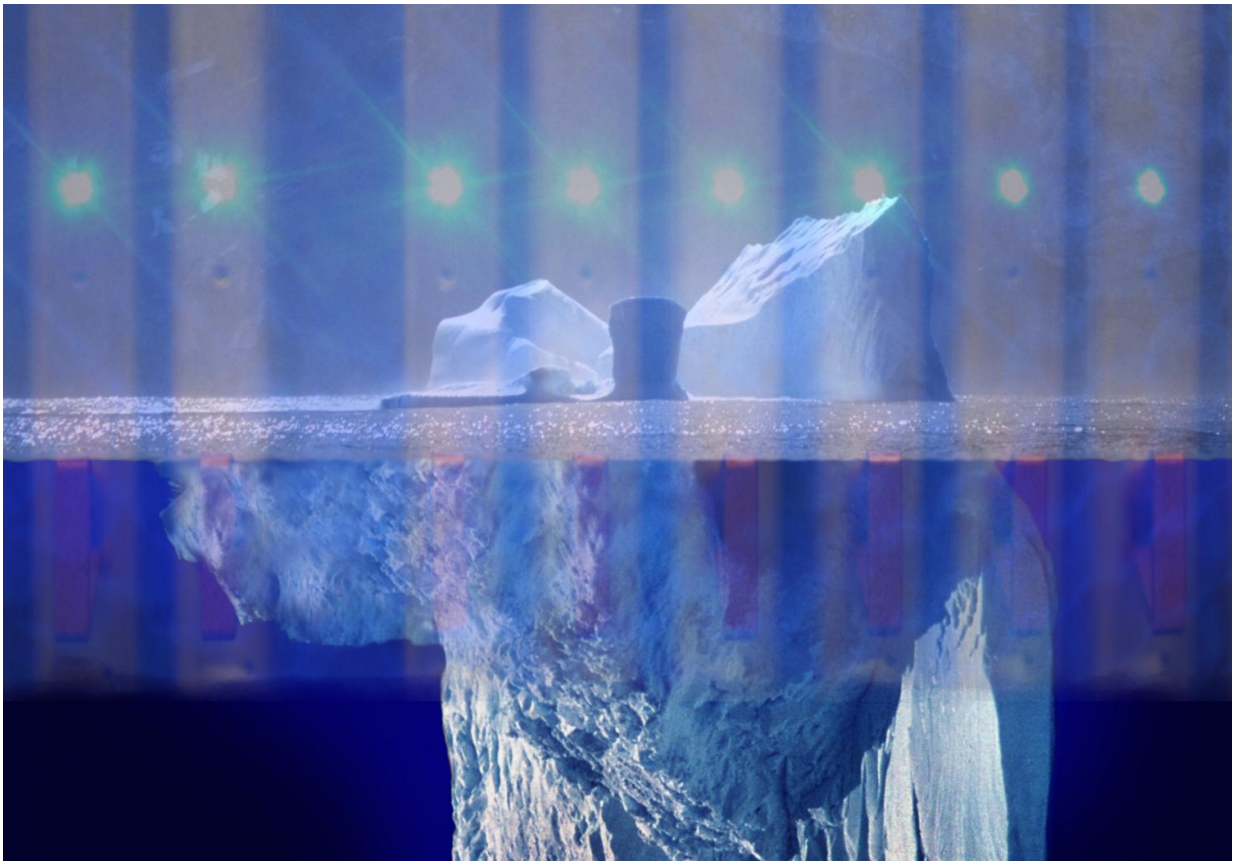
When data's deep, dark places need to be illuminated

February 6 2017



The World Wide Web is like an iceberg. Most of its data lies hidden below the surface in the deep web. The data-intensive Wrangler supercomputer at TACC is helping researchers get meaningful answers from the hidden data of the public web. Credit: TACC

Much of the data of the World Wide Web hides like an iceberg below the surface. The so-called 'deep web' has been estimated to be 500 times bigger than the 'surface web' seen through search engines like Google. For scientists and others, the deep web holds important computer code and its licensing agreements. Nestled further inside the deep web, one finds the 'dark web,' a place where images and video are used by traders in illicit drugs, weapons, and human trafficking. A new data-intensive supercomputer called Wrangler is helping researchers obtain meaningful answers from the hidden data of the public web.

