REPORT FROM THE 4th WORKSHOP ON EXTREMELY LARGE DATABASES

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

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

ABSTRACT

Academic and industrial users are increasingly facing the challenge of petabytes of data, but managing and analyzing such large data sets still remains a daunting task. The 4th Extremely Large Databases workshop was organized to examine the needs of communities under-represented at the past workshops facing these issues. Approaches to big data statistical analytics as well as emerging opportunities related to emerging hardware technologies were also debated. Writable extreme scale databases and the science benchmark were discussed. This paper is the final report of the discussions and activities at this workshop.

Keywords: Analytics, Database, Petascale, Exascale, VLDB, XLDB

1 EXECUTIVE SUMMARY

The 4th XLDB workshop (XLDB4) focused on challenges and solutions in the oil/gas, finance and medical/bioinformatics communities, as well as several cross-domain big data topics.

The three domain-specific panels expressed similar concerns about an explosion of data and limits of the current state of the art, despite having different applications and analyses. All three communities (and others present) were struggling with these challenges: integrating disparate data sets including unstructured or semi-structured data; noise and data cleansing; and building and deploying complex analytical models in rapidly changing environments. The oil/gas exploration and production business analyzed petascale seismic and sensor data using both proprietary rendering algorithms and common scientific techniques like curve fitting, usually with highly summarized data. The refining and chemicals business had terabyte, but growing, datasets. Most processing of historical financial transaction data was offline, highly parallelizable, and used relatively simple summarization algorithms, although the results often fed into more complex models. Those models may then be applied, especially by credit card processors, to real-time transactions using extremely low-latency stream processing systems. High-throughput sequencing and other laboratory techniques as well as increasingly electronic medical records (including images) produced the large datasets in the medical/bioinformatics field. Applications here included shape searching, similarity finding, disease modeling, and fault diagnosis in drug production. The medical community was striking for its non-technical issues including strict regulation and minimal data sharing.

Progress was made on the science benchmark that was conceived at previous XLDB workshops. This benchmark was created to provide concrete examples of science needs for database providers and to drive solutions for current and emerging needs. Its specifications and details have now been published. The next iteration will go beyond processing of images and time series of images to include use cases from additional science domains.

Statistical analysis tools and techniques were reportedly insufficient for big, distributed data sets. First, statistical tools should be developed to scale efficiently to big data sizes. Second, approximating and sampling techniques should be used more often with large data sets, since they can reduce the computational cost dramatically. Finally, existing statistical tools should be made easier to use by non-specialists.

New hardware developments have made big data computation more accessible though uncertain in some ways. Power was the biggest issue and one that would drive the future of hardware as well as analysis. Regarding

performance, more evidence of the potential speedup from GPU computing was shown through examples of complex computations within SQL databases. Attendees were enthusiastic about new storage technologies like solid state disks but disagreed on whether they would displace older, proven technologies. They similarly disagreed on whether many-core processing (hundreds to thousands) would begin to replicate core+memory on-chip rather than the current, simpler model.

True updates were rarely needed in extreme-scale databases since the vast majority of big data sets were immutable. Wherever data updates were necessary, attendees preferred data versioning and history over updating data in-place.

The next workshop is expected to convene in the fall of 2011. Reference case studies, high performance extreme-scale visualization, data simulation, and cloud computing were among most demanded topics.

2 ABOUT THE WORKSHOP

The Extremely Large Databases (XLDB) Workshops provide a forum for topics related to databases of terabyte through petabyte to exabyte scale. The 4th workshop[1] (XLDB4) in this series was held at SLAC in Menlo Park, CA on October 5, 2010. The main goals of the workshop were to:

- reach out to the communities under-represented at the past workshops, in particular oil/gas, finance and medical/bioinformatics,
- review special topics in big data analytics: approaches to big data statistical analytics, opportunities from emerging hardware technologies, and writable extreme-scale databases, and
- discuss the advancement of the state of the art of XLDB.

