Interactive Data Science at Scale

David A. Bader

@Prof_DavidBader

http://www.cs.njit.edu/~bader

NJIT New Jersey Institute of Technology

NJIT New Jersey Institute of Technology

Primarily seeking research students interested in:

- 3-credit MS Project
- 6-credit MS Thesis
- Hourly wage

David A. Bader

Distinguished Professor and
Director, Institute for Data Science

- IEEE Fellow, SIAM Fellow, AAAS Fellow
- Recent Service:
  - White House's National Strategic Computing Initiative (NSCI) panel
  - Computing Research Association Board
  - NSF Advisory Committee on Cyberinfrastructure
  - Council on Competitiveness HPC Advisory Committee
  - IEEE Computer Society Board of Governors
  - IEEE IPDPS Steering Committee
  - Editor-in-Chief, ACM Transactions on Parallel Computing
  - Editor-in-Chief, IEEE Transactions on Parallel and Distributed Systems
- Over $185M of research awards
- 250+ publications, ≥ 11,000 citations, h-index ≥ 61
- National Science Foundation CAREER Award recipient
- Directed: Facebook AI Systems
- Directed: NVIDIA GPU Center of Excellence, NVIDIA AI Lab (NVAIL)
- Directed: Sony-Toshiba-IBM Center for the Cell/B.E. Processor
- Founder: Graph500 List benchmarking “Big Data” platforms
- Recognized as a “RockStar” of High Performance Computing by InsideHPC in 2012 and as HPCwire’s People to Watch in 2012 and 2014.

New Jersey Institute of Technology

High Performance Algorithms for Interactive Data Science at Scale

(PI: Bader) 3/2021 – 2/2022, NSF CCF-2109988

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. This award will design and implement fundamental algorithms for high performance computing solutions that enable the interactive large-scale data analysis of massive data sets.

This project focuses on these three important data structures for data analytics:
1) suffix array construction,
2) 'treap' construction, and
3) distributed memory join algorithms,

useful for analyzing large scale strings, implementing random search in large string data sets, and generating new relations, respectively.

To evaluate and show the effectiveness of the proposed algorithms, these algorithms will be implemented in and contribute to an open source NumPy-like software framework that aims to provide productive data discovery tools on massive, dozens-of-terabytes data sets by bringing together the productivity of Python with world-class high performance computing.

NJIT New Jersey Institute of Technology

NJIT New Jersey Institute of Technology

Alumni Giving MyNJIT Login Tools

Search NJIT

Admissions Academics Research About Life at NJIT Athletics I AM A... APPLY NOW

Institute for Data Science Aims to Democratize Supercomputing With NSF Grant

Written by: Evan Koblentz

Published: Wednesday, March 17, 2021

New algorithms from at NJIT can make supercomputer power available to almost anyone

Ordinary people could soon have greater ability to analyze massive amounts of information, based on new algorithms and software tools being designed at NJIT, intended to simplify

https://news.njit.edu/institute-data-science-aims-democratize-supercomputing-nsf-grant

NJIT New Jersey Institute of Technology

Data-Quad

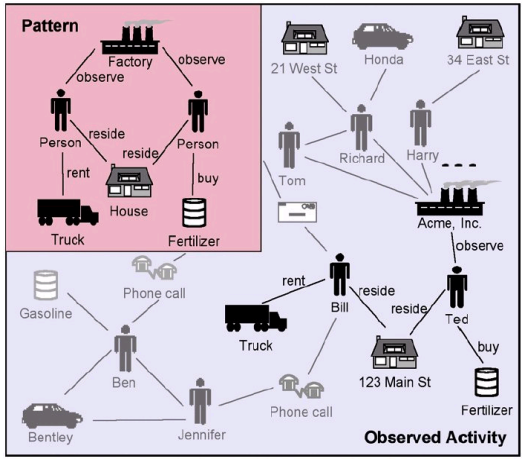
Unknown

Known

Objects

Unknown

Patterns



Known

New Jersey Institute of Technology

Graph Data Science: Real-world challenges

All involve exascale streaming graphs:

- **Health care** → disease spread, detection and prevention of epidemics/pandemics (e.g. SARS, Avian flu, H1N1 "swine" flu)
- **Massive social networks** → understanding communities, intentions, population dynamics, pandemic spread, transportation and evacuation
- **Intelligence** → business analytics, anomaly detection, security, knowledge discovery from massive data sets
- **Systems Biology** → understanding complex life systems, drug design, microbial research, unravel the mysteries of the HIV virus; understand life, disease,
- **Electric Power Grid** → communication, transportation, energy, water, food supply
- **Modeling and Simulation** → Perform full-scale economic-social-political simulations

