# MRPrintStatistics Function Analysis

## Function: MRPrintStatistics(inputPoints, C)

This function processes an RDD of (point, group) pairs to compute, for each centroid I, the number of points from groups ‘A’ and ‘B’ assigned to it.

* InputPoints: Array of tuples (point, group), where point is a d-dimensional tuple and group is a character.
* C: Set of centroids computed using the K-Means algorithm.

## Memory and Distributed Computation Costs

The total number of points is N, but since we have applied the random partitioning technique with l partitions, each RDD has points.

### Round 1

* Map Phase:
* Each point is processed to compute the index of the closest centroid. . This transformation requires N/L space for each RDD à
* Reduce Phase:
* Let , where count is the sum of all the 1’s in . Since we have K clusters and two groups, the maximum local memory used is

### Round 2

* Map Phase:
* This phase changes the key of the tuples: . This transformation does not increase the number of records (2K), memory cost:
* Reduce Phase:
* Records are grouped by the centroid indexes i in the shuffle phase so the reduce phase does not do any transformation. Memory cost:

### Total Bound

The resulting RDD contains dictionaries mapping group labels to count, with an impact of memory of:

The last equation follows from the fact that, in the worst case, we could have one cluster per point, so K will be of the same order as N. As we can see, supposing that a point space, the result does not depend on the dimension of the points.