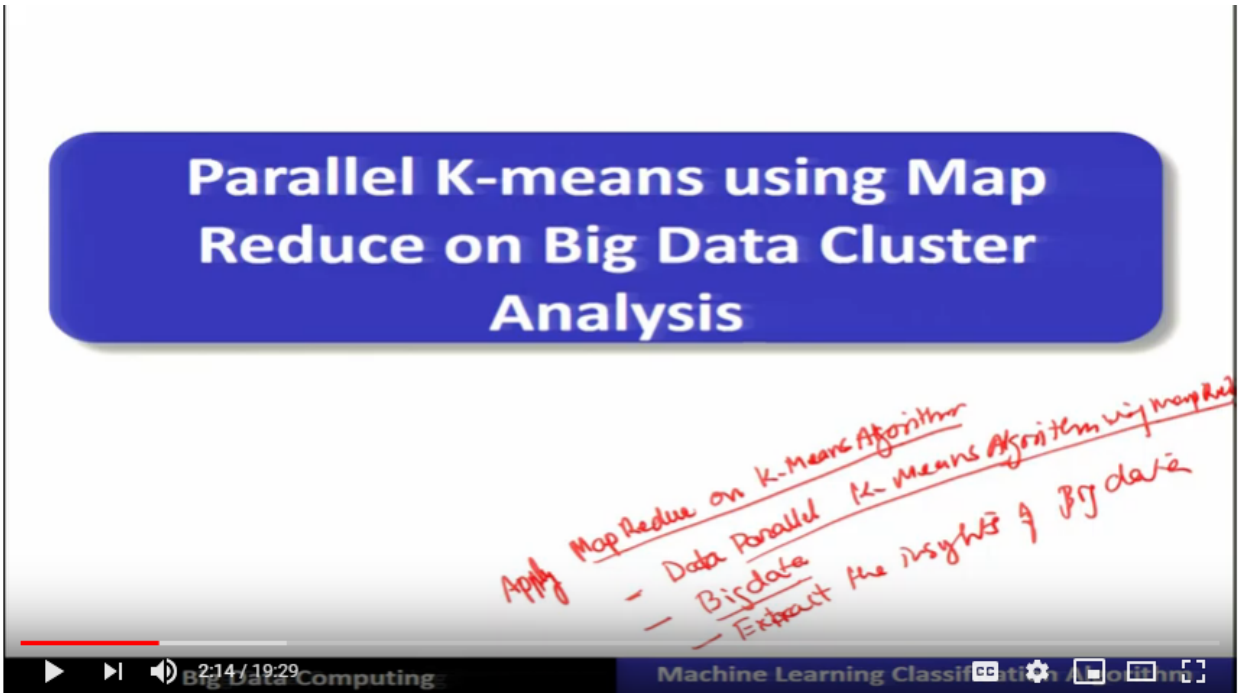
Lecture - 26
Parallel K-means using Map Reduce on Big Data Cluster Analysis

