BigDat 2017 MIDAS and SPIDAL Tutorial

Geoffrey Fox, Indiana University February 10 2017
On behalf of
Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang,
Indiana University, Rutgers, Kansas University, Arizona State, Utah, Virginia Tech, Stony Brook.

Abstract

Two major trends in computing systems are the growth in high performance computing (HPC) with an international exascale initiative, and the big data phenomenon with an accompanying cloud infrastructure of well publicized dramatic and increasing size and sophistication. This tutorial weaves these trends together using some key building blocks. The first is HPC-ABDS, the High Performance Computing (HPC) enhanced Apache Big Data Stack. (ABDS). Here we aim at using the major open source Big Data software environment but develop the principles allowing use of HPC software and hardware to achieve good performance. We give several examples of software (for example Hadoop and Heron) and algorithms implemented in this software. The second building block is the SPIDAL library (Scalable Parallel Interoperable Data Analytics Library) of scalable machine learning and data analysis software. We give examples including clustering, topic modeling and dimension reduction and their visualization. The third building block is an analysis of simulation and big data use cases in terms of 64 separate features (varying from data volume to "suitable for MapReduce" to kernel algorithm used). This allows an understanding of what type of hardware and software is needed for what type of exhibited features; it allows a one to either unify or distinguish applications across the simulation and Big Data regimes. The final building block is DevOps and Software defined Systems. These allow one to package software so it runs across a variety of hardware (albeit with varying performance) with just a mouse click. These building blocks are finally linked together as a proposed convergence of Big Data and Exascale Computing.
This tutorial builds on work of a collaboration funded as NSF14-43054 started October 1, 2014. It contains descriptive material and several explicit hands-on tutorials. Much open source software is available.

Tutorial plan is at
https://docs.google.com/document/d/17PLTeYUjB_0JmUtdfN9KCYU4wnr3hScvMVUlYzLPvic/edit# or http://dsc.soic.indiana.edu/publications/SPIDALTutorialProgram-Feb2017.pdf or DOI

Poster http://dsc.soic.indiana.edu/presentations/NSF1443054_CIF21DIBBS_Poster_v2.pdf
Report http://dsc.soic.indiana.edu/publications/SPIDAL-DIBBSreport_July2016.pdf
HPC-ABDS http://hpc-abds.org/kaleidoscope/

Tutorial Winter School http://grammars.grlmc.com/BigDat2017/
http://grammars.grlmc.com/BigDat2017/

Component Presentations

1. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Tutorial Overview: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

2. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Master Presentation: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

3. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Big Data Use Cases: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

4. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Big Data Use Case Examples: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

5. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Deterministic Annealing Algorithms for Analytics: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

6. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "WebPlotViz: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

7. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "General Discussion of HPC-ABDS: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

8. Saliya Ekanayake (Virginia Tech); Geoffrey Fox (Indiana University), "SPIDAL Java Optimization: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

9. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "SPIDAL Analytics Performance: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

10. Ioannis Paraskevakos, Andre Luckow, Shantenu Jha (Rutgers); Oliver Beckstein (Arizona State); presented by Geoffrey Fox, "MIDAS- Molecular Dynamics Analysis Tutorial", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

11. Langshi Chen, Bingjing Zhang, Peng Bo, Judy Qiu, Indiana University; presented by Geoffrey Fox, "Harp HPC for Big Data: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

12. David Crandall (Indiana University); Fusheng Wang (Stony Brook); "Image & Model Fitting Abstractions", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

13. David Crandall (Indiana University); John Paden (Kansas); "Polar Science Applications", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

14. Jun Kong (Emory University); Fusheng Wang (Stony Brook); presented by Geoffrey Fox, "2D/3D Pathology Image and Spatial Analysis", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy

February 13-14 2017

15. Fusheng Wang (Stony Brook); presented by Geoffrey Fox, "Integrative Big Spatial Data Analytics for Public Health Studies: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

16. Oliver Beckstein (Arizona State); presented by Geoffrey Fox, "MDAnalysis and Biomolecular Simulations", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

17. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "HPC Cloud Convergence: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

18. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Software Defined Systems: February 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

