Scientific Data Management at the Johns Hopkins Institute for Data Intensive Engineering and Science

Yanif Ahmad, Randal Burns, Michael Kazhdan, Charles Meneveau,
Alex Szalay, Andreas Terzis
{yanif,randal,misha,terzis}@cs.jhu.edu, {meneveau,szalay}@jhu.edu
Johns Hopkins University, Baltimore MD 21218, USA.
http://idies.jhu.edu

1. INTRODUCTION

Scientific computing has long been one of the deep and challenging applications of computer science and data management, from early endeavors in numerical simulation, to recent undertakings in the life sciences, such as genome assembly. Complex computational problems abound and their solutions transform our understanding of the physical world. The data management community's interest in scientific applications has grown over the last decade due to the commoditization of parallelism, diminishing system administration costs, and a search for relevance beyond enterprise applications.

Research in scientific computing raises non-technical challenges, such as overcoming the paucity of resources needed for experimentation, and establishing a collaborative research agenda that fosters a mutual appreciation of problems, results in a concerted effort to develop software tools, and makes all researchers successful in their respective fields. In light of this, we report on a recently formed institute at the Johns Hopkins University to further the interaction between computer science, and science and engineering. We describe ongoing projects at the institute and our collaboration experiences.

2. IDIES

The Institute for Data Intensive Engineering and Science (IDIES) was founded in April 2009 to bring together data intensive computing in a variety of science and engineering disciplines. The mission of IDIES is to foster the interdisciplinary development of tools and methods that derive knowledge from the massive datasets that today's instruments, experiments, and simulations generate at exponentially growing rates.

2.1 Research and Academic Model

IDIES includes faculty members drawn from a variety of disciplines ranging from Physics and Astronomy, Mechanical Engineering, the Sheridan Libraries, Applied Math, School of Medicine, School of Public Health and the Human Language Technology Center of Excellence. The authors of this paper make up the Computer Science members of IDIES and two of their long-standing collaborators.

The establishment of a data intensive computing center has numerous benefits in addition to providing an umbrella for data-driven research. It acts as a beacon for the remainder of the Hopkins community and beyond, where our members are the first port of call for anyone with big data problems from academia, industry, and government agencies in the vicinity and nationwide, such as the Johns Hopkins Applied Physics Laboratory, a not-for-profit research lab housing 3,000 engineers and scientists.

The Johns Hopkins University provides an extremely conducive environment for these efforts. The demographics of the university leads to an outward-looking faculty within departments. The model of research being conducted through interdisciplinary centers is the norm rather than the exception (e.g. the Human Language Technology Center of Excellence). This facilitates application-driven research, combining scientific domain knowledge with core computer systems research expertise.

From an academic standpoint, the interdisciplinary model of IDIES percolates through to the students, engaging students in an application and dataset-driven research agenda. Our agenda fosters a breadth of knowledge beyond data management and computer science, and encourages a hands-on approach to system development. This results in tangible, usable components that further the underlying scientific problem, as well as promoting an experimental approach to data management challenges.

This research model leads to the development of communities around individual datasets and the curation of these datasets both in terms of data cleaning and functionality supported on top of the dataset. Most often, this process produces public resource for widespread academic usage.

2.2 Scientific Data Workflow at IDIES

Traditionally, under the scientific method, basic research involves the formation of hypotheses and theories, designing experiments for their validation, collecting data by realising experimentation, and analysing data to guide new insights for further research through iteration of this workflow. All stages of this workflow loop exhibit heavy computational and data management needs, particularly experimentation, simulation and analysis. The emergence of data-driven science has front-loaded much of this iterative workflow, where an intense period of experimentation results in a substantially larger data volume for processing, followed by an intense period of exploration and analysis of the data. Both modes must address large dataset challenges.

Many of these challenges appear individually in other domains, but are greatly exacerbated in scientific computing by the diversity of complex, instance-specific semantics. The end result is an application domain with extremely high overheads and barriers to entry through the combination of limited software reuse, and the brittle infrastructures arising from ad-hoc composition of tools that are difficult to parameterize and generalize.

