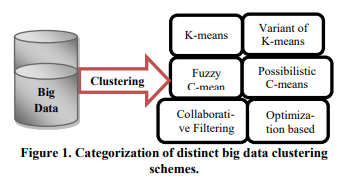
“Big Data” is an array of procedures and technologies that requires a reformed way of unity to uncover the huge volume of data, which is unseen, varied and complex. In order to process and make it simple for further research and development various data mining methods are available. Clustering is one such method which is used for data mining where we group data sets having similar features together which are known as clusters. “Big Data” is data engineering and clustering is the process of discovering similarities in the data. Though there are challenges in applying clustering techniques to big data. Therefore, we have following clustering techniques which are used for handling Big data:

1. Hierarchical Clustering
2. Partitioning Method
3. Density Based
4. General Techniques

Now discussing when and why these above-mentioned methods are used:

1. Hierarchical Clustering is used to overcome the quantity definition issue of a cluster. This is an extension of theory decision Big data model and is outlined based on the risk calculation. It is used where the data set is huge, and the data is highly dimensional.
2. Partitioning is used where we must partition a data set which is having different modes and the data sets are stable. Like PFClust is used for problems which are having multi-dimensional features and it is effective on large scale data sets too.
3. Density based methods are used for problems based on real time. The interesting fact about this approach is that it is very accurate for figuring out clusters where we have different densities and can be used where Map Reduce can be executed in parallel.
4. General Techniques includes which mainly focusses on efficiency. Like for example k-mean method is founded on the concepts of logically rotating the data based on similarity in the data set. This algorithm is based on 2 cycles: first one is representing data clustering and second one is ranking to denote out of boundary similar data sets. Overall we determine the value of k which denotes grouping based on similarity of features and adding values to the clusters based on similarity. The complex part is calculating the values of distance function and its very time consuming process.

There are few other theories as well regarding the clustering techniques in Big data. This can be generalized as follows:



References :

1. <https://www.researchgate.net/publication/322096571_Clustering_methods_in_Big_data>
2. <https://www.ijeat.org/wp-content/uploads/papers/v8i2s2/B10400182S219.pdf>