BIRCH
Clustering Algorithm

BIRCH : Balanced Iterative Reducing and Clustering Using Hierarchies

Minju Kim

2022. 1. 5. WED

What is BIRCH?

“ BIRCH stands for

Balanced Iterative Reducing and Clustering Using Hierarchies.”

Pros

- BIRCH can cluster large datasets by first generating a small and compact summary of the the large dataset
- BIRCH is often used to complement other clustering algorithms by creating a summary of the dataset that the other clustering algorithm can now use

Cons

- BIRCH can only process metric attributes. (no categorical attributes)

The BIRCH Cluster Algorithm's 2 stages

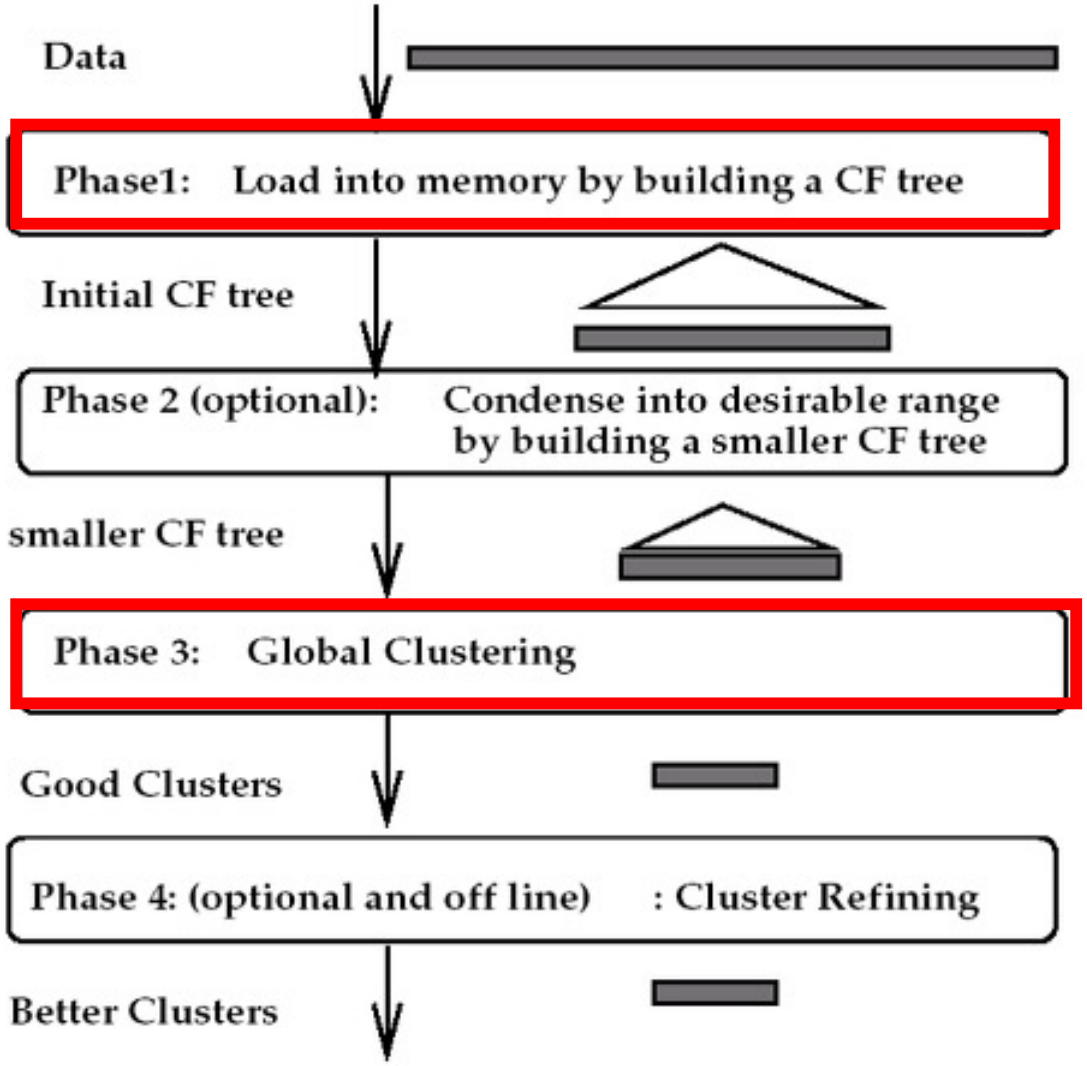
1. Building the CF Tree

- BIRCH summarizes large datasets into smaller, dense regions called Clustering Feature (CF) entries.

2. Global Clustering

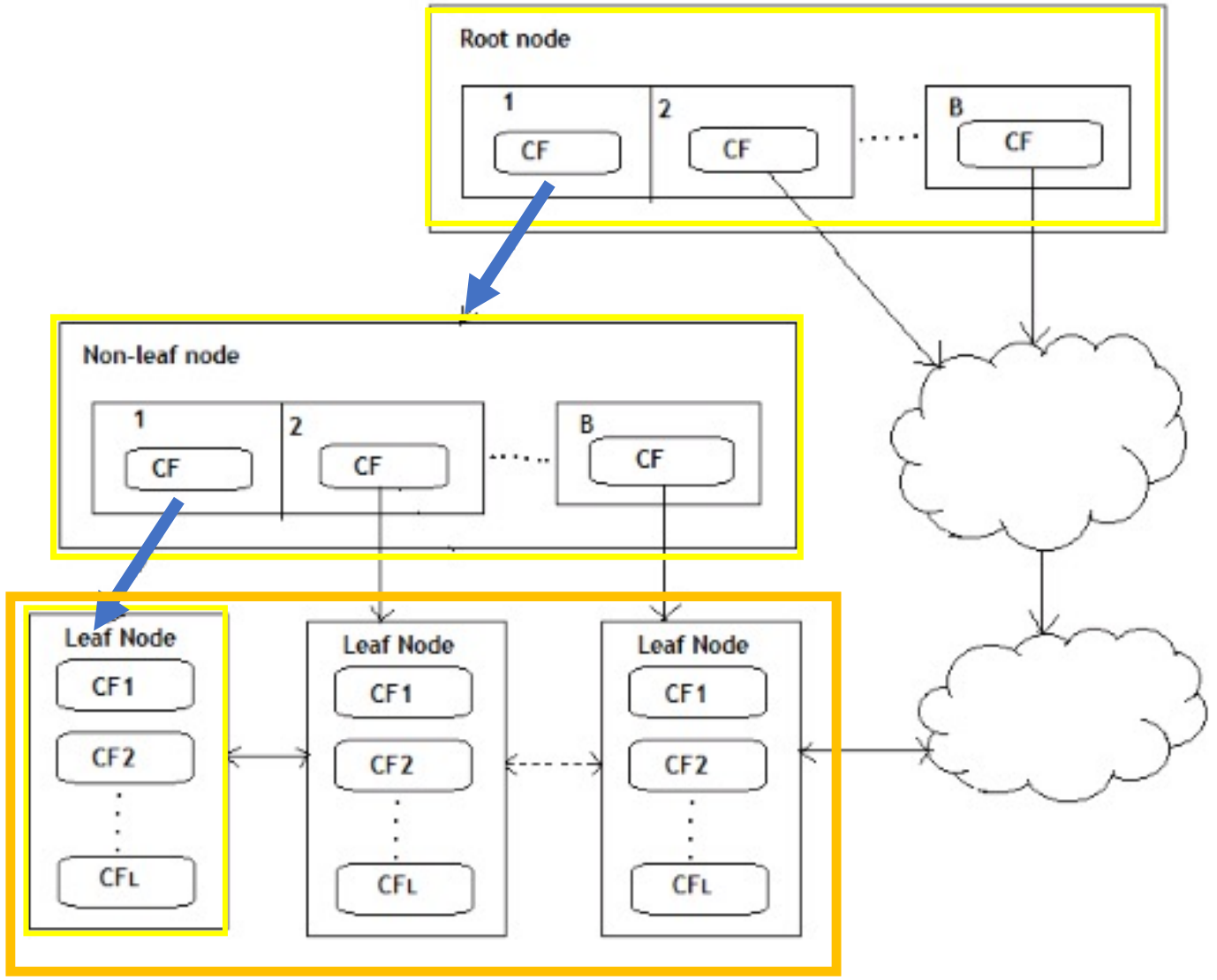
- Applies an existing clustering algorithm on the leaves of the CF tree.

=> Two Step Clustering



How does it work?

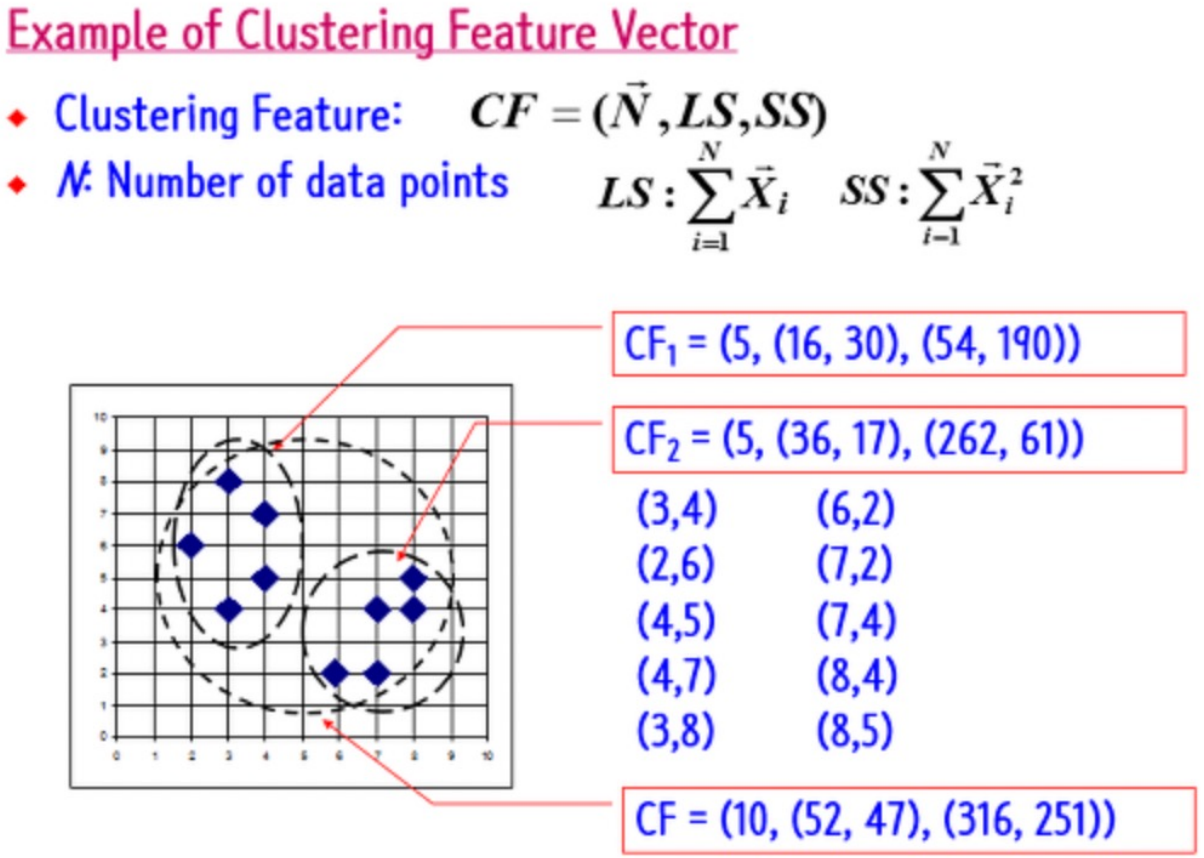
- BIRCH uses a tree structure to create a cluster
- CF Tree : Clustering Feature Tree
- Each node of this tree is composed of several Clustering features (CF).



More about Cluster Feature

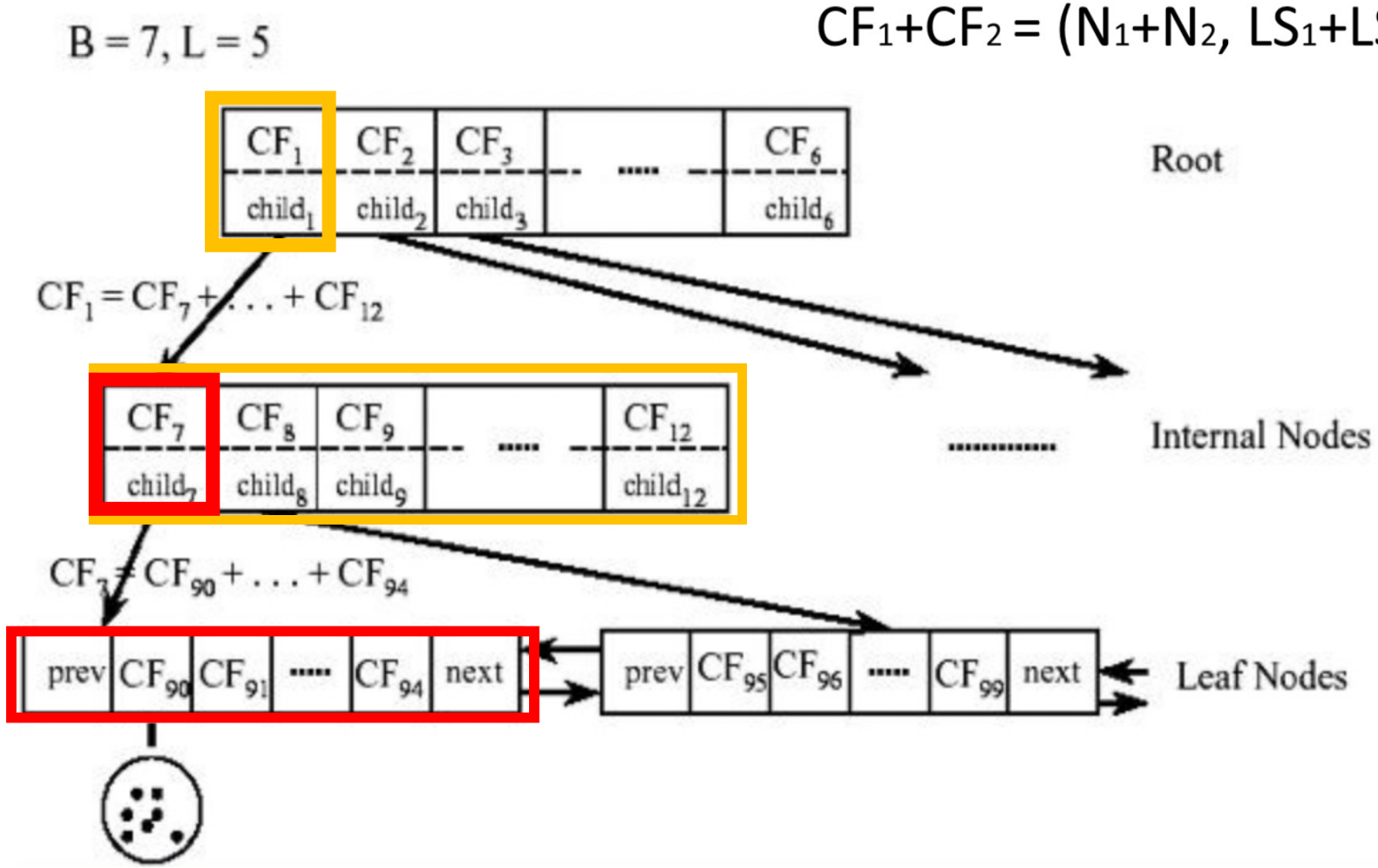
- Each CF is a triplet, which can be represented by (N, LS, SS).
- CF = (N, LS, SS)
  - N : the number of sample points
  - LS : the linear sum of the feature dimensions of the sample points
  - SS : the square sum of the feature dimensions of the sample points

- Together, the linear sum and the squared sum are equivalent to the mean and variance of the data point.

