REPORT FROM THE 6th WORKSHOP ON EXTREMELY LARGE DATABASES

Daniel Liwei Wang[1] , Jacek Becla[2*], Kian-Tat Lim[3]

SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
[1] *Email:* danielw@slac.stanford.edu
[*2] *Email:* becla@slac.stanford.edu
[3] *Email:* ktl@slac.stanford.edu

ABSTRACT

Petascale data management and analysis remain one of the main unresolved challenges in today's computing. The 6th Extremely Large Databases workshop was convened alongside the XLDB conference to discuss the challenges in the health care, biology, and natural resources communities. The role of cloud computing, the dominance of file-based solutions in science applications, in-situ and predictive analysis, and commercial software use in academic environments were discussed in depth as well. This paper summarizes the discussions of this workshop.

Keywords: Analytics, Database, Petascale, Exascale, XLDB, Big data, Extreme-scale

1 EXECUTIVE SUMMARY

The 6th XLDB workshop (held during XLDB-2012) focused on the health care and biology communities, in-situ and predictive analytics, cloud computing, commercial software use in academic environments, and a new community (natural resources).

The workshop began with an unplanned discussion on why scientists overwhelmingly store and manage data as files while databases are mostly unused. Files and hierarchical directory structures have become common computer knowledge while databases remain cryptic, cumbersome, and unpredictable. A divide between data producers and users further complicates the situation: producers usually do not provide an integrated platform for analysis, and nearly all downstream software operate on files (as retrieved from producers) not databases. The database community should recognize that files are the canonical forms of data and deal with it.

The biology and health care attendees identified six main data-related problems endemic within their community: sociological impediments to integration, high velocity and variance of the data, a lack of single-pass algorithms, insufficient understanding of data, poor provenance, and the absence of a widely-known repository for computational use cases. Attendees agreed that establishing evaluation criteria for measuring and comparing various implementations and proposed solutions, building prototype database-solutions based on existing public data, constructing a benchmark, and open sharing of computing and data management practices would be helpful.

Predictive and in-situ analyses will be required for exascale computing, but participants did not know of a vision for integrating them into the workflow of simulation-based experiments. Ideally, such scientific projects would follow the example of the Sloan Digital Sky Survey by engaging top science and computing experts as champions who would help them overcome the resistance to data sharing.

Cloud computing discussions centered on four themes: economics, novelty, deployment, and reliability and control. Overall, participants felt that cloud computing, whether public or private, had significant advantages, especially for smaller organizations and groups without sufficient backfill load and for applications designed with cloud considerations. Calculating the benefits of using cloud resources is very complex and involves accounting for reliability, security, and entire infrastructure costs. Nobody questions the trade-offs: privacy, security concerns, and complications with post-intrusion forensics. Attendees expect cloud computing to become more cost-competitive over time and funding agencies to revise their policies to be more compatible with cloud computing.

Vendors asked why so few academic scientists and institutions adopt proprietary software and hardware. Reasons mentioned ranged from tight, short-term budgets and geographically-distributed collaboration to uncertain corporate product lifetimes and future licensing costs. Debuggability, documentation, and product support were cited as particular areas where science needs and a company's target use case diverge. Participants noted that nearly every new DBMS is open-source while every old one is closed-source.

XLDB participants from the natural resources community said their rapidly-growing large-scale data are dominated by information collected by equipment instrumentation and now-ubiquitous sensing. The resource extraction industry faces the challenge of integrating data from disparate sources (e.g., satellite sensors and climate models) and different formats. However, the huge scale of many mining operations means that relatively minor improvements in data-driven efficiency and management have translated into multimillion-dollar cost savings.

Participants found the 6th XLDB workshop to be useful and not needing major changes. One suggested tweak was to choose one or two topics to be re-visited regularly for updates. New participants found the written reports especially useful.

2 ABOUT THE WORKSHOP

Since 2007, the annual Extremely Large Database (XLDB) workshop has hosted influential discussions on topics related to databases of terabyte through petabyte to exabyte scale. The 6th workshop (http://www-conf.slac.stanford.edu/xldb2012/Workshop.asp) in this series was held at Stanford University in Stanford, California, on September 13, 2012. The main goals of the workshop were to:

- reach out to a new community – natural resources – and explore more deeply the health care and biology communities,
- review community and hybrid cloud computing and commercial software in academic environments, and
- discuss a specific use case: in-situ analytics at one national lab.

