insideBIGDATA

InsideBIGDATA Guide to

Scientific Research

by Daniel D. Gutierrez

BROUGHT TO YOU BY

Big Data for Scientific Research – An Overview

The rapid evolution of big data technology in the past few years has changed forever the pursuit of scientific exploration and discovery. Along with traditional experiment and theory, computational modeling and simulation is a third paradigm for science. Its value lies in exploring areas of science in which physical experimentation is unfeasible and insights cannot be revealed analytically, such as in climate modeling, seismology and galaxy formation. More recently, big data has been called the "the fourth paradigm" of science. Big data can be observed, in a real sense, by computers processing it and often by humans reviewing visualizations created from it. In the past, humans had to reduce the data, often using techniques of statistical sampling, to be able to make sense of it. Now, new big data processing techniques will help us make sense of it without traditional reduction.

Jim Gray, the late U.S. computer scientist from Microsoft in 2007 described a major shift going on in the area of scientific research as — "fourth paradigm" for scientific exploration and discovery.

He predicted that the collection, analysis, and visualization of increasingly large amounts of data would change the very nature of science. One of the goals of big data discussed in the book *The Fourth Paradigm*[1] is to make the scientific record a first-class scientific object. Fast forward to 2015 and we see distinct evidence for how the big data technology stack is facilitating this change.

This technology guide is geared toward scientific researchers working at universities and other research institutions (e.g. NASA, JPL, NIH, etc.) who may benefit from learning more about how big data is meaningfully transformative in the way it can be applied to the data collection and analysis part of their projects. Further, we'll illustrate how Dell big data technology solutions powered by Intel are actively helping scientists who are focused on their data, on their models and on their research results.

Here is a short-list of several scientific areas currently using or planning to use big data technology solutions to manage the influx of unparalleled amounts of data:

- **Astronomy –** the proposed Large Synoptic Survey Telescope (LSST) in Chile is expected to create 12.8 gigabytes of data every 39 seconds, for a sustained data rate of 330 megabytes per second. Over a ten-hour winter night, LSST will thus collect up to 13 terabytes.

[1] Tony Hey, Stewart Tansley, and Kristin Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery (Redmond, Wash.: Microsoft Research, 2009).

Contents

- **Genomics –** the Wellcome Trust Sanger Institute in Cambridge, UK can store 18 petabytes of data. All labs need to manipulate data to yield research results. As prices drop for high-throughput instruments such as automated genome sequencers, small biology labs can become big data generators. Biological data are much more heterogeneous than those in other scientific fields. They stem from a wide range of experiments that yield many types of information, such as genetic sequences, interactions of proteins or findings in Electronic Medical Records (EMRs). A single sequenced human genome is around 140 gