# Higgs: Bringing the power of data-mining to astrophysics, biology

September 19 2012, by Nick Statt


A data tape storage robot at Brookhaven's RHIC/ATLAS Computing Facility.

(Phys.org)—The art of data mining is about searching for the extraordinary within a vast ocean of regularity. This can be a painful process in any field, but especially in particle physics, where the amount of data can be enormous, and 'extraordinary' means a new understanding about the fundamental underpinnings of our universe. Now, a tool first conceived in 2005 to manage data from the world's largest particle accelerator may soon push the boundaries of other disciplines. When repurposed, it could bring the immense power of data mining to a variety

of fields, effectively cracking open the possibility for more discoveries to be pulled up from ever-increasing mountains of scientific data.

Advanced data management tools offer scientists a way to cut through the noise by analyzing information across a vast network. The result is a searchable pool that software can sift through and use for a specific purpose. One such hunt was for the Higgs boson, the last remaining elementary particle of the Standard Model that, in theory, endows other particles with mass.

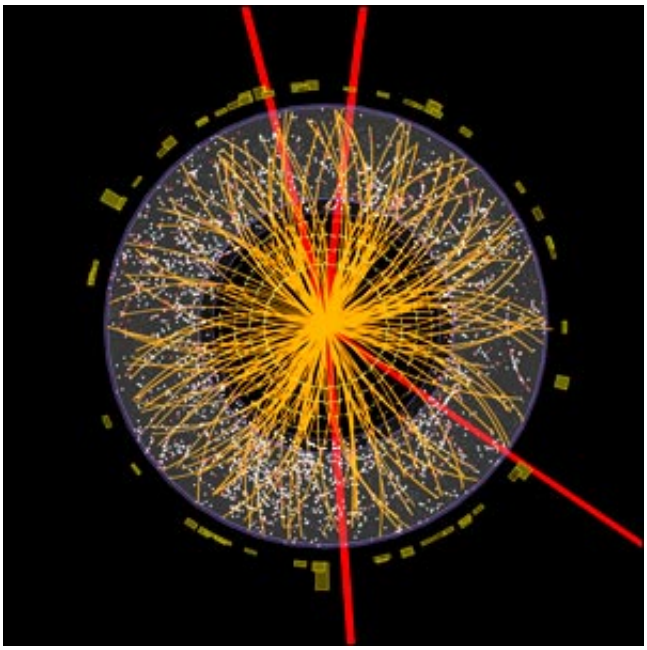
With the help of a system called PanDA, or Production and Distributed Analysis, researchers at CERN's Large Hadron Collider (LHC) in Geneva, Switzerland discovered such a particle by slamming protons together at relativistic speeds hundreds of millions of times per second. The data produced from those trillions of collisions—roughly 13 million gigabytes worth of raw information—was processed by the PanDA system across a worldwide network and made available to thousands of scientists around the globe. From there, they were able to pinpoint an unknown boson containing a mass between 125–127 GeV, a characteristic consistent with the long-sought Higgs.

The sheer amount of data arises from the fact that each particle collision carries unique signatures that compete for attention with the millions of other collisions happening nanoseconds later. These must be recorded, processed, and analyzed as distinct events in a steady stream of information.

Kaushik De, one of the two architects of the PanDA system and director of the University of Texas at Arlington's Center of Excellence for High Energy Physics, said it was extremely useful in the LHC experiments. "We could not have found the Higgs without it," he explained.

After the PanDA system proved invaluable in high energy and particle

physics experiments, scientists began considering an upgraded, universal version that could prove useful in other areas, like searching the sea of mysteries hidden in the cosmos. From the anomalies across the universe's roughly 14-billion year history to the elusive and invisible dark matter that constitutes about 84 percent of all matter, a data tool like PanDA could help scientists uncover something as revolutionary as the Higgs in the fields of astrophysics and cosmology.



This dense image of an ATLAS collision shows the individual particle signatures that indicated the discovery of something very close to the long-sought Higgs boson. Credit: ATLAS Experiment 2012 CERN.

"We are physicists, and we developed this software to be used by a high-energy physics experiment," said Alexei Klimentov, a physicist at Brookhaven Lab and the ATLAS Distributed Computing Coordinator. "But very soon we realized that it could be used in other areas."

So Klimentov and De, representing the computing partnership between Brookhaven Lab and UT Arlington that first developed the PanDA system in 2005, are heading up this new initiative with $1.7 million of U.S. Department of Energy Office of Science funding.

The current PanDA system is a based on a "data grid" concept in which it acts as part of a worldwide network – the LHC Computing Grid. The grid is strung together by a system of tiered data centers, ranging from Tier-0, the CERN computer center that receives the initial raw data, to 11 Tier-1 centers located across three continents where the data is reprocessed. (Brookhaven Lab is the sole tier-1 data center in the U.S. that distributes and analyzes data from the LHC's ATLAS experiment). From there, more than a hundred smaller Tier-2 centers store parts of the enormous data set and allow it to be accessed and analyzed.

All in all, the grid encompasses over 150 petabytes of disk space across 34 countries, which are all wired together through both private fiber optic networks and high-speed portions of the public Internet. At the CERN computing center, data can be transferred at the astounding speed of 10 gigabytes per second.

Once everything is stored and in place, PanDA comes into play. Researchers request certain parameters they would like to hone in on, and the job is sent to the Tier-1 and Tier-2 centers. The PanDA system sits in the center of the grid, able to communicate with those data centers and then transfer the requested data across the entire network.

If such a system were to be put in the thick of millions of gigabytes of astronomical telescope data, it could become a remarkable tool for probing the countless mysteries of space, offering insights into dark matter and the growing base of knowledge on extrasolar planets and the possibilities of extraterrestrial life.

Officials from the International Space Station's Alpha-Magnetic Spectrometer (AMS) experiment have also expressed interest in using the PanDA system. The AMS experiment works primarily by measuring cosmic rays and determining what they tell us about dark matter, antimatter and other forms of unusualness floating in space and invisible to the human eye. Since its installation in May of last year, the AMS has already recorded over 18 billion cosmic rays, all of which could use the support of an advanced data management tool. That means that with the help of PanDA, the study of the universe's origins, evolution and increasingly complex make-up might yield a major discovery sooner than scientists thought possible.

Biomedical applications are also promising, especially given the prominence of supercomputing and data mining in the fields of genetics and medicine. Applied to those fields, PanDA could yield another discovery like the recent finding that what was thought to be "junk" DNA actually contains integral gene switches that may have lasting effects on everything from cancer treatment to understanding the human genome. Such fundamental discoveries lie in the heart of big data sets, and PanDA gives scientists the power to efficiently move beyond the computing chaos and pull out the key points in the data.

But there are still challenges that need to be solved over the course of the three-year DOE grant to UT Arlington and Brookhaven Lab. One of their biggest tasks in the upgrade, said Klimentov, is equipping the PanDA system with current cloud computing technology. Because the existing grid system involves sending requests and transferring recreations of the data across high-speed wired networks, it is inefficient when compared with the possibilities of a worldwide cloud.

Provided by Brookhaven National Laboratory