TEMPLE UNIVERSITY®

Fall 2019 Colloquium

Department of Computer and Information Sciences

Enabling High-performance Sampling for Big Data Processing

Dr. Jun Wang

Professor of Computer Engineering

Director of the Computer Architecture and Storage Systems (CASS) Laboratory

University of Central Florida, Orlando, FL, USA

Wednesday, October 22nd, 11 AM, SERC 306

Abstract: In this talk, we aim to demonstrate how to perform sampling in today's big data processing platforms. We enable both efficient and accurate approximations on arbitrary sub-datasets of a large dataset. Due to the prohibitive storage overhead of caching offline samples for each sub-dataset, existing offline sample based systems provide high accuracy results for only a limited number of sub-datasets, such as the popular ones. On the other hand, current online sample based approximation systems, which generate samples at runtime, do not take into account the uneven storage distribution of a sub-dataset. They work well for uniform distribution of a sub-dataset while suffer low sampling efficiency and poor estimation accuracy on unevenly distributed sub-datasets.

To address the problem, 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 and make good use of such information to facilitate online sampling. We have implemented Sapprox into Hadoop ecosystem as an example system and open sourced it on GitHub. Our comprehensive experimental results show that Sapprox can achieve a speedup by up to a factor of 20 over the precise execution.

Bio: Dr. Jun Wang is a Professor of Computer Engineering; and Director of the Computer Architecture and Storage Systems (CASS) Laboratory at the University of Central Florida, Orlando, FL, USA. He has conducted extensive research in the areas of Computer Systems and Data-Intensive Computing. His specific research interests include massive storage and file Systems in a local, distributed and parallel systems environment. Dr. Wang is the recipient of the National Science Foundation Early Career Award 2009 and Department of Energy Early Career Principal Investigator Award 2005. He has authored over 120 publications in premier journals such as IEEE Transactions on Computers, IEEE Transactions on Parallel and Distributed Systems, and leading HPC and systems conferences such as VLDB, HPDC, EuroSys, IPDPS, ICS, Middleware, FAST. He has graduated 13 Ph.D. students who upon their graduations were employed by major US IT corporations (e.g., Google, Microsoft, etc). He has been serving on the editorial board for the IEEE transactions on parallel and distributed systems, and IEEE transactions on cloud computing.