ASC: Improving Spark Driver Performance with SPARK Automatic Checkpoint

*School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China
Zw19880717@sjtu.edu.cn, chen-hp @sjtu.edu.cn ,hu.fei @sjtu.edu.cn

Abstract*— Many great big data processing platforms, for example Hadoop Map Reduce, are keeping improving large-scale data processing performance which make big data processing focus of IT industry. Among them Spark has become increasingly popular big data processing framework since it was presented in 2010 first time. Spark use RDD for its data abstraction, targeting at the multiple iteration large-scale data processing with reuse of data, the in-memory feature of RDD make spark faster than many other non-in-memory big data processing platform. However in-memory feature also bring the volatile problem, a failure or a missing RDD will cause Spark to recompute all the missing RDD on the lineage. And a long lineage will also increasing the time cost and memory usage of Driver analyzing the lineage. A checkpoint will cut off the lineage and save the data which is required in the coming computing, the frequency to make a checkpoint and the RDDs which are selected to save will significantly influence the performance. In this paper, we are presenting an automatic checkpoint algorithm on Spark to help solve the long lineage problem with less influence on the performance. The automatic checkpoint will select the necessary RDD to save and bring an acceptable overhead and improve the time performance for multiple iteration.

Key words: Spark, automatic checkpoint, lineage, distributed computing, big data.

Wei Zhu. He received his Bachlor degree of Computer science and technology, Chongqing University in 2011 Chongqing China. And now his is working for his Master degree of software engineering in Shanghai Jiao Tong University, Shanghai, China.
He is interested in fields of distributed system and big data processing.

Haopeng Chen. He received his Ph.D degree from Department of Computer Science and Engineering, Northwestern Polytechinal University, Xi'an, Shanxi Province, China in 2001. He has worked in School of Software, Shanghai Jiao Tong University since 2004 after he finished his two-year postdoctoral research job in Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He got the position of Associate Professor in 2008. In 2010, he studied and researched in Georgia Institute of Technology as a visiting scholar. His research group focuses on Distributed Computing and Software Engineering. They have kept researching on Web Services, Web 2.0, Java EE, .NET, and SOA for several years. Recently, they are also interested in cloud computing and researching on the relevant areas, such as cloud federation, resource management, dynamic scaling up and down, and so on.

Fei Hu. He recieved his Bachelor degree from Department of computer software, Northwest University, Xi'an, Shanxi Province, China in 1990 and received his Master degree of computer science and engineering and Ph.D of Precision Guidance and Control both from Northwest Polytechnical University, Xi'an, Shanxi Province, China in 1993 and 1998.
He has worked in Department of Computer Science and Engineering , Northwestern Polytechnical University lecturer , from 1993 to 2006. From 2006/ 9 to now he has worked in School of Software, Shanghai Jiao Tong University. His Publications are as follows: Zhiyang Zhang, Fei Hu and Jian Li, "Autonomous Flight Control System Designed for Small-Scale Helicopter Based on Approximate Dynamic Inversion," The 3rd IEEE International Conference on Advanced Computer Control (ICACC 2011), 18th to 20th January 2011, Harbin, China.