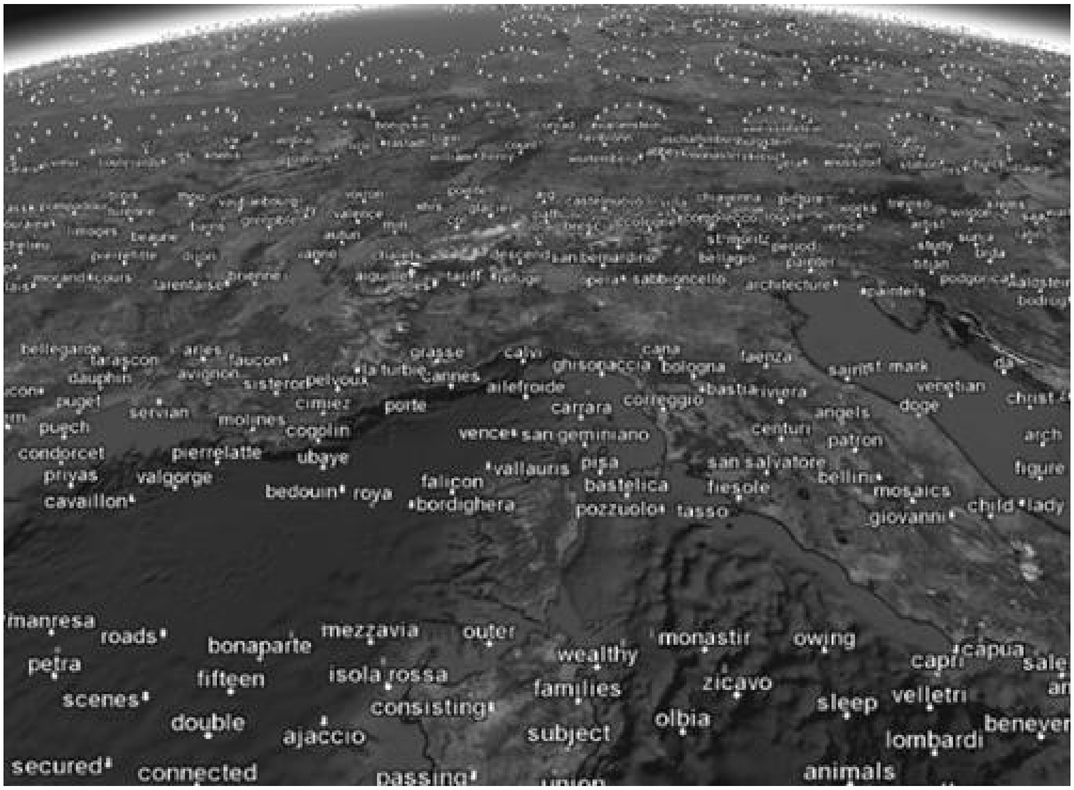
PHYS.ORG

Hip-hip-Hadoop: Data mining for science

May 28 2013



This is the word distributions from the RTM model in Google Earth created using Hadoop on the Longhorn cluster at the Texas Advanced Computing Center. The map shows words strongly associated with a region. Credit: Texas Advanced Computing Center

The model of distributed calculations, where a problem is broken down into distinct parts that can be solved individually on a computer and then recombined, has been around for decades. Divide-and-conquer techniques allow scientists to predict complex phenomenon from tornado formation to the qualities of nanomaterials to tomorrow's weather forecast.

But when Google developed the MapReduce algorithm, it added a distinct wrinkle to this method of distributed computing and opened new doors for commercial and scientific endeavors.

Apache Hadoop is an open-source software framework that evolved from Google's MapReduce algorithm. Many Internet giants—Facebook, Yahoo, eBay, Twitter—rely on Hadoop to crunch data across thousands of computer servers in order to quickly identify and serve customized data to consumers.