REQUIRES PREDICTING / INFLUENCE CHANGE IN REAL-TIME AT SCALE

New Jersey Institute of Technology

Streaming Analytics move us from reporting the news to predictive analytics

Traditional HPC

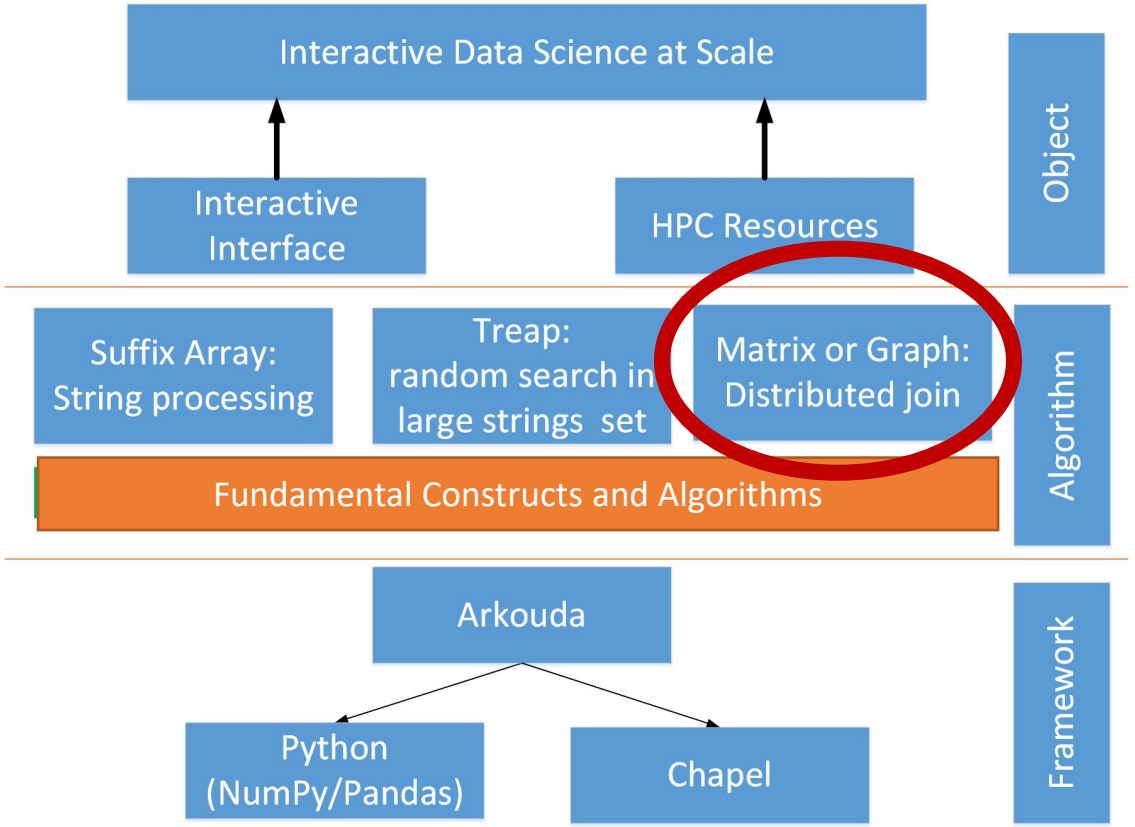
- Great for “static” data sets.
- Massive scalability at the cost of programmability.
- Great for dense problems.
  - Sparse problems typically underutilize the system.

Streaming Analytics

- Requires specialized analytics and data structures.
- Rapidly changing data.
- Low data re-usage.
  - Focused on memory operations and not FLOPS.