This XLDB workshop was followed by a 1.5 day open conference which was attended by 150 people. This report covers only the workshop portion. Information about the conference sessions, including the presentations, can be found at the conference website[2].

2.1 Participation

Like its predecessors, XLDB4 was invitational in order to keep the group both small enough for interactive discussions and balanced for good representation among communities. XLDB4's attendance numbered 55 people representing science and industry database users, academic database researchers, and database vendors. Industrial user representation was greater compared to past workshops. Further attendance details are on the website.

2.2 Structure

Continuing the XLDB tradition, XLDB4 was composed of interactive discussions. The first set were panel discussions on domain-specific challenges and solutions. Next were discussions focused on specific cross-domain big data topics. The concluding discussions reviewed the science benchmark and plans for the next XLDB.

3 USER COMMUNITIES' PERSPECTIVES

XLDB4 involved several new user communities in the areas of oil/gas, finance and medical/bioinformatics. The first two of these three had never been represented or discussed at past workshops. Bioinformatics was discussed at XLDB3, but the topic was expanded to include new areas such as medical informatics and biosecurity.

Many similarities to other domains raised in the past were noted, including needs for incremental scalability, full automation of operations, fault tolerance, approximate results for queries, and software simplicity. Regarding the issue of using a database, particularly a relational one, versus analysis with a map/reduce framework, it was noted that adding SQL-like features to non-database solutions such as Hadoop increases users' interest in these solutions, in some cases by as much as a hundredfold.

[1] Workshop website: http://www-conf.slac.stanford.edu/xldb10/Workshop.asp
[2] Conference website: http://www-conf.slac.stanford.edu/xldb10/

3.1 Oil / Gas

The oil/gas panel consisted of representatives from two large multinational corporations — Exxon Mobil and Chevron — and TechnoImaging, a company recently spun-off from the University of Utah's CEMI[3]. Some panel members were from the *upstream*, or oil exploration and production, portion of the business and others came from the *downstream*, or oil refinement and chemistry portion.

The industry has had petascale data for some time now, having "more bits than barrels" as one panelist put it. The largest data sets are from upstream users and consist of long-term seismic measurements. They are used to build 3-D earth models at the highest possible resolution. Smaller, but rapidly growing, data sets come from instrumented wells, production facilities, refineries, and chemical plants. These sensor readings have typical sampling rates on the order of 1 per minute except when detection of certain transient features requires higher rates (~100Hz); their total data volumes may be in the dozens of terabytes. The growth in data volumes comes from faster and cheaper sensors and more wells that need monitoring. In some cases, e.g., passive sensing, data are unavoidably noisy and their sources poorly known, and a large amount of metadata is required to make them useful. Recorded data are typically de-sampled and compressed immediately after collection.

Common analyses include stacking multiple images together to achieve higher resolution, creating depth images through reflection coefficients, and curve fitting. These are similar to techniques used in astronomy and other sciences. Pattern-matching is used to explore new regions by comparing them to well-understood regions. Seismic data rendering is reportedly similar to movie industry rendering done by Pixar or DreamWorks, but these are difficult to compare since algorithms in both industries are proprietary. Most complex analyses on larger data sets are hard-coded as fixed pipelines; ad-hoc querying is highly desired but thought to be too difficult.

Industry insiders considered their data analytics practices to be "stone age" (in some cases dating back to the 1980s). For example, disk is often still viewed as a precious resource, so data analyzers receive only very highly summarized data. Existing tools make it hard to extract, analyze and visualize data. The community relies mostly on off-the-shelf software; among the tools mentioned were Paradigm[4], Schlumberger's Petrel[5], Halliburton's OpenWorks[6], Spotfire[7], Apache Tomcat Application Server, Matlab, Oracle, and SQL Server. Legal considerations are strong barriers to evaluating, let alone deploying, new software.

