Merge or Split?

Chaitan Baru, Associate Director, Data Initiatives, SDSC
(currently on assignment at National Science Foundation)
Related Activities

• **Panels at SC, VLDB**
  • Organized by NITRD High-End Computing and Big Data Groups

• **At SC 2015**
  • *Supercomputing and Big Data: From Collision to Convergence*
  • Panelists: David Bader (GaTech), Ian Foster (Chicago), Bruce Hendrickson (Sandia), Randy Bryant (OSTP), George Biros (U.Texas), Andrew W. Moore (CMU)

• **At VLDB 2015**
  • *Exascale and Big Data*
  • Panelists: Peter Baumann (Jacobs University), Paul Brown (SciDB), Michael Carey (UC Irvine), Guy Lohman, (IBM Almaden), Arie Shoshani (LBL)
Merge from Big Data to HPC

• Adapting Big Data software stacks for HPC is probably more fruitful than other way around – viz., adapting HPC software to handle Big Data needs

• Because
  • HPC: well-established software ecosystem, highly sensitive to performance, established codebases
  • Big Data: Rapidly evolving and emerging software ecosystem, evolving applications needs, price/performance is more relevant
Merge vs Split

- **HPCBD:** Focus on performance of the HPCBD software stack (+ implicitly the hardware)
- **But there could be multiple stacks**
  - Not 100’s, or 10’s, but perhaps >5, <10 ?
  - E.g. stream processing; genomic processing; geospatial data processing; deep learning with image data; …
- **Can we enumerate a few stacks, based on functionality?**
  - Do we need reference datasets for each stack?
- **Could we run a workshop to identify stacks and how stack-based benchmarking would work**
  - Can we develop “reference stacks”…how should that be done?
  - Streaming data processing will be big…