In 2010, the Texas Advanced Computing Center (TACC) at The University of Texas at Austin formally began experimenting with Hadoop to test the technology's applicability for scientific problems. TACC Research Associate Weijia Xu and University of Texas Professors Matthew Lease and Jason Baldridge won a Longhorn Innovation for Technology Fund (LIFT) grant to build a Hadoop-optimized cluster on Longhorn, a remote visualization system that TACC built in 2009 with support from the National Science Foundation.

Beyond industry applications, Hadoop is a popular platform for data intensive scientific discovery, particularly as a means of mining dense and large datasets for important connections and meaningful trends. However, at the time the TACC project started, there was no available infrastructure at U.S. high-performance computing centers where researchers and students could experiment with how Hadoop and supercomputers could be used together. The project's initial goal was to enable data intensive computing research and education at the university and, eventually, across the nation.

The infrastructure for a Hadoop cluster differs slightly from what supercomputing centers typically deploy. "In most high-performance computing systems, storage and analysis are separate. But Hadoop requires you to store your data locally on the compute node," Xu said. "The LIFT grant let us add local drives and storage to enable researchers to do experimental Hadoop-style studies on a current production system."

This system offers researchers a total of 48, eight-processor nodes on TACC's Longhorn cluster to run Hadoop in a coordinated way with accompanying large-memory processors. A user on the system can request all 48 nodes for a maximum of 96 terabytes (TB) of distributed storage. What's special about the Longhorn cluster at TACC isn't simply the beefed-up hardware for running Hadoop; rather it's the ability for researchers to leverage the vast compute capabilities of the center, including powerful visualization and data analysis systems, to further their investigations. The end-to-end research workflow enabled by at TACC could not be done anywhere else, and as a bonus, researchers get access to the full suite of tools available at the center to do computational research.

"The best part is that Hadoop is easy to use without requiring users to be experts," said Xu. "It handles a lot of the low-level computing behavior, so people don't need to have a lot of knowledge about I/O or memory structures to get started."

Researchers who are not C++ or Fortran programmers can quickly harness the power of Hadoop on Longhorn to query massive collections and databases in new ways, using more intuitive languages like R, Python and Matlab. In this way, the Hadoop system allows researchers to focus on the specifics of their research questions while allowing the system to handle the complexities of managing large parallel queries and job management.

"Hadoop provides researchers with the first major tool for doing groundbreaking research in the era of Big Data," said Niall Gaffney, TACC's director of data intensive computing. "I am very excited to see its early and fruitful adoption amongst researchers as well as the explorations into how it can be used to take advantage of the world class supercomputing resources TACC provides."

Biomarkers and Bookmarks

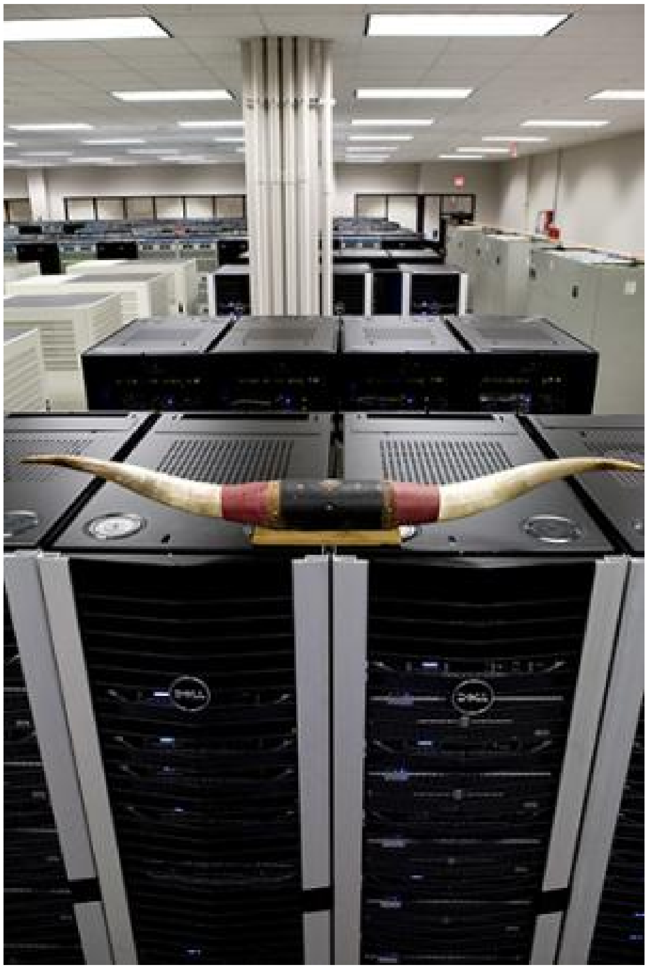
Since coming online in mid-November 2010, more than 65 researchers and students have used TACC's Hadoop system to perform more than one million hours of data intensive computations on 19 different projects, enabling dozens of papers and presentations. The projects range from natural language processing to detecting haloes in astronomical datasets, but share a reliance on data mining tools and a need for large, parallel computing systems.