The Wrangler supercomputer got its start in response to the question, can a computer be built to handle massive amounts of I/O (input and output)? The National Science Foundation (NSF) in 2013 got behind this effort and awarded the Texas Advanced Computing Center (TACC), Indiana University, and the University of Chicago $11.2 million to build a first-of-its-kind data-intensive supercomputer. Wrangler's 600 terabytes of lightning-fast flash storage enabled the speedy reads and writes of files needed to fly past big data bottlenecks that can slow down even the fastest computers. It was built to work in tandem with number crunchers such as TACC's Stampede, which in 2013 was the sixth fastest computer in the world.

While Wrangler was being built, a separate project came together headed by the Defense Advanced Research Projects Agency (DARPA) of the U.S. Department of Defense. Back in 1969, DARPA had built the ARPANET, which eventually grew to become the Internet, as a way to exchange files and share information. In 2014, DARPA wanted something new - a search engine for the deep web. They were motivated to uncover the deep web's hidden and illegal activity, according to Chris Mattmann, chief architect in the Instrument and Science Data Systems Section of the NASA Jet Propulsion Laboratory (JPL) at the California Institute of Technology.

"Behind forms and logins, there are bad things. Behind the dynamic portions of the web like AJAX and Javascript, people are doing nefarious things," said Mattmann. They're not indexed because the web crawlers of Google and others ignore most images, video, and audio files. "People are going on a forum site and they're posting a picture of a woman that they're trafficking. And they're asking for payment for that. People are going to a different site and they're posting illicit drugs, or weapons, guns, or things like that to sell," he said.

Mattmann added that an even more inaccessible portion of the deep web called the 'dark web' can only be reached through a special browser client and protocol called TOR, The Onion Router. "On the dark web," said Mattmann, "they're doing even more nefarious things." They traffic in guns and human organs, he explained. "They're basically doing these activities and then they're tying them back to terrorism."

In response, DARPA started a program called Memex. Its name blends 'memory' with 'index' and has roots to an influential 1945 Atlantic magazine article penned by U.S. engineer and Raytheon founder Vannevar Bush. His futuristic essay imagined making all of a person's communications - books, records, and even all spoken and written words - in fingertip reach. The DARPA Memex program sought to make the deep web accessible. "The goal of Memex was to provide search engines the information retrieval capacity to deal with those situations and to help defense and law enforcement go after the bad guys there," Mattmann said.

Karanjeet Singh is a University of Southern California graduate student who works with Chris Mattmann on Memex and other projects. "The objective is to get more and more domain-specific (specialized) information from the Internet and try to make facts from that information," said Singh said. He added that agencies such as law enforcement continue to tailor their questions to the limitations of search

engines. In some ways the cart leads the horse in deep web search. "Although we have a lot of search-based queries through different search engines like Google," Singh said, "it's still a challenge to query the system in way that answers your questions directly."

Once the Memex user extracts the information they need, they can apply tools such as named entity recognizer, sentiment analysis, and topic summarization. This can help law enforcement agencies like the U.S. Federal Bureau of Investigations find links between different activities, such as illegal weapon sales and human trafficking, Singh explained.

"Let's say that we have one system directly in front of us, and there is some crime going on," Singh said. "The FBI comes in and they have some set of questions or some specific information, such as a person with such hair color, this much age. Probably the best thing would be to mention a user ID on the Internet that the person is using. So with all three pieces of information, if you feed it into the Memex system, Memex would search in the database it has collected and would yield the web pages that match that information. It would yield the statistics, like where this person has been or where it has been sited in geolocation and also in the form of graphs and others."

"What JPL is trying to do is trying to automate all of these processes into a system where you can just feed in the questions and and we get the answers," Singh said. For that he worked with an open source web crawler called Apache Nutch. It retrieves and collects web page and domain information of the deep web. The MapReduce framework powers those crawls with a divide-and-conquer approach to big data that breaks it up into small pieces that run simultaneously. The problem is that even the fastest computers like Stampede weren't designed to handle the input and output of millions of files needed for the Memex project.

The Wrangler data-intensive supercomputer avoids data overload by

virtue of its 600 terabytes of speedy flash storage. What's more, Wrangler supports the Hadoop framework, which runs using MapReduce. "Wrangler, as a platform, can run very large Hadoop-based and Spark-based crawling jobs," Mattmann said. "It's a fantastic resource that we didn't have before as a mechanism to do research; to go out and test our algorithms and our new search engines and our crawlers on these sites; and to evaluate the extractions and analytics and things like that afterwards. Wrangler has been an amazing resource to help us do that, to run these large-scale crawls, to do these type of evaluations, to help develop techniques that are helping save people, stop crime, and stop terrorism around the world."