The workshop was held on the last of XLDB-2012's four days. Information regarding the tutorials and conference portions of XLDB-2012, which were held during the first three days, can be found at the conference website (http://www-conf.slac.stanford.edu/xldb2012/). Information about past XLDB workshops, including the reports, can be found at http://xldb.org/events.

2.1 Participation

Attendance at the 6th workshop was, like its predecessors, by invitation. This keeps the group small enough for interactive discussions and the represented communities balanced. Fifty-seven attended, representing science and industry database users, academic database researchers, and database vendors. Industrial representation continued to grow. Further attendance details are on the workshop website.

2.2 Structure

Maintaining XLDB tradition, the workshop's loose structure stimulated candid but productive discussions on the most important topics for XLDB practitioners. It began with an unplanned debate on storing and analyzing data in file systems versus in databases. Subsequent sessions covered data usage in biology, integrated analysis at a national lab, practical cloud computing for large data, natural resource extraction, and the use of commercial software in academic environments. The workshop concluded with an update on the state of the XLDB use case collection and planning for the next XLDB event.

3 BIOLOGY AND HEALTH CARE

Discussion about biology and health-care began at the 5th XLDB (see *Data Science Journal*, *Vol. 11*: https://www.jstage.jst.go.jp/article/dsj/11/0/11_012-010/_article). Biology and health-care representatives attending the 6th XLDB workshop described the state of data management in their disciplines' communities as universally poor. The biggest problems are undisciplined data practices, reduction algorithms, and non-standardized data that are difficult to integrate. Real scalability is rarely the primary problem, but the size of raw datasets is growing beyond what many organizations can currently store and manage affordably. It is difficult to predefine biology-relevant schemas because large databases are relatively new in biology, and it is not clear what information should be saved for later processing. As a result, the biology and health-care communities require flexible systems that allow relatively free-form tagging and associations.

Workshop attendees described the information culture in biological disciplines as being detail-focused, data-possessive, fragmented, and suspicious of computer technologies. One biologist complained about the difficulty of managing data locked away in many spreadsheets. Other participants said that the simple, loose model of spreadsheets is an advantage because it facilitates sharing. Moreover, scalability is not difficult because spreadsheets are relatively small—the interface discourages inputting larger amounts of data.

Because the spreadsheet model is so prevalent in biology, data-analysis tools must integrate with spreadsheet interfaces. Figure 1 shows where spreadsheets are used in a common process for producing science from data. Attendees considered cloud-hosted spreadsheets (e.g., Google Docs) to be an advance due to their built-in collaborative features. Some attendees suggested using Tableau (http://www.tableausoftware.com/), a popular data analytics package, for its power and scalability. But others noted that this tool did not seem to be a good fit with biological data and was best used for operational metrics and cube-structured data.

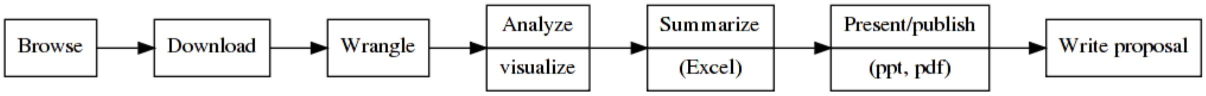

Figure 1. Generalized data-focused science production

With human genomics data, scalability is an impending concern. While a sequence for a single genome can be described in “only” about 30 GB, sequencing the genomes of 100 million people quickly adds up to 3 exabytes. Furthermore, doctors may need to take multiple sequences to get the information needed for individual cancer patients because a tumor's genetic fingerprint is often unique and may evolve rapidly. One biologist pointed out that effective compression techniques may be very helpful in reducing genomic data storage needs because the human genome is 99 percent identical among individuals. Multiple sequences for one person should be highly compressible while genomic information from different people may be more difficult to compress. Another biologist countered that funding agencies are applying considerable pressure on scientists to archive the raw measurement data from sequencing machines, even though such noisy and poorly compressible data are rarely used. Since genetic sequence data is still relatively new to biologists, some have considerable concern about discarding data before exhausting their analyses. Fortunately, funding agencies are beginning to understand the tradeoff between keeping all data versus having more data storage capacity available to collect more sequences and do more science. Although sequencing data has significant uncertainties, genomic theory has placed an upper bound on the information content in each sample.