We have assembled a critical mass of researchers at IDIES with expertise spanning many aspects of the scientific research workflow: Randal Burns (storage systems and transaction processing), Andreas Terzis (sensor and wireless networks), Michael Kazhdan (computer graphics, mesh processing and geometric retrieval) and Yanif Ahmad (data management and declarative languages). We are extremely fortunate to work with scientists and engineers who have developed a deep appreciation of data management throughout their careers, including Alex Szalay (physics, astronomy), and Charles Meneveau (mechanical engineering, turbulence).

3. THE SLOAN DIGITAL SKY SURVEY

The focus on data-intensive computating at JHU started when the Department of Physics and Astronomy joined the Sloan Digital Sky Survey project (SDSS), the astronomy equivalent of the "Human Genome Project." The SDSS created a high resolution multi-wavelength map of the Northern Sky with 2.5 trillion pixels of imaging. The results of the image segmentation were placed in a relational database that has grown to 12 terabytes over the last ten years and has created a new way to do astronomy. Based upon citation statistics, the database of the SDSS has been the most used astronomy facility in the world. A large fraction of the world's astronomy community has learned to use SQL to formulate their research questions and can run observations instantly, rather than waiting for months to get their turn at a telescope.

The original database was built in close collaboration between Alex Szalay's team at JHU and Jim Gray of Microsoft. The database is based on SQL Server, contains about 70 tables of data and descriptive metadata, and incorporates a substantial amount of astronomy code in the form of User Defined Functions. Among other things, we have built a high precision GIS system for astronomy, based on the Hierarchical Triangular Mesh, implemented as a set of class libraries, wrapped into C# functions, callable from T-SQL.

Today about 30% of the world's professional astronomy community has a server side database in this system. The two-stage parallel loading environment and the whole framework is now in use by many groups, both at JHU and all over the world, and has been repurposed from astronomy to many other disciplines, such as environmental science and radiation oncology. Most recently, the Pan-STARRS project, on the way to its first of many petabytes of data, builds upon a scaled-out version of the SDSS design.

4. APPLICATIONS AND DATASETS

IDIES research spans a wide variety of applications and datasets. We present a sample here.

4.1 Turbulence

Hydrodynamic turbulence is a formidably difficult and pressing problem. However, direct numerical simulations of turbulence for the flow conditions prevalent in most engineering, geophysics, and astrophysics applications are impractical. Also, most critical scientific problems, such as predicting the Earth's climate and developing energy sources, require reliable simulations of turbulent flows.

Typically, individual researchers perform large simulations that are analyzed during the computation with only a small subset of data stored for subsequent analysis. The majority of the time evolution is discarded. As a result, the same simulations must be repeated after new questions arise that were not initially obvious; most breakthrough concepts cannot be anticipated in advance. Thus, a new paradigm is emerging that creates large and easily accessible databases that contain the full space-time history of simulated flows.

In IDIES, we have embarked on building such databases. The JHU public turbulence database houses a 27 TB database that contains the entire time history of a 1024^3 mesh point pseudo-spectral

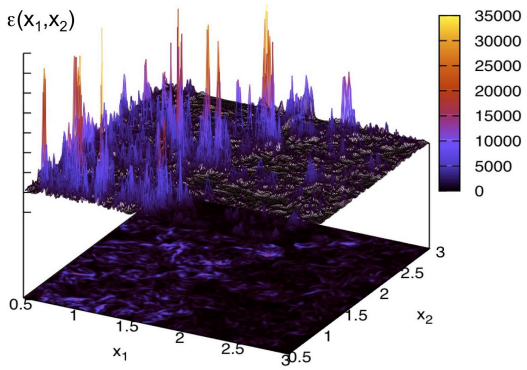

Figure 1: Contour and surface plot of dissipation field in isotropic turbulence as determined from a 2D cut through the 4D data from the JHU public turbulence database.

direct numerical simulation of forced isotropic turbulence [15, 12]. 1024 time-steps have been stored, covering a full "large-eddy" turnover time. A Web service fulfills user requests for velocities, pressure, and various space-derivatives and interpolation functions (see [7]). The 1024^4 isotropic turbulence database has been in operation continuously since 2008 and has served more than 50 billion point queries for research and teaching purposes.

