**Local Space Analysis ML**

The MRPrintStatistics() method calculates statistics on the clusters generated by the K-means algorithm, considering the distribution of points between the demographic groups and The goal of this analysis is to determine an upper bound for the local space indicator in the MapReduce context, justifying each transformation used.

**Method Phases**

1. Input and Output:
   * Input: A distributed RDD of pairs where `Vector` represents a point and `Character` the demographic group .
   * Output: A list of triplets where:
     + : Index of the centroid.
     + : Number of points from group assigned to the centroid.
     + : Number of points from group assigned to the centroid.
2. Computation Phases:

* Map Phase:
  + Each partition processes the assigned points and calculates the index of the closest centroid for each point.
  + Generates pairs or based on the demographic group.
* Reduce Phase:
  + The pairs are aggregated to sum the counts of and for each centroid.

**Upper Bound Calculation for Local Space**

The function MRPrintStatistics() calculates statistics on clusters generated by the KMeans algorithm in a MapReduce context. The local space bound is given by:

1. Space for points in a partition:

* Each partition processes at most points, where is the total number of points and is the number of partitions.
* The local space required is proportional to the number of points in the partition, hence .

1. Space for intermediate results:

* Each partition maintains at most clusters, where is the number of centroids.
* Since the are partitions, the total space for intermediate results is proportional to .

1. Space for centroids:

* The centroids are stored in an array of size , where each centroid has dimensions
* The space required to store the centroids is therefore proportional to

**Transformations used in the function**

1. Round 1:

* **Map Phase**

Each point of vector dimension needs to find its closest centroid, computing the distance to each of centroids. Each distance computations is , so for a single point the cost is .

* **Reduce Phase**

Each reduce task will receive all the partial counts for a given centroid from each partition, which produces up to tuples.

1. Round 2:

* **Map Phase**
* **Reduce Phase**

Each centroid has up to contributions from the partitions. These values are merged and sorted (by the function sortByKey()) to produce the final output.

**Conclusion**

The upper bound for the local space required by the MRPrintStatistics() method is: