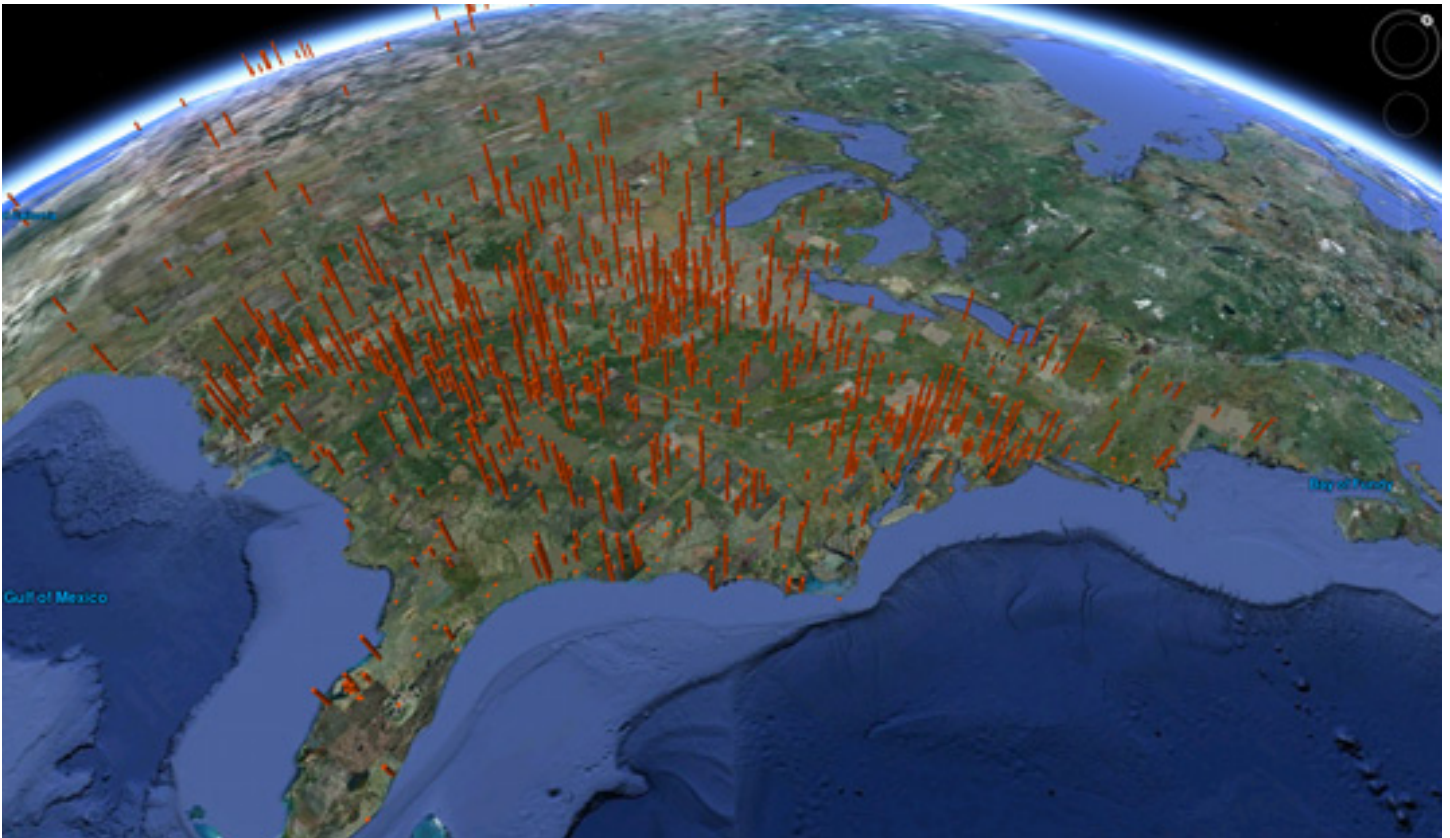


Enabling Data-Intensive Research via Cloud Computing

Longhorn Innovation Fund for Technology (LIFT) supports research into emerging computing environment



Geographical association of words from Memoirs of the Union's Three Great Civil War Generals by David Widger, from Dr. Jason Baldrige's TextGrounder project.

Information overload may have met its match. With Enabling Data-Intensive Research and Education at The University of Texas at Austin via Cloud Computing, a team of researchers and educators are not only examining the potential of processing vast quantities of data quickly, they are discovering how the use of large datasets can lead to all kinds of interesting questions.