Singh and Mattmann don't just use Wrangler to help fight crime. A separate project looks for a different kind of rule breaker. The Distributed Release Audit Tool (DRAT) audits software licenses of massive code repositories, which can store hundreds of millions of lines of code and millions of files. DRAT got its start because DARPA needed to audit the massive code repository of its national-scale 100-million-dollar-funded presidential initiative called XDATA. Over 60 different kinds of software licenses exist that authorize the use of code. What got lost in the shuffle of XDATA is whether developers followed DARPA guidelines of permissive and open source licenses, according to Chris Mattmann.

Mattmann's team at NASA JPL initially took the job on with an Apache open source tool called RAT, the Release Audit Tool. Right off the bat, big problems came up working with the big data. "What we found after running RAT on this very large code repository was that after about three or four weeks, RAT still hadn't completed. We were running it on a supercomputer, a very large cloud computer. And we just couldn't get it to complete," Mattmann said. Some other problems with RAT bugged the team. It didn't give status reports. And RAT would get hung up checking binary code - the ones and zeroes that typically just hold data

such as video and were not the target of the software audit.

Mattmann's team took RAT and tailored it for parallel computers with a distributed algorithm, mapping the problem into small chunks that run simultaneously over the many cores of a supercomputer. It's then reduced into a final result. The MapReduce workflow runs on top of the Apache Object Oriented Data Technology, which integrates and processes scientific archives.

The distributed version of RAT, or DRAT, was able to complete the XDATA job in two hours on a Mac laptop that previously hung up a 24-core, 48 GB RAM supercomputer at NASA for weeks. DRAT was ready for even bigger challenges.

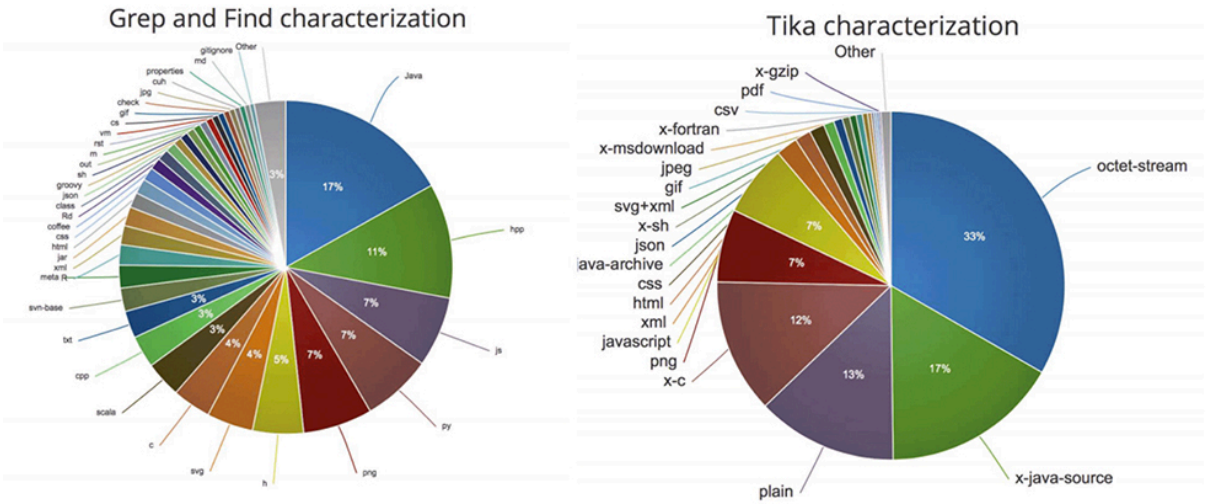
"A number of other projects came to us wanting to do this," Mattmann said. The EarthCube project of the National Science Foundation had a very large climate modeling repository and sought out Mattmann's team. "They asked us if all these scientists are putting licenses on their code, or whether they're open source, or if they're using the right components. And so we did a very big, large auditing for them," Mattmann said.

"That's where Wrangler comes in," Karanjeet Singh said. "We have all the tools and equipment on Wrangler, thanks to the TACC team. What we did was we just configured our DRAT tool on Wrangler and ran distributedly with the compute nodes in Wrangler. We scanned whole Apache SVN repositories, which includes all of the Apache open source projects."

The project Mattmann's team is working on early 2017 is to run DRAT on the Wrangler supercomputer over historically all of the code that Apache has developed since its existence - including over 200 projects with over two million revisions in a code repository on the order of hundreds of millions to billions of files.

