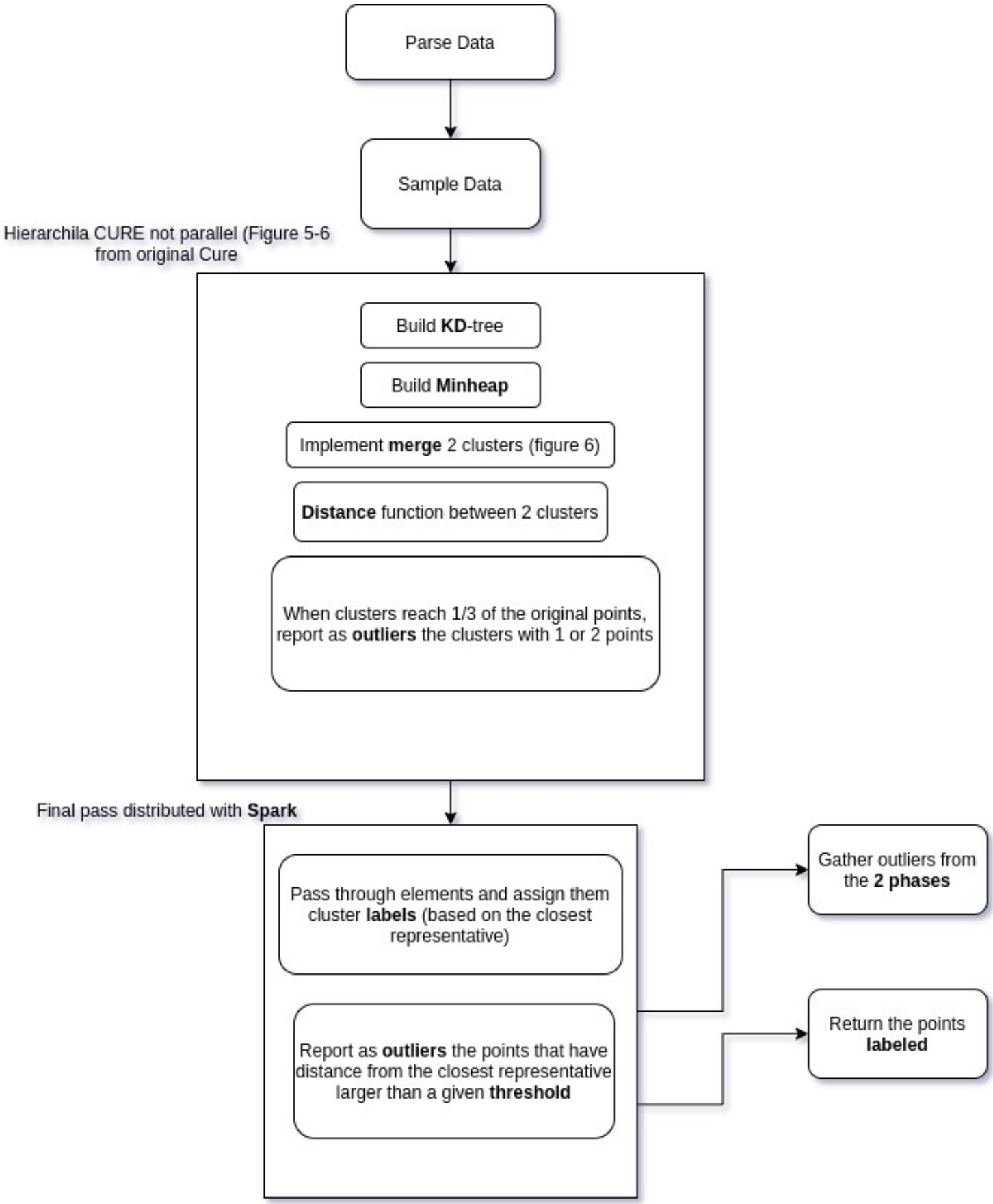
Parse Data
Sample Data
Hierarchila CURE not parallel (Figure 5-6 from original Cure
Build **KD**-tree
Build **Minheap**
Implement **merge** 2 clusters (figure 6)
Distance function between 2 clusters
When clusters reach 1/3 of the original points, report as **outliers** the clusters with 1 or 2 points
Final pass distributed with **Spark**
Pass through elements and assign them cluster **labels** (based on the closest representative)
Report as **outliers** the points that have distance from the closest representative larger than a given **threshold**
Gather outliers from the **2 phases**
Return the points **labeled**

KD-Tree
It will be used to find in logn the closest cluster. Check citation [Sam90] from original paper. It should support add/delete node and closest_cluster. There are implementation for scala https://github.com/thesamet/kdtree-scala and https://github.com/justinian336/trees .

MinHeap (Priority Queue)
It will be used to get the cluster that has the minimum distance from the closest cluster. It is supported in the standard library of scala in https://www.scala-lang.org/api/current/scala/collection/mutable/PriorityQueue.html we will only need to add the relocate when 2 clusters are merged.

Algorithm will take the Points, parameter a (0,1) and a threshold to calculate the outliers and return the outliers found along with the labeled data (in a file with the format x,y,label).