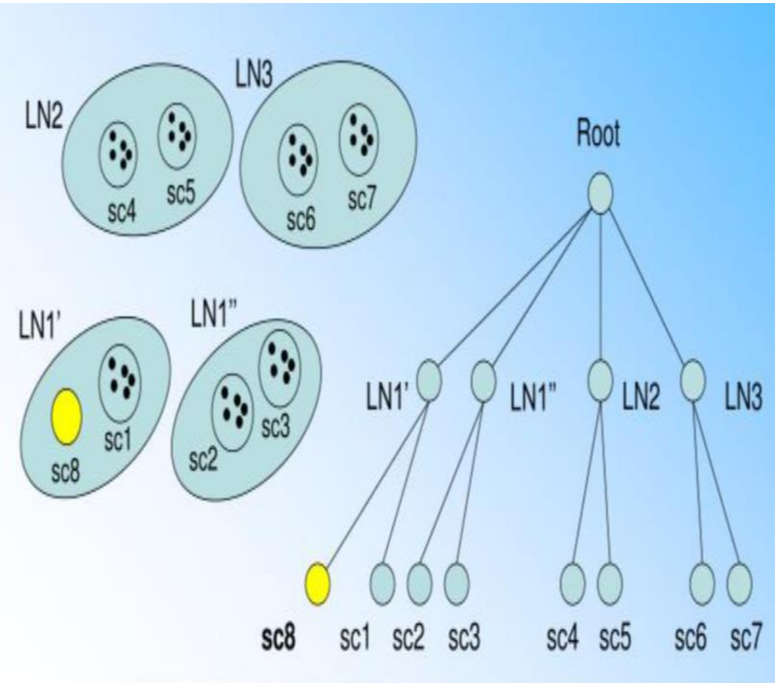
# BIRCH CLUSTERING

## Balanced Iterative Reducing And Clustering Using Hierarchies - (BIRCH)

Clustering algorithms like K-means clustering do not perform clustering very efficiently and it is difficult to process large datasets with a limited amount of resources. So, regular clustering algorithms do not scale well in terms of running time and quality as the size of the dataset increases. This is where BIRCH clustering comes in.

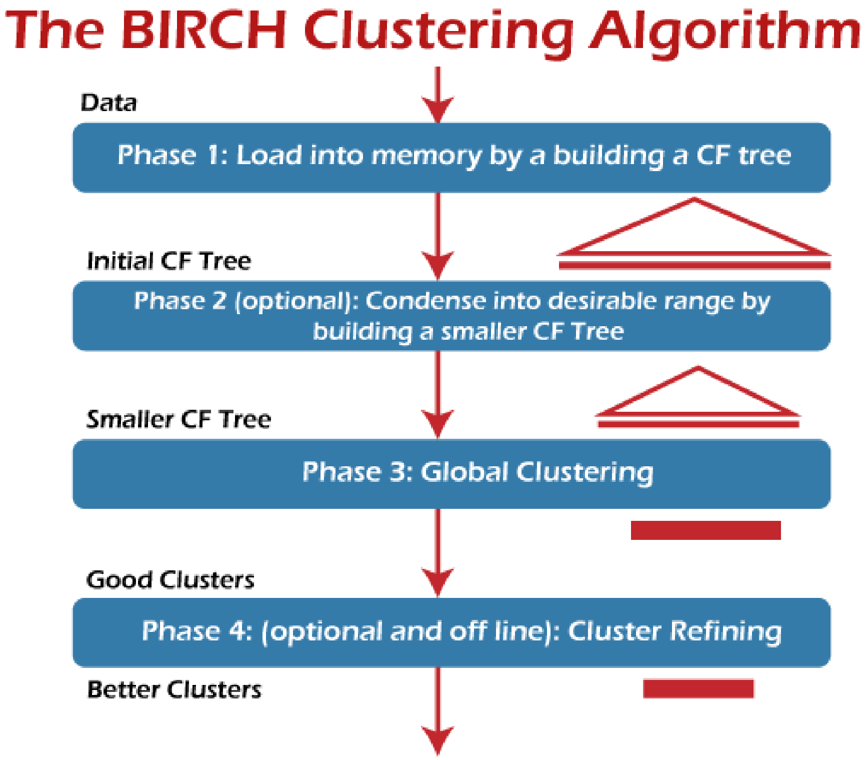
BIRCH clustering algorithm can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset.



# THE BIRCH CLUSTERING ALGORITHM CONSISTS OF TWO STAGES:

1. **Building the CF Tree:** BIRCH summarizes large datasets into smaller, dense regions called Clustering Feature (CF) entries. Formally, a Clustering Feature entry is defined as an ordered triple (N, LS, SS) where 'N' is the number of data points in the cluster, 'LS' is the linear sum of the data points, and 'SS' is the squared sum of the data points in the cluster. A CF entry can be composed of other CF entries. Optionally, we can condense this initial CF tree into a smaller CF.

2. **Global Clustering:** Applies an existing clustering algorithm on the leaves of the CF tree. A CF tree is a tree where each leaf node contains a sub-cluster. Every entry in a CF tree contains a pointer to a child node, and a CF entry made up of the sum of CF entries in the child nodes. Optionally, we can refine these clusters.



The BIRCH Clustering Algorithm

Data

Phase 1: Load into memory by a building a CF tree

Initial CF Tree

Phase 2 (optional): Condense into desirable range by building a smaller CF Tree

Smaller CF Tree

Phase 3: Global Clustering

Good Clusters

Phase 4: (optional and off line): Cluster Refining

Better Clusters

# Advantage And Disadvantage

## Advantages

- ✓ BIRCH is useful for performing precise Clustering on large datasets
- ✓ An main advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multidimensional metric data points to produce the best quality clustering for a given set of resources (memory and time constraints). In most cases, BIRCH only requires a single scan of the database.

## Disadvantages

- ✓ BIRCH has one major drawback, it can only process metric attributes. A metric attribute is an attribute whose values can be represented in Euclidean space, i.e., no categorical attributes should be present