**Project 6: Twister K-Means**

**K-Means:**

K-means ia an un-supervised learning algorithm. It is used for clustering of data pionts. In statistics and machine learning, K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters where each observation belongs to the cluster with the nearest mean

**Twister:**

Twister is a light-weight Map – Reduce programming model that uses iterative approach to extend the functionality of normal Map – Reduce paradigm.

**Twister K-Means:**

Introduction:

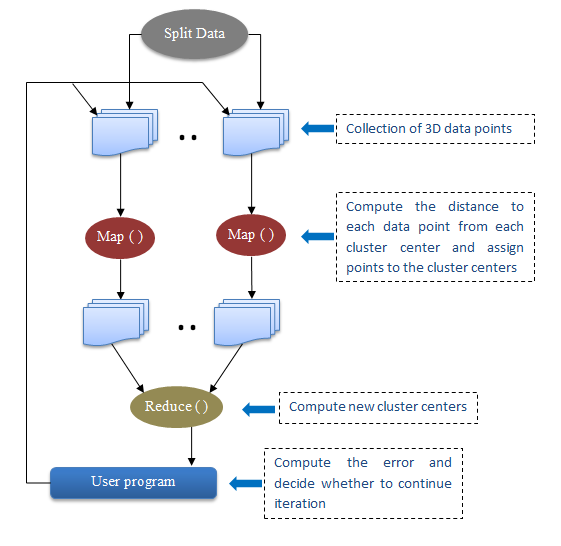
In each iteration all the map task get the same input data (current cluster centers) and each computes a partial cluster center by going through the corresponding 3D data points. A reduce task computes the average of the partial cluster centers and produces the cluster centers for the next step. Once the main program receives these new cluster centers, it calculates the difference between the old and new, then determines if it needs to execute another cycle of MapReduce computations. 

Fig. The workflow of K-means application with Twister MapReduce framework

Program Sketch:

* Each Map task gets a set of 3D points. These data points are static i.e. they do not change over the execution of the code and Map task access them in each iteration.
* Due to invariant nature of the data points, they are marked as static in Twister. So, loaded only once in the memory for the entire execution.
* The cluster centers are computed in each iteration are dynamic data, as they tend to change over the iterations. Also, the calculated cluster centers of a particular iteration serve as the input for the next if they don't satisfy the required conditions.

Data State:

Data points: Static

Centroids: Dynamic.

Distance formula used: Euclidean