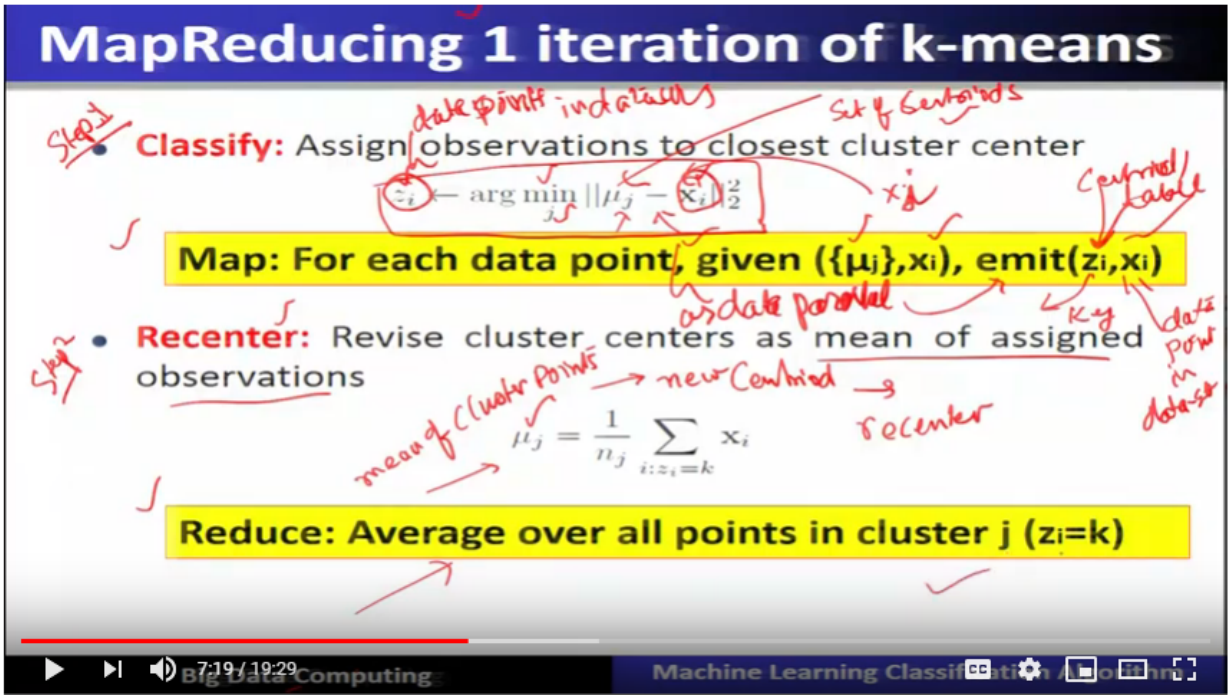
Parallel k-means using MapReduce, on big data cluster analysis, we have seen in this particular section, you will see the parallel k-means algorithm that means we will use the MapReduce.

Refer slide time :(0:36)



We will apply, the MapReduce apply, the matter MapReduce on, k-means algorithm. Now important thing here is that, this particular method of applying the MapReduce on, k-means will make this data parallel algorithm. Now after doing this data parallel k-means algorithm, using MapReduce, now we can use this particular algorithm, for the analysis of a big data. So, we can use this, particular algorithm that is called,' Parallel k', means algorithm using MapReduce with the big data. And we can extract the insights of big data. So, let us see the more details of, how MapReduce, can be applied on the k-means to achieve this parallel k-means using, MapReduce which is used for the big data analysis.

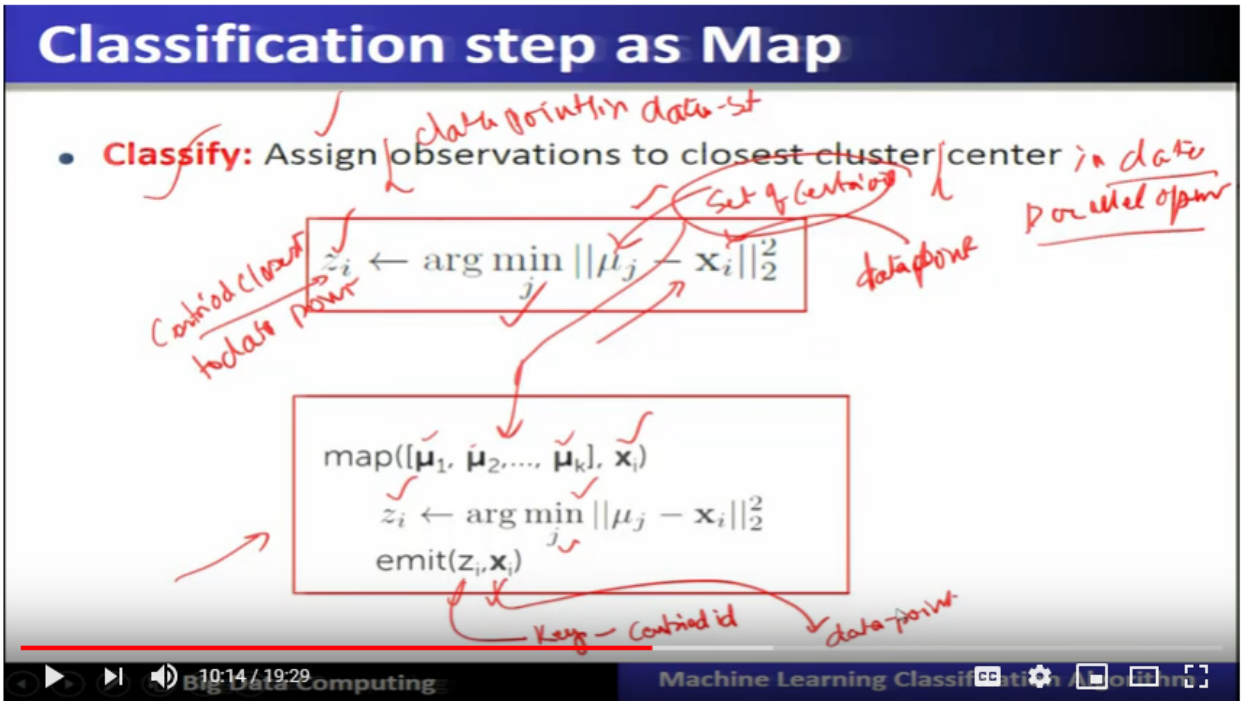
Refer slide time :(02:19)



