spark/applicationHistory/loc
al-1525325917818
18/05/03 05:38:39 INFO spark.SparkContext: Registered listener com.cloudera.spar
k.lineage.ClouderaNavigatorListener
18/05/03 05:38:39 INFO storage.MemoryStore: Block broadcast_0 stored as values i
n memory (estimated size 303.4 KB, free 529.7 MB)
18/05/03 05:38:39 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as
b
ytes in memory (estimated size 26.1 KB, free 529.7 MB)
18/05/03 05:38:39 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in
mem
ory on localhost:34836 (size: 26.1 KB, free: 530.0 MB)
18/05/03 05:38:39 INFO spark.SparkContext: Created broadcast 0 from textFile at
SparkTask3.java:23
18/05/03 05:38:39 INFO mapred.FileInputFormat: Total input paths to process : 7
18/05/03 05:38:39 INFO spark.SparkContext: Starting job: sortByKey at SparkTask3
.java:49
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Registering RDD 7 (mapToPair at
S
parkTask3.java:33)
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Got job 0 (sortByKey at SparkTask
3.java:49) with 46 output partitions
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (sortB
yKey at SparkTask3.java:49)
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Parents of final stage: List(Shuf
fleMapStage 0)
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapS
tage 0)
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0
(Map
PartitionsRDD[7] at mapToPair at SparkTask3.java:33), which has no missing paren
ts
18/05/03 05:38:40 INFO storage.MemoryStore: Block broadcast_1 stored as values i
n memory (estimated size 6.6 KB, free 529.7 MB)
18/05/03 05:38:40 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as
b
ytes in memory (estimated size 3.4 KB, free 529.7 MB)

18/05/03 05:38:40 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in mem
ory on localhost:34836 (size: 3.4 KB, free: 530.0 MB)
18/05/03 05:38:40 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1004
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Submitting 46 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[7] at mapToPair at SparkTask3.java:33) (firs
t 15 tasks are for partitions Vector(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1
3, 14))
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 46 tasks
18/05/03 05:38:40 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, ANY, 2305 bytes)
18/05/03 05:38:40 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, localhost, executor driver, partition 1, ANY, 2305 bytes)
18/05/03 05:38:40 INFO scheduler.TaskSetManager: Starting task 2.0 in stage 0.0 (TID 2, localhost, executor driver, partition 2, ANY, 2305 bytes)
18/05/03 05:38:40 INFO scheduler.TaskSetManager: Starting task 3.0 in stage 0.0 (TID 3, localhost, executor driver, partition 3, ANY, 2305 bytes)
18/05/03 05:38:40 INFO executor.Executor: Running task 1.0 in stage 0.0 (TID 1)
18/05/03 05:38:40 INFO executor.Executor: Running task 2.0 in stage 0.0 (TID 2)
18/05/03 05:38:40 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
18/05/03 05:38:40 INFO executor.Executor: Running task 3.0 in stage 0.0 (TID 3)
18/05/03 05:38:40 INFO executor.Executor: Fetching spark://10.0.0.229:37893/jars
/testproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar with timestamp
15253259177
73
18/05/03 05:38:40 INFO spark.ExecutorAllocationManager: New executor driver has registered (new total is 1)
18/05/03 05:38:40 INFO util.Utils: Fetching spark://10.0.0.229:37893/jars/testpr
oject-0.0.1-SNAPSHOT-jar-with-dependencies.jar to /tmp/spark-4cb0292e-cbab-
454b-
a52f-e1cbfb4a6d81/userFiles-d1db28dd-5240-4165-95f4-
51d6fb548f52/fetchFileTemp63
9457861714034847.tmp
18/05/03 05:38:40 INFO executor.Executor: Adding file:/tmp/spark-4cb0292e-cbab-4
54b-a52f-e1cbfb4a6d81/userFiles-d1db28dd-5240-4165-95f4-
51d6fb548f52/testproject
-0.0.1-SNAPSHOT-jar-with-dependencies.jar to class loader

```
18/05/03 05:38:40 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:402653184+134217728
18/05/03 05:38:40 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:0+134217728
18/05/03 05:38:40 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:268435456+134217728
18/05/03 05:38:40 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:134217728+134217728
18/05/03 05:38:40 INFO Configuration.deprecation: mapred.tip.id is deprecated. I
nstead, use mapreduce.task.id
18/05/03 05:38:40 INFO Configuration.deprecation: mapred.task.id is deprecated.