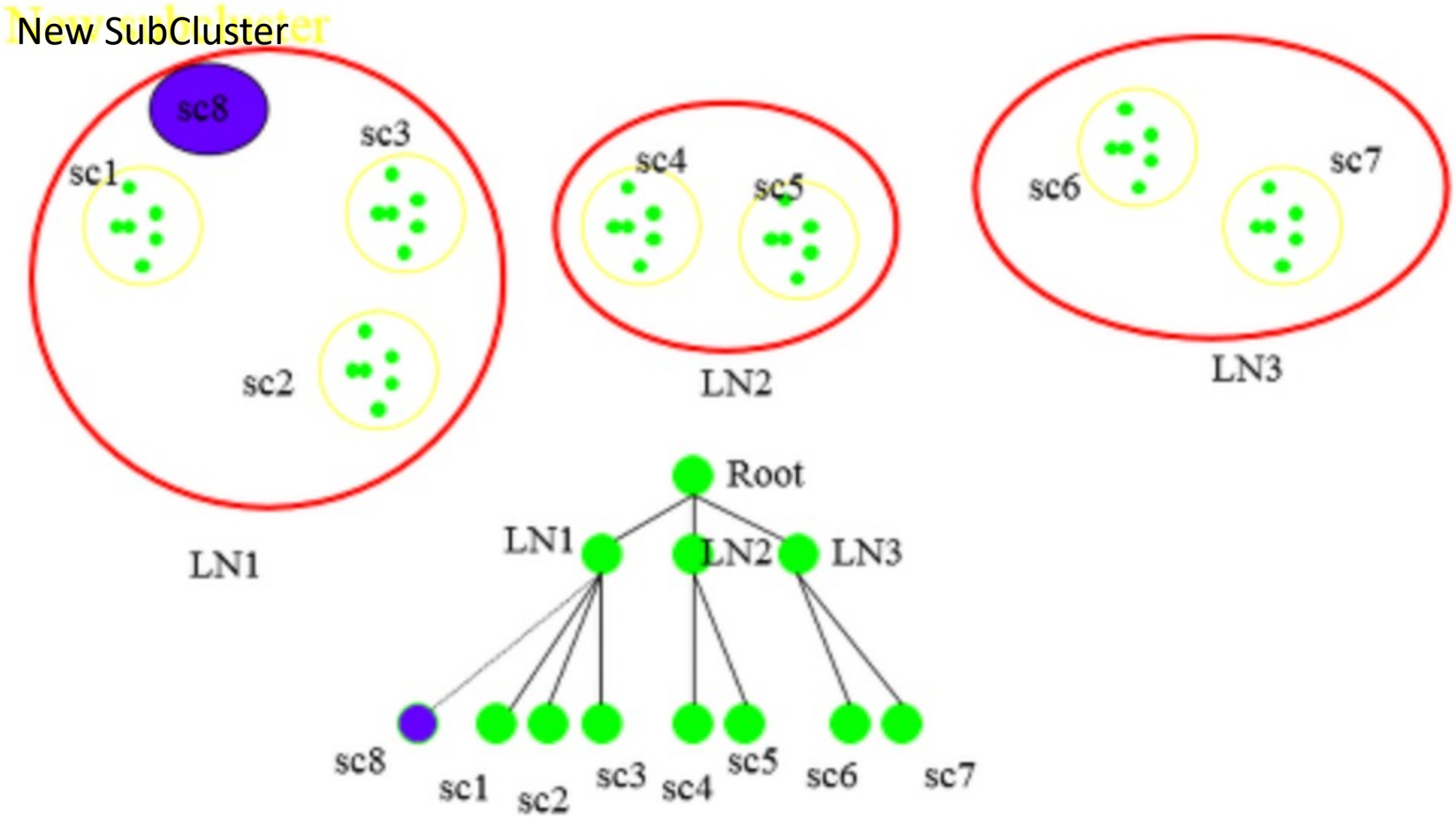


More about Cluster Feature(CF Tree)