Overview of Research

- Objective:
  - One-stop solution for non-HPC experts with massive data sets. As graphs become huge, with billion or trillions of edges, regular computing at a laptop level becomes more difficult, time consuming, and may be impossible!
- Developmental focus:
  - Data structures and algorithms for graph and other problems.
- Framework:
  - Arkouda

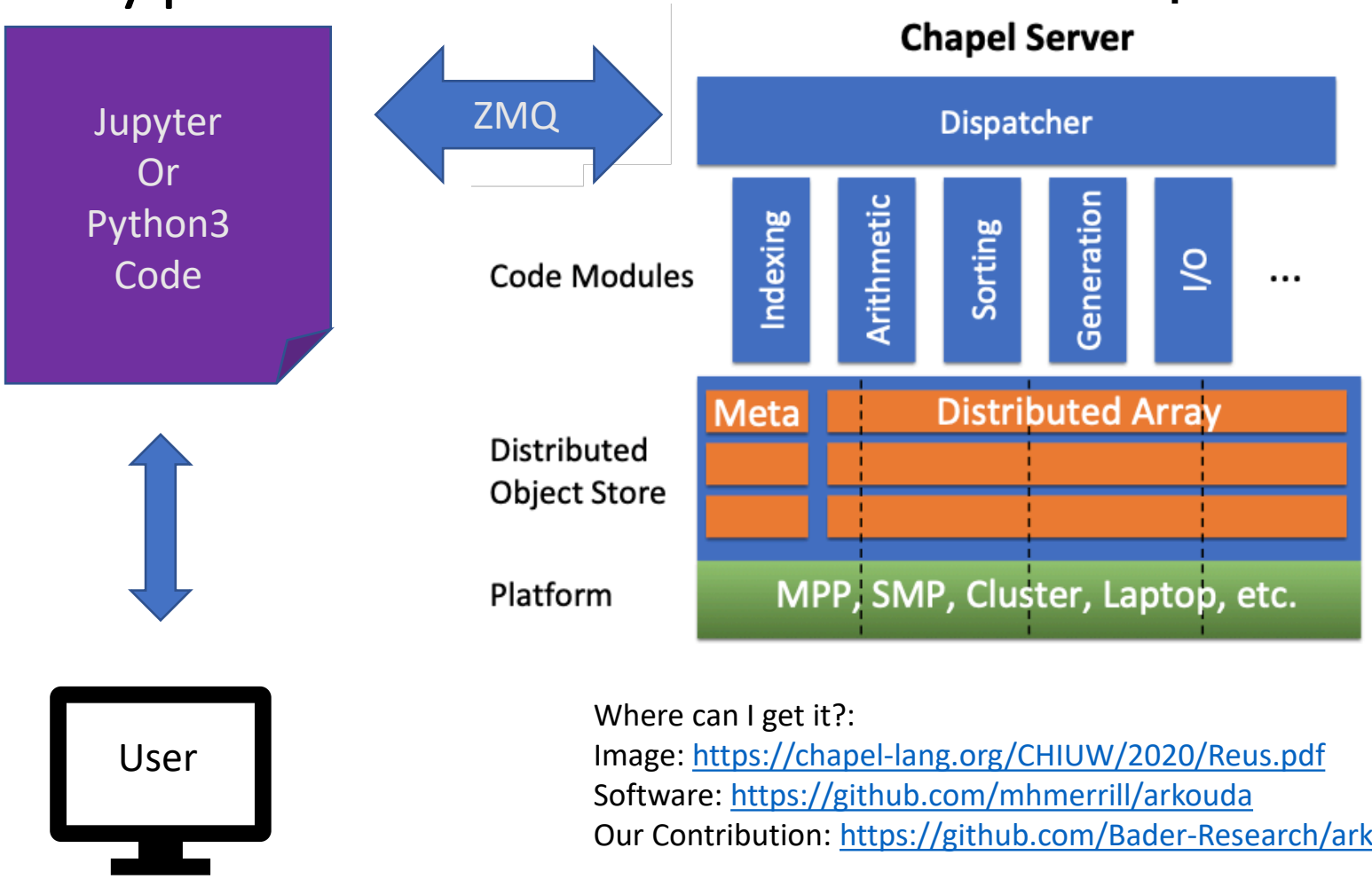


New Jersey Institute of Technology

Productivity + Performance

New Jersey Institute of Technology

Typical Environment Set-Up



Python3 Implementation:

- Pdarray class
- Rely on Python to reduce complexity
- Integrate with and use NumPy

Server Implementation:

- High-level language with C-comparable performance
- Great parallelism handling
- Great distributed array support
- Portable code: laptop --> HPC

Where can I get it?:
Image: https://chapel-lang.org/CHIUW/2020/Reus.pdf
Software: https://github.com/mhmerrill/arkouda
Our Contribution: https://github.com/Bader-Research/arkouda/tree/streaming

New Jersey Institute of Technology