Arachne: An Open-Source Framework for Interactive Massive-Scale Graph Analytics

David A. Bader
Distinguished Professor
Director, Institute for Data Science
New Jersey Institute of Technology
https://davidbader.net/

Friday, January 16 at 3pm – IRB 4105

Abstract

A real-world challenge in data science is to develop interactive methods for quickly analyzing new and novel data sets that are potentially of massive scale. In this talk, Bader will discuss his development of graph algorithms in the context of Arkouda, an open-source NumPy-like replacement for interactive data science on tens of terabytes of data. Massive-scale analytics is an emerging field that integrates the power of high-performance computing and mathematical modeling to extract key insights and information from large-scale data sets. Productivity in massive-scale analytics entails quick interpretation of results through easy-to-use frameworks, while also adhering to design principles that combine high-performance computing and user-friendly simplicity. However, data scientists often encounter challenges, especially with graph analytics, which require the analysis of complex data from various domains, such as the cybersecurity, natural and social sciences. To address this issue, we introduce Arachne, an open-source framework that enhances accessibility and usability in massive-scale graph analytics. Arachne offers novel algorithms and implementations of graph kernels for efficient data analysis, such as connected components, breadth-first search, triangle counting, k-truss, among others. The high-performance algorithms are integrated into a back-end server written in HPE/Cray's Chapel language and can be accessed through a Python application programming interface (API).

In this talk, Bader presents an overview of the algorithms his research group has implemented into Arachne and, if applicable, the algorithmic innovations of each. Further, Bader will discuss improvements to our graph data structure to store extra information such as node labels, edge relationships, and node and edge properties. Arachne is built as an extension to the open-source Arkouda framework and allows for graphs to be generated from Arkouda dataframes. The open-source code for Arachne can be found at https://github.com/Bears-R-Us/arkouda-njit. This is joint work with Oliver Alvarado Rodriguez, Zhihui Du, Joseph Patchett, Naren Khatwani, Fuhuan Li, Bader is supported in part by the National Science Foundation award CCF-2109988.

Biography

David A. Bader is a Distinguished Professor and founder of the Department of Data Science and inaugural Director of the Institute for Data Science at New Jersey Institute of Technology. Prior to this, he served as founding Professor and Chair of the School of Computational Science and Engineering, College of Computing, at Georgia Institute of Technology. Dr. Bader is a Fellow of the IEEE, ACM, AAAS, and SIAM; a recipient of the IEEE Sidney Fernbach Award; and the 2022 Innovation Hall of Fame inductee of the University of Maryland's A. James Clark School of Engineering. He advises the White House, most recently on the National Strategic Computing Initiative (NSCI) and Future Advanced Computing Ecosystem (FACE). Bader's interests are at the intersection of high-performance computing and real-world applications, including cybersecurity, massive-scale analytics, and computational genomics, and he has co-authored over 300 scholarly papers and has best paper awards from ISC, IEEE HPEC, and IEEE/ACM SC. Dr. Bader has served as a lead scientist in several DARPA programs including High

Productivity Computing Systems (HPCS) with IBM, Ubiquitous High Performance Computing (UHPC) with NVIDIA, Anomaly Detection at Multiple Scales (ADAMS), Power Efficiency Revolution For Embedded Computing Technologies (PERFECT), Hierarchical Identify Verify Exploit (HIVE), and Software-Defined Hardware (SDH). Recently, Bader received an NVIDIA AI Lab (NVAIL) award, and a Facebook Research AI Hardware/Software Co-Design award. Dr. Bader is Editor-in-Chief of the ACM Transactions on Parallel Computing, and General Co-Chair of IPDPS 2021, and previously served as Editor-in-Chief of the IEEE Transactions on Parallel and Distributed Systems. He serves on the leadership team of Northeast Big Data Innovation Hub as the inaugural chair of the Seed Fund Steering Committee. ROI-NJ recognized Bader as a technology influencer on its 2021 inaugural and 2022 lists. In 2012, Bader was the recipient of University of Maryland's Electrical and Computer Engineering Distinguished Alumni Award. In 2014, Bader received the Outstanding Senior Faculty Research Award from Georgia Tech. Bader has also served as Director of the Sony-Toshiba-IBM Center of Competence for the Cell Broadband Engine Processor and Director of an NVIDIA GPU Center of Excellence. In 1998, Bader built the first Linux supercomputer that led to a high-performance computing (HPC) revolution, and Hyperion Research estimates that the total economic value of Linux supercomputing pioneered by Bader has been over $100 trillion over the past 25 years. The Computer History Museum recogizes Bader for developing the first Linux-based supercomputer which became the predominant architecture for all major supercomputers in the world. Bader is a cofounder of the Graph500 List for benchmarking "Big Data" computing platforms. He is recognized as a "RockStar" of High Performance Computing by InsideHPC and as HPCwire's People to Watch in 2012 and 2014.