Announcing the Final Examination of Xuhong Zhang for the degree of Doctor of Philosophy

Time & Location: October 27, 2017 at 10:30 AM in HEC 438
Title: Reducing the Overhead of Memory Space, Network Communication and Disk I/O for Analytic Frameworks in Big Data Ecosystem

To facilitate big data processing, many distributed analytic frameworks and storage systems such as Apache Hadoop, Apache Hama, Apache Spark and Hadoop Distributed File System (HDFS) have been developed. Currently, many researchers are conducting research to either make them more scalable or enable them to support more analysis applications. In my PhD study, I conducted three main works in this topic, which are minimizing the communication delay in Apache Hama, minimizing the memory space and computational overhead in HDFS and minimizing the disk I/O overhead for approximation applications in Hadoop ecosystem. Specifically, in Apache Hama, communication delay makes up a large percentage of the overall graph processing time. While most recent research has focused on reducing the number of network messages, we add a runtime communication and computation scheduler to overlap them as much as possible. As a result, communication delay can be mitigated. In HDFS, the block location table and its corresponding maintenance could occupy more than half of the memory space and 30% of processing capacity in master node, which severely limits the scalability and performance of master node. We propose Deister that uses deterministic mathematical calculations to eliminate the huge block location table and its corresponding maintenance. My third work proposes to enable both efficient and accurate approximations on arbitrary sub-datasets of a large dataset. Existing offline sampling based approximation systems are not adaptive to dynamic query workloads and online sampling based approximation systems suffer from low I/O efficiency and poor estimation accuracy. Therefore, we develop a distribution aware method called Sapprox. Our idea is to collect the occurrences of a sub-dataset at each logical partition of a dataset (storage distribution) in the distributed system at a very small cost, and make good use of such information to facilitate online sampling.

Major: Computer Engineering

Educational Career:
Bachelor’s of Software Engineering, BS, 2011, Harbin Institute of Technology
Master's of Computer Science, MS, 2013, Georgia State University

Committee in Charge:
Jun Wang, Chair, Electrical and Computer Engineering
Mingjie Lin, Department of Electrical and Computer Engineering
Shaojie Zhang, Department of Computer Science
Deliang Fan, Department of Electrical and Computer Engineering
Fei Liu, Department of Computer Science

Approved for distribution by Jun Wang, Committee Chair, on October 5, 2017.

The public is welcome to attend.