The oil/gas community reported its biggest problem to be poor data integration. Data sets originate from multiple sources world-wide and often cannot leave their countries of origin. Data sources and schemas are typically completely disjoint (e.g., brought in through acquisitions and never properly merged). However, upper management is pushing for data to be cleaned and better integrated since the cost of building new wells is increasing and the payoff of more educated decisions is apparent. Another reason for better data standardization is data exchange (driven by cost-cutting) between contract parties, like government agencies (such as USGS) and the oil/gas companies.

Other problems were poorly or inconsistently formatted data, difficulties in synthesizing different types of data, poor tools, and poor tracking of past work. Important data is often unstructured and/or in forms difficult to analyze, such as PowerPoint or XML, and analysis involves "art in interpretation." Analysis is further complicated by the variety of available aerial models (electric, acoustic, atomic, magnetic, seismic and others). Additionally, the lack of appropriate tools and procedures to capture, preserve and query data provenance results in unnecessary repeats of similar experiments.

3.2 Finance

The finance community was represented by users from JP Morgan Chase and VISA. They revealed different processing models, one dominated by low-latency lightweight transaction processing (credit card processing)

[3] Consortium for Electromagnetic Modeling and Inversion, http://cemi-dt-13.gg.utah.edu/~wmcemi/
[4] http://www.pdgm.com
[5] http://www.slb.com/services/software/geo/petrel.aspx
[6] http://www.halliburton.com/ps/Default.aspx?navid=210&pageid=852
[7] http://spotfire.tibco.com/

and another dominated by big offline data analysis (banking). Petascale data sets are not uncommon, especially where historical data must be kept for regulatory reasons (e.g., within banks). Data sets are naturally divisible, typically into geographical "zones" containing ~1 petabyte each. Zones are further subdivided into ~20 terabyte pieces to bypass limitations of existing off-the-shelf systems used for analyzing the data.

Computation at credit card processors was dominated by extremely low-latency stream processing on many small pieces of data. In these systems, each transaction performs ~400 jobs ("joins") that process encrypted data (including hashes of names and account numbers) against recent (within 1 year) customer data without a database. Each transaction operates in a stream that processes up to a thousand transactions per second, and streams are typically grouped in 100 stream units. Because of tight latency requirements, these streams operate only on a terabyte, keeping everything in RAM. Card processors also performed offline analysis on larger, petabyte data sets that include longer time ranges (up to 10 years) for tasks like building neural nets and risk models, but "anything that makes money" is a stream process. Processing at banks, on the other hand, was dominated by heavy offline, pro-active analysis on cumulative historical data: continuous risk calculation, fraud detection, and pattern analysis. Compared to card processors, banks kept more data of a larger number of different types (e.g., binaries like check images).

Offline analysis is highly parallelizable — simultaneous runs of >100 streams are typical. Time-based analyses (monthly, daily, 10 min, 30 sec batches) are the most common. A typical process relies on basic commands such as `cat`, `grep`, `awk`, or `sort`. 80% of processing can be classified as simple grouping and sorting, independent of the type of analyses run.

The finance community reported that the required performance is usually achieved through expensive brute force in both hardware (e.g. hardware accelerators) and software (high end, vendor managed). For security reasons only private clouds tightly sealed with firewalls are used. Banking data centers tend to be as large as those run by web companies (O(100K) nodes), although the number of them is much smaller.

A typical analysis software stack includes dozens of different off-the-shelf programs ranging from Hadoop, through R, to Oracle, DB2, and Teradata. Custom C++ code is also prevalent, although it is gradually being replaced by Python equivalents (observing a 40:1 reduction in lines of code). R is used primarily with smaller (few gigabytes) data sets because of its complexity when used in full-stream processing. Different geographical areas are analyzed independently and never merged together.

The main problem cited by banking was just accessing data quickly—not just large scale data, but data buried in spreadsheets too. Another tough challenge cited was the construction and deployment of reliable models, e.g., those for fraud detection. Models are manually built in-house by modelers who typically do not truly understand what is modeled *and* how to model it. Their deployment is complicated when models exhibit anomalous behavior when using live versus test data. Models help the system do the right thing despite unique conditions where the best customers look like the worst customers — frequent travelers trigger many false fraud alerts, but use credit cards the most. Other important needs included high availability, hot-hot failover, as well as role-based and row-level access.