Now this particular, iterations of the k-means, which is the step number one, which we have seen in the previous slide that is to classify, that means all the we have to assign, the, the data set, the data points, in the data set, to the closest, centroid or to the closest cluster Center, which is represented by this particular equation let us understand this particular equation. Now if we are given, a set of centroids, then for a particular data item, X I. We are calculating the similarity measure and we have to identify that particular, centroid which is having that minimum distance. So, this particular arm in, this particular formula, will identify that particular cluster, which is having that closures, closest or which is very similar or which is having the least distance or a minimum distance, with that data set, compared to the other plus other centroids, hence this particular J is the label, which will be identified, for a particular X I. Now we are now going to get, Z I as the label or as a centroid for, this data point X I. Now the next step, in the k-means algorithm which we have seen in the previous slides is called,' Recenter', that means after identifying the clusters, that means after classifying the data set points, into the different clusters. Now for every cluster we are going to calculate the mean, of that particular cluster. So, for every cluster we are going to compute, the mean and that will become the new centroid, for every cluster this is called,' Recenter'. And these two step is going to iterate. So, this these two step will form the first iteration, of the MapReduce k-means algorithm. Now let us use the map function and reduce function, which can be applied now on k-means algorithm. Now the map function would be here for each data point, we have to we will identify, this particular centroid which is closest to this a point XI and let us assume that ZI is the closest centroid, which is being identified, for a data point X I and the map function, will perform this operation, in a data parallel operation, as a data parallel operation. And it will emit this key value pair. So, key will be key will be the, the centroid or you can also call it as a label and the value will be that data point, data point is in the data set or the data sample. Now as far as the step number 2 of the first iteration which is called,' Computing', the recenter, which is nothing but identifying, the mean of that particular cluster, mean of the cluster points. So, mean of a cluster point, will become the new centroid, which is nothing but we call it as recenter. Now as far as the reduced function, for this operation, will be

