Part 3

Question 1&2

Based on the SimpleExamle Java program, develope two java program one with cache RDD one without.

LogCount.java

count the hits to '/assets/img/loading.gif' and '/assets/js/lightbox.js' without using cache.

```
Dataset<String> logData = spark.read().textFile(logFile);
```

Load RDD Dataset twice in counting jobs

```
long numAs = logData.filter(s -> s.contains("/assets/img/loading.gif")).count();
```

```
long numBs = logData.filter(s -> s.contains("/assets/js/lightbox.js")).count();
```

Running time estimation:

```
import java.util.concurrent.TimeUnit; ...
// counting jobs
//Running time in nano seconds
```

```
long startTime = System.nanoTime();
long endTime = System.nanoTime();
long timeElapsed = endTime - startTime;
```

Submit jar file using spark-submit with cluster mode:

```
student@CC-AM-29:~/spark/test/LogCount$ ~/spark/bin/spark-submit --class "LogCount" \
> --master yarn \
> --deploy-mode cluster \
> --driver-memory 1g \
> --executor-memory 1g \
> --executor-cores 1 \
> --queue default \
> target/logcount*.jar
```

```
2019-03-16 00:10:29 INFO  Client:54 -
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-27
         ApplicationMaster RPC port: 43787
         queue: default
         start time: 1552694997098
         final status: SUCCEEDED
         tracking URL: http://CC-AM-29:8088/proxy/application_1550625818410_0063/
         user: student
```

Output:

As shown in the figure below:

Hits to the ....loading.gif: 294

Hits to the ....lightbox.js: 297

Running time without cache RDD: 32270140 (Microseconds (μs))

```
1. student@CC-AM-29: ~/hadoop/logs/userlogs/application_1550625818410_0055/container_1550625818410_0055_01_000001 (ssh)
2019-03-15 22:46:58 INFO  DAGScheduler:54 - running: Set()
2019-03-15 22:46:58 INFO  DAGScheduler:54 - waiting: Set(ResultStage 3)
2019-03-15 22:46:58 INFO  DAGScheduler:54 - failed: Set()
2019-03-15 22:46:58 INFO  DAGScheduler:54 - Submitting ResultStage 3 (MapPartitionsRDD[11] at count at LogCount.java:16), which has no missing parents
2019-03-15 22:46:58 INFO  MemoryStore:54 - Block broadcast_5 stored as values in memory (estimated size 7.1 KB, free 365.6 MB)
2019-03-15 22:46:58 INFO  MemoryStore:54 - Block broadcast_5_piece0 stored as bytes in memory (estimated size 3.8 KB, free 365.6 MB)
2019-03-15 22:46:58 INFO  BlockManagerInfo:54 - Added broadcast_5_piece0 in memory on CC-AM-29:39783 (size: 3.8 KB, free: 366.2 MB)
2019-03-15 22:46:58 INFO  SparkContext:54 - Created broadcast 5 from broadcast at DAGScheduler.scala:1161
2019-03-15 22:46:58 INFO  DAGScheduler:54 - Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[11] at count at LogCount.java:16) (first 15 tasks a
re for partitions Vector(0))
2019-03-15 22:46:58 INFO  YarnClusterScheduler:54 - Adding task set 3.0 with 1 tasks
2019-03-15 22:46:58 INFO  TaskSetManager:54 - Starting task 0.0 in stage 3.0 (TID 9, CC-AM-27, executor 2, partition 0, NODE_LOCAL, 7756 bytes)
2019-03-15 22:46:58 INFO  BlockManagerInfo:54 - Added broadcast_5_piece0 in memory on CC-AM-27:41125 (size: 3.8 KB, free: 366.2 MB)
2019-03-15 22:46:58 INFO  MapOutputTrackerMasterEndpoint:54 - Asked to send map output locations for shuffle 1 to 138.197.97.219:36632
2019-03-15 22:46:58 INFO  TaskSetManager:54 - Finished task 0.0 in stage 3.0 (TID 9) in 156 ms on CC-AM-27 (executor 2) (1/1)
2019-03-15 22:46:58 INFO  YarnClusterScheduler:54 - Removed TaskSet 3.0, whose tasks have all completed, from pool
2019-03-15 22:46:58 INFO  DAGScheduler:54 - ResultStage 3 (count at LogCount.java:16) finished in 0.166 s
2019-03-15 22:46:58 INFO  DAGScheduler:54 - Job 1 finished: count at LogCount.java:16, took 4.565502 s
Hits to loading.gif: 294
Hits to lightbox.js: 297
Execution time without cache RDD: 32270140
2019-03-15 22:46:58 INFO  AbstractConnector:318 - Stopped Spark@1f585190{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-15 22:46:58 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-29:40443
2019-03-15 22:46:58 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-15 22:46:58 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-15 22:46:58 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-15 22:46:58 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-15 22:46:58 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-15 22:46:58 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-15 22:46:58 INFO  BlockManager:54 - BlockManager stopped
2019-03-15 22:46:58 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-15 22:46:58 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-15 22:46:58 INFO  SparkContext:54 - Successfully stopped SparkContext
2019-03-15 22:46:58 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-15 22:46:59 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-15 22:46:59 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-15 22:46:59 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-29:9000/user/student/.sparkStaging/application_1550625818410_0055
2019-03-15 22:46:59 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-15 22:46:59 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/appcache/application_1550625818410_005
5/spark-cb518bc7-e3b1-4635-a962-4e37417b8fc4
```