Background

The increasing ability to generate vast quantities of data presents technical challenges for both researchers and educators as data storage and transfer approaches critical mass and the exchange of large data sets puts established information practices to the test. Increasingly, institutions of higher education have to plan strategically for this fundamental shift in how scientific data analysis can be done. The paradigm of “cloud computing” provides an environment that is up to the task of doing data-intensive computation for research and educational purposes. In this context, cloud computing is a distributed computing paradigm that enables

large datasets to be sliced and assigned to available computer nodes where the data can be processed locally, avoiding network-transfer delays. This makes it possible for researchers to query tables with trillions of rows of information or search across all the servers in a data center.

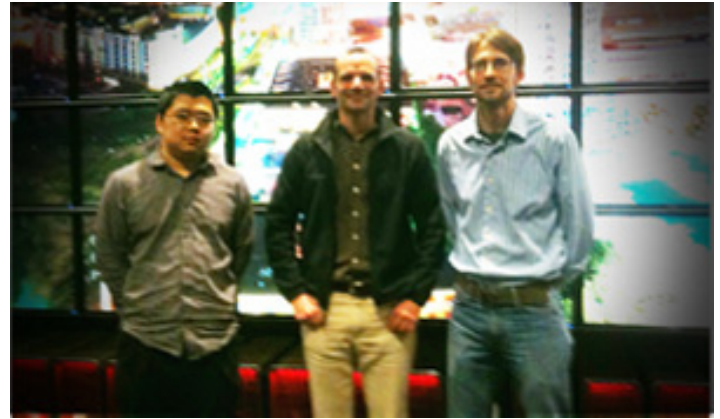
With its combination of internationally recognized research faculty and the world-class computing systems of the Texas Advanced Computing Center (TACC), The University of Texas at Austin is uniquely positioned to become a pioneer in this developing field of data-intensive research and education. Jason Baldrige, Assistant Professor, Department of Linguistics, College of Liberal Arts, Matthew Lease, Assistant Professor, School of Information and Weijia Xu, Research Associate, Texas Advanced Computing Center are part of an interdisciplinary team focused on enabling innovative solutions for existing research projects and adding new data-intensive computing content and courses to the University's academic and professional development offerings.

Progress

Funding from the Longhorn Innovation Fund for Technology (LIFT) has made it possible for their project, “Enabling Data-Intensive Research and Education at UT Austin via Cloud Computing” to purchase and install 304 hard drives with a total storage capacity of 112 terabytes on TACC’s Longhorn computer cluster and begin “Hadooping” at the University. Apache Hadoop is open source software used for reliable, scalable, distributed computing. Hadoop is an implementation of the MapReduce programming paradigm originally developed by Google. It enables the creation of applications capable of processing huge quantities of data on large clusters of computing nodes and do it quickly.

Computational Linguistics

In collaboration with Lease and Xu, Baldrige is using Hadoop for his research in computational linguistics to analyze large datasets of texts for geographical and temporal references. Expanding on work done under an award from The New York Community Trust to develop software called TextGrounder, Baldrige is conducting geo-referencing analysis of texts to ground language to place and time. Examples include geolocating multilingual Wikipedia pages and Civil War era texts, as well as working with the UT Libraries’ Human Rights Documentation Initiative to analyze testimonies from the Rwandan genocide (in English, French, and Kinyarwanda). Baldrige observes: “Hadoop lets you ask interesting questions based on large data sets. It allows the text to speak in new ways.” The information gathered is used to visualize these texts using geobrowsers like Google Earth; current visualizations available on the TextGrounder wiki show how the system connects language to time and space.



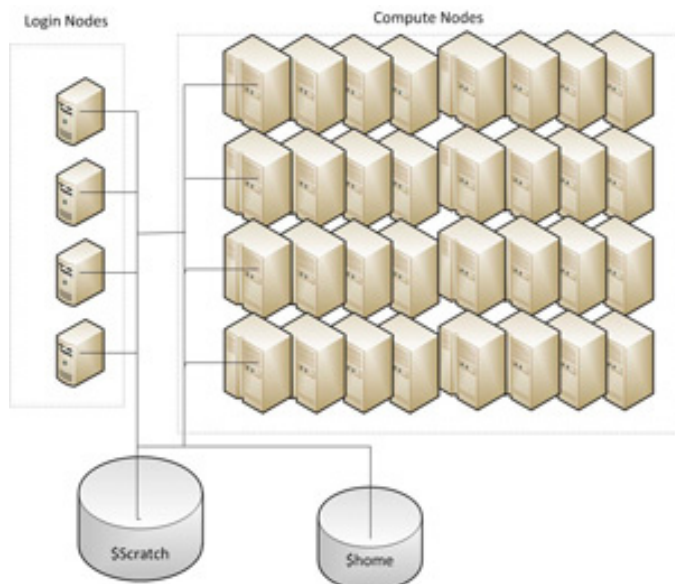
Weijia Xu (TACC), Matthew Lease (iSchool), and Jason Baldrige (UTCL) are driving the testing and documentation for conducting research using the Hadoop cluster.

