that node. Even if load balance is theoretically impossible, our model and algorithm could minimize the load difference between nodes and achieve basically the load balance.

The optimized model further reduced the overall storage pressure. According to the variation in the node storage observed in each experiment, the model optimization greatly lowered the node storage required to save the distributed files.

## 7. CONCLUSIONS

This paper probes into the all-to-all comparison of large dataset, and gives a formal mathematical description of the problem. Then, a multi-objective file distribution model was constructed based on the LP, aiming to localize the data, balance node storage and loads, minimize the storage occupation, and control the occupied storage within the storage limit of each node. To save storage space, the established model was further optimized, and the file distribution algorithm was designed for the distributed environment. Finally, our model and algorithm were proved valid through several experiments.

## ACKNOWLEDGMENT

The work is funded in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61462070, the Doctoral research fund project of Inner Mongolia Agricultural University under Grant No. BJ09-44 and the Inner Mongolia Autonomous Region Key Laboratory of big data research and application for agriculture and animal husbandry.

## REFERENCES

[1] Zhang, Y.F., Tian, Y.C., Kelly, W., Fidge, C., Gao, J. (2015). Application of simulated annealing to data distribution for all-to-all comparison problems in homogeneous systems. International Conference on Neural Information Processing, Springer, Cham, 683-691. https://doi.org/10.1007/978-3-319-26555-1_77

[2] Baert, Q., Caron, A.C., Morge, M., Routier, J.C. (2018). Fair task allocation for large data sets analysis. Revue d'Intelligence Artificielle, 31(4): 401-426. https://doi.org/10.3166/RIA.31.401-426

[3] Zhang, Y.F., Tian, Y.C., Kelly, W., Fidge, C. (2014). A distributed computing framework for all to all comparison problems. Proceedings of IECON'14. Washington D. C, USA: IEEE Press, 2499-2505. https://doi.org/10.1109/IECON.2014.7048857

[4] Li, L.X., Gao, J., Mu, R. (2019). Optimal data file allocation for all-to-all comparison in distributed system: A case study on genetic sequence comparison. International Journal of Computers, Communications & Control, 14(2): 199-211. https://doi.org/10.15837/ijccc.2019.2.3526

[5] Shen, X., Choudhary, A. (2003). A distributed multi-storage resource architecture and I/O performance prediction for scientific computing. Cluster Computing, 6(3): 189-200. https://doi.org/10.1109/HPDC.2000.868631

[6] Song, A., Zhao, M., Xue, Y., Luo, J. (2016). MHDFS: A memory-based hadoop framework for large data storage. Scientific Programming, 2016: 1808396. http://dx.doi.org/10.1155/2016/1808396

[7] Chen, F., Liu, J., Zhu, Y. (2017). A real-time scheduling strategy based on processing framework of hadoop. 2017 IEEE International Congress on Big Data (BigData Congress), 2017: 321-328. https://doi.org/10.1109/BigDataCongress.2017.48

[8] Lin, W. (2012). An improved data placement strategy for Hadoop. Journal of South China University of Technology (Natural Science Edition), 40(1): 152-158.

[9] Li, X., Zhang, H., Hu, Q., Huang, X. (2017). Research on power customer segmentation based on big data of intelligent city. 2017 29th Chinese Control and Decision Conference (CCDC), 3207-3211. http://dx.doi.org/10.1109/CCDC.2017.7979059

[10] Ghemawat, S., Gobioff, H., Leung, S.T. (2003). The google file system.

[11] Lin, X., Lin, P., Huang, P., Chen, L., Fan, Z., Huang, P. (2015). Modeling the task of google mapreduce workload. 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 1229-1232. https://doi.org/10.1109/CCGrid.2015.104

[12] Guo, Y., Rao, J., Cheng, D., Zhou, X. (2016). iShuffle: Improving hadoop performance with shuffle-on-write. IEEE Transactions on Parallel and Distributed Systems, 28(6): 1649-1662. https://doi.org/10.1109/TPDS.2016.2587645

[13] Jiao, X., Mu, J., He, Y.C., Chen, C. (2017). Efficient ADMM decoding of LDPC codes using lookup tables. IEEE Transactions on Communications, 65(4): 1425-1437. https://doi.org/10.1109/TCOMM.2017.2659733

[14] Qin, Z., Liu, X., Cao, B. (2016). Multi-level linear programming subject to max-product fuzzy relation equalities. International workshop on Mathematics and Decision Science, Springer, Cham, 220-226. https://doi.org/10.1007/978-3-319-66514-6_23

[15] Sinha, S.B., Sinha, S. (2004). A linear programming approach for linear multi-level programming problems. Journal of the Operational Research Society, 55(3): 312-316. https://doi.org/10.1057/palgrave.jors.2601701

[16] El-Bakry, M. (2010). Using linear programming models for minimizing harmonics values in cascaded multilevel inverters. 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 696-702. https://doi.org/10.1109/AIM.2010.5695713