4.2 Life Under Your Feet

Lack of field measurements, collected over long periods of time and at biologically significant spatial granularity, hinders scientific understanding of the effects of environmental conditions to the soil ecosystem. Wireless Sensor Networks (WSNs) promise to address ecologists' predicaments through a fountain of readings from low-cost sensors deployed with minimal disturbance to the monitored site.

In the fall of 2005 we built a proof-of-concept WSN to validate this claim. The end-to-end *Life Under Your Feet* (LUYF) system includes motes that collect environmental parameters such as soil moisture and temperature, static and mobile gateways that periodically download collected measurements through the Koala reliable transfer protocol [13], a database that stores collected measurements, access tools for analyzing the data, a Web site that serves the data, and tools to monitor the network. LUYF has been deployed in multiple forests in networks whose sizes range from ten to fifty nodes, deployed from a couple of weeks to more than two years. The LUYF database contains more than 120 million measurements, data derivatives, and the provenance of stored data.

4.3 DC Genome

The IT industry is the one of the fastest growing sectors of the U.S. economy in terms of its energy consumption. According to a 2007 EPA report, U.S. data centers consumed 61 billion kWh in 2006—enough energy to power 5.8 million households. Under conservative estimates, IT energy consumption is projected to double by 2011. Only a fraction of the electricity consumed powers IT equipment. The rest is used by environmental control systems such as air conditioning, (de-)humidifiers, and water chillers, or is lost during delivery.

A key reason for inefficiency is the lack of visibility into the data center operating conditions. Conventional wisdom dictates that IT equipment need excessive cooling to operate reliably, so the AC systems are set very cool and fan speeds set high, to reduce the danger of hot spots. Also, when servers issue thermal alarms, data centers have limited means to diagnose and remedy the problem other than further decreasing the temperature.

Given the data center's complex airflow and thermodynamics, dense and real-time environmental monitoring is necessary to improve energy efficiency. The data collected can help operators troubleshoot thermal alarms, make intelligent decisions on rack layout and server deployments, and innovate on facility management.

Wireless sensor network technology is an ideal candidate for this monitoring task as it is low-cost, nonintrusive, can provide wide coverage, and can be easily repurposed. The Data Center (DC) Genome project—a collaborative effort with Microsoft Research—aims to understand how energy is consumed in data centers as a function of facility design, cooling supply, server hardware, and workload distribution through data collected from large-scale sensor networks, and to use this understanding to optimize and control data center resources.

4.4 Data Conservancy

Research projects have finite lifetimes and the data products they produce need to persist long after the community of researchers have disbanded and funding has ceased. The preservation of scientific data makes scientific discovery available to future researchers and preserves our investment in the sciences. However, the data management community has yet to define self-sustaining preservation data management platforms that are widely available and many critical data sets are being lost. For example, the completed SDSS sky survey represents 8 years of telescope time and has produced more than 4 TB of curated data. As of now, the project is complete and scientists and staff are moving.

At JHU, we have undertaken the long-term preser-

vation of scientific data in the *Data Conservancy* project: an NSF-funded Sustainable Digital Data Preservation and Access Network (DataNets) project. Preservation targets include earth science, biology, and social science data in addition to astronomy and the SDSS. The Data Conservancy will create a preservation environment that retains not only the data, but the metadata and queries that express how scientists interact with the data. Scientists now and in the future will be able to query SDSS data to repeat and validate previous findings and continue the exploration of the universe.

5. RESEARCH PROJECTS

We now survey the data management techniques inspired by the above applications.

