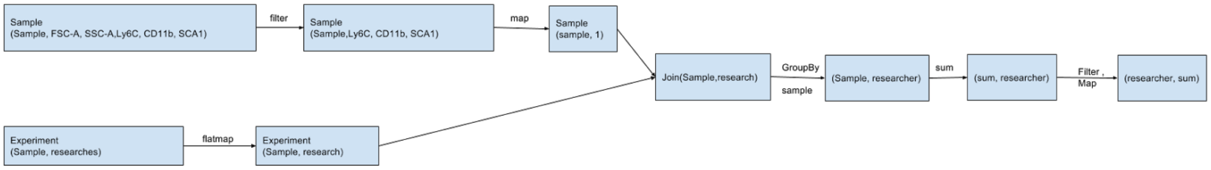
COMP5349 Assign2

Task1

In this task, the measures files(like files in cytmoetry/large directory) are read as csv in flink, followed by a filter operation to ignore data with invalid FSC-A and SSC-A value. After that, a map function is applied to emit experimentID and an Integer counter only.
The experiment file(like file share/cytometry/experiments) is read as csv in flink either, meanwhile a flatmap transformation is used to split different researchers who cooperated on same measures.
A join operation is executed upon two dataset loaded above based on the same key - experiment-sample-id(the first column) followed by an groupBy transformation and on the same column after which a sum aggregation is put on. Finally, sortPartions are introduced to sort out the researcher with most measures and sorted by alphabetically order.



Performance:

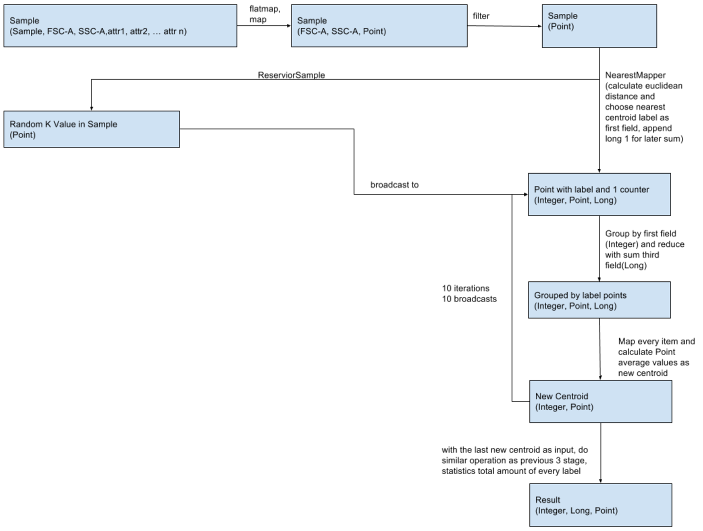
Dataset	Runtime	Run Command On Large Dataset
Small	2452 ms	flink run -m yarn-cluster -yn 3 \ -c assign2.Task1 \ comp5349assign2-1.0-SNAPSHOT.jar \ --output hdfs:///user/cshe6391/assign2/task1 \ --sample-dir hdfs:///share/cytometry/large/ \ --experiment-path hdfs:///share/cytometry/experiments.csv
Large	4690 ms	

Task2

In this task, the measures files(like files in cytmoetry/large directory) are read as csv in flink, followed by a filter operation to ignore data with invalid FSC-A and SSC-A value just the same as Task 1. On the second stage, k random points are set upon the parameter k as the initial centroid. Because this centroid dataset will be used repeatedly, IterativeDataSet is applied to boost performance.

To calculate the distance and choose the nearest centroid, two customized class and one customized rich map function is designed as follows: the Point Class is designed for store sample point in this task. Ly6C, CD11b and SCA1 are set as default values chosen to fill Point class. CenterPoint extends Point Class which own extra field to mark the cluster label. The NearestMapper is the rich map function designed for calculate euclidean distance between points and label the center point. BroadcastSet is used for every iteration on new calculated centroid.

In every iteration, each sample point is used to calculate distance between every centroid and label the correct nearest centroid with flink map transformation. In the final stage, here comes a statistics operation starting with a reduceBy transformation based on centroid label(range from zero to k). Point associated with respective centroid are summarised helping to solve the new centroid.



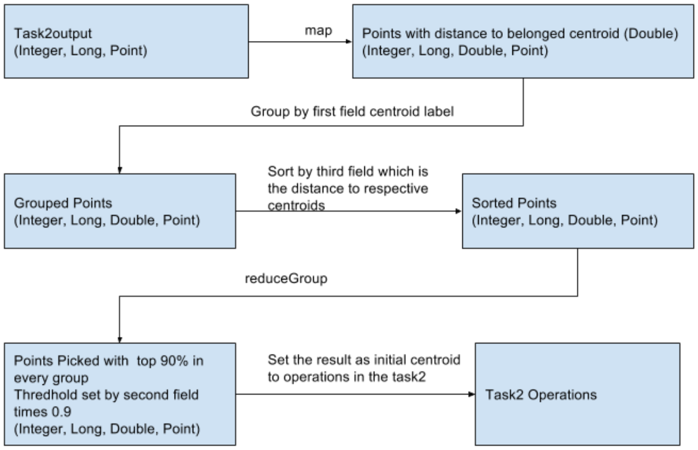
Performance:

Dataset	Runtime	Run Command On Large Dataset
Small	6568 ms	flink run -m yarn-cluster -yn 3 \ -c assign2.Task2 \ comp5349assign2-1.0-SNAPSHOT.jar \ --k 3 \ --iterator 10 \ --sample-dir hdfs:///share/cytometry/large/ \ --output hdfs:///user/cshe6391/assign2/task2 \ --mask 00110001000000
Large	28739 ms	

--iterator and --mask is optional, their default value is 10 and 0011000100000 respectively.

Task3

This task does similar jobs as task2 in the beginning but it remain the distance between itself and its centroid. The distance is used to sort and emit the first 90% related points to form a new sample dataset. At the end, we repeat the same operation as task2 and output the final result.


Task2output
(Integer, Long, Point)
map
Points with distance to belonged centroid (Double)
(Integer, Long, Double, Point)
Group by first field centroid label
Grouped Points
(Integer, Long, Double, Point)
Sort by third field which is the distance to respective centroids
Sorted Points
(Integer, Long, Double, Point)
reduceGroup
Points Picked with top 90% in every group
Thredhold set by second field times 0.9
(Integer, Long, Double, Point)
Set the result as initial centroid to operations in the task2
Task2 Operations

Performance:

Dataset	Runtime	Run Command On Large Dataset
Small	13035 ms	flink run -m yarn-cluster -yn 3 \ -c assign2.Task3 \ comp5349assign2-1.0-SNAPSHOT.jar \ --k 3 \ --sample-dir hdfs:///share/cytometry/large/ \ --task2out-dir hdfs:///user/cshe6391/assign2/task2/ \ --output hdfs:///user/cshe6391/assign2/task3/ \ --mask 00110001000000
Large	80461 ms	

--iterator and --mask is optional, their default value is 10 and 0011000100000 respectively.

Appendix

HDFS Location:

Task 1	/user/cshe6391/assign2/task1
Task 2	/user/cshe6391/assign2/task2
Task 3	/user/cshe6391/assign2/task3