3.3 Medical / Bioinformatics

The medical/bioinformatics panel was represented by institutional users from NIH (molecule screening), two children's hospitals (medical records, analyzing proteins), U.S. Department of Energy (cybersecurity, genomics), and NASA (early disease detection). The medical and bioinformatics communities reported a data explosion similar to other domains, caused by similar reasons: cheaper and higher-resolution instruments. All reported some degree of unpreparedness for the scope and scale of emerging data volumes.

Examples of medical and bioinformatics analytics include shape searching, similarity finding, disease modeling and analysis, and fault diagnosis in drug production. The last requires detailed provenance tracking of data from highly distributed sources, and was reported as the most demanding provenance-related use case so far.

The most striking difference between these two communities and the rest is (lack of) research cohesion. Both communities reported wide dispersion and fragmentation, with many small groups competing instead of

collaborating. Data is rarely shared due to ethical concerns and extreme regulations, as well as due to a desire to protect research that might yield valuable publications. As some noted, most problems are related to humans rather than software. They complained of the culture of cutthroat competition and non-sharing rather than collaboration but did not blame the researchers, agreeing that it seemed necessary for survival in their funding and leadership structures.

Data quality is a big problem. Because hardware and processing practices at wet-laboratories change frequently, sometimes from one month to the next, there is little time to achieve production and process stability. The arrival of new hardware means new proprietary formats and changes to the nature and form of collected data. Even if the instruments and formats were stable, there is considerable variation in its collection. In many cases, in particular when medical records are involved, data is observational in nature — it is noisy and collected inconsistently without any enforced standards. Thus data are *not* recorded with science and analytics in mind. Later analysis is complicated by missing data that is biased in statistically significant ways — for example, a doctor might determine that patient is not sick enough and not collect certain data. "Negative results" are often not collected because they are useless for publication, even though they would be valuable for future statistical analysis.

4 SCIENCE BENCHMARK

The concept of a science benchmark was first introduced at XLDB2. The idea behind it is to capture the essence of science data processing and analysis, including not only querying processed data but the processing or "cooking" of raw data itself. The benchmark would highlight areas that are not well-served by traditional RDBMSes and would serve both as a repository of abstract, general, multidisciplinary use cases and as a spur to database developers to provide features that are useful to science and science-like industry.

The current version of the benchmark covers one such area: processing of images and time series of images. The data and queries are based on astronomy and in particular the LSST project, but they are designed to be general enough to represent the needs of similar domains such as geoscience and medical imaging, although the exact alignment between the benchmark and those domains needs to be determined and may require adjustment. Example queries include detecting objects in images, resampling an image onto a different pixel grid, and finding intersections of object trajectories with regions of space and time. The benchmark can be scaled to different levels of computational difficulty and data size, enabling measurement on systems from a single computer up to a large cluster.

The team working on the benchmark submitted a paper to ICDE'11, and the paper was made publicly available shortly after the workshop through the XLDB4 website[8]. The data generator and a sample implementation on top of MySQL are also available on the XLDB wiki[9].

At least three providers (SciDB, Greenplum, MonetDB) expressed interest in trying to run the benchmark. Broadening the buy-in from science and engaging more science disciplines were discussed as the most important steps to make the benchmark more useful. Other possibilities included a text-oriented benchmark emphasizing UDFs and including more of the overall data management process. Building a TPC-like organization that would coordinate the benchmark effort was considered but was deemed premature at the current level of momentum.

5 APPROACHES TO BIG DATA STATISTICAL ANALYTICS

During this session, attendees discussed statistical analytics in a big data context, discussing the main problems, current practices, and future directions. Statistical methods are widely used in many areas like forecasting, bio-defense, and web user modeling. A few truths seemed obvious: (a) the current methods should be more scalable, simpler, and more accessible to non-experts; (b) use of analytics is widespread and becoming more so; and (c) the data volumes for analytics have long grown past the point where simple hardware upgrades are sufficient for large data sizes — new techniques must be used.

