Advanced Analytics Environments For Large Data Collections

Mark Watson
Federal Reserve Bank of Kansas City
Background

• Research environment started in the early 1990’s
  • Adopted HPC for computational platform in 2002
  • Added MPP technologies in 2009 for data hosting
• Took on large micro data management 2008
• Took on confidential large data micro management 2010

Currently we support 400+ HPC & 500+ data users at 12 locations across 6 business lines hosting 90 terabytes of data
What we learned during the financial crisis

• Traditional software products do not maximize the performance of the hardware
• Traditional methods for statistical and computational analysis of large datasets can be expensive and challenging
• The issues related to managing large data are analytical and not transactional
• There is no silver bullet that can provide a perfect computing experience for all users.
Computing support for large micro data

Needs to provide:

• The ability to incrementally scale infrastructure to meet capacity demands (processing or volume);
• Parallel processing technology as well as serial processing; and
• The ability to work with collected data in a compressed format
Data warehousing support for large micro data

Needs to provide:
• Unencumbered access to the data
  • Fully expose data to provide ad-hoc query access
• SQL queries to pull data samples
  • Graphical SQL clients or command-line
• In-database analytics (pushing code to the data)
  • Sample preparation or analysis prior to pulling data
• MapReduce framework with PL-languages for analytics
• 3rd party libraries to run in-database
Our solution for research

• Integrated HPC and storage systems into one ecosystem
• Conscious thought went into making sure our infrastructure can scale by using parallel solutions and removing as many bottlenecks as possible.
  • Parallel Storage (Panasas & Ceph)
  • Parallel Compute (HPC and job scheduler)
  • Parallel Database (MPP)
  • 10G Interconnect
  • Parallel Tools (pbzip2, pigz, etc)
Typical research use of a HPC

• Used to solve complex, computationally intensive or large data computing problems:
  • Serial processing
  • Parallel processing
  • Leveraging accelerators (GPU’s)

• User access methods:
  • Web browser (graphical) or command-line (non-graphical)
  • Start Linux sandbox sessions
  • Job submissions via web browser, command-line, or local applications
How an HPC and data storage work together

- Limit the movement of data by interconnecting HPC and data warehouse
- Querying data from cluster is almost peer to peer when pulling data
- Training/evolving users to push code to data and take advantage of in-database analytics
**Typical software stack**

### Cluster Programming Languages

- Matlab\(^1\)
- STATA SE & MP\(^1\)
- SAS & SAS Studio\(^1\)
- R-Project/Rstudio\(^2\)
- PERL, Python, Java, Julia, CUDA and Etc.\(^2\)
- Intel C/C++/FORTRAN compilers libraries and development tools\(^1\)
- GNU C/C++ & Fortran compilers and libraries\(^2\)
- RATS\(^1\)
- Mathematica\(^1\)
- Maple\(^1\)
- Octave\(^2\)
- Gretl\(^2\)
- WEKA\(^2\)
- Anaconda\(^2\)
- OpenACC\(^1\)

### Cluster Tools & Libraries

- Linux graphical desktop and environmental tools\(^2\)
- GNOME, shells, editors, LibreOffice and Etc.\(^2\)
- Pgadmin/PSQL/SquirrelSQL\(^2\)
- Stat Transfer\(^1\)
- Spark\(^2\)
- QGIS\(^2\)
- IMSL\(^1\)
- Intel libraries & MPI\(^1\)
- GNU MPI & OpenMP\(^2\)

### Data Tools

- Prod, Dev MPP and PostgreSQL database\(^1,2\)
- MapReduce\(^1\)
- MADlib libraries\(^2\)
- PostGIS\(^2\)
- PSQL/in-SQL analytics functions\(^2\)
- Cluster User Utilities
- MatterMost\(^1\)
- OSGeo-Live/Geoserver\(^2\)
- MapServer\(^2\)
- Connector/R-Shiny\(^1\)
- GitLab\(^1\)

1 Commercial software
2 Open source software
Technology choices across activities

Different types of analytics are better in different environments

Research Analytic Environment

Operationalized Research Analytic Environment

Enterprise Analytic and BI Environment

Fluid Dynamic Large Choice of Tools Emerging Technologies Self Service

Structured Controlled Limited Choice of Tools Established Technologies Managed
Cost depends on where you are

• Significant cost difference based on what is being built

• Affects staffing model and technology choices
  • Research oriented: low hardware/software costs, high personnel costs – both support and commitment from users
  • Enterprise oriented: high hardware/software costs, low personnel costs – GUI based, possible contract support
Questions?

Thank you for your time.

Mark Watson
mark.a.watson@kc.frb.org
(816) 881 - 2238