Brainstorming on “Big Data” Projects

29 May 2014

What is “Big Data” to us?

There nothing “small” about what we do today!

Physics production/analyses are independent-event based processing using Root with positioning of data to jobs (either before hand or just in time)

We’ve spent over a decade optimizing our systems (data management, grid processing, data caches, archives) for this analysis approach

This work is not done - e.g. smarter caching, more efficient protocols, better use of networks, use of cloud systems – all deserve work

What is our “Big Data” lacking?

Deficiencies of the current approach:

- Cannot do large scale interactive processing (batch processing typically has long turn-around times)
- Difficult to do “global” analyses on huge data sets
- Need heavy-weight and optimized data management (positioning/moving/streaming data is key for performance)
- Need large bandwidth wide-area-networks (the fact we have these makes current approach doable)

Interesting question: How much more science can we extract if we could easily “play” with huge amounts of data interactively or in near-real time?

Our Vision

Taking HEP data analysis to big data and beyond

- Extract more science quicker and easier
- Providing an interactive framework for HEP analysis using *modern* computing and storage resources
- Taking advantage of new architectures & clouds

What is “Big data” to others?

4V’s - Volume, Velocity, Veracity, Variety

e.g. map-reduce “appliances” (data is loaded into large system with many nodes and ***indexed*** - each node applies map function to its ***local*** data satisfying a query, results from all are “reduced” to a result). Indexing and locality makes things fast! Think why your Google search for “The Pacific Northwest Tree Octopus” doesn’t take a week to return!

Our current approach lacks the index and locality (and probably a lot of other smarts that Google does)

Explore projects that involve these ideas...

Potential Projects

o Determine physics or other topics that would benefit from other processing approaches. Is there science we are not doing today because we cannot process the necessary data? For example, NOvA template analysis, treating g-2 and LAr data as signal processing, mining monitoring data

o Explore indexing technologies (make “virtual skims”, avoid copying the data;
interestingly we had this in Tevatron Run I, can we do this for big data sets?)

o Explore data “appliance techniques” (e.g. keeping most/all data resident on nearby disk for fast local and repeated access). Perhaps a project using Fermicloud? Amazon?

o Explore new architectures (e.g. multicores) - special requirements for i/o?

o Continue to strengthen current approach (make data management and components smarter and easier to use, and usable for clouds) while looking to the future