Interactive Large-Scale Data and Graph Analytics

Oliver Alvarado Rodriguez, Naren Khatwani, Zhihui Du, and David Bader
Department of Data Science
New Jersey Institute of Technology
Newark, NJ, USA
{oaa9,nk88,zd4,bader}@njit.edu

Abstract

There is an ever-growing need for data analytical tools that can handle massive data sets. Arkouda is a Python framework with a Chapel back-end created with the intention to scale NumPy operations at scale for datasets that exceeds tens of terabytes in size. The Python front-end allows for data scientists to utilize the functionality of Arkouda to carry out expensive high-performance computing (HPC) kernels that require the usage of large distributed arrays. Arkouda is not designed with the intention to totally replace libraries like Pandas or NumPy, but rather provide the capability to handle datasets that are massive in size in a highly-scalable environment. The goal is to create an environment that is beneficial for exploratory data and graph analysis (EDA) while staying simple enough for all data scientists to be able to pick up without an issue. Recently, our group at NJIT has created a new graph analysis library based off Arkouda under the name Arachne. The purpose of this tutorial is to provide a comprehensive view of typical pipelines that can be built and integrated with Arkouda. We will first begin by introducing an overview of Arkouda for and then move to Arachne. Examples will be provided with the questions and problems data scientists may want to answer and how Arkouda and Arachne can fit in to solve said problems. We will conclude with questions and further work that our group is planning for Arachne. Both Arkouda and Arachne are open-source and found on GitHub.

Prerequisites

1. A general understanding of data exploration may be needed but not required.
2. Basic knowledge of a programming language, need not be Python, but the tutorial will be using Python.

Important Links

1. Arkouda GitHub page – https://github.com/Bears-R-Us/arkouda
2. Arachne GitHub page – https://github.com/Bears-R-Us/arkouda-njit
3. Tutorial GitHub page – https://github.com/njit-hpc-initiative/tutorial-arkouda-njit (more to be added)

Tutors

1. Oliver Alvarado Rodriguez - oaa9@njit.edu

   Oliver Alvarado Rodriguez is currently a computer science Ph.D. student at New Jersey Institute of Technology in Newark, NJ. He performs research under the supervision of Dr. David Bader. He received his B.S. in computer science with a minor in mathematics from William Paterson University in Wayne, NJ in May 2020 with summa cum laude honors. During his undergraduate studies, he was a member

of the Honors College, a part of the Upsilon Pi Epsilon honor society for computing and information disciplines, and was also awarded the Omicron Omega award for excellence in computer science. His research interests involve the design and implementation of algorithms in the areas of high-performance analytics, machine learning, and graph theory. He has also dabbled with some cryptographical and computer security research during his undergraduate studies. He was recently awarded a best paper presentation award at the 2020 BDML/ICAIP conference for his presentation on the paper titled "A Study of Machine Learning Inference Benchmarks" done in collaboration with Dev Dave and under the tutelage of Dr. Weihua Liu and Dr. Bogong Su. Oliver recently served as the student keynote speaker at the Spring 2022 meeting of the Academic Data Science Alliance, where he presented the keynote talk: "Enabling Exploratory Large Scale Graph Analytics through Arkouda."

2. Naren Khatwani - `nk88@njit.edu`

   Naren Khatwani is a Graduate Student majoring in Computer Science at NJIT in Newark, NJ. He has been working under the supervision of Dr David Bader's Research Group as a Research Assistant. Naren has completed his B.E in Computer Engineering from University of Mumbai, India. His research interests lie in the domain of High Performance Computing and Data Analytics.

3. David A. Bader - `bader@njit.edu`

   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, and a recipient of the IEEE Computer Society Sidney Fernbach Award. He advises the White House, most recently on the National Strategic Computing Initiative (NSCI) and Future Advanced Computing Ecosystem (FACE). Dr. Bader is a leading expert in solving global grand challenges in science, engineering, computing, and data science. His 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.

4. Zhihui Du - `zd4@njit.edu`

   Zhihui Du received the BE degree in 1992 in computer department from Tianjian University. He received the MS and PhD degrees in computer science, respectively, in 1995 and 1998, from Peking University. From 1998 to 2000, he worked at Tsinghua University as a postdoctor. From 2001 to 2019, he worked at Tsinghua University as an associate professor in the Department of Computer Science and Technology. In 2008, he visited Georgia Tech for one year. His research areas include cluster system design, parallel algorithm design, task and message scheduling, resource and QoS management in grid and cloud computing.

   He has authored/co-authored two books, translated two books and edited three books in parallel computing or related fields. As the PI, he has finished more than 10 parallel computing related projects and published more than 100 parallel computing or related papers. As a major contributor, he designed and built the "DeepSuper- 21C" supercomputer which was included in the top500 list (Nov. 2003, Rank 163). His book on MPI programming is widely used in China in the parallel programming fields. He has served as the Vice Chair/PC member of more than 10 parallel processing or related conferences. He is an IEEE/ACM member.