5.1 Instrumenting the Real World with Sensor Networks

Sensor networks deployed to collect scientific data (e.g., [14, 16, 19]) have been shown to be plagued with measurement faults. These faults must be detected to prevent pollution of the experiment and waste of network resources. At the same time, networks should autonomously adapt to sensed events, for example by increasing sampling rates or raising alarms. Here, events are measurements that deviate from "normal" data patterns, yet represent features of the underlying phenomenon, such as rain events in the case of soil moisture measurements.

However, detection algorithms tailored to specific types of faults lead to false positives when exposed to multiple types of faults [6]. More importantly, algorithms which classify measurements that deviate from the recent past as faulty tend to misclassify *events* as faults [6]. This misclassification is particularly undesirable because measurements from episodic events are invaluable data for domain scientists and should be given the highest priority.

In our recent work we unified fault and event detection under a more general *anomaly detection* framework, in which online algorithms classify measurements that significantly deviate from a learned model of data as anomalies [5]. We avoid misclassification by including punctuated, yet infrequent events in the training set, thus allowing the system to distinguish faults from events of interest. Specifically, we developed an anomaly detection frameworks based on Echo State Networks as well as Bayesian Networks and implemented these frameworks on a mote-class device. We showed that learning-based techniques are more sensitive to subtler faults and generate fewer false positives than the rule-based fault detection techniques.

5.2 Incremental and Model-Based Continuous Queries

Data-driven stream processing arises naturally in scientific applications given continuous data acquisition. Here at Johns Hopkins we are studying the incremental foundations of stream processing, and investigating the use of mathematical representations of data in model-based stream engines.

The DBToaster [9] project investigates incremental query processing of large dynamic data workloads. While stream processors answer queries over recent, contiguous windows of data streams, a dynamic data management system has to handle large, arbitrarily long-lived, non-contiguous state. A dynamic data management system is well-suited for streaming analysis, combining streaming data with persistent data for example, in scientific applications, to detect stream outliers with the aid of a historical database.

DBToaster transforms continuous queries to be evaluated as incrementally as possible. DBToaster applies the concept of using delta queries, as found in incremental view maintenance, recursively, yielding higher-order delta queries much as with higher-order derivatives from calculus. By intelligently materializing and reusing higher-order deltas, we avoid repeated computation of delta queries.

The Pulse [2] project investigates the use of piecewise polynomials, in a traditional stream processing engine. The input data stream is represented as a piecewise polynomial, while queries expressed in a declarative first-order logic based language such as SQL are transformed into a series of systems of equations. Query processing then involves solving equation systems. Streams represented by mathematical models can interpolate missing data, for example from sensor and network failures, or extrapolate data for predictive query processing and "what-if" queries. Furthermore, polynomial streams are a highly compact data representation and can often be processed extremely efficiently.

5.3 Batch Processing: Data-Driven Exploration and Analysis

Publicly-available data-intensive scientific services experience a tragedy of the commons. User workloads are I/O bound and concurrent workloads interfere with each other, creating congestion. As our scientific data services become more popular, the user experience inevitably degrades. Jim Gray expressed this best, in the early days of the Sloan Digital Sky Survey, when he said, "The only things that we have to fear are success and failure."

The Sloan Digital Sky Survey uses data replica-

tion, workload classification, and server storage to handle long-running, data-intensive queries. The Catalog and Archive Server Jobs (CasJobs) [11] system defined a separate instance of the SDSS database for asynchronous queries that allowed unrestricted SQL access to the SDSS and stored query results on server local storage for subsequent analysis. This insulated power-users from casual, interactive use and made long-running queries faster and more reliable: jobs transfer results to local storage and job completion does not rely on client-server connectivity.

Our subsequent efforts focused on increasing the throughput of concurrent jobs based on I/O sharing among queries. I/O cannot be provisioned incrementally to meet increased demand, e.g. by adding servers, nor is caching effective for data-intensive scientific queries that scan large data sets. Instead, we identify queries that share data, e.g. scan the same relation, execute the queries concurrently. For declarative Astronomy queries in which data may be processed in any order, the LifeRaft scheduler [20] co-schedules queries against an ordering of the data that maximizes data sharing. In LifeRaft, each data region may be accessed a single time to compute the partial results of all queries that use that data. The Job-Aware Workload (JAWS) scheduler [21] extends LifeRaft in order to support workflows in which queries exhibit data dependencies. This is typical of queries to the Turbulence Database Cluster in which the results derived from one timestep of simulation are used as input to the next timestep. Both LifeRaft and JAWS avoid query starvation through adaptive and incremental trade-offs between query throughput and response time. Data-driven batch scheduling typically improves throughput by a factor of four or more.