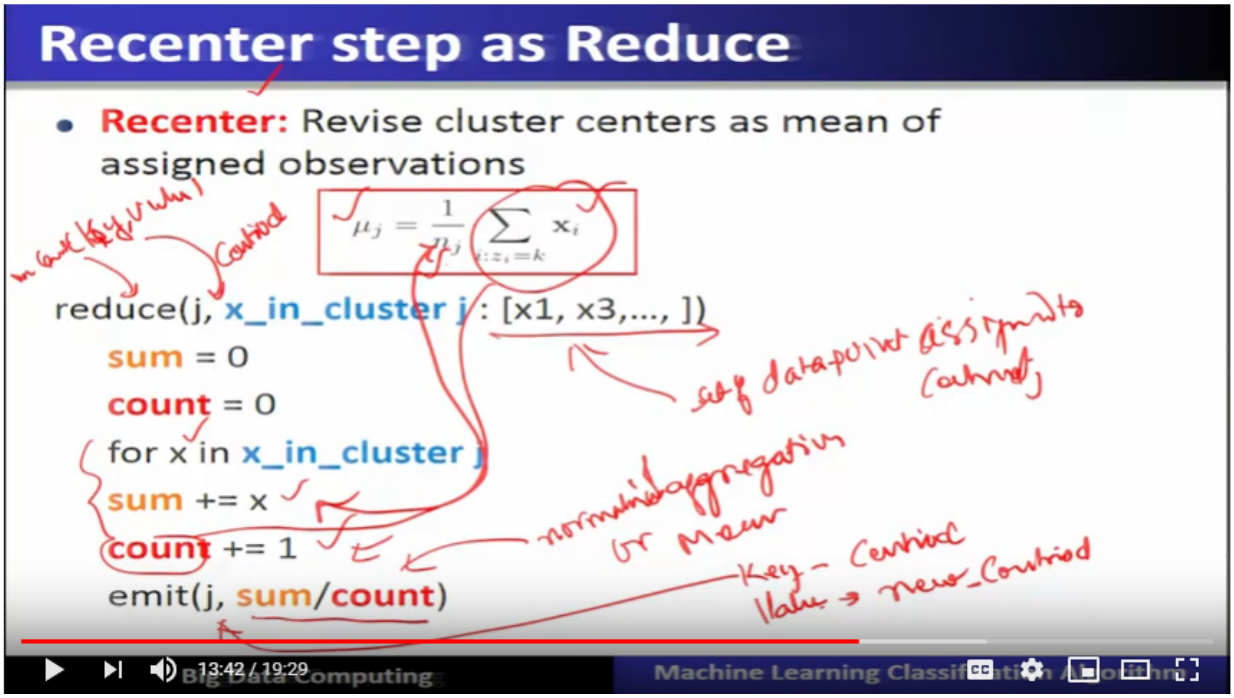
performed, by taking this key value pair, from the previous MapReduce and for that particular that means it will group by that particular centroid, all data points and perform the average. So, average of all data points, for that but your cluster, which is Zl.

Refer slide time :(07:24)



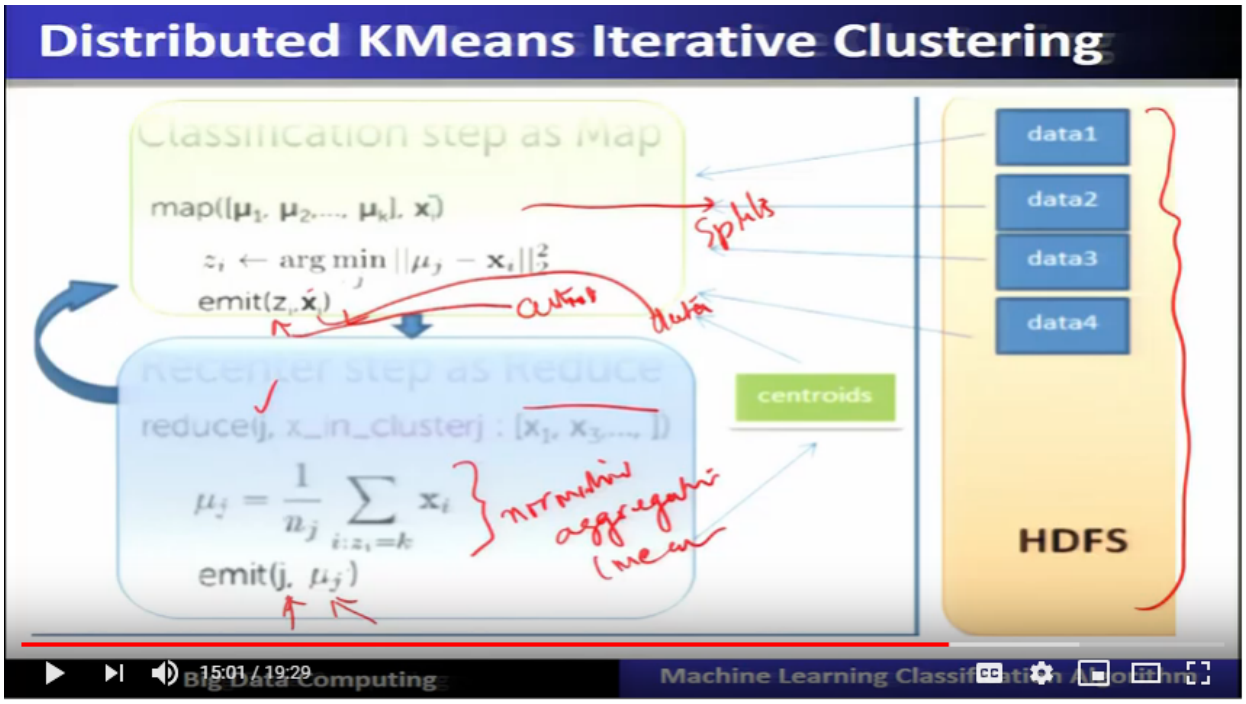
Now let us see in more details, how this map and reduce function is being implemented. So, the first step is called classification or a classifying, classify the data points, data points in the data set, to the closest, centroid in a data parallel operation. So, in this particular equation, you can see that if you are given, the set of centroids. So, for a given data point, it will calculate, the distance measure, the differences and the minimum of that particular difference, will be that data will be that centroid, which is very closest to the data point, compared to all other one and that will become, as the label or the centroid, which is closest, to the data point. So, this way if we perform this activity for, all data sets in data parallel operation, then we can specify using my function. So, map function will have the input, as the set of all data all, all centroids. So, this is the set of all centralized, u1 u2these are all let us seek a different, centroids are given in the map function and X I is the data point. Now this X I will be computing the, the similarity measure or that the distance measure with all the centroids and the minimum, distance centroid ,let us assume it is Z, which will be identified which will be assigned to Z I and the map function, will omit that Zl and this X I. So, the key which this map function will emit is the centroid ID, which is the closest to this data point and the value will be that data point.