19. Geoffrey Fox, David Crandall, Judy Qiu, Gregor Von Laszewski, Shantenu Jha, John Paden, Oliver Beckstein, Tom Cheatham, Madhav Marathe, Fusheng Wang, "Status and Challenges: January 2017", BigDat 2017 MIDAS and SPIDAL Tutorial Bari Italy February 13-14 2017

Organization

See http://dsc.soic.indiana.edu/presentations/SPIDAL-Tutorial-Feb2017.pptx which organizes tutorial and http://dsc.soic.indiana.edu/presentations/SPIDAL-TutorialOverview-Feb2017.pptx which summarizes this

- **Introduction**
  http://dsc.soic.indiana.edu/presentations/SPIDAL-TutorialOverview-Feb2017.pptx

- **NIST Big Data Use Case Analysis Ogres/Diamonds**
  http://dsc.soic.indiana.edu/presentations/SPIDAL-UseCase-Feb2017.pptx
  Optional additional material
  http://dsc.soic.indiana.edu/presentations/SPIDAL-UseCaseExamples-Feb2017.pptx

- **Examples of HPC Analytics and applications**: Clustering, dimension reduction, visualization
  - http://dsc.soic.indiana.edu/presentations/SPIDAL-DAAlgorithmsAnalytics-Feb2017.pptx covers Clustering and Dimension Reduction
  - http://dsc.soic.indiana.edu/presentations/SPIDAL-Webplotviz-Feb2017.pptx covers WebPlotViz
  - Tutorial in Master. See https://dsc-spidal.github.io/tutorials/

- **HPC-ABDS** with performance of High Performance Computing and the rich functionality of the commodity Apache Big Data Stack

    - HPC-ABDS Concept
      http://dsc.soic.indiana.edu/presentations/SPIDAL-GeneralHPCABDS-Feb2017.pptx
    - SPIDAL Java
      http://dsc.soic.indiana.edu/presentations/SPIDAL-Java-Optimized-Feb2017.pptx
    - Core Analytics Performance
      http://dsc.soic.indiana.edu/presentations/SPIDAL-AnalyticsPerformance-Feb2017.pptx
    - HPC Flink/Spark and parallel/distributed computing in ABDS in Master

- **MIDAS examples**
        - MIDAS and Biomolecular Simulations
          http://dsc.soic.indiana.edu/presentations/SPIDAL-MIDAS-BioSim-Feb2017.pptx
        - Harp: HPC Hadoop
          http://dsc.soic.indiana.edu/presentations/SPIDAL-Harp%20HPC%20for%20Big%20Data-Feb2017.pptx includes machine learning system architecture and algorithms such as LDA
        - Streaming Applications in HPC-ABDS (HPC Storm) see master (http://dsc.soic.indiana.edu/presentations/SPIDAL-Tutorial-Feb2017.pptx)

- **SPIDAL Algorithms**
    - General discussion and Graphs in Master

    - **Image Based Analytics and their applications**
      http://dsc.soic.indiana.edu/presentations/SPIDAL-ImageModelAbstractions-Feb2017.pptx

- **SPIDAL Motivating Applications**
    - **Polar Science**
      http://dsc.soic.indiana.edu/presentations/SPIDAL-RadarImaging-Feb2017.pptx
    - **Pathology including SPIDAL image analysis**
      http://dsc.soic.indiana.edu/presentations/SPIDAL-PathologySpatial-Feb2017.pptx
    - **Geospatial studies (public health)**
      http://dsc.soic.indiana.edu/presentations/SPIDAL-PublicHealth-Feb2017.pptx

    - **Biomolecular Simulation Data Analysis**
        - Application
          http://dsc.soic.indiana.edu/presentations/SPIDAL_MDAnalysis_Biomolecular_Simulations_Feb2017.pptx
        - Initial libraries at Utah/Arizona State (Utah: cpptraj https://github.com/Amber-MD/cpptraj ; ASU: MDAnalysis http://mdanalysis.org)

        - See earlier MIDAS tutorial http://dsc.soic.indiana.edu/presentations/SPIDAL-MIDAS-BioSim-Feb2017.pptx

