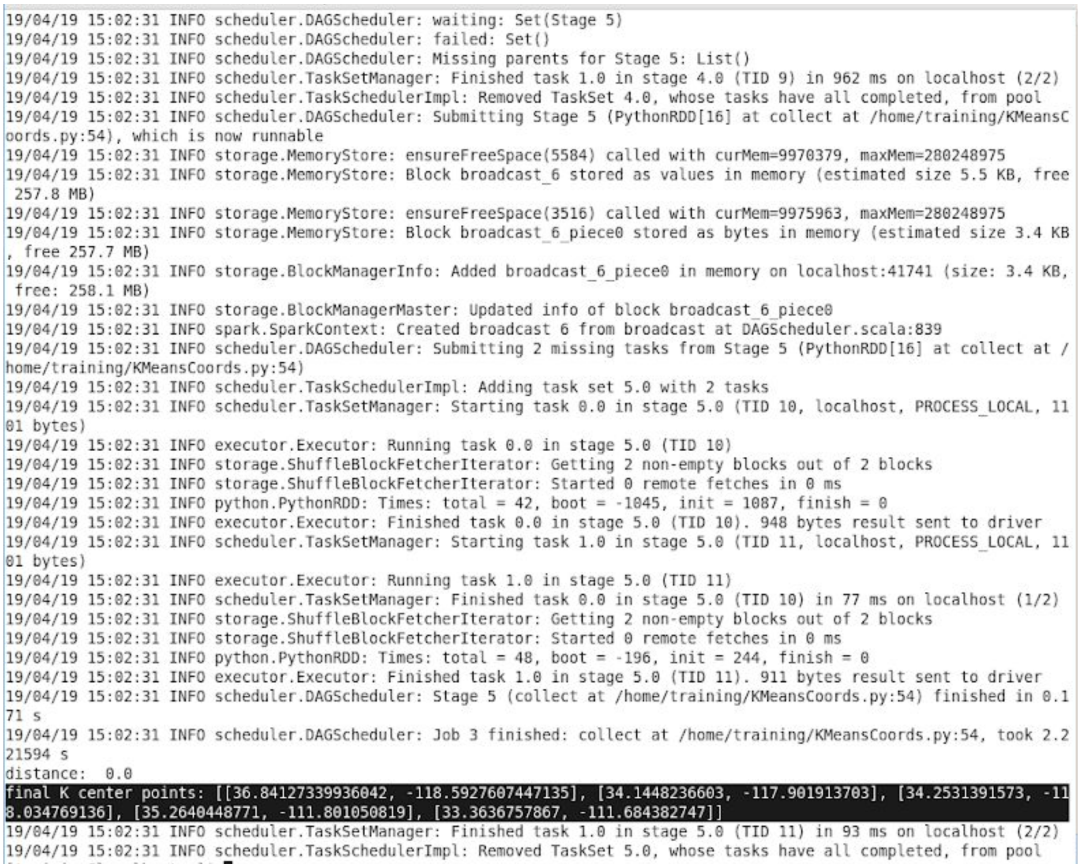
Matthew Kramer

Implement an Iterative Algorithm with Spark

RDD Transformation Procedure

1. Find the best index of a particular point x in an array of points.
    a. The return index represents the closest point to x.
2. Loop through the array of points, calculating the distance from each point to x in order to find the closest one.
3. Calculate a new set of K means by finding occurences where the total distance between means is less than the threshold distance between iterations (converging distance).
4. Map each coordinate point to the index of the point that it's closest to in the array of points.
5. Reduce the result by adding the longitudes and latitudes for every point closest to the center in addition to the total amount of closest points.
6. Map each point to a new center by finding the average latitude and longitude for each pair of closest points and save them in a new array.
7. Find the difference between the current and original distance of each center.