LogCountcache.java

Using RDD cache, load the data into spark memory once

```
Dataset<String> logData = spark.read().textFile(logFile).cache();
```

```
student@CC-AM-29:~/spark/test/LogCount$ ~/spark/bin/spark-submit --class "LogCountcache" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memo
ry 1g --executor-cores 1 --queue default target/logcount*.jar
```

Output:

Running time with cache RDD: 25658202 (Microseconds (μs))

```
1. student@CC-AM-27: ~/hadoop/logs/userlogs/application_1550625818410_0063/container_1550625818410_0063_01_000001 (ssh)
2019-03-16 00:10:28 INFO  DAGScheduler:54 - waiting: Set(ResultStage 3)
2019-03-16 00:10:28 INFO  DAGScheduler:54 - failed: Set()
2019-03-16 00:10:28 INFO  DAGScheduler:54 - Submitting ResultStage 3 (MapPartitionsRDD[18] at count at LogCountcache.java:16), which has no missing parents
2019-03-16 00:10:28 INFO  MemoryStore:54 - Block broadcast_4 stored as values in memory (estimated size 7.1 KB, free 365.9 MB)
2019-03-16 00:10:28 INFO  MemoryStore:54 - Block broadcast_4_piece0 stored as bytes in memory (estimated size 3.8 KB, free 365.9 MB)
2019-03-16 00:10:28 INFO  BlockManagerInfo:54 - Added broadcast_4_piece0 in memory on CC-AM-27:43275 (size: 3.8 KB, free: 366.3 MB)
2019-03-16 00:10:28 INFO  SparkContext:54 - Created broadcast 4 from broadcast at DAGScheduler.scala:1161
2019-03-16 00:10:28 INFO  DAGScheduler:54 - Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[18] at count at LogCountcache.java:16) (first 15 ta
sks are for partitions Vector(0))
2019-03-16 00:10:28 INFO  YarnClusterScheduler:54 - Adding task set 3.0 with 1 tasks
2019-03-16 00:10:28 INFO  TaskSetManager:54 - Starting task 0.0 in stage 3.0 (TID 9, CC-AM-28, executor 2, partition 0, NODE_LOCAL, 7756 bytes)
2019-03-16 00:10:28 INFO  BlockManagerInfo:54 - Added broadcast_4_piece0 in memory on CC-AM-28:34509 (size: 3.8 KB, free: 103.3 MB)
2019-03-16 00:10:28 INFO  MapOutputTrackerMasterEndpoint:54 - Asked to send map output locations for shuffle 1 to 159.65.241.149:58742
2019-03-16 00:10:29 INFO  TaskSetManager:54 - Finished task 0.0 in stage 3.0 (TID 9) in 391 ms on CC-AM-28 (executor 2) (1/1)
2019-03-16 00:10:29 INFO  YarnClusterScheduler:54 - Removed TaskSet 3.0, whose tasks have all completed, from pool
2019-03-16 00:10:29 INFO  DAGScheduler:54 - ResultStage 3 (count at LogCountcache.java:16) finished in 0.399 s
2019-03-16 00:10:29 INFO  DAGScheduler:54 - Job 1 finished: count at LogCountcache.java:16, took 1.376649 s
Hits to loading.gif: 294
Hits to lightbox.js: 297
Execution time with cache RDD: 25658202
2019-03-16 00:10:29 INFO  AbstractConnector:318 - Stopped Spark@c1c94e9{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-16 00:10:29 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-27:36849
2019-03-16 00:10:29 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-16 00:10:29 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-16 00:10:29 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-16 00:10:29 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-16 00:10:29 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-16 00:10:29 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-16 00:10:29 INFO  BlockManager:54 - BlockManager stopped
2019-03-16 00:10:29 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-16 00:10:29 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-16 00:10:29 INFO  SparkContext:54 - Successfully stopped SparkContext
2019-03-16 00:10:29 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-16 00:10:29 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-16 00:10:29 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-16 00:10:29 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-29:9000/user/student/.sparkStaging/application_1550625818410_0063
2019-03-16 00:10:29 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-16 00:10:29 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/appcache/application_1550625818410_006
3/spark-f3afdd20-1400-42c9-a542-c4b7ac0b6a75
```