Refer slide time :(10:18)



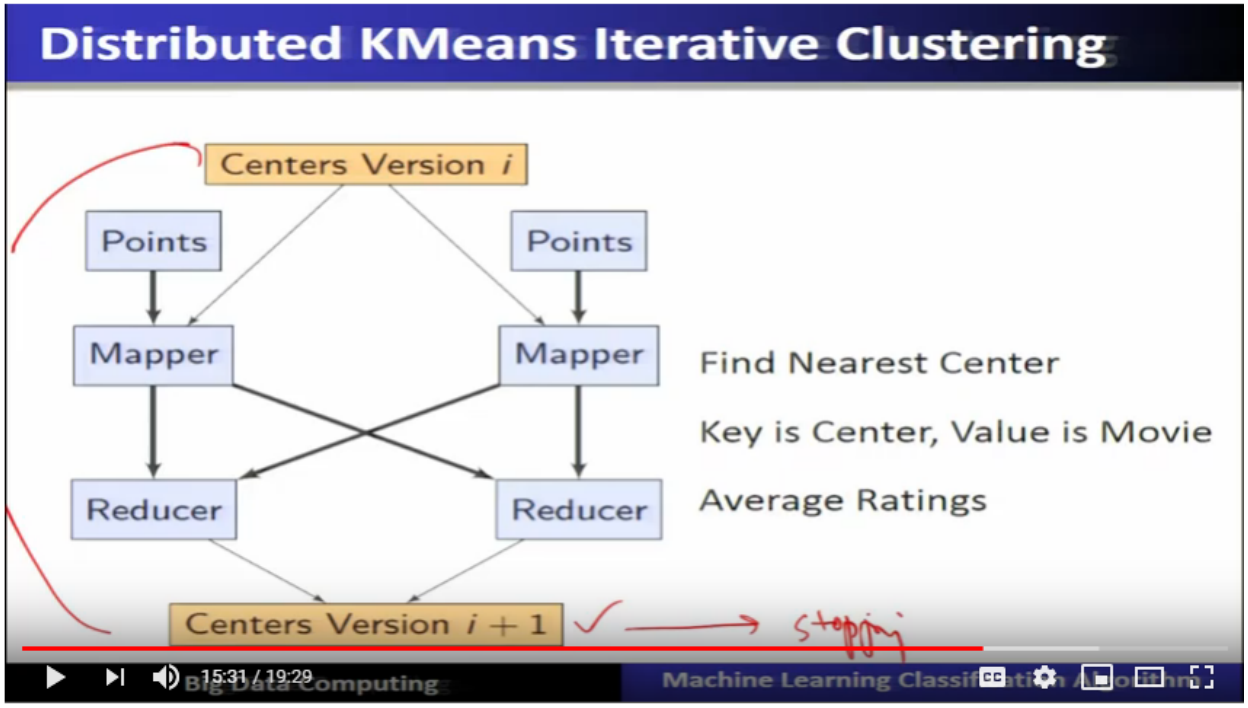
Now let us see the reduce function, which is nothing but computing the recenter. Now after doing the mean or it will identify the mean of all the data points, within a particular, data within a particular cluster and after doing the normalization, summation and the normalization it will identify the new centroid. So, for all the cluster ID all the clusters, there will be are center or computing the, the summation, of all the data points and normalizing, it that is nothing but the calculating the mean for all things. Let us see how it will be done, performed using a reduced function. Now as we have seen from the map function that there will be key value pair which will be emit, from the map. So, map will emit the key value pair. So, key will be that centroid ID, let us say this Is the centroid and this value, will be the set of all set of data points, which are assigned, to the centroid J. so, what it will do is it will compute, it will find out the total it will find out the summation, of all data points X, within it within the clusters and then it will do the normalization. So, here for that there will be a for loop. So, it will iterate through all the all the data points, in that particular set of that cluster and it will do this particular summation that means it will do the summation, of it this particular equation it will operate, in this way and then it will also do a counting and finally the count value, will contain the total number of points which are doing which are which are being summed up for normalization. So, now it will do this mean or the normalization, normalized aggregation is being done, normalized aggregation or it will be computed the mean for every cluster. And it will emit this is the key, key means it will emit the centroid and value will be the normalized value, which is a new centroid. So, after calculating the new centroid, now again it will go to the next iteration and this iteration, will repeat. So, we have seen that how the map and reduce has been implementing this k-means algorithm.