Final K Center Points

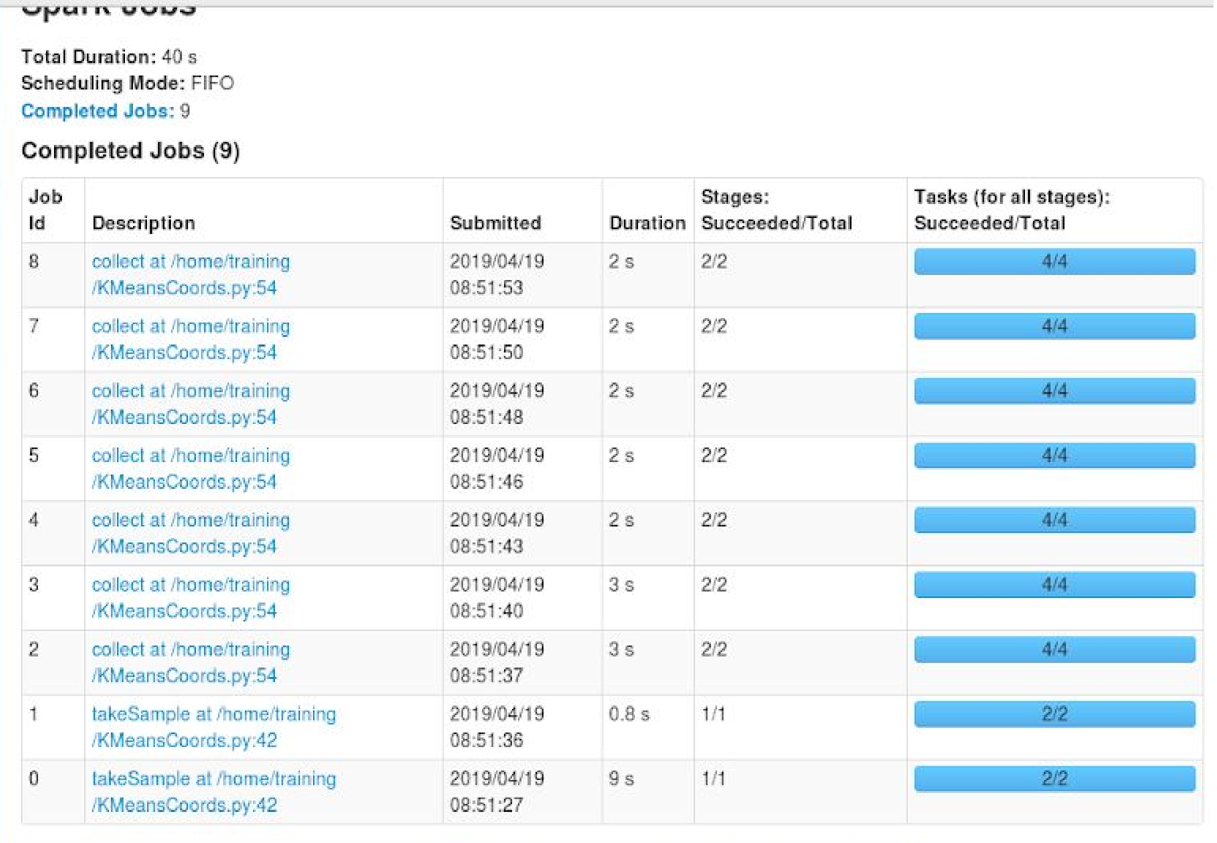


```
19/04/19 15:02:31 INFO scheduler.DAGScheduler: waiting: Set(Stage 5)
19/04/19 15:02:31 INFO scheduler.DAGScheduler: failed: Set()
19/04/19 15:02:31 INFO scheduler.DAGScheduler: Missing parents for Stage 5: List()
19/04/19 15:02:31 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 4.0 (TID 9) in 962 ms on localhost (2/2)
19/04/19 15:02:31 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool
19/04/19 15:02:31 INFO scheduler.DAGScheduler: Submitting Stage 5 (PythonRDD[16] at collect at /home/training/KMeansC
oords.py:54), which is now runnable
19/04/19 15:02:31 INFO storage.MemoryStore: ensureFreeSpace(5584) called with curMem=9970379, maxMem=280248975
19/04/19 15:02:31 INFO storage.MemoryStore: Block broadcast_6 stored as values in memory (estimated size 5.5 KB, free
 257.8 MB)
19/04/19 15:02:31 INFO storage.MemoryStore: ensureFreeSpace(3516) called with curMem=9975963, maxMem=280248975
19/04/19 15:02:31 INFO storage.MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 3.4 KB
, free 257.7 MB)
19/04/19 15:02:31 INFO storage.BlockManagerInfo: Added broadcast_6_piece0 in memory on localhost:41741 (size: 3.4 KB,
 free: 258.1 MB)
19/04/19 15:02:31 INFO storage.BlockManagerMaster: Updated info of block broadcast_6_piece0
19/04/19 15:02:31 INFO spark.SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:839
19/04/19 15:02:31 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from Stage 5 (PythonRDD[16] at collect at /
home/training/KMeansCoords.py:54)
19/04/19 15:02:31 INFO scheduler.TaskSchedulerImpl: Adding task set 5.0 with 2 tasks
19/04/19 15:02:31 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 10, localhost, PROCESS_LOCAL, 11
01 bytes)
19/04/19 15:02:31 INFO executor.Executor: Running task 0.0 in stage 5.0 (TID 10)
19/04/19 15:02:31 INFO storage.ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
19/04/19 15:02:31 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
19/04/19 15:02:31 INFO python.PythonRDD: Times: total = 42, boot = -1045, init = 1087, finish = 0
19/04/19 15:02:31 INFO executor.Executor: Finished task 0.0 in stage 5.0 (TID 10). 948 bytes result sent to driver