The rapid development of gene-sequencing technologies is also producing challenges familiar to those who worked in the early days of computing: new sequencing machines are continually being developed that are dramatically reducing the cost of genome sequencing but delivering their data in constantly changing forms. Different machines employ different chemistries, different precisions, different probabilities, and different error models. Even if the community agreed upon a standard data format, integration would still remain difficult because similar measurements from different machines have slightly different data models. This diversity of instruments and data makes integration difficult for even data-management-savvy biologists. Subjective inconsistency and variability in medical attributes (for example, the different ways that individual doctors annotate electrocardiograms) also complicate defining and adopting standards. One attendee described how diverse data was stored using MongoDB, archived to Amazon's Glacier, and retrieved only when needed for analysis. This approach kept data safe and accessible but preserved their diversity in form, effectively deferring integration to each scientist's analysis.

Sequence data across non-human species is another significant problem that is likely to increase as human sequencing matures. Soil genomics (and more generally, population genomics) was one example. Because they are concerned with genes in entire populations rather than a single species (e.g., human), soil geneticists encounter problems of scale that are qualitatively and numerically different—compression is less obviously effective for genomes of countless species that likely undergo far greater rates of mutation.

Provenance is a growing problem in biology. Historically, scientists shared data in papers and described their provenance in detailed prose. As new instruments and equipment have enabled scientists to collect more and different types of data, provenance has not caught up, except at the most sophisticated, large institutes, such as the National Center for Biotechnology Information. Wet lab results can vary drastically, so recording their provenance, as well as that of the samples, is essential. For example, the same sample processed at two different labs will often yield answers that differ by up to 20 percent while processing the same sample at three places might result in a spread as high as 40 percent.

Biologists want to migrate from hypothesis-driven science to hypothesis-free science, but the hardware, software, and cultural infrastructure are not yet ready. Too often, developing a new algorithm leads to re-reading the entire dataset, which many attendees deemed impractical.

Overall, biology and health care attendees identified six main problems. One, sociological not technological factors are impeding integration among multiple data sets. Two, incoming data come too quickly (velocity) and in too many forms (variance) to be handled easily. Three, the lack of online, single-pass algorithms causes significant delays in analysis and stymies iterative analysis and discovery. Four, a poor understanding of what data are needed for reanalysis results in bloated and unnecessarily wasteful data archival. Five, with a few exceptions, poor provenance practices make reproducing experiments difficult, which in turn leads to unnecessary repetition of past work. Six, there is no widely-known repository for computational biology use cases.

Attendees agreed that one way to improve the computing and data management in biology and health care is to establish evaluation criteria for measuring and comparing various implementations and proposed solutions. One suggestion is to prototype a database-based solution by ingesting existing public data (such as 1000Genomes: http://www.1000genomes.org/) into a database and reproducing the analysis described in an ENCODE: http://genome.ucsc.edu/ENCODE/) paper. This has not happened yet partly because the reference data supporting these papers are typically not archived publicly. Although papers are required to provide URLs to their data, the authors' and publishers' neglect has left the linked data unusable. Another reason is that the extract-transform-load (ETL) process is inherently expensive, amounting to about 70 percent of the cost of installing a data warehouse, according to one participant. Despite these difficulties, attendees believed that prototyping such a solution as a benchmark would have significant value in guiding more effective data use in biology. Open sharing of scientists' and institutions' computing and data management practices would also be helpful.

The discussion also underlined two fundamentally different perspectives on data: scientists manage data as samples and files while the database community thinks of data in terms of algebra, structures, and models.

4 IN-SITU ANALYSIS AND PREDICTIVE COMPUTING

“In-situ analysis” is defined as analysis performed alongside or embedded within modeling or simulation code. Such analysis immediately computes derived products (possibly including visualizations) from the raw simulation products to avoid storing raw data determined to be of low value. In-situ analysis provides scientists with a shorter feedback delay while they explore parameter spaces of thousands of variables in their predictive models. This type of immediate analysis will be required for exascale computing, where I/O bandwidths will likely be insufficient to dump out all the raw data for later analysis offline.