At present, we are collaborating with Yahoo! to combine LifeRaft and JAWS with their work on shared scans for Hadoop! and Pig workloads [1].

5.4 Data-Intensive Architectures

In deploying scientific databases, IDIES has designed and built several data-intensive clusters. The GrayWulf system [18] represents the evolution of the SDSS architecture to a scalable, multi-tenant database cluster. The name pays homage to Jim Gray who defined this class of computing and references BeoWulf clusters. GrayWulf provides cluster management tools to manage workflows, localize computation to data, monitor systems status, and recover from faults. The GrayWulf system won the Supercomputing Storage Challenge in 2008 and, as part of that competition, demonstrated a sustained data rate of 70 GB/s for a parallel SQL workload.

Figure 2: Visualization of the raw data returned by 3D scanners (left) and a surface reconstruction fit to the data (right).

The cost and power consumption of the Graywulf scales linearly with storage size and therefore will soon face a *power consumption wall* as scientific data sets continue to increase in size. To resolve this challenge, we recently proposed an alternative architecture comprising large number of so-called *Amdahl blades* that combine energy-efficient CPUs with solid state disks to increase sequential read I/O throughput by an order of magnitude while keeping power consumption constant [17]. The same preliminary study also showed that while keeping the total cost of ownership constant, Amdahl blades offer five times the throughput of the Graywulf system.

5.5 Fitting Models to Data: Large Surface Reconstruction

Using state-of-the-art laser range scanners, it is has now become possible to acquire 3D data at remarkably high resolution. These advances in scanning technology enable sub-millimeter resolution scans from Stanford's Digital Michelangelo Project [10], allowing art historians to make out the details of individual chisel marks, and reason about sculpting techniques without having to go to Florence. Similarly, with the ten centimeter resolution LIDAR fly-by over of New York City, environmentalists can plan the city's solar power capacity.

One of the challenges here is transforming the data returned by the 3D scanner into a coherent 3D model. Specifically, fitting a water-tight surface to the set of disjoint point-samples returned by the scanner (see Figure 2.) Research at Johns Hopkins, in collaboration with Microsoft Research, addresses this challenge by reducing fitting a surface to the scan data to solving a 3D Poisson equation [8].

To provide a tractable solution for large models, we have developed a new multigrid technique that supports the solution of Poisson equations formu-

lated over an octree adapted to the scanned points. This solver retains the benefits derived from solving a global system (providing robustness in the presence of missing data and noise), is capable of reconstructing models at the resolution of the input data, and has a space/time complexity that is linear in the size of the input. We have extended the approach with a partial ordering of octree nodes along an axis to provide both out-of-core [3] and distributed processing [4]. We have reconstructed models of up to one billion points (e.g. the David dataset from [10]) on a distributed cluster in just half a day.

6. SUMMARY

Data-intensive scientific computing has emerged as a theme around which we can engage in outreach for data management. We have enjoyed a rewarding intellectual experience here at IDIES, and champion this mode of research across the broader computer science research community. We encourage any science or engineering students interested in studying, or researchers seeking collaborations on data intensive challenges to explore the institute's webpage (http://idies.jhu.edu) and authors' webpages, and to contact an author via email.