Flow Cytometry

Flow cytometry (FCM) is a biomedical research technology widely used by immunologists, cancer biologists and infectious disease investigators to distinguish cell types based on the expression of distinct combinations of protein markers. However, the emerging scale of the data produced by flow cytometry is forcing researchers to consider new approaches for data analysis and interpretation.

For Yu Qian and Richard Scheuermann at the J. Craig Venter Institute, Hadoop offered the ability to expand successful research without having to rewrite a large community code that took several years to write.

"In the last decade, flow cytometry has experienced dramatic technical advances in both instrumentation and reagent development. The standard methods for analyzing FCM data have not kept pace with these advances in laboratory technologies. The increased complexity of these data has made it difficult to identify and compare cell populations using traditional manual gating strategies," the authors wrote in a recent manuscript. "An emerging solution is running automated analysis methods on large cyberinfrastructures."



Longhorn, a 256-node Dell visualization cluster, is designed for remote interactive visualization and data analysis. Credit: Texas Advanced Computing Center

ParaFlow, a software system for parallelizing flow cytometry data analysis, was implemented and tested on the Longhorn Hadoop cluster. The Hadoop cluster automatically creates and schedules parallel tasks based on the user job specification. Researchers only need to change the programs and their parameters in the application layer when they want to parallelize different analytical pipelines.

"Before they could only do this type of analysis on a small scale," Xu said. "Now, they can easily do a lot of samples at the same time. This generates a large quantity of data and helps pinpoint what type of virus or disease they're dealing with."

The researchers have begun to expand this methodology to Stampede, TACC's 10-petaflop supercomputer, which was deployed in January 2013. Stampede, a larger and faster cluster than Longhorn, potentially allows automated analysis of tens of thousands of FCM data files generated under different conditions, and their comparative study for identifying novel marker expression patterns and cell types, which was previously a "mission impossible" without this kind of cyberinfrastructure.

Computational Linguistics

For Jason Baldridge, professor of Linguistics at The University of Texas at Austin, access to the Longhorn Hadoop cluster allowed him to conduct large-scale, geo-referencing analyses of texts to ground language in place and time. In one research project, Baldridge applied a software tool called TextGrounder to map words from a 10 million word corpora to geographic locations. The words were drawn from the Perry-Castañeda Library Travel corpus, a collection of ninety-four British and American books on world travel and history from the late nineteenth and early twentieth centuries. Through the analysis on Longhorn, they were able to show that "bonaparte" is strongly associated with Corsica and Sardinia, while "glacier" and "chalet" correlated with locations in the Alps. The results were published in a special issue of Texas Studies in Literature and Language: Linguistics and Literary Studies: Computation and Convergence in Fall 2012.

Other examples of Baldridge's work on Longhorn 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). Baldridge transforms the Hadoop-generated information into visualizations using geobrowsers like Google Earth to illustrate how language is connected across time and space.

"Hadoop lets you ask interesting questions based on large data sets," Baldridge observed. "It allows the text to speak in new ways."

