NSF 1443054: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science PI: Geoffrey C. Fox

See http://dsc.soic.indiana.edu/publications/SPIDAL-DIBBSreport_July2016.pdf and http://spidal.org

Proposal Challenge: Deliver on objectives of proposal

1) **Big Data Application Analysis** identifies features of data intensive applications that need to be supported in software and represented in benchmarks. This analysis was started for proposal and has been extended to support HPC-Simulations-Big Data convergence. The project is a collaboration between computer and domain scientists in **application areas** in Biomolecular Simulations, Network Science, Epidemiology, Computer Vision, Spatial Geographical Information Systems, Remote Sensing for Polar Science and Pathology Informatics.
2) **HPC-ABDS** as Cloud-HPC interoperable software with performance of HPC (High Performance Computing) and the rich functionality of the commodity Apache Big Data Stack was a bold idea developed for proposal. We have successfully delivered and extended this approach, which is one of ideas described in Exascale Big Data report.
   http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec2016pathways-16Nov16-b.pdf
3) **MIDAS** integrating middleware that links HPC and ABDS now has several components including an architecture for Big Data analytics, an integration of HPC in communication and scheduling on ABDS; it also has rules to get high performance Java scientific code.
4) **SPIDAL** (Scalable Parallel Interoperable Data Analytics Library) now has 20 members with domain specific (general) and core algorithms.
5) **Benchmarks**. We reached out to database community with keynote and paper at WBDB2015 Benchmarking Workshop.
6) Streaming Analytics and Systems is a new opportunity identified in 2 workshops (http://streamingsystems.org).

Current Challenge: software development and integration

We have made significant progress in all aspects of this project but now we need to pull this together with a software engineering and integration task. As well as uniform packaging and testing of MIDAS and SPIDAL, we need to design a good API for the SPIDAL members (the current ABDS libraries such as MLlib have a poor API as well as low performance!) so that we have a SPIDAL library and a MIDAS middleware product which either an application developer or an XSEDE resource provider can download. Fragility of software ecosystem and performance are pervasive challenges while churn in data-intensive platform is very high.

Future Challenge: Deployment, Training and Outreach of MIDAS Software and SPIDAL Library

Even as the middleware and analytics are being developed and properly packaged we need to turn to deployment and training and a proactive outreach to users and service providers. We need to consider traditional library as well as PaaS and SaaS deployments. Fox will give a 6 hour tutorial on MIDAS and SPIDAL at a European winter school in February 2017 and this will increase our work in this area.