"This is something that's only done incrementally and never done at that sort of scale before. We were able to do it on Wrangler in about two weeks. We were really excited about that," Mattmann said.

Apache Tika formed one of the key components to the success of DRAT. It discerns Multipurpose Internet Mail Extensions (MIME) file types and extracts its metadata, the data about the data. "We call Apache Tika the 'babel fish,' like 'The Hitchhiker's Guide to the Galaxy,'" Mattmann said. "Put the babel fish to your ear to understand any language. The goal with Tika is to provide any type of file, any file found on the Internet or otherwise to it and it will understand it for you at the other end...A lot of those investments and research approaches in Tika have been accelerated through these projects from DARPA, NASA, and the NSF that my group is funded by," Mattmann said.



File type breakdown of XDATA. Credit: Chris Mattmann

"A lot of the metadata that we're extracting is based on these machine-learning, clustering, and named-entity recognition approaches. Who's in this image? Or who's it talking about in these files? The people, the places, the organizations, the dates, the times. Because those are all very important things. Tika was one of the core technologies used - it was one of only two - to uncover the Panama Papers global controversy of hiding money in offshore global corporations," Mattmann said.

Chris Mattmann, the first NASA staffer to join the board of the Apache Foundation, helped create Apache Tika, along with the scalable text search engine Apache Lucerne and the search platform Apache Solr. "Those two core technologies are what they used to go through all the leaked (Panama Papers) data and make the connections between everybody - the companies, and people, and whatever," Mattmann said.

Mattmann gets these core technologies to scale up on supercomputers by 'wrapping' them up on the Apache Spark framework software. Spark is basically an in-memory version of the Apache Hadoop capability MapReduce, intelligently sharing memory across the compute cluster. "Spark can improve the speed of Hadoop type of jobs by a factor of 100 to 1,000, depending on the underlying type of hardware," Mattmann said.

"Wrangler is a new generation system, which supports good technologies like Hadoop. And you can definitely run Spark on top of it as well, which really solves the new technological problems that we are facing," Singh said.

Making sense out of big data guides much of the worldwide efforts behind 'machine learning,' a slightly oxymoronic term according to computer scientist Thomas Sterling of Indiana University. "It's a somewhat incorrect phrase because the machine doesn't actually understand anything that it learns. But it does help people see patterns

and trends within data that would otherwise escape us. And it allows us to manage the massive amount and extraordinary growth of information we're having to deal with," Sterling said in a 2014 interview with TACC.

One application of machine learning that interested NASA JPL's Chris Mattmann is TensorFlow, developed by Google. It offers open source commodity-based access to very large-scale machine learning. TensorFlow's Inception version three model trains the software to classify images. From a picture the model can basically tell a stop sign from a cat, for instance. Incorporated into Memex, Mattmann said Tensorflow takes its web crawls of images and video and looks for descriptors that can aid in "catching a bad guy or saving somebody, identifying an illegal weapon, identifying something like counterfeit electronics, and things like this."

"Wrangler is moving into providing TensorFlow as a capability," Mattmann said. "One of the traditional things that stopped a regular Joe from really taking advantage of large-scale machine learning is that a lot of these toolkits like Tensorflow are optimized for a particular type of hardware, GPUs or graphics processing units." This specialized hardware isn't typically found in most computers.

"Wrangler, providing GPU-types of hardware on top of its petabyte of flash storage and all of the other advantages in the types of machines it provides, is fantastic. It lets us do this at very large scale, over lots of data and run these machine learning classifiers and these tool kits and models that exist," Mattmann said.

What's more, Tensorflow is compute intensive and runs very slowly on most systems, which becomes a big problem when analyzing millions of images looking for needles in the haystack. "Wrangler does the job," Singh said. Singh and others of Mattmann's team are currently using Tensorflow on Wrangler. "We don't have any results yet, but we know

that - the tool that we have built through Tensorflow is definitely producing some results. But we are yet to test with the millions of images that we have crawled and how good it produces the results," Singh said.

"I'm appreciative," said Chris Mattmann, "of being a member of the advisory board of the staff at TACC and to Niall Gaffney, Dan Stanzione, Weijia Xu and all the people who are working at TACC to make Wrangler accessible and useful; and also for their listening to the people who are doing science and research on it, like my group. It wouldn't be possible without them. It's a national treasure. It should keep moving forward."

Provided by University of Texas at Austin