In reinventing predictive, or simulation-based, computing, scientists aim to modernize scientific computing with the same bleeding-edge technology popular in technology companies. What participants felt is lacking, however, is a vision of how to incorporate in-situ analysis and visualization into the overall workflow of simulation-based experiments, including a UI for experimenters to control and direct the process. Ideally, scientific projects would mimic the close collaboration with computer scientists that was prevalent in the Sloan Digital Sky Survey. Participants identified two reasons for Sloan's success: (a) the collaboration of arguably the world's foremost database expert (J. Gray) with a recognized science domain expert (A. Szalay), who championed the use of advanced computation (and became an expert programmer in the process), and (b) the open-access model and distribution of Sloan's data – 4,900 of the 5,000 published papers based on Sloan data came from outside the Sloan collaboration. In attempting to replicate this success in other scientific communities, major challenges will be identifying domain experts (equivalent to Szalay) and overcoming the traditional resistance to data sharing that unfortunately pervades several scientific disciplines.

5 PRACTICAL CLOUD COMPUTING

This session's moderator identified four areas for discussing cloud computing. The first was economics: how important is the time-share-style cost savings of an available elastic computation resource compared with provisioning for rare peak demands? The next was novelty: how much of the current interest is from aggressive marketing as opposed to recent technological advances? The third was deployment: what problems or environments demand private cloud deployments instead of public cloud services? The last was reliability and control: how important are potential problems with reliability (i.e., unexpected failures) and privacy (e.g., unintended data or information leakage) that are inherent in current cloud implementations, both public and private?

Regarding economics, participants agreed that while in-house resources are often cheaper than cloud resources, the calculation is not simple. A tally of in-house costs includes much more than those of just the computing equipment and administration staff. Costs associated with the physical building, land, operations, data replication, and data movement (including application performance and metered bandwidth) must also be included. In many cases, the high reliability offered by cloud providers is not necessary, and lower reliability at a lower cost is preferred. Some funding agencies, such as the NIH, do not allow project data to be stored in a public cloud. The privacy required for HIPAA compliance is also difficult in public clouds. Cloud-infrastructure is also difficult to “turn over” to forensic investigators in the event of intrusion. Still, participants claimed that cloud providers are likely to have superior security compared to most individual users and institutions. Cloud computation is also not an automatic solution for managing load spikes. The cost of moving data to/from cloud providers is often significant. Some computing loads are also difficult to port to cloud environments. That is, they would exhibit poor performance on cloud resources due to less-than-optimal generic implementation assumptions.

In general, the sentiment of XLDB participants is that cloud-aware applications are substantially more efficient on cloud hardware than applications designed without cloud considerations. The latter are usually slower than expected and occasionally too slow to be worthwhile. One cloud-specific feature example is the storage API. Legacy applications are typically written to read/write out of a shared file system for overall job input and output while clouds typically provide specific APIs for storing “blobs” of data.

One participant claimed that over the long term, operating a private cloud or batch farm in-house would cost one-third that of AWS, as long as an institution has enough computation work to occupy it fully. Cloud providers are for-profit businesses and naturally have to charge cost-plus-profit. Spikes in demand are treated as high-priority tasks/allocations, and providers assume they will have enough low-priority tasks to “backfill” during low-demand periods. Several participants questioned the claim that in most situations there would be enough low-priority tasks to fill in the “valleys of demand” but agreed that backfill would not be a problem in larger institutional environments.

Participants agreed that cloud computing's advantages are not strictly marketing hype. A key advantage is standardized machine management, which converts system administration and server operations to merely script management. Software and workflows can also be published as machine images, thus simplifying reproduction. A key disadvantage is the difficulty of moving data although one cloud provider has mitigated this network bandwidth problem by allowing shipments of physical disks for data ingestion.

Some advantages are less obvious. One participant claimed that cloud resources are especially useful for debugging because allocating jobs among virtual machines could provide insights into whether problems are related to hardware, software, or the network. Clouds' elasticity enables some to use them as on-demand prototyping platforms and others to use them as deployment platforms that naturally handle load spikes. Finally, upcoming features, such as guaranteed IOPS, can further reduce current disadvantages.

Overall, participant sentiment was that cloud computing, whether public or private, has significant advantages although there are always trade-offs. Smaller organizations and groups that did not have enough backfill should be better served by sharing private clouds rather than running their own clusters. Participants thought it would be challenging to deploy certain combinations of cloud and other computing resources. Complexity is especially problematic for scientific computing because a heavy fraction of legacy code is involved, and because scientific computing is a low-revenue market for cloud providers, less support is available for its unique issues. Also, algorithms might have to be adapted to run with the weaker consistency guarantees and slower communication found in a cloud environment.

Some users were concerned that a public cloud might have poor security compared to a private cloud or private system but were challenged by one participant, “Amazon has better security than you.” In addition, using a public cloud can make sharing of data and even complete analysis setups easier than with a private cloud.

