centroids weighted by local cluster sizes. These new global centroids are, then again, broadcasted back to the slaves for the next iteration of K-means. In this manner, the process continues until the centroids converge. In this implementation, the data is not written to the disk but the primary bottleneck lies in the communication when MPI is used with peer-to-peer networks since aggregation is costly and the network performance will be low.

6.3 K-means on GPU

The pseudo code for k-means clustering algorithm on MPI is given below:

K-means: GPU
Input: Data points D, number of clusters k
Step 1: Do until global centroids converge
Step 2: Upload data points to each multiprocessor and centroids to the shared memory
Step 3: Multiprocessor works with one data vector at a time and associate it with the closest centroid.
Step 4: Centroid recalculation is done on CPU.

In case of K-means, each processor is given a small task i.e. assigning a data point to a centroid. Centroid recalculation is done on the CPU as a single core of GPU is not powerful. Therefore, centroids are uploaded to the shared memory of the GPU, and the data points are classified and uploaded into each multiprocessor. These multiprocessors work on one data vector at a time and associate it with the closest centroid. Once all the points are assigned to the centroids, CPU recalculates the centroids and again will upload the new centroids to the multiprocessors. This process is repeated until the centroids converge. Another aspect to consider here is the density of the data. If the data is sparse, many multiprocessors will stop due to scarcity of data vectors to compute, which will eventually degrade the performance. In a nutshell, the performance of GPUs will be the best when the data is relatively denser and when the algorithm is carefully modified to take advantage of processing cores [1].

7. Comparison of Different Platforms

Features	*MapReduce*	*MPI*	*GPU*
Fault Tolerance	Have efficient in-built fault tolerant mechanisms	Don't have any fault tolerance mechanisms	Have fault tolerance mechanisms
Iterative Processing	Not suitable	Suitable	Highly suitable
Scalability	Highly scalable	Highly scalable	Less scalable
Data Size	Suitable for processing large data sets	Suitable for processing large data sets	Not suitable for processing large data sets
Limitation	Disk access is a major limitation which significantly degrades the performance.	Primary limitation lies in the communication when MPI is used with peer-to-peer networks since aggregation is costly.	The primary limitation is the limited memory because when the data size is more than the size of the GPU memory, the performance decreases.

8. Conclusion

This paper describes big data and 4v's (challenges of big data). Nowadays, the volume of information exchange involves a huge amount of data processing. So through this paper we try to focus on implementation of k-means algorithm on different big data analytics platforms for efficient data processing. K-means algorithm was chosen because of its iterative nature. The future of big data analytic involves implementing various algorithms like nearest neighbor, decision tree, page rank, etc. on different platforms. One can decide to choose the particular platform for specific application based upon its features and limitations. Combination of platforms can be used for better performance.

9. Acknowledgement

We would like to thank our guide, Prof. Sudhakar Jadhav for his guidance and support, which has helped us, complete this research paper successfully.

References

[1] Dilpreet Singh and Chandan K Reddy(A survey on platforms for big data analytics)
[2] http://www.techopedia.com/definition/13816/mapreduce
[3] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
[4] http://www.webopedia.com/TERM/B/big_data.html
[5] http://www.slideshare.net/nagaritwikindugu/big-data-analyticsintrohadoop-map-reducemahoutkmeans-clusteringhbase
[6] http://www.techopedia.com/2/30575/trends/big-data/big-datas-key-challenges
[7] http://www.sas.com/en_us/insights/analytics/big-data-analytics.html
[8] http://www.slideshare.net/nasrinhussain1/big-data-ppt-31616290
[9] https://en.wikipedia.org/wiki/Message_Passing_Interface
[10] https://www.hpcvl.org/faqs/programming/mpi-message-passing-interface
[11] http://www.onmyphd.com/?p=k-means.clustering
[12] https://en.wikipedia.org/wiki/Graphics_processing_unit
[13] http://www.fuzzyl.com/products/gpu-analytics/
[14] http://www.datacenterknowledge.com/archives/2012/09/12/a-look-into-the-big-data-battleground-analyzing-the-market/
[15] http://bigdata-tuts.blogspot.in/
[16] https://en.wikipedia.org/wiki/Big_data