Computational Journalism

Lease is developing new methods for finding information in massive datasets by applying large-scale distributed computation to perform media analytics. His work in this new area of computational journalism focuses on using rich computational tools to analyze news articles, blogs and user comments and find ways to support journalists in coping with the massive amount of online information. By decomposing huge volumes of information into text “snippets,” Lease can track how a single idea or concept “flows” from its originating source across multiple information providers, redistributors, and consumers. By following the way these text snippets evolve, it is possible to identify the creation and dissemination of specific ideas and information. His research is part of a project called “REACTION (Retrieval, Extraction and Aggregation Computing Technology for Integrating and Organizing News), a collaborative effort between the UT Austin Portugal program and several Portuguese news media companies and universities.

Data Mining

Xu is using Hadoop to investigate data mining methods for large-scale data visualizations. The goal of his work is to advance data-driven knowledge discovery by enabling interactive visualizations on terascale datasets. Using data mining in the field of astronomy is a good example. As more advanced simulations bring in ever increasing amounts of data, large scale computational research can filter the data and enable the researcher to analyze a more manageable amount of information. Xu is also using Hadoop as part of his work on a cooperative research agreement conducted by TACC and the National Center for Advanced Systems and Technologies (NCAST) at the National Archives and Records Administration. This project is designed to assist archivists in their work by developing methods for receiving and analyzing large amounts of electronic records.



Hadoop on the Longhorn cluster has brought immediate benefits to research, teaching and learning around campus. Here, an illustration that shows the Longhorn cluster and its file system.

An essential part of the project team's vision is to promote education and training on the latest developments in data-intensive computing across campus. Lease and Baldrige are currently developing a class they will co-teach in the 2011 fall semester. LIN386M/INF385T – Data-Intensive Text Processing with MapReduce will give students with graduate standing the opportunity to learn about Hadoop and gain valuable experience in data-intensive computing. Xu has already added a lecture about using Hadoop on the Longhorn cluster to his course Visualization and Data Analysis for Scientists and Engineers (SSC374E/SSC394E.) The team has also created a wiki and is documenting their initial experiences with Hadoop. They are hopeful that other UT researchers will contribute to the wiki and discussion forums as a means of building the Hadoop community on campus. Xu is also adding a tutorial on using Hadoop on the Longhorn cluster to the regular schedule of TACC training workshops. Additional research talks, course offerings and online documentation will help grow campus-wide cloud computing expertise for faculty, students and staff at the University.

April 27, 2011

**Betsy Busby
Communications Coordinator
Information Technology Services**

Benefits

Introduction of Hadoop on the Longhorn cluster has brought immediate benefits to research, teaching and learning around campus. As initial adopters, Baldrige, Lease and Xu are driving the testing and documentation for conducting research using the Hadoop cluster. Their initial results are informing the work of a broad cross section of committed adopters across campus and sparking interest and inquiry from the larger external research community. The addition of new content and courses related to data-intensive computing provides students with the opportunity to acquire cutting edge skills. This not only gives them a competitive edge in the job market but positions them to make significant research contributions at UT Austin and beyond. The expectation is that the current investment in cloud computing education provides critical professional training that will lead to scientific discovery based on massive data analysis.

Next Steps

Now that implementing the Hadoop cluster and testing the software is complete, next steps in the project focus on promoting large scale data analysis research projects across campus. Developing education and training that will help faculty, students and staff acquire skills in data-intensive computing is ongoing. The investment in hardware will continue to enable the development environment for data-intensive computing after the project is over and TACC has agreed to update the Hadoop software and documentation on how to use it.

Before the end of the funded period, the team plans to compile usage statistics on data-intensive computations enabled by the project. To date, six projects have been launched; this number is expected to grow as the project moves out of the setup and testing phase and users discover that they can run large data analysis tasks using Hadoop through a remote login to the Longhorn cluster.