[8] http://www-conf.slac.stanford.edu/xldb10/docs/SSDB_benchmark.pdf
[9] https://confluence.slac.stanford.edu/display/XLDB/SS-DB+Benchmark

Computational statistics, like most computational applications, is still in transition to software that can scale and deal with the practicalities of big data. Most methods assume that data sets fit in a machine's main memory, and are only applicable to big, distributed data sets through much pain. Success requires an awareness (perhaps an intimate one) of implementation details like data partitioning schemes (big data sets are invariably partitioned), infrastructure details (e.g, topology, memory size), and runtime hardware failure. Few who have such an awareness/skill are also statisticians, and it does not seem practical in the long term for statisticians to worry about implementation details. Good solutions must be built, and in the near term, statisticians need to get involved in new areas, such as databases, visualization, or exotic hardware technologies. Solutions for some hard-to-parallelize algorithms, like machine learning, will be difficult to build, but solutions for most algorithms should be tractable.

The off-the-shelf statistics software systems used by statisticians, in general, have not embraced big data and are difficult if not impossible to use with large data sets. Attendees wished for more statistics capabilities in languages familiar to non-statisticians (like SQL), but cautioned that education was necessary because powerful tools are "dangerous" in the wrong hands. Still, statisticians are reluctant to learn new, more scalable methods because they are "stuck" in software systems such as R, SAS, and MATLAB that took extraordinary effort to master but that are extremely productive on desktop-sized datasets. To save human time, analytics need to be as automated as possible, and statistics functions need to be more widely available (e.g., in scalable tools). Attendees repeatedly called for merging the power of statistical tools with the scalability of Hadoop.

The use of analytics is so widespread that large organizations (especially in industry) now perform "analytics of analytics" to share knowledge, avoid duplication of effort, optimize resource usage ("avoid 15 identical jobs each touching the same petabyte of data"), and connect clusters of people doing similar analytics, in internal LinkedIn-style social networks.

Some have dealt with big data volumes by not persisting them. Instead, they perform continuous analytics on live data streams and visualize the results directly without persisting them. Certain data characteristics — variability, for example — are difficult to visualize, however.

To reduce the data and computational intensity, some participants pushed for more exploration of approximate results because they can be computed so much more quickly and because perfect results are nearly impossible in the presence of faults and "messy data". Others cautioned that approximate results (from probabilistic algorithms or sampling strategies) can be misleading and could easily be interpreted incorrectly, warning of a "slippery slope." Yet it is clear that, at least in some cases, computational costs can be reduced by using simpler algorithms, especially with bigger data volumes. Attendees cited anecdotal evidence that simpler models generalize and produce better results, noting that real data is messy and additional variables add big human costs to understanding. Research into new models and algorithms is hampered, however, by the limited availability of large, freely-distributable data sets.

6 EMERGING HARDWARE TECHNOLOGIES

Undeniably, the appetite for data is "growing faster than memory gets cheaper". Flash memory-based solid state disks (SSDs), with their fast random access and potential for high bandwidth at a relatively low price, are attempting to satisfy the need for increased performance while keeping up with the hunger for larger sizes. They are quickly appearing in production systems, but opinions are highly divided on whether or not flash memory or other storage technologies like memristors, can "change the curve" of computing by displacing old, cheap, proven technologies. Opinions are also divided regarding the future direction of general-purpose multi-core processing — many argue that replicating units of cores and memory is more likely than continuing the expansion of the number of cores per physical CPU package.

After demonstrations of their use and effectiveness by several disciplines, GPUs are now commonly considered as a means to accelerate data processing and analysis. Dividing tasks into small, highly-parallel units executed on GPUs has the potential to drastically speed up many applications, including complex computations within SQL databases, as a team from JHU demonstrated. In their case, moving computation to GPUs meant the use of

different algorithms; they pointed out that the tree-based algorithms commonly found in database processing were inefficient at bin sizes small enough to fit in GPU implementations.