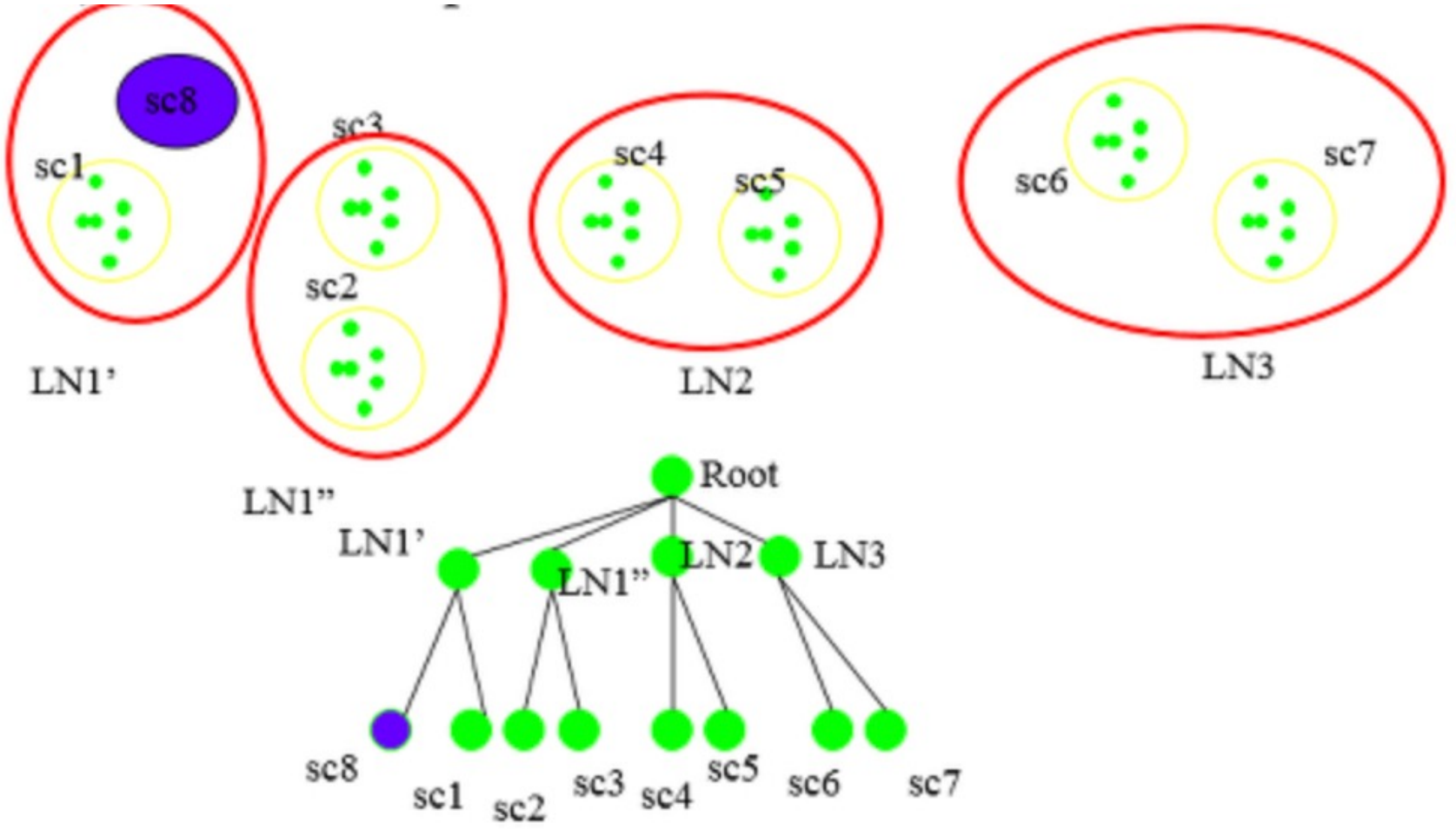
$CF_1+CF_2 = (N_1+N_2, LS_1+LS_2, SS_1+SS_2)$



