Research Data Management and Data Analytics

Our research focuses on accelerating scientific discovery by making it easier for researchers to discover, understand, organize, move, analyze, and publish data. In addition to computer science, our research engages fields as diverse as materials science, biology, archeology, and social sciences. We have a number of projects available to participants of the 2017 BigDataX REU program in the following areas: research data management, elastic cloud provisioning, and information extraction and analytics. Students are also welcome to propose projects related to shared areas of interest. Additional information can be found on our website: `labs.globus.org`.

Research data management

The changing landscape of science has created a world in which significant amounts of scientific data is distributed across a number of disparate and heterogeneous storage locations. It is now common for data to be generated, analyzed, shared, published, and archived in different locations, each of which may even expose a different storage interface (e.g., Posix, object stores, etc.). Given the increasing amount, heterogeneity, and distribution of research data it is fast becoming burdensome for researchers to efficiently manage their data. To address these challenges we have developed Globus [1]—a hosted service that provides a collection of core research data management capabilities to researchers. Globus allows researchers to easily manage, transfer, share, and publish large amounts of distributed data. It currently supports more than 40,000 registered users and has moved over 230 PB of data.

Data search: We are developing methods to index vast amounts of scientific data irrespective of its location and the particular storage system in which it is stored. We aim to provide a scalable search index that allows researchers to find, browse, and discover disparate scientific data based upon file system metadata and metadata buried within scientific data formats. In this project, students will explore methods for indexing large amounts of distributed data. The resulting index is intended to be applicable both to tightly coupled distributed file systems and also loosely coupled distributed storage systems such as those managed by Globus. Students will develop models for extracting an array of file system and file-based metadata, investigate methods for efficiently indexing large amounts of data to support common query models (e.g., free-text search), and analyze the performance of their approaches under a range of scenarios and using real data. Students may also explore scalable methods for identifying and extracting scientific metadata as well as algorithms for synchronizing the index with frequently changing, distributed data.

Transfer performance prediction: Globus has been used to conduct more than 3 million transfers on behalf of users. As a result, we have assembled a large database of historical information from which we can explore methods for predicting various future conditions. Examples include predicting endpoint usage by specific users and predicting transfer performance (e.g., throughput, errors, etc.) for specific transfers. Students will explore potentially useful features by analyzing historical data. They will then develop and evaluate heuristics for predicting events based on features derived from their analysis of historical data. Finally they will explore prediction techniques using, for example, collaborative filtering and neural networks, to predict future events and conditions. We have conducted a preliminary study of endpoint prediction and demonstrated high accuracy [2].

Automated cloud provisioning

While the benefits of cloud computing are obvious, there are significant technical challenges associated with scaling analyses efficiently and cost-effectively. These challenges are particularly important to scientists, as efficient and cost-effective scaling enables order of magnitude larger sample sizes from which to derive results. We are researching autonomic approaches for provisioning cloud infrastructure and scheduling workloads. Our approach leverages three core components: an automated provisioning service, a profiling service, and a collection of cost-aware scheduling algorithms. Our stand-alone autonomic cloud provisioning service is designed to dynamically provision cloud infrastructure for executing high throughput computing workloads (e.g., using HTCondor) [3, 4]. The service makes decisions based on projected execution time and cost, real-time economic information, and current and projected workload. Our profiling service automates the creation of tool profiles: concise descriptions of the performance and CPU, memory, network, and disk requirements of a supplied tool under different cloud environments and scenarios [5]. Building upon this provisioning model we have developed cost-aware and deadline

constrained scheduling algorithms to efficiently allocate workload over provisioned instances [6]. We have deployed these approaches in Globus Genomics [7] and observed significant cost and execution benefits [3].

Application profiling: Many systems that make use of cloud infrastructure have rich historical usage logs. For example, Globus Genomics records information such as submission time, script/application executed, completion time, instance type used, and even resource usage at various points in time. In this project students will develop methods that leverage this information to create or augment profiles. Students will first analyze historical logs to identify important features related to application performance. They will then develop algorithms (e.g., based on input data size, cloud environment, application settings, etc.) and explore the ability of statistical models and machine learning algorithms to predict execution time and resource usage. We will leverage these inferred models in two ways: 1) as a method for creating partial profiles that can be directly used for provisioning; and 2) as an indicator for where subsequent profiling should be focused (e.g., where there are gaps in the model).

Provisioning service integration: The second project aims to apply the provisioning service to improve cloud acquisition processes in the Cloud Kotta data enclave [8]. Cloud Kotta provides secure data storage and analytics capabilities to a large number of social scientists and students. It uses elastic cloud computing resources to satisfy the needs of user-driven computational social science anlaytics, however it does not use cost-aware provisioning models to acquire resources. In this project students will extend the provisioning service to integrate with Cloud Kotta, adding support for the Amazon Web Services queue and extending the Cloud Kotta acquisition software. They will then investigate methods to profile several social science analysis applications (either from logs or using the profiling service). Finally, students will leverage profiles and market prediction models to optimize the choices made by Cloud Kotta.

Information extraction and data analytics

There is a wealth of valuable information locked within unstructured scientific data (e.g., papers, files, network graphs, etc.). If harnessed, this information can offer new scientific insight and enable new avenues of investigation. Towards this goal, we are working on several projects that aim to improve our ability to extract knowledge from unstructured publications and scientific data formats as well as exploring methods for analyzing the information available within an institution to better understand the research ecosystem.