Some claimed that increasing network bandwidth speeds would be a game-changer for analytics, as the currently available 10Gbps (or emerging 40Gbps) bandwidth is close to the speed of local storage. Fast networks will certainly enable better virtualization and streaming.

The biggest issue for everyone was power. The most common techniques mentioned to limit consumption were: (a) eliminating computing (thinking before computing, eliminating useless queries), (b) optimizing computing (building power-optimized software), and (c) optimizing hardware (lower-power CPUs, GPUs instead of CPUs, SSDs instead of spinning disks). Power considerations were the most likely to decide the shapes of future analysis and hardware.

7 WRITABLE EXTREME SCALE DATABASES

The session on writable extreme scale databases exposed a limited need for true updates on extreme scale databases. Participants noted that the vast majority of data was immutable, primarily because published results must always be reproducible — nobody dared to update raw data. Derived data products, however, often require updates. For example, LSST will need to update some portion of its derived data products daily while tracking fast-moving and fast-changing astronomical objects. In most cases where updates are necessary, projects are choosing to append and track lineage instead of updating data in-place.

All agreed that guaranteeing true consistency at a large scale is too hard and too expensive, and thus users have increasingly accepted weaker data consistency, relying on provenance to recover from the unexpected.

In summary, the session underscored the needs for tracking versions, history, and provenance reliably and did not expose any new big challenges related to updating large data sets.

8 NEXT STEPS

As in the past, a small portion of the workshop was devoted to future planning.

The future of the science benchmark was discussed. The next steps include publishing the ICDE'11 submission (done immediately after the workshop), publishing the benchmark along with an explanation of how to synthesize the input data, incrementally improving the benchmark using community feedback, and aligning the benchmark with additional science disciplines. Finding similarities between the benchmark and industrial needs, in particular from big areas such as health care, was viewed as a positive step forward. Increasing awareness of the benchmark should encourage vendor competition to support scientific needs.

Participants were once again overwhelmingly satisfied with the value of the workshop and thought it should continue. They agreed that the next workshop should be held in the fall of 2011 in the San Francisco Bay Area and that the current format and length should be left unchanged. They suggested that XLDB5 cover reference cases, high performance visualization for extreme scale data, analytical (including extract-transform-load) tools, simulation data in science (e.g., climatology) and industry (e.g., automotive industry), and cloud computing. The “reference cases” were particularly highly demanded; participants envisioned an in-depth examination of at least two concrete examples of built systems, one where data is federated, and one where data is kept in a single instance. They hoped to learn about the architecture, fault tolerance and data replication strategies, and the tactics of getting daily analytical jobs done. The attendees suggested including representatives from health care, pharmaceutical research, the movie industry, the automotive industry, the census, and national intelligence.

ACKNOWLEDGMENTS

The XLDB4 workshop was sponsored by SciDB and Zetics, Greenplum, IBM, eBay, LSST, and Aster Data.

The XLDB4 was organized by a committee consisting of Magdalena Balazinska (University of Washington), Jacek Becla (SLAC, chair), Jeff Hammerbacher (Cloudera), Peter Fox (Tetherless World Constellation), Kian-Tat Lim (SLAC), Raghu Ramakrishnan (Yahoo!), and Arie Shoshani (LBNL).

GLOSSARY

CEMI – Consortium for Electromagnetic Modeling and Inversion
ETL – extract-transform-load
DOE – Department of Energy
GPU – Graphics Processing Unit
ICDE – International Conference on Data Engineering
JHU – Johns Hopkins University
NASA – National Aeronautics and Space
netCDF – Network Common Data Form
NIH – National Institutes of Health
RDBMS – Relational Data Base Management System
SKA – Square Kilometer Array
SSD – Solid State Disk
TPC – Transaction Processing Performance Council
UDF – User Defined Function
USGS – U.S. Geological Survey
XLDB – eXtremely Large Data Base