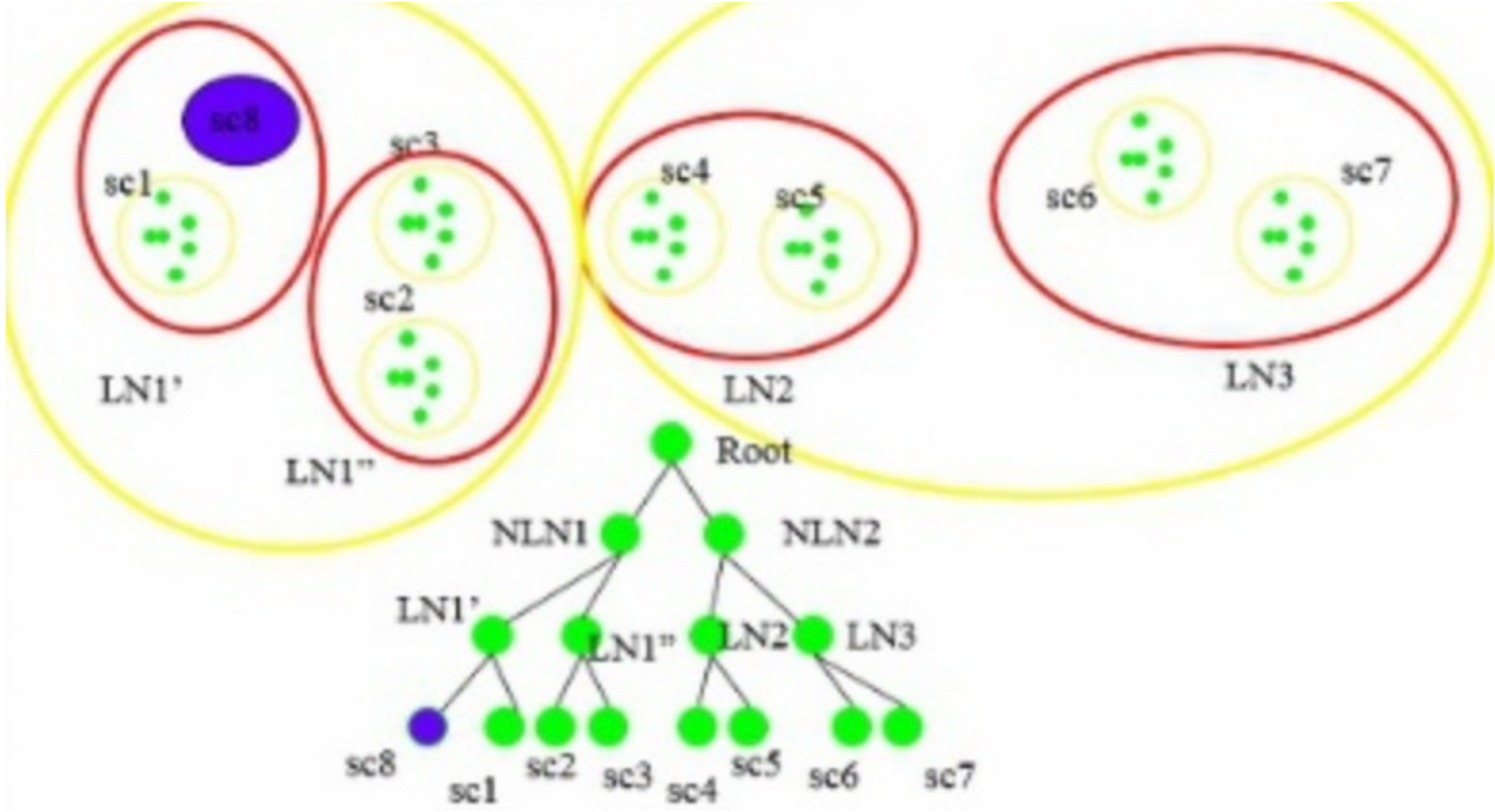
Example of BIRCH Algorithm



If the branching factor of a leaf node can not exceed 3, then LN1 is split



If the branching factor of a non-leaf node can not exceed 3,
then the root is split and the height of the CF Tree increases by one.



Summary

- BIRCH uses hierarchical methods to cluster and reduce data
- BIRCH is local (instead of global). Each clustering decision is made without scanning all data points or currently existing clusters.
- BIRCH algorithm uses a tree structure to create a cluster, which is called the Clustering Feature Tree (CF Tree)

References

- https://morioh.com/p/c23e0d680669
- https://www2.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf
- https://medium.com/geekculture/balanced-iterative-reducing-and-clustering-using-hierarchies-birch-1428bb06bb38
- https://t-lab.tistory.com/25
- https://www.geeksforgeeks.org/ml-birch-clustering/
- https://scikit-learn.org/stable/modules/clustering.html#birch

Thank you