Testing Hadoop with Emerging Hardware

Researchers at TACC and in academia are not the only ones interested in exploring the potential of Hadoop for data driven application. Intel has also been working with TACC to assess the impact of new hardware the company has developed on the performance of Hadoop applications.

In a recent white paper, Intel researchers and TACC staff described experiments using the Intel 10GBASE-T network adapters on Hadoop. They asked: How do faster interconnects within a Hadoop clusters influence the performance of common scientific workflows?

To answer this question, TACC's Data Mining and Statistics Group ran a variety of common workloads and saw speed-ups for the majority of the applications. For file transfers—which are common in MapReduce applications—the amount of time spent on the network was 77% less than using 10GbE (Gigabit Ethernet) than with a 1GbE, resulting in an overall time reduction of 25% for the analysis.

"We are excited to see Intel Ethernet networking technologies accelerating the cutting-edge work TACC is doing with Hadoop," said Steve Schultz, director of product marketing, Intel Networking Division. "We look forward to continuing our collaborations with the researchers at TACC as they harness the power of the latest Intel hardware and software for their data-intensive scientific workloads."

Training Data Scientists

Deploying a new cluster with important, but largely untested technology for scientists is a great first step. But you also have to identify and build a community to take advantage of these emerging tools. TACC has been a leader in education and outreach to the public, offering training, tutorials and university-level instruction on Hadoop as it relates to high-performance parallel computing.

In Fall 2011 and 2012, Xu introduced Hadoop to students in the Visualization and Data Analysis course he co-teaches in the Division of Statistics and Scientific Computing at the university. In addition, Baldridge and Lease jointly designed a new course, "Data-Intensive Computing for Text Analysis," which was offered in Fall 2011, that involved significant use of TACC's Hadoop resources. Interestingly, the course attracted a multi-disciplinary group with 16 computer science students, four iSchool students, three linguistics students, and two electrical and computer engineering students.

At the end of May 2013, Xu will chair a workshop on Benchmarks, Performance Optimization, and Emerging Hardware of Big Data Systems and Applications in conjunction with 2013 IEEE International Conference on Big Data.

Which of the host of new heterogeneous hardware and software technologies available for high-performance clusters are best suited for data-intensive applications? And how can HPC systems be optimally designed to solve big data problems? These are the questions that TACC's Hadoop R&D seeks to answer.

Continuing Research and Future Plans

Applied research continues on TACC's Hadoop Longhorn system. Working with Yan Zhang from UT's iSchool, Xu is applying data mining and machine learning techniques to study health communication.

Online health communities allow users to share experiences and exchange information with peers with similar medical conditions. They have become a valuable source for patients and caregivers for informational and emotional support.

"While connecting users to those whom they may never be able to connect to otherwise, online communities present a new information environment that does not operate under the old publishing paradigm. This creates new challenges for users to access and evaluate information," Zhang explained. "In response to these challenges, better system functions should be designed to facilitate information access and evaluation."

Zhang and Xu are currently working with an online forum for autism, named Autism Support Network, to design a visualized browsing system. The new system would help users quickly find any aspect of the information that they need about autism, such as treatment, medications and challenges at school, and find peer-users that share common problems or interests with them.

"We are employing data visualization techniques to make the relationships among topics explicit to users, with an expectation that this will help them find the information and social support they need more quickly," Zhang said.

As the pace at which we generate data accelerates, efforts to develop new analysis tools and their timely adoption within the nation's massive HPC cyberinfrastructure becomes much more important. TACC is building on its efforts in this arena. In the coming months and years, TACC will offer more dedicated hardware, software and a growing research group to allow the flood of data to accelerate the rate of scientific discovery.

"Hadoop is the first of what will be many new powerful tools developed for the era of Big Data," Gaffney said. "I look forward to the game-changing research that will come from its adoption at TACC as we continue to help researchers wrangle results from these massive collections of data."

Provided by University of Texas at Austin

Citation: Hip-hip-Hadoop: Data mining for science (2013, May 28) retrieved 25 April 2024 from https://phys.org/news/2013-05-hip-hip-hadoop-science.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.