NJIT | INSTITUTE FOR DATA SCIENCE

Data Science Seminar Series

Arkouda: Interactive Supercomputing for Data Science

William Reus, Ph.D.
Data Scientist
U.S. Department of Defense

Date: Wednesday, March 24th, 2021
Time: 4:00 PM – 5:00 PM EDT
Location: Zoom Virtual Room
Web Link: https://njit-institute-for-data-science.eventbrite.com

Data science and high-performance computing (HPC) should be a great match: with datasets growing well beyond the memory of a single node and computations becoming ever more communication-intensive, the need for HPC in data science seems clear. And yet, there remains a frustrating gap between the practice of data science and HPC technology. One major reason is that data science is an interactive sport -- data scientists overwhelmingly gravitate towards interactive platforms (e.g. Jupyter notebooks) and interpreted languages (e.g. Python) -- whereas the culture of HPC tends to eschew interactivity in favor of compiled programs and batch jobs. While HPC practitioners prize computational efficiency, data scientists live by the very different maxim of rapid hypothesis testing and have demonstrated that they are willing to ignore HPC technologies entirely rather than give up interactivity. Bridging this gap entails a change in thinking about the purpose of an HPC and how it should be used.

This talk motivates and demonstrates the interactive use of up to hundreds of HPC nodes in exploratory data science workflows that lie upstream of the kinds of graph and machine learning algorithms that HPC codes typically implement. To support this vital but oft-overlooked activity, we developed an open-source package called Arkouda, which exposes massively parallel, distributed NumPy-like arrays and Pandas DataFrame-like functionality to a Jupyter notebook running Python 3. We have chosen the NumPy and Pandas abstractions and Jupyter and Python as front-end technologies in order to conform to interfaces familiar to data scientists. Additionally, because Arkouda arrays can be constructed from and exported to NumPy arrays, users can perform heavy computations on hundreds of HPC nodes and bring back small sets of results for rich introspection in a single-node Python environment, thereby combining the strengths of both environments. Data scientists are currently using Arkouda to explore and transform many terabytes of data in interactive sessions. As the community grows, we envision these same capabilities being used to construct graphs or extract features and pass these data structures, in memory, to specialized HPC codes. In short, we see Arkouda as a potential framework for integrating the burgeoning ecosystem of HPC software with the interactive workflows of data scientists.

Dr. Reus is a physical chemist by training, having earned his Ph.D. from Harvard in the field of molecular electronics. Since graduate school, he has been cross-training in statistics and parallel computing in order to apply his scientific expertise to problems in cyberdefense. Dr. Reus lives near the Chesapeake Bay with his wife and three children.