Scalable Clustering Algorithm based on Apache Spark Framework for Handling Big Data

Neha Bharill, Dr. Aruna Tiwari
Computer Science and Engineering
Indian Institute of Science and Technology Indore

Objective: To efficiently analyze the enormous amount of data collected every day from various emerging technologies. We proposed novel clustering algorithms called Scalable Random Sampling with Iterative Optimization Fuzzy C-Means (SRSIO-FCM) and Scalable Literal Fuzzy C-Means (SLFCM) implemented on Apache Spark Big Data processing Framework for effective clustering of Big Data.

Proposed Scalable Random Sampling with Iterative Optimization Fuzzy C-Means (SRSIO-FCM)

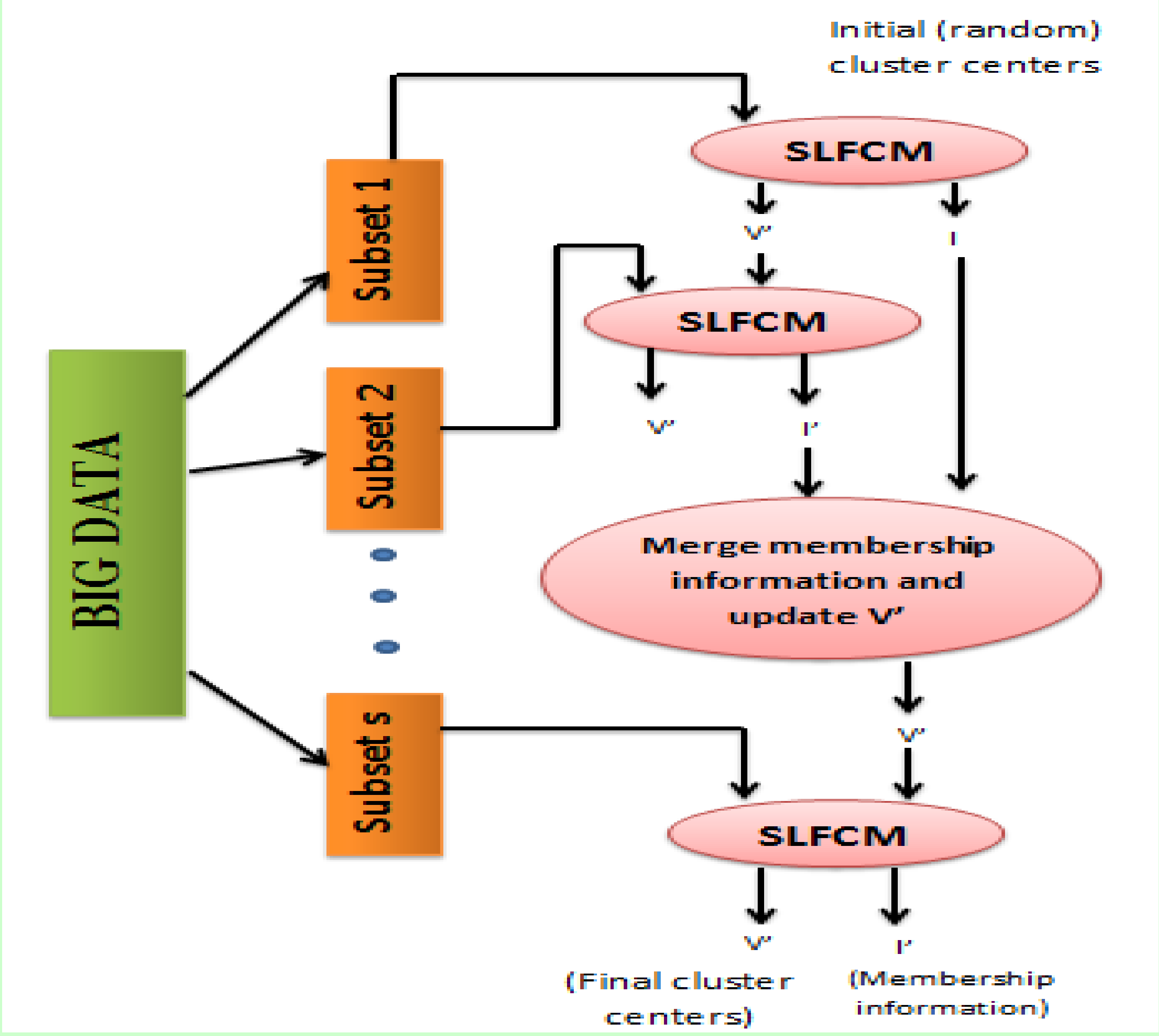

Fig. 1 Workflow of SRSIO-FCM

Working of SLFCM on Apache Spark

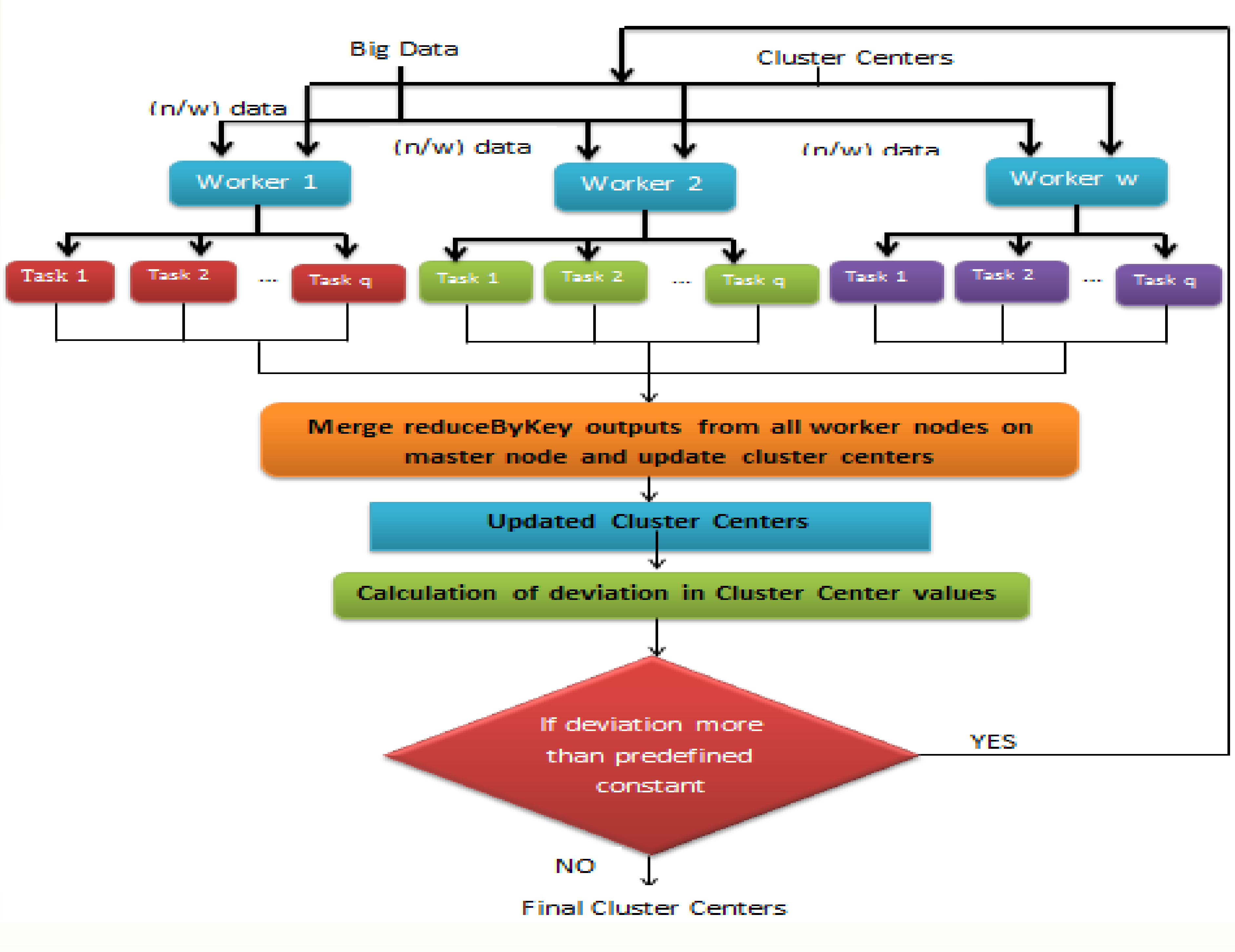

Fig. 3 Flow chart showing working of SLFCM on Apache Spark

Proposed Scalable Literal Fuzzy C-Means

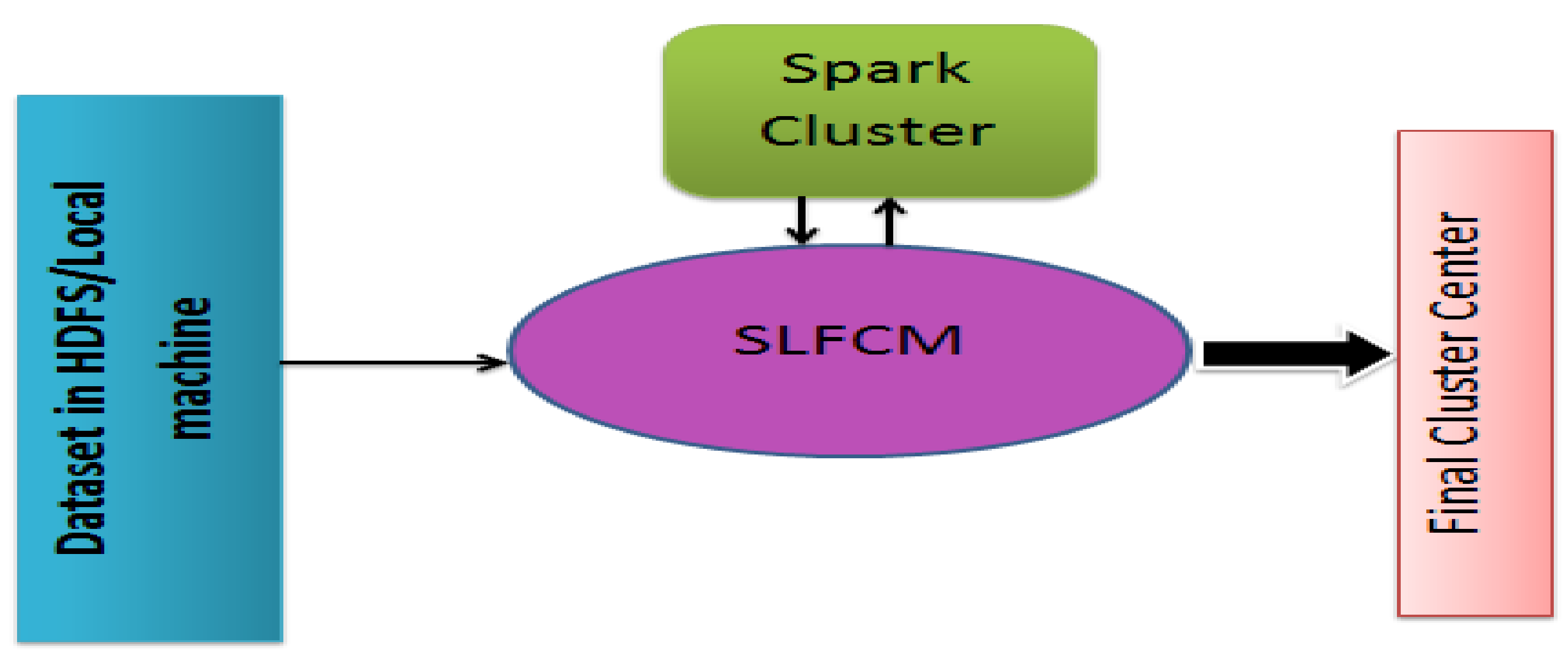

Fig. 2 Workflow of SLFCM

Experimental Results

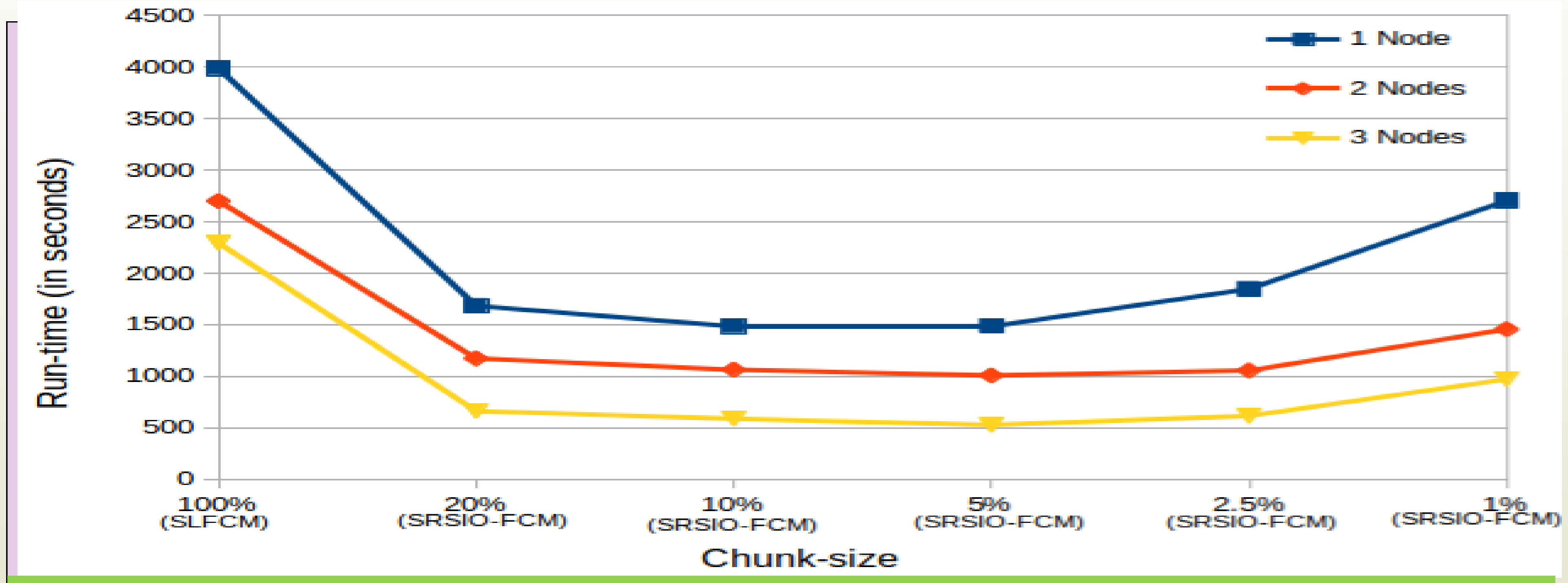

Fig. 4 Run-time comparison on MINST8m dataset

Conclusion

1) The proposed novel clustering algorithms SRSIO-FCM and SLFCM is designed to deal with the challenges associated with fuzzy clustering for handling Big Data.

2) Experimental results evaluated in terms of speed-up, size-up and scale-up in comparison of SLFCM shows significant reduction in run-time.

References

1)J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh,and A. H. Byers, “Big data: The next frontier for innovation,competition, and productivity,” pp. 1–137, 2011

2) Y. Wang, L. Chen, and J.-P. Mei, “Incremental fuzzy clustering with multiple medoids for large data,” *IEEE Transactions on FuzzySystems, vol. 22, no. 6, pp. 1557–1568, 2014.*

Publications

1) Neha Bharill, Aruna Tiwari, Aayushi Malviya, “Fuzzy Based Clustering Algorithms to Handle Big Data with Implementation on Apache Spark” IEEE BigDataService, IEEE Computer Society Conference, Exeter College, Oxford, UK, March 29- April 1, 2016., pp. 95-104..

2) Neha Bharill, Aruna Tiwari, “ Enhanced Cluster Validity Index for the Evaluation of Optimal Number of Clusters for Fuzzy c-Means Algorithm”, in Proceeding of IEEE International Conference on Fuzzy Systems, IEEE World Congress on Computational Intelligence- WCCI, Beijing International Convention Center, China, July 6-11,2014, pp. 1526-1523.