Instead, use mapreduce.task.attempt.id
18/05/03 05:38:40 INFO Configuration.deprecation: mapred.task.is.map is deprecat
ed. Instead, use mapreduce.task.ismap
18/05/03 05:38:40 INFO Configuration.deprecation: mapred.task.partition is depre
cated. Instead, use mapreduce.task.partition
18/05/03 05:38:40 INFO Configuration.deprecation: mapred.job.id is deprecated. I
nstead, use mapreduce.job.id
18/05/03 05:38:40 ERROR executor.Executor: Exception in task 0.0 in stage 0.0 (T
ID 0)
java.lang.ArrayIndexOutOfBoundsException: 9
        at com.spark.Assignment3.SparkTask3.lambda$main$88a05c0$2(SparkTask3.jav
a:31)
        at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.appl
y(JavaPairRDD.scala:1015)
        at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
        at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
        at org.apache.spark.util.collection.ExternalSorter.insertAll(ExternalSor
ter.scala:194)
        at org.apache.spark.shuffle.sort.SortShuffleWriter.write(SortShuffleWrit
er.scala:64)
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:73)
```

```
	at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:41)
	at org.apache.spark.scheduler.Task.run(Task.scala:89)
	at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
	at java.lang.Thread.run(Thread.java:748)
18/05/03 05:38:40 INFO scheduler.TaskSetManager: Starting task 4.0 in stage 0.0
(TID 4, localhost, executor driver, partition 4, ANY, 2305 bytes)
18/05/03 05:38:40 INFO executor.Executor: Running task 4.0 in stage 0.0 (TID 4)
18/05/03 05:38:40 INFO rdd.HadoopRDD: Input split: hdfs://ip-10-0-0-229.ap-south
-
1.compute.internal:8020/user/root/spark_assignment/input_dataset/yellow_tripdat
a_2017-07.csv:536870912+134217728
18/05/03 05:38:40 WARN scheduler.TaskSetManager: Lost task 0.0 in stage 0.0 (TID
0, localhost, executor driver): java.lang.ArrayIndexOutOfBoundsException: 9
	at com.spark.Assignment3.SparkTask3.lambda$main$88a05c0$2(SparkTask3.jav
a:31)
	at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.appl
y(JavaPairRDD.scala:1015)
	at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
	at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
	at org.apache.spark.util.collection.ExternalSorter.insertAll(ExternalSor
ter.scala:194)
	at org.apache.spark.shuffle.sort.SortShuffleWriter.write(SortShuffleWrit
er.scala:64)
	at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:73)
	at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:41)
	at org.apache.spark.scheduler.Task.run(Task.scala:89)
	at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
	at java.lang.Thread.run(Thread.java:748)
```

```
18/05/03 05:38:40 ERROR scheduler.TaskSetManager: Task 0 in stage 0.0 failed 1 t
imes; aborting job
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Cancelling stage 0
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Stage 0 was cancelled
18/05/03 05:38:40 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (mapToPair at
S
parkTask3.java:33) failed in 0.793 s due to Job aborted due to stage failure: Ta
sk 0 in stage 0.0 failed 1 times, most recent failure: Lost task 0.0 in stage 0.
