

Benjamin Morrison
November 24, 2019
Assignment 7
Hadoop

Implement an Iterative Algorithm with Spark

RDD transformation procedure

Helper functions are created in order to assist the RDD transformation process. These are `closestPoint()`, `addPoints()` and `distanceSquared()`. We begin by splitting the line with the comma as a delimiter. Following that a new RDD is created which grabs only the 3rd and 4th fields from the file. We then create another RDD which ignores cases where 0 is present. We then create a k-length array and take random samples of location points. Following this, we iterate to find the best points, add the points and then calculate the distance using the helper functions. Finally, we print them out.

Below is a screenshot of the final output showing the final center points.

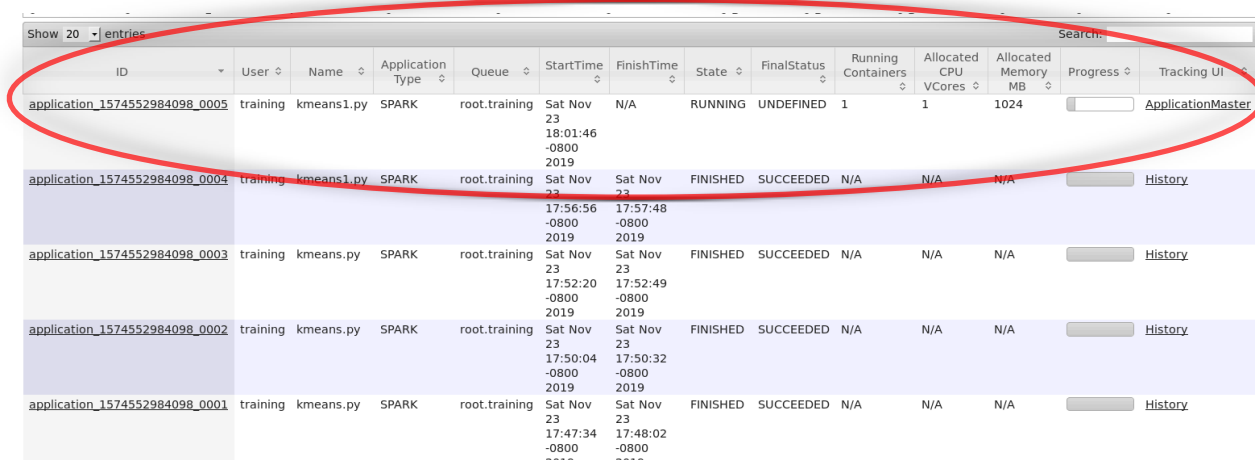
```
19/11/23 17:57:47 INFO scheduler.DAGScheduler: Job 8 finished: collect at /home/training/training_materials/dev1/exercises/spark-iterati
ve/kmeans1.py:65, took 2.143981 s
Distance between iterations: 0.0332419852833
Final center points: [[34.49158712091043, -118.21165071858718], [38.16939508093724, -121.21805924888717], [35.0852504610273, -112.574893
586628], [33.760054946070575, -116.56902791276211], [43.96551873516047, -121.3930092136969]]
19/11/23 17:57:47 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 15.0 (TID 31) in 84 ms on localhost (2/2)
19/11/23 17:57:47 INFO cluster.YarnScheduler: Removed TaskSet 15.0, whose tasks have all completed, from pool
[training@localhost spark-iterative]$
```

Tracking the job

Below is the command I used to submit this job to spark utilizing python

```
spark-submit --master yarn-client kmeans1.py /loudacre/
devicestatus_etl/*
```

We can track the progress of the job created from the above command by navigating to the URL <http://localhost:8088> from the VM. Below is a screenshot from this address that shows the process from the above screenshot after it has finished.



ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU V-Cores	Allocated Memory MB	Progress	Tracking UI
application_1574552984098_0005	training	kmeans1.py	SPARK	root.training	Sat Nov 23 18:01:46 -0800 2019	N/A	RUNNING	UNDEFINED	1	1	1024	<div></div>	ApplicationMaster
application_1574552984098_0004	training	kmeans1.py	SPARK	root.training	Sat Nov 23 17:56:56 -0800 2019	Sat Nov 23 17:57:48 -0800 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1574552984098_0003	training	kmeans.py	SPARK	root.training	Sat Nov 23 17:52:20 -0800 2019	Sat Nov 23 17:52:49 -0800 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1574552984098_0002	training	kmeans.py	SPARK	root.training	Sat Nov 23 17:50:04 -0800 2019	Sat Nov 23 17:50:32 -0800 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1574552984098_0001	training	kmeans.py	SPARK	root.training	Sat Nov 23 17:47:34 -0800 2019	Sat Nov 23 17:48:02 -0800 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History