- **HPC Simulation, Big Data basic Convergence**
    - HPC Cloud 1.0 http://dsc.soic.indiana.edu/presentations/SPIDAL-Convergence-Feb2017.pptx
    - DevOps and Cloudmesh -- Software defined Systems and their deployment leading to HPCCloud 2.0 http://dsc.soic.indiana.edu/presentations/SPIDAL-SoftwareDefinedSystems-Feb2017.pptx
    - Above implies Event-based function as a service as natural SPIDAL deployment. HPCCloud 3.0

- **Futures**
    - Status http://dsc.soic.indiana.edu/presentations/SPIDAL-Status-Feb2017.pptx
    - Apache Beam for orchestration in Master

Links

- https://github.com/DSC-SPIDAL Indiana University Software
- Tutorials on Indiana University SPIDAL software https://dsc-spidal.github.io/tutorials/ or if you are at Indiana University see https://docs.google.com/document/d/1HBqQpgr-9YMD01BJzizZAwAIF-OcA7FvlUWDaFOVMwQ/
- http://dsc.soic.indiana.edu/presentations/SPIDAL-SoftwareDefinedSystems-Feb2017.ppt has detailed tutorial instructions for running Cloud 2.0 software automation on Amazon AWS and Azure
- https://github.com/radical-cybertools/MIDAS-tutorial Rutgers and Biomolecular
- simulation software
- http://hpc-abds.org/kaleidoscope/ Many HPC-ABDS links including our publications
- http://bigdataopensourceprojects.soic.indiana.edu/#section1 Unit 1. Classic Introduction to Big Data from INFO I-524 Class Spring 2015
- Geoffrey Fox, "Big Data Applications & Analytics Motivation: Big Data and the Cloud; Centerpieces of the Future Economy", Classic Introduction to Big Data Applications and Analytics INFO I-523 Spring 2015 class.

- Geoffrey Fox, MOOC discussion of NIST use cases is section 5 of https://bigdatacoursespring2015.appspot.com/preview

    - Geoffrey Fox, "Overview of NIST Big Data Public Working Group (NBD-PWG) Process and Results", Spring 2014 I523 Part I of lectures on NIST Big Data use cases

    - Geoffrey Fox, "51 Big Data Use Cases", Spring 2014 I523 Part II of lectures on NIST Big

Data use cases

- Geoffrey Fox, "Features of 51 Big Data Use Cases", Spring 2014 I523 Part III of lectures on NIST Big Data use cases

- http://bigdataopensourceprojects.soic.indiana.edu/#section1 Unit 2. Parts A to C cover (the ~10) data access patterns
- http://bigdataopensourceprojects.soic.indiana.edu/#section1 Unit 2. Parts D to G summarize HPC-ABDS software stack
- http://bigdataopensourceprojects.soic.indiana.edu/#section3 for lectures covering the Spring 2015 HPC-ABDS at level of around 1 slide for each ~300 (then) members

Theses

- Saliya Ekanayake, "Towards a Systematic Study of Big Data Performance and Benchmarking", Indiana University Ph.D Dissertation Defense, September 28th, 2016
- Yang Ruan SCALABLE AND ROBUST CLUSTERING AND VISUALIZATION FOR LARGE-SCALE BIOINFORMATICS DATA Indiana University PhD defense 18 August 2014
- Thilina Gunarathne SCALABLE PARALLEL COMPUTING ON CLOUDS: EFFICIENT AND SCALABLE ARCHITECTURES TO PERFORM PLEASINGLY PARALLEL, MAPREDUCE AND ITERATIVE DATA INTENSIVE COMPUTATIONS ON CLOUD ENVIRONMENTS Indiana University PhD defense April 21 2014
- Zhenhua Guo HIGH PERFORMANCE INTEGRATION OF DATA PARALLEL FILE SYSTEMS AND COMPUTING: OPTIMIZING MAPREDUCE Indiana University PhD August 31 2012
- Seung-Hee Bae SCALABLE HIGH PERFORMANCE MULTIDIMENSIONAL SCALING Indiana University PhD January 17 2012
- Jong Youl Choi Unsupervised Learning Of Finite Mixture Models With Deterministic Annealing For Large-scale Data Analysis Indiana University PhD January 12 2012
- Jaliya Ekanayake ARCHITECTURE AND PERFORMANCE OF RUNTIME ENVIRONMENTS FOR DATA INTENSIVE SCALABLE COMPUTING Indiana University PhD December 20 2010