Refer slide time :(13:46)



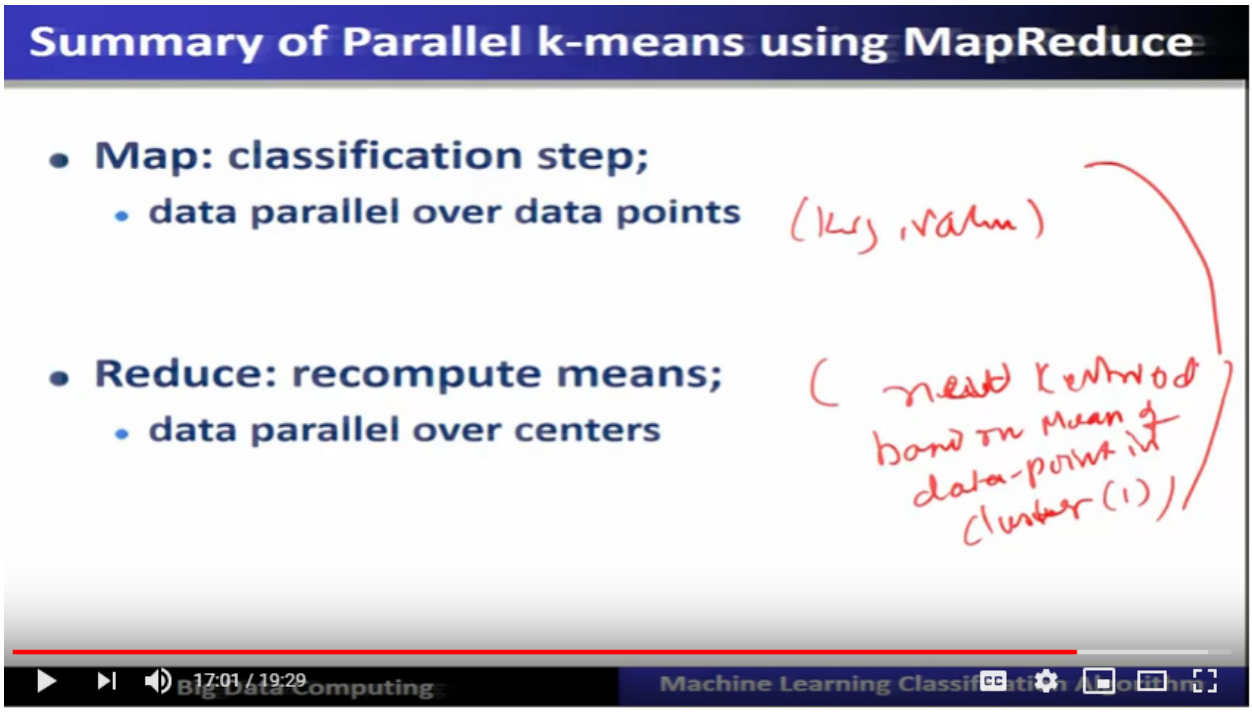
Let us see the same thing in the form of a picture. So, you can see that this is the data set and this particular map function, will be performed on different data plates. So, map function will be applied, on a different data splits and this will map will emit the key value pair. So, key will be the centroid and for all the values, of this data points, it will emit reduce function, what it will do it will now group by key. And it will collect all such data points which belong to a particular cluster. And it will form the normalized, aggregation that is nothing but it will calculate the mean of it. So, for that particular centroid J it will compute the mean that is nothing but the new centroid and this particular, way this k-means algorithm will iterate to the next step until it converges.