0 (TID 0, localhost, executor driver): java.lang.ArrayIndexOutOfBoundsException:
9
        at com.spark.Assignment3.SparkTask3.lambda$main$88a05c0$2(SparkTask3.jav
a:31)
        at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.appl
y(JavaPairRDD.scala:1015)
        at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
        at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
        at org.apache.spark.util.collection.ExternalSorter.insertAll(ExternalSor
ter.scala:194)
        at org.apache.spark.shuffle.sort.SortShuffleWriter.write(SortShuffleWrit
er.scala:64)
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:73)
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:41)
        at org.apache.spark.scheduler.Task.run(Task.scala:89)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
        at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
18/05/03 05:38:40 INFO executor.Executor: Executor is trying to kill task 1.0 in
stage 0.0 (TID 1)
18/05/03 05:38:40 INFO executor.Executor: Executor is trying to kill task 2.0 in
stage 0.0 (TID 2)
18/05/03 05:38:40 INFO executor.Executor: Executor is trying to kill task 3.0 in
stage 0.0 (TID 3)
18/05/03 05:38:40 INFO executor.Executor: Executor is trying to kill task 4.0 in
stage 0.0 (TID 4)
```

18/05/03 05:38:40 INFO executor.Executor: Executor killed task 3.0 in stage 0.0 (TID 3)
18/05/03 05:38:40 INFO executor.Executor: Executor killed task 2.0 in stage 0.0 (TID 2)
18/05/03 05:38:40 INFO executor.Executor: Executor killed task 1.0 in stage 0.0 (TID 1)
18/05/03 05:38:40 INFO executor.Executor: Executor killed task 4.0 in stage 0.0 (TID 4)
18/05/03 05:38:40 WARN scheduler.TaskSetManager: Lost task 3.0 in stage 0.0 (TID 3, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose t
asks have all completed, from pool
18/05/03 05:38:40 WARN scheduler.TaskSetManager: Lost task 4.0 in stage 0.0 (TID 4, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose t
asks have all completed, from pool
18/05/03 05:38:40 WARN scheduler.TaskSetManager: Lost task 2.0 in stage 0.0 (TID 2, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose t
asks have all completed, from pool
18/05/03 05:38:40 WARN scheduler.TaskSetManager: Lost task 1.0 in stage 0.0 (TID 1, localhost, executor driver): TaskKilled (killed intentionally)
18/05/03 05:38:40 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose t
asks have all completed, from pool
18/05/03 05:38:40 INFO scheduler.DAGScheduler: Job 0 failed: sortByKey at SparkT
ask3.java:49, took 0.972470 s
Exception in thread "main" org.apache.spark.SparkException: Job aborted due to s
tage failure: Task 0 in stage 0.0 failed 1 times, most recent failure: Lost task
0.0 in stage 0.0 (TID 0, localhost, executor driver): java.lang.ArrayIndexOutOf
BoundsException: 9
	at com.spark.Assignment3.SparkTask3.lambda$main$88a05c0$2(SparkTask3.jav
a:31)
	at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.appl
y(JavaPairRDD.scala:1015)
	at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
	at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)

at org.apache.spark.util.collection.ExternalSorter.insertAll(ExternalSor
ter.scala:194)
at org.apache.spark.shuffle.sort.SortShuffleWriter.write(SortShuffleWrit
er.scala:64)
at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:73)
at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:41)
at org.apache.spark.scheduler.Task.run(Task.scala:89)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DA
GScheduler$$failJobAndIndependentStages(DAGScheduler.scala:1457)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(D
AGScheduler.scala:1445)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(D
AGScheduler.scala:1444)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.
scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala
:1444)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$
1.apply(DAGScheduler.scala:799)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$
1.apply(DAGScheduler.scala:799)
at scala.Option.foreach(Option.scala:236)
at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGSchedu
ler.scala:799)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.doOnReceive(D
AGScheduler.scala:1668)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAG
Scheduler.scala:1627)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAG
Scheduler.scala:1616)

```
	at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:48)
	at org.apache.spark.scheduler.DAGScheduler.runJob(DAGScheduler.scala:620
)
	at org.apache.spark.SparkContext.runJob(SparkContext.scala:1862)
	at org.apache.spark.SparkContext.runJob(SparkContext.scala:1875)
	at org.apache.spark.SparkContext.runJob(SparkContext.scala:1888)
	at org.apache.spark.SparkContext.runJob(SparkContext.scala:1959)
	at org.apache.spark.rdd.RDD$$anonfun$collect$1.apply(RDD.scala:927)
	at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.s
cala:150)
	at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.s
cala:111)
	at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
	at org.apache.spark.rdd.RDD.collect(RDD.scala:926)
	at org.apache.spark.RangePartitioner$.sketch(Partitioner.scala:264)
	at org.apache.spark.RangePartitioner.<init>(Partitioner.scala:126)
	at org.apache.spark.rdd.OrderedRDDFunctions$$anonfun$sortByKey$1.apply(O
rderedRDDFunctions.scala:62)
	at org.apache.spark.rdd.OrderedRDDFunctions$$anonfun$sortByKey$1.apply(O
rderedRDDFunctions.scala:61)
	at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.s
cala:150)
	at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.s
cala:111)
	at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
	at org.apache.spark.rdd.OrderedRDDFunctions.sortByKey(OrderedRDDFunction
s.scala:61)
	at org.apache.spark.api.java.JavaPairRDD.sortByKey(JavaPairRDD.scala:902
)
	at org.apache.spark.api.java.JavaPairRDD.sortByKey(JavaPairRDD.scala:872
)
	at org.apache.spark.api.java.JavaPairRDD.sortByKey(JavaPairRDD.scala:862
)
	at com.spark.Assignment3.SparkTask3.main(SparkTask3.java:49)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
	at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:62)
	at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
	at java.lang.reflect.Method.invoke(Method.java:498)
```

```
	at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSub
mit$$runMain(SparkSubmit.scala:730)
	at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:18
1)
	at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)
	at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)
	at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.lang.ArrayIndexOutOfBoundsException: 9
	at com.spark.Assignment3.SparkTask3.lambda$main$88a05c0$2(SparkTask3.jav
a:31)
	at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.appl
y(JavaPairRDD.scala:1015)
	at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
	at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
	at org.apache.spark.util.collection.ExternalSorter.insertAll(ExternalSor
ter.scala:194)
	at org.apache.spark.shuffle.sort.SortShuffleWriter.write(SortShuffleWrit
er.scala:64)
	at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:73)
	at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scal
a:41)
	at org.apache.spark.scheduler.Task.run(Task.scala:89)
	at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:242)
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.
java:1149)
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
.java:624)
	at java.lang.Thread.run(Thread.java:748)
18/05/03 05:38:40 INFO spark.SparkContext: Invoking stop() from shutdown hook
18/05/03 05:38:40 WARN spark.ExecutorAllocationManager: No stages are running,
b
ut numRunningTasks != 0
18/05/03 05:38:41 INFO ui.SparkUI: Stopped Spark web UI at http://10.0.0.229:404
0
18/05/03 05:38:41 INFO spark.MapOutputTrackerMasterEndpoint:
MapOutputTrackerMas
terEndpoint stopped!
18/05/03 05:38:41 INFO storage.MemoryStore: MemoryStore cleared
18/05/03 05:38:41 INFO storage.BlockManager: BlockManager stopped
```

18/05/03 05:38:41 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
18/05/03 05:38:41 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinator Endpoint: OutputCommitCoordinator stopped!
18/05/03 05:38:41 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shuttin
g down remote daemon.
18/05/03 05:38:41 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote
daemon shut down; proceeding with flushing remote transports.
18/05/03 05:38:41 INFO spark.SparkContext: Successfully stopped SparkContext
18/05/03 05:38:41 INFO util.ShutdownHookManager: Shutdown hook called
18/05/03 05:38:41 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-4
cb0292e-cbab-454b-a52f-e1cbfb4a6d81
18/05/03 05:38:41 INFO Remoting: Remoting shut down
18/05/03 05:38:41 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remotin
g shut down.
[root@ip-10-0-0-229 ~]#