Attendees expected the decision calculus for cloud computing to change over time. The spot market for AWS resources, for example, was created by Amazon to sell excess capacity of the resources partially purchased because third-party demand exceeded the available backfill capacity of resources to fulfill its own business operations. Some believed that while cloud costs are declining more slowly than Moore's Law, they should eventually become much more competitive as the cloud services industry matures and computing technology improvements fall short of the Moore's Law rate. Participants expected that funding agencies will eventually revise their policies to be more compatible with cloud computing. One participant noted that the NIH wants to stop funding private clusters and servers because biologists are terrible at running clusters.

6 COMMERCIAL SOFTWARE IN ACADEMIC ENVIRONMENTS

Vendors attending XLDB pressed for specific reasons for academics' limited adoption of proprietary software or hardware. Several reasons were discussed.

In academic environments, the standard practice favors trading *capex* (capital expenditures) for *opex* (operating expenditures). In other words, because funding is always tight and short-term (grants are typically for less than three years), the pressure is to minimize equipment and software purchase costs relative to operating expenses (i.e., graduate students at below-market rates). All agreed, however, that scientific output would likely increase greatly if grad students did not have to perform system administration or other tasks outside their field.

Projects are often distributed both administratively and geographically, which means there is no centralized entity to share licensing or support costs. Some project participants are the lone collaborators at their institutions. Open-source community support is often more personal, more direct, and more relevant. Several presented examples of users fixing bugs themselves although this was not particularly common.

Open-source software also encourages community involvement, whereas commercial software is viewed as a black box full of proprietary secrets. Due to the long service life of large scientific experiments, the future cost of commercial licenses is also always a worry. Guaranteeing *perpetually* low pricing for academic/non-profit use was mentioned as a worthwhile incentive that vendors should consider. The price would not have to be fixed. Rather, a simple company statement guaranteeing an x-percent discount from the regular commercial license fee would be a large step forward.

Companies also offer no guarantees that they will survive for the duration of these long projects. In contrast, open-source support communities invariably persist. Useful developer documentation is also typically more available for open-source software than for commercial programs, for which a source code dump may be all that is offered. Software escrow was not seen as a sufficient solution as scientists would eventually need to pick up support.

Historically, database vendors have not paid sufficient attention to the needs of scientific users. One participant said scientists have been burned badly by commercial software. He cited a case where there was a problem with a top-tier commercial database used by a group of scientists. The vendor acknowledged the problem but refused to fix it, explaining that the scientists' use was not the company's target use case. Scientists feel this experience was not unexpected.

Tradition is another reason for limited adoption of commercial software. Scientists have been historically poor, and only in recent years have commercial software vendors offered preferentially low pricing for academic/non-profit use. As a result, very few science educators and mentors have the background to teach commercial software. Funding agencies rarely permit line-item billing for support costs although some participants felt that this could change soon. Funding agencies are now taking data curation more seriously. They seem increasingly willing to pay for long-term data archiving and access to an organization that would guarantee its own survival.

Some participants questioned the proposition that proprietary systems would lower costs. Proprietary systems generally assume some intellectual IT infrastructure that generally does not exist in academia. Usually, each scientist is her own IT staff. Participants also agreed that enterprise software companies typically assume the existence of an enterprise IT department that would provide some insulation between users and vendors.

Cooperation with commercial vendors in testing and support was raised as one possible solution. Participants noted that scientific communities are generally prompt and detailed in their feedback, but the litmus test would be whether a company accepts fixes from its user community. Some attendees resented the idea of being free beta testers for vendors, however, especially before the vendors have proved the advantages of their proprietary software.

Transparency in open-source software communities is a key advantage that usually outweighs the often-unfinished functionality of the program itself. Open-source users share usage, problems, and solutions freely in open communities. In contrast, discussions in commercial software communities are typically filtered and limited if the companies are even willing to share their usage at all. Participants noted that we are now observing a dramatic shift: almost every new DBMS is open source while every old one is closed source.

