SUMMARY OF DIFFERENCE

RUI REN, CHUNGHSUAN WU, ZHOUWANG FU, TAO SONG, YANQIANG LIU,
ZHENGWEI QI, AND HAIBING GUAN

This paper is the extension of our conference paper at PPoPP '18 [1]. In this paper, we provide a new data processing solution for the industrial big data based on our conference paper. Besides, we use the credible benchmark to demonstrate that our solution is effective in the big data analysis. This paper makes the following distinct contributions compared with the conference paper.

(1) We propose a new performance model called *Framework Resources Quantification* (FRQ) model. The FRQ model quantifies computing and I/O resources by five input parameters. Furthermore, the FRQ model visualizes the resource scheduling strategies of the DAG frameworks in the time dimension. In this paper, we use the model to evaluate the deficiencies of different resources scheduling strategies and then optimize them.
(2) In the conference paper, we only implement SCache on Spark. In this paper, we also implement SCache on Hadoop MapReduce, since Spark and Hadoop MapReduce are the two most distributed computing frameworks using in industrial big data analysis. This result also demonstrates the compatibility and adaptability of SCache as a cross-framework plug-in.
(3) We append two parts to the evaluation section. Firstly, we evaluate the FRQ model in both our in-house environment and Amazon EC2 environment. The error between the FRQ's calculated value and the experimental value is mainly below 10%. Secondly, we evaluate the performance of Hadoop MapReduce with SCache. We use the same Amazon EC2 environment as the conference paper (50-nodes m4.xlarge cluster). According to the experiments, Hadoop MapReduce with SCache optimizes job completion time by up to 15% and an average of 13%.

In summary, we propose a new performance model to provide a theoretical basis for performance optimization and improve evaluations by adding experiments on Hadoop MapReduce. We believe that the added content makes a sufficient contribution to this journal submission.

References

[1] Z. Fu, T. Song, Z. Qi, and H. Guan, "Efficient shuffle management with scache for dag computing frameworks," in *Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. ACM, 2018, pp. 305–316.