Refer slide time :(15:06)



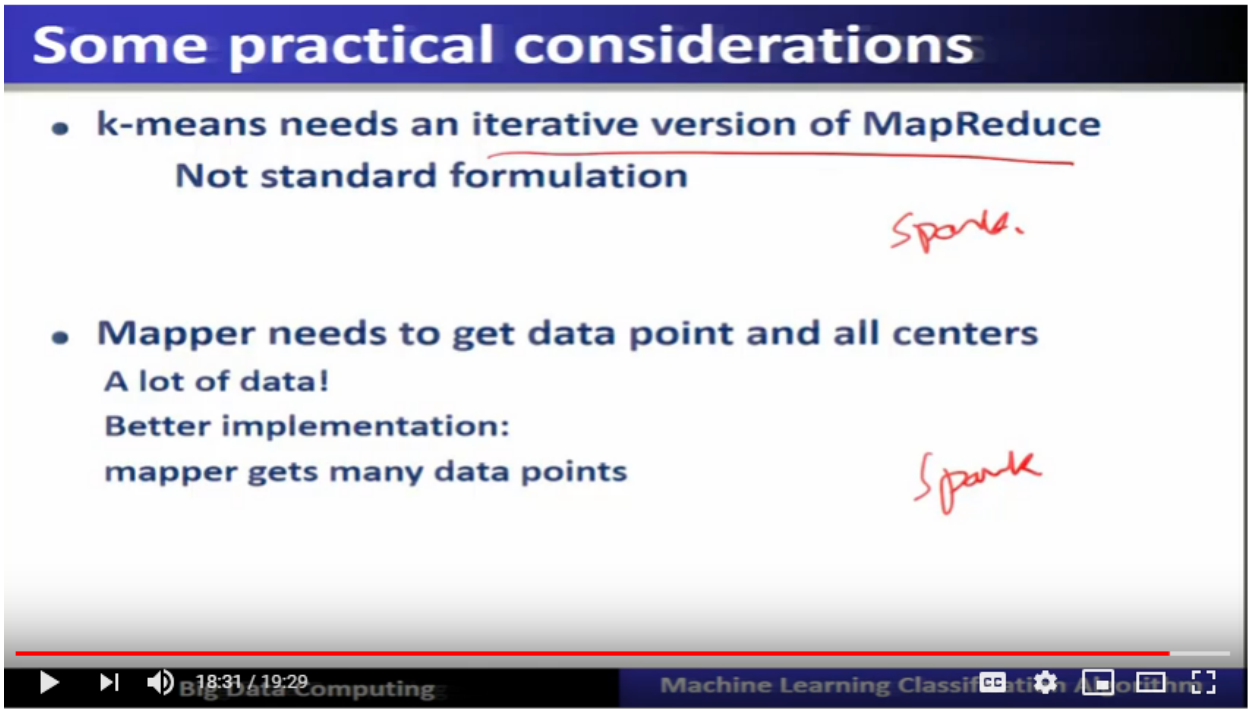
So, this particular new data centers, centers once it is being identified, then again it will go and classify, all the data set accordingly. And so on this particular iterations keeps on iterating unless, there is a stopping criteria and once the stopping criteria, is met then these iterations will stop and the algorithm.