19/04/19 15:02:31 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 5.0 (TID 11, localhost, PROCESS_LOCAL, 11
01 bytes)
19/04/19 15:02:31 INFO executor.Executor: Running task 1.0 in stage 5.0 (TID 11)
19/04/19 15:02:31 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 10) in 77 ms on localhost (1/2)
19/04/19 15:02:31 INFO storage.ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
19/04/19 15:02:31 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
19/04/19 15:02:31 INFO python.PythonRDD: Times: total = 48, boot = -196, init = 244, finish = 0
19/04/19 15:02:31 INFO executor.Executor: Finished task 1.0 in stage 5.0 (TID 11). 911 bytes result sent to driver
19/04/19 15:02:31 INFO scheduler.DAGScheduler: Stage 5 (collect at /home/training/KMeansCoords.py:54) finished in 0.1
71 s
19/04/19 15:02:31 INFO scheduler.DAGScheduler: Job 3 finished: collect at /home/training/KMeansCoords.py:54, took 2.2
21594 s
distance:  0.0
final K center points: [[36.84127339936042, -118.5927607447135], [34.1448236603, -117.901913703], [34.2531391573, -11
8.034769136], [35.2640448771, -111.801050819], [33.3636757867, -111.684382747]]
19/04/19 15:02:31 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 5.0 (TID 11) in 93 ms on localhost (2/2)
19/04/19 15:02:31 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
```

Tracking the Job

- Open Mozilla Firefox
- Go to:

  localhost:4040



Spark Jobs

Total Duration: 40 s
Scheduling Mode: FIFO
Completed Jobs: 9

Completed Jobs (9)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
8	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:53	2 s	2/2	4/4
7	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:50	2 s	2/2	4/4
6	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:48	2 s	2/2	4/4
5	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:46	2 s	2/2	4/4
4	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:43	2 s	2/2	4/4
3	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:40	3 s	2/2	4/4
2	collect at /home/training /KMeansCoords.py:54	2019/04/19 08:51:37	3 s	2/2	4/4
1	takeSample at /home/training /KMeansCoords.py:42	2019/04/19 08:51:36	0.8 s	1/1	2/2
0	takeSample at /home/training /KMeansCoords.py:42	2019/04/19 08:51:27	9 s	1/1	2/2