Information extraction: We are investigating methods to extract information (e.g., scientific facts, methodologies, and discoveries) from published literature [9, 10]. This is a particularly challenging task as information is represented in heterogeneous and unstructured formats (e.g., free-text, equations, figures, and tables). We have developed techniques to automatically extract information from publications, exploit expert crowds to review extracted values, and make curated values available via a web service. Our approach leverages machine learning techniques to classify and rank both publications and the items (e.g., figures, tables, equations) within them.

In this project we aim to extend these capabilities to extract a wider range of "facts" from publications and potentially adapt these approaches to extract structured information from scientific files. Students will leverage natural language processing (NLP) and machine learning techniques to identify pertinent information and extract values. They will explore various approaches such as using rule-based methods, classification models, supervised machine learning, and potentially even crowdsourcing. As part of this project students will investigate methods for scaling the extraction process using containers and elastic cloud infrastructure.

Institutional analytics: We are developing an extensible analytics platform with the goal of extracting actionable insights from a variety of information sources available within an institution. This project is comprised of two components: first, developing techniques for analyzing heterogeneous data sources; and second, developing a scalable service to make processed data and recommendations accessible to users.

In the first project students will explore methods to capture and map institutional collaboration and domain expertize, thus creating a complex graph that links researchers, outputs, competencies, and funding opportunities. They will implement methods for processing text-based data sources (e.g., publications), classifying objects according to topics using techniques such as topic modeling and word embedding, and deriving linkages between entities. Students will then investigate models for deriving metrics from the resulting graphs, focusing on areas such as focus and influence. Finally, we will explore approaches for producing customized recommendations based on analysis of large amounts of aggregated information.

The second project will build upon the analytics engine developed in the first project to create a scalable pipeline for ingesting and analyzing new data. Students will also develop a user-oriented service and explore visu-

alization techniques for making complex, multi-dimensional data accessible to users. We will focus on developing scalable cloud-based analysis techniques and exposing information via a hosted REST service and HTML-based user interface.

Mentor

Dr. Kyle Chard is a Senior Researcher and Fellow in the Computation Institute at the University of Chicago and Argonne National Laboratory. He received his Ph.D. from the department of Engineering and Computer Science at Victoria University of Wellington in 2011. His research focuses on applying computational and data-intensive approaches to solve scientific problems. He is particularly interested in the application of autonomic computing and cost-aware provisioning to make use of on-demand cloud infrastructure; service oriented science, as part of the Globus project; and information extraction and analytics. Kyle's doctoral dissertation focused on federating disparate resource providers using a co-operative allocation model in which management components are hosted on participating providers' resources and secure economic protocols enable trustworthy and privacy preserving allocations.

References

[1] K. Chard, S. Tuecke, and I. Foster, "Efficient and secure transfer, synchronization, and sharing of big data," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 46–55, 2014.

[2] W. Agnew, M. Fischer, I. Foster, and K. Chard, "An ensemble-based recommendation engine for scientific data transfers," in *7th International Workshop on Data-Intensive Computing in the Clouds*, 2016.

[3] R. Chard, K. Chard, K. Bubendorfer, L. Lacinski, R. Madduri, and I. Foster, "Cost-aware cloud provisioning," in *Proceedings of the 11th IEEE International Conference on e-Science (e-Science)*, 2015, pp. 136–144.

[4] R. Chard, K. Chard, K. Bubendorfer, L. Lacinski, R. Madduri, and I. Foster, "Cost-aware elastic cloud provisioning for scientific workloads," in *Proceedings of the 8th IEEE International Conference on Cloud Computing (CLOUD)*, 2015, pp. 971–974.

[5] R. Chard, K. Chard, B. Bg, K. Bubendorfer, A. Rodriguez, R. Madduri, and I. Foster, "An automated tool profiling service for the cloud," in *IEEE International Symposium on Cluster Computing and the Grid (CCGrid)*, 2016, Manuscript submitted for publication.

[6] V. Arabnejad, K. Bubendorfer, B. Ng, and K. Chard, "A deadline constrained critical path heuristic for cost-effectively scheduling workflows," in *Accepted to the 8th IEEE/ACM International Conference on Utility and Cloud Computing*, 2015.

[7] R. Madduri, D. Sulakhe, L. Lacinski, B. Liu, A. Rodriguez, K. Chard, U. Dave, and I. Foster, "Experiences building Globus Genomics: A next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 13, pp. 2266–2279, 2014.

[8] Y. N. Babuji, K. Chard, A. Gerow, and E. Duede, "Cloud kotta: Enabling secure and scalable data analytics in the cloud," in *IEEE International Conference on Big Data*, 2016.

[9] R. Tchoua, J. Qin, D. Audus, K. Chard, I. Foster, and J. de Pablo, "Blending education and polymer science: Student contributions to a thermodynamic property database," *Journal of Chemical Education*, 2015, Manuscript submitted for publication.

[10] R. Tchoua, K. Chard, J. de Pablo, I. Foster, J. Qin, and M. Rasic, "ChiDB: An automated framework for extracting materials properties from literature," in *the 4th Greater Chicago Area Systems Research Workshop (GCASR)*, 2015.