Refer slide time :(15:36)



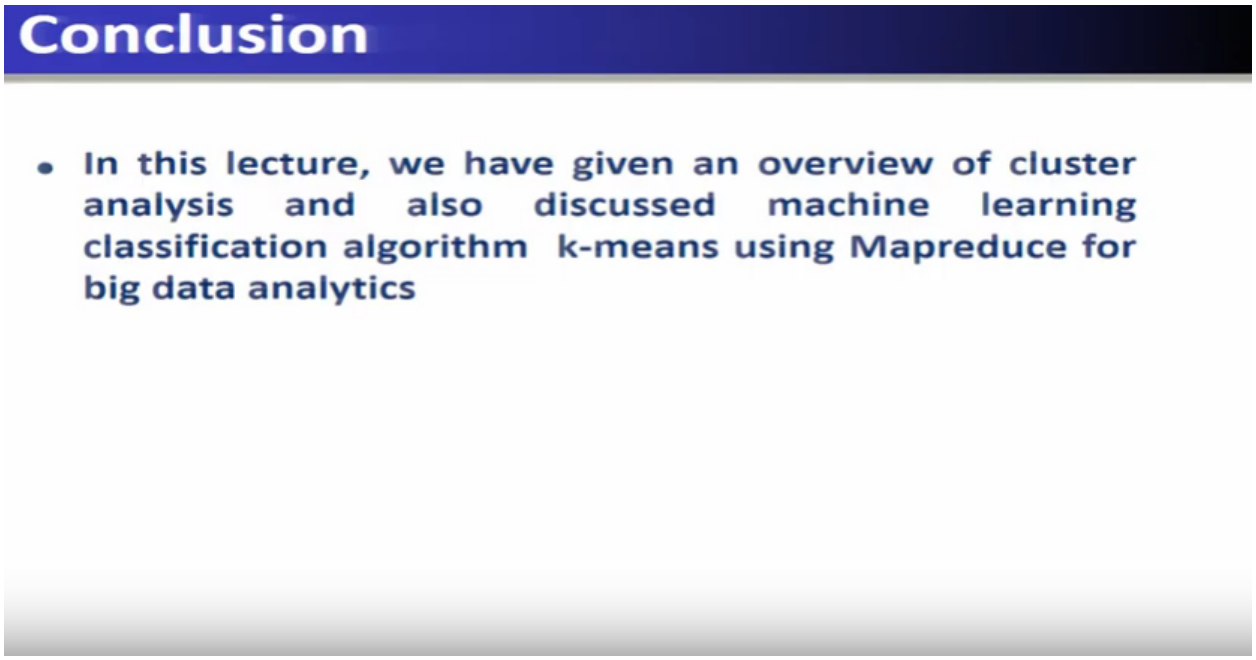
Will terminate, in summary we have seen, the parallel k-means algorithm, which is nothing but an implementation of MapReduce, for k-means algorithm. So, we have also seen that the two steps that is the using map and reduce. We have implemented, we have seen the implementation using map function for the classification step wherein the data parallel operation, is performed or different data points, to classify the data points, into the key value pair. Now second operation which is called a,' Reduce Operation', which is nothing but to come to recomputed, the means are the new centroids, the new centroids are computed based on, the mean of data points, in the clusters, which is identified in the step number one. So, after identifying it then it will go to the next iteration and so on. So, these iteration these two operations, that is the classification and recomputation, we have seen how using map and reduce is being implemented.

Refer slide time :(17:05)



Now we see that this k-means algorithm is an iterative version is an iterative algorithm and which requires an iterative version of a MapReduce. And in the, the previous one chance of Hadoop and MapReduce version 1, we have we have there were several new implementation, of iterative MapReduce. Now with the new version of Hadoop, which internally supports the iterative MapReduce, in his park, this particular iteration is not going to be a problematic and is going to be very straight forward that we will see in further lectures, more detail that this is not that we can use it iterative version of MapReduce, using some other means such as is Park. Now another practical consideration is about the mapper, which needs to get the data points and all the data and all the center's, lot of data is going to be moved, in the old or in the previous version of MapReduce, but now in a newer implementation like in spark and all. So, in memory configuration is being used, to store the data and over the iterations, all these are implemented and handled for better, optimization and efficient implementation.

Refer slide time :(18:36)



Conclusion in this lecture: We have given an overview of the parallel k-means algorithm as an implementation of Map Reduce. And which is one of the very important, unsupervised machine learning, for different big data analytics, application. Thank you.