

Part 3: Different Parallelism

Please try different levels of parallelism: 2, 3, 4, 5

Design

To try out different levels of parallelism, the configuration for spark-submit was set by:

```
--conf spark.default.parallelism = 2
```

It can be observed that with higher level of parallelism ($\rightarrow 5$), a convergence is achieved. While when parallelism is lower (2 or 3), no convergence was achieved until the maximum iteration was reached.

In addition, with a parallelism of 5 rather than 4, the convergence is achieved faster. (0.529500s vs 0.549806s)

Parallelism = 2:

```
2019-04-23 15:58:54 INFO DAGScheduler:54 - Job 43 finished: collect at /spark-examples/labs/lab2/part3/./nba-kmeans-3.py:86, took 0.429484 s
Iteration - 40 round
Update: [[23.876, 5.0318, 16.3141], [4.0666, 2.786, 17.9298], [7.6774, 2.6829, 8.4829], [21.7573, 4.2738, 5.7392]]
```

Parallelism = 3:

```
2019-04-23 16:07:54 INFO DAGScheduler:54 - Job 43 finished: collect at /spark-examples/labs/lab2/part3/./nba-kmeans-3.py:86, took 0.507689 s
Iteration - 40 round
Update: [[22.1216, 4.2449, 5.478], [23.8189, 5.0402, 16.2664], [8.3097, 2.836, 8.5697], [4.0173, 2.768, 17.891]]
```

Parallelism = 4:

```
2019-04-23 16:14:15 INFO DAGScheduler:54 - ResultStage 35 (collect at /spark-examples/labs/lab2/part3/./nba-kmeans-3.py:86) finished in 0.194 s
2019-04-23 16:14:15 INFO DAGScheduler:54 - Job 19 finished: collect at /spark-examples/labs/lab2/part3/./nba-kmeans-3.py:86, took 0.549806 s
Converge at the 15 iteration

Final Centroids: [[6.4024, 2.5723, 9.1307], [23.7756, 4.9935, 16.1945], [21.2841, 4.2365, 5.6852], [4.1076, 2.8058, 18.3176]]
```

Parallelism = 5:

```
2019-04-23 16:16:52 INFO TaskSchedulerImpl:54 - Removed TaskSet 51.0, whose tasks have all completed, from pool
2019-04-23 16:16:52 INFO DAGScheduler:54 - ResultStage 51 (collect at /spark-examples/labs/lab2/part3/./nba-kmeans-3.py:86) finished in 0.200 s
2019-04-23 16:16:52 INFO DAGScheduler:54 - Job 27 finished: collect at /spark-examples/labs/lab2/part3/./nba-kmeans-3.py:86, took 0.529500 s
Converge at the 23 iteration

Final Centroids: [[8.3097, 2.836, 8.5697], [23.8189, 5.0402, 16.2664], [22.1216, 4.2449, 5.478], [4.0173, 2.768, 17.891]]
2019-04-23 16:16:52 INFO AbstractConnector:318 - Stopped Spark@380f39d6{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```