Acknowledgements. We thank the Whiting School of Engineering and the Krieger School of Arts and Sciences at the Johns Hopkins University in supporting the mission of IDIES. We also thank Microsoft, the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, NVIDIA and Yahoo! This material is based upon work supported by the National Science Foundation under Grant Nos. AST-0939767, OCI-1040114, CMMI-0941530, CCF-0937810, CMMI-0923018, OCI-0830876, AST-0428325, IIS-0430848, CNS-0546648, OCI-0937947, AST-0225645, AST-0122449, CNS-0720730, CCF-0746039. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] P. Agrawal, D. Kifer, and C. Olston. Scheduling Shared Scans of Large Data Files. In *VLDB*, 2008.
[2] Y. Ahmad, O. Papaemmanouil, U. Çetintemel, J. Rogers. Simultaneous equation systems for query processing on continuous-time data streams. In *ICDE*, 2008.
[3] M. Bolitho, M. Kazhdan, R. Burns, H. Hoppe. Multilevel streaming for out-of-core surface reconstruction. In *Symposium on Geometry Processing*, 2007.
[4] M. Bolitho, M. Kazhdan, R. Burns, H. Hoppe. Parallel poisson surface reconstruction. In *International Symposium on Visual Computing*, 2009.
[5] M. Chang, A. Terzis, and P. Bonnet. Mote-Based Online Anomaly Detection using Echo State Networks. In *DCOSS*, 2009.
[6] J. Gupchup, A. Sharma, A. Terzis, R. Burns, A. Szalay. The Perils of Detecting Measurement Faults in Environmental Monitoring Networks. In *DCOSS*, 2008.
[7] JHU Turbulence Database. http://turbulence.pha.jhu.edu.
[8] M. Kazhdan, M. Bolitho, H. Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing*, 2006.
[9] O. Kennedy, Y. Ahmad, C. Koch. DBToaster: Agile views for a dynamic data management system. In *CIDR*, 2011.
[10] M. Levoy et al. The digital Michelangelo project: 3d scanning of large statues. In *SIGGRAPH*, 2000.
[11] N. Li and A. Thakars. CasJobs and MyDB: A batch query workbench. *Computing in Science and Engineering*, 10(1), 2008.
[12] Y. Li, E. Perlman, M. Wang, Y. Yang, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink. A public turbulence database cluster and applications to study lagrangian evolution of velocity increments in turbulence. *Journal of Turbulence*, 9(31), 2008.
[13] R. Musaloiu-E., C.-J. Liang, and A. Terzis. Koala: Ultra-low power data retrieval in wireless sensor networks. In *IPSN*, 2008.
[14] R. Musăloiu-E., A. Terzis, K. Szlavecz, A. Szalay, J. Cogan, J. Gray. Life Under your Feet: A WSN for Soil Ecology. In *EmNets Workshop*, 2006.
[15] E. A. Perlman, R. C. Burns, Y. Li, C. Meneveau. Data exploration of turbulence simulations using a database cluster. In *Supercomputing*, 2007.
[16] L. Selavo, A. Wood, Q. Cao, T. Sookoor, H. Liu, A. Srinivasan, Y. Wu, W. Kang, J. Stankovic, D. Young, and J. Porter. LUSTER: Wireless Sensor Network for Environmental Research. In *ACM SenSys*, 2007.
[17] A. Szalay, G. Bell, H. Huang, A. Terzis, and A. White. Low-power amdahl-balanced blades for data intensive computing. In *Proceedings of the 2nd Workshop on Power Aware Computing and Systems (HotPower '09)*, 2009.
[18] A. S. Szalay et al. GrayWulf: Scalable clustered architecture for data intensive computing. In *HICSS*, 2009.
[19] G. Tolle, J. Polastre, R. Szewczyk, N. Turner, K. Tu, P. Buonadonna, S. Burgess, D. Gay, W. Hong, T. Dawson, D. Culler. A Macroscope in the Redwoods. In *ACM SenSys*, 2005.
[20] X. Wang, R. Burns, and T. Malik. Liferaft: Data-driven, batch processing for the exploration of scientific databases. In *CIDR*, 2009.
[21] X. Wang, E. Perlman, R. Burns, T. Malik, T. Budavári, C. Meneveau, and A. Szalay. JAWS: job-aware workload scheduling for the exploration of turbulence simulations. In *Supercomputing*, 2010.