Tutorial Outline

In this section we outline the major to topics that will be covered under Arkouda and some use cases Arkouda can be utilized for. End users have been using pandas and NumPy in Python for all of there large data set exploratory needs. Arkouda can work as massive scale replacement for NumPy and provide pandas-like analysis of data sets for exploratory data analysis.

1. Introduction (45min)

(a) Tutor Introductions – We will each give a quick personal introduction and talk about our prior and current research.

(b) Why Chapel? – Introduction to the productivity and simplicity of the Chapel language during development.

(c) Why Arkouda? – The power of Arkouda for distributed memory processing will be presented.

(d) Why Arachne? – Arachne will be presented with steps on how Arkouda can be used to further increase the analytical power of graph results.

(e) Server Setup – We will show the tutorial attendees the setup on how the installation process goes. However, they would not have to do this setup themselves.

(f) User Interface – A sample Jupyter notebook interface will be demonstrated with diagrams on how the connection interface works.

2. Break (10min)

3. Data Analytics (1hr)

   (a) Use of CSV in Pandas and HDF5 in Arkouda – Showing the difference in the input file format used for analyzing datasets in Arkouda contrary to other libraries.

   (b) Arithmetic Functions in Arkouda – Demonstrating the use of Arithmetic functions supported by Arkouda which are supported by NumPy and Pandas.

   (c) Not Restricting the use of Pandas or NumPy – Explaining how Arkouda can enhance the functionalities of both Pandas and NumPy.

   (d) Example Dataset for analysis using Arkouda – A sample analysis of a Food Dataset from Kaggle that demonstrates the use of basic arithmetic and data analysis functions in Arkouda.

4. Break (10min)

5. Graph Analytics (1hr)

   (a) Graph Definitions – We will define what a graph is and how it can facilitate data processing.

   (b) Simple Graph Queries – There are simple graph queries that can be easily performed such as printing out edges by weight order, vertex labels, etc.

   (c) Graph Analytics Deep-Dive – All the algorithms developed and implemented will be presented with concrete examples on possible usages.

6. Break (10min)

7. Conclusion (30min)

8. Q&A (15min)

Past Work

At the writing of this proposal, no previous major tutorials have been presented at PPoPP or other large similar conferences for Arkouda. A tutorial that will be used for inspiration is the Arkouda Hack-a-Thon tutorial presented at New Jersey Institute of Technology. We will follow a similar structure with some more emphasis on the mathematical operations that are taking place and the graph analytical functions we have implemented. Dr. William Reus has demonstrated exploratory data analysis on the NYCTaxi Dataset (this dataset has been widely used for presenting exploratory data analysis over the last few years) with examples of interoperating between Pandas and Arkouda at a small scale. Users can find the notebooks with the example use of Arkouda in the repository.

Publicity

We will use our Twitter accounts to tweet with the PPoPP hashtag to promote the tutorial. Further, leading up to the tutorial, we will be updating our central GitHub repository with resources and documentation for the attendees. All these updates will be announced as they are completed. Further, we will also host a tutorial website with all pertinent information and slides including links to our GitHub page for code samples.

Duration and Expected Audience Size

We expect 50+ participants and anticipate our tutorial to only take half a day, 4 hours.

References

[1] Zhihui Du, Oliver Alvarado Rodriguez, David A. Bader, Michael Merrill, and William Reus. "Exploratory Large Scale Graph Analytics in Arkouda". In: *The 8th Annual Chapel Implementers and Users Workshop (CHIUW)*. June 2021.

[2] Zhihui Du, Oliver Alvarado Rodriguez, Joseph Patchett, and David A. Bader. "Interactive Graph Stream Analytics in Arkouda". In: *Algorithms* 14.8 (2021). ISSN: 1999-4893. DOI: 10.3390/a14080221. URL: https://www.mdpi.com/1999-4893/14/8/221.

[3] Joseph T. Patchett, Zhihui Du, Fuhuan Li, and David A. Bader. "Triangle Centrality in Arkouda". In: *The 26th Annual IEEE High Performance Extreme Computing Conference (HPEC), Virtual, September 19-23, 2022*. 2022.

[4] William Reus. "CHIUW 2020 Keynote Arkouda: Chapel-Powered, Interactive Supercomputing for Data Science". In: *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. 2020, pp. 650–650.

[5] Oliver Alvardo Rodriguez, Zhihui Du, Joseph T. Patchett, Fuhuan Li, and David A. Bader. "Arachne: An Arkouda Package for Large-Scale Graph Analytics". In: *The 26th Annual IEEE High Performance Extreme Computing Conference (HPEC), Virtual, September 19-23, 2022*. 2022.