7 FILE SYSTEMS VS. DATABASES

The workshop began with an unplanned discussion on why scientists overwhelmingly store and manage data as files in a file system while databases are mostly unused. Several factors were identified. Probably the biggest factor is a continuing lack of database training in science curricula, whereas files and hierarchical directory structures are common computer knowledge. Some projects place their experiments' metadata in databases, but this practice is only adopted by the most advanced scientists. Most scientists instead encode metadata using strict file and directory naming conventions. An attendee explained that while file naming is cumbersome, it is also transparent and that it is better to teach people how to ask better questions than to teach them how to find files using a database. Another factor is the belief that files are the “truth” and databases would only be a cache placed on top. Scientists consider files much easier to read and manipulate, whereas databases are mysterious and unpredictable. Files can be moved, exchanged, manipulated, and viewed while mobile and organized in obvious ways while databases are perceived to require professional management. Files have well-understood failure modes while database problems are cryptic. Another attendee said database use is further deterred by their incompatibility with specialized analysis tools that are irreplaceable for many scientists.

Some researchers have recognized the value in database approaches for organization and have eliminated human-understandable directory- and file-naming in favor of databases containing file metadata that refer to UUID-named files. Still, one scientist insisted that while metadata and location information could be placed in a database, he would never place “primary data” there because there was no added value in doing so. He admitted that databases and SQL have good capabilities for searching for data, but scientists are not well-trained in taking advantage of such features. Furthermore, SQL semantics are not well-suited for complex scientific data analysis. Hybrid approaches, where subsets of scientific algorithms are pushed into the database, are significantly more complex and not guaranteed to outperform.

Another reason for scientists' reliance on files is a divide between data producers and data users. The common practice is for a data producer, e.g., a satellite operations center, to make data available for download but not to provide an integrated platform for analysis. Thus the integration problem is pushed to individual users and groups who are likely to be even less equipped to integrate data into a managed platform, such as a database. Thus nearly all (if not all) downstream software tools operate on files not databases.

The representatives of the database community attending the workshop admitted that their larger community should "*recognize that files are critical pieces of data and deal with it.*"

8 NEW COMMUNITY: NATURAL RESOURCES

Three new participants engaged the 6th XLDB workshop with a new technical community: natural resources. They were: a representative from IBM's Smarter Planet campaign, a researcher who uses water management and monitoring for recreation, weather, and climate prediction, and a scientist studying snowpack and remote sensing. Overall, commercial interests have use cases, methods, and difficulties similar to those in the geoscience community, which was introduced at a past XLDB workshop. Commercial interests seem to be focused on resource extraction rather than resource management, however. Data problems in extraction are characteristic of a community just beginning to integrate data-driven practices. Large-scale data in this community are dominated by information collected by equipment instrumentation and new sensing technologies, which are growing rapidly.

As in other disciplines, the natural resource representatives acknowledged their need for a framework for accessing and integrating multiple disparate data sources. These might include multiple satellites or sensors, a combination of climate models and remote sensing, or data presented in different formats (e.g., sensor data and human-recorded logs). This is not simply a "data integration" challenge because properly combining information from multiple sources may be analysis-specific and require significant user input. Time series processing is used to build predictive models from historical data.

The natural resource extraction representative noted that data analytics has already been shown to produce dramatic cost savings in his industry. The scale of mining operations and their equipment means that even relatively simple improvements in efficiency and equipment management can easily translate into multimillion-dollar cost savings. There were two examples. The data-driven management of spare parts saved one mining company $200 million annually. Another $100 billion-a-year company saved $2 billion by unifying and optimizing its supply and production data. While new sensing technologies and forms of data are not trivial to manage, there was little evidence of intractable scale in the problems described. Instead, most of the effort was spent coping with new applications of existing data and new forms of data never measured before, such as sensor data from struts of huge ore hauling trucks and poor or inconsistent data quality of human-recorded data (e.g., from equipment inspections between shifts).

The institutions providing remote sensing data do not expend significant effort on integration. Instead, they are concerned with the diversity of analysis that their users desire. Data movement is a significant problem as well. Scientists also preferred HDF5 or NetCDF files as "transferable wads of information" over perhaps-more-integrated databases.

9 OTHER TOPICS

Data management associated with large physics collaborations, such as the LHC, was mentioned as far more cost-efficient despite the huge price tags. The efficiency comes from sharing knowledge and resources.

Some attendees argued that billing scientists for their real CPU and I/O usage at finer granularity would create incentives for devising the efficient, scalable algorithms needed for exascale computing. They said the current model encourages inefficient programming by focusing on maximizing the output of fixed-cost purchased resources (e.g., trying to scale out to X cores, rather than maximizing the effectiveness of each CPU cycle). Scalability is still important, however, because all new "big iron" petascale and exascale machines are, in fact, "scaled-out" clusters of simpler nodes, rather than more powerful "scaled-up" single machines.

Vendors asked science representatives which software languages they used and needed, noting that tools or languages that do not match problems are irrelevant. Attendees replied that there is no small subset of languages used across all scientific communities. Each community has its own traditions for using general-purpose languages and tools, some of which are domain-specific. Participants speculated that scientists have too little time to risk experimenting with tools not already used within their communities.

Database vendors admitted that their community has largely ignored UI design. Science representatives agreed that a good UI could trigger greater adoption, noting that web browsers were adopted quickly without much resistance.

A recurring theme in the workshop was the problem of poorly organized data. While a big problem regardless of scale, poor data organization causes exceptionally difficult problems at huge scales and seems to be steadily worsening. Participants noted the common problem of data "hidden" in poorly organized files. One attendee related the challenge of dealing with data that a retiring technician had kept in 30,000 spreadsheets, innumerable emails, and even some screenshots. Solutions suggested for this problem range from providing a search facility for unstructured data to creating and requiring a comprehensive system that imposes structure on the data from the moment it is created.

A few topics were suggested for future discussion:

- Computer memory hierarchies. In current implementations, main memory (traditionally DRAM) is now very far from the CPU, a reality that is at odds with the logical computing model of CPUs executing instructions read from and manipulating data stored in main memory.
- Estimation of data value and mobility. A conference panelist noted that data is stored in cheap, but slow, storage, from which it is analyzed until its value (a per-byte quantity) is great enough to duplicate and write in higher-performance locations, such as databases. Participants noted that it would be useful to have metrics that would quantify and identify when it would be beneficial to move data from storage to databases, even when large amounts are involved. The panelists did not describe a means of estimating such value, nor did workshop participants have any answers.
- Data publishing and providing open access to data. Involvement of government representatives was requested for this topic.
- Data in files versus data in databases. Participants wanted further exploration, specifically addressing the questions of what data could be placed in databases and how databases could support data files.
- Integration of databases and domain-specific tools, languages, and APIs. Participants felt a collaboration of data scientists and database researchers could make considerable progress in integration, especially if facilitated by XLDB.
- In-browser data analytics. Data analysts and database researchers need to discuss what is needed from each of their communities to enable in-browser data analytics.

10 NEXT STEPS

Participants found the 6th XLDB workshop to be useful and not needing major changes. The format, length, location, and timing were perfect. One suggested tweak was to choose one or two topics that would be re-visited regularly at workshops and process updates. The written reports were found to be useful, in particular by new participants who are orienting themselves with XLDB. While most want the workshop to continue to be loosely structured, one said that a stronger structure might reduce unnecessarily-repeated discussions and would be more sustainable.

11 ACKNOWLEDGMENTS

The XLDB-2012 event was sponsored by MetaScale, Vertica/HP, LinkedIn, ExxonMobil, IBM, InterSystems, Chevron, Cloudera, SciDB&Paradigm4, Hypertable, and MonetDB. Additional funding was donated by the Gordon and Betty Moore Foundation.

The XLDB-2012 event was organized by a committee consisting of Jacek Becla, (SLAC, chair), Vivien Bonazzi (NHGRI/NIH), John Chambers (Stanford University), Scott Klasky (ORNL), Steve Lacy (ExxonMobil Chemical), Carol McCall (GNS Healthcare), Chris Mentzel (Gordon and Betty Moore Foundation), and Daniel L. Wang (SLAC).

The report was edited by Michael Ross.

This work was supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515.

GLOSSARY

API – Application Programming Interface
AWS – Amazon Web Services
CPU – Central Processing Unit
DOE – Department of Energy
DRAM – Dynamic Random-Access Memory
ENCODE – Encyclopedia of DNA Elements
GPL – GNU General Public License
HIPAA – The Health Insurance Portability and Accountability Act of 1996 (P.L.104-191)
HDF5 – Hierarchical Data Format 5
I/O – Input/Output
IOPS – Input/Output operations Per Second
IT – Information Technology
LHC – Large Hadron Collider
NetCDF – Network Common Data Form
NHGRI – National Human Genome Research Institute
NIH – National Institutes of Health
SQL – Structured Query Language
UI – User Interface
UUID – Universally Unique Identifier
XLDB – eXtremely Large DataBases

(Article history: Received 12 March 2013, Accepted 19 March 2013, Available online 8 May 2013)