The amount of generated and stored data has been growing rapidly. It is estimated that 2.5 quintillion bytes of data are generated every day, and 90% of the data in the world today has been created in the last two years. How to solve these big data issues has become a hot topic in both industry and academia.

Due to the complexity of big data platforms, we stratify it into four layers: storage layer, resource management layer, computing layer, and methodology layer. This dissertation proposes novel approaches to address the performance issues of big data platforms like Hadoop and Spark in all the four layers.

We first present an improved HDFS design called SMARTH in the storage layer. It utilizes asynchronous multi-pipeline data transfers instead of a single pipeline stop-and-wait mechanism. Secondly, we propose an optimized Hadoop extension called MRapid for the resource management layer, which significantly speeds up the execution of short jobs. It is completely backward-compatible to Hadoop and imposes negligible overhead. Thirdly, we introduce an efficient 3-level performance model, called Hedgehog, and focus on improving resource utilization and performance for the computing layer. This design is the first white-box performance model for Spark. Fourthly, we optimize the current implementation of SGD in Spark’s MLlib for methodology layer by reusing data partition for multiple times within a single iteration to find better candidate weights in a more efficient way. At last, we present a scalable and distributed geographic information system for methodology layer, called Dart, based on Hadoop and HBase. Dart provides a hybrid table schema to store spatial data in HBase so that the Reduce process can be omitted for operations like calculating the mean center and the median center. It employs reasonable pre-splitting and hash techniques to avoid data imbalance and hot region problems. It also supports massive spatial data analysis like K Nearest Neighbors (KNN) and Geometric Median Distribution.

Major: Computer Science

Educational Career:
Bachelor’s of Mathematics and Applied Mathematics, BS, 2006, Hebei University
Master’s of Computer Science, MS, 2015, University of Wyoming

Committee in Charge:
Liqiang Wang, Chair, Computer Science
Damla Turgut, Computer Science
Jun Wang, Electrical and Computer Engineering
Shunpu Zhang, Statistics

Approved for distribution by Liqiang Wang, Committee Chair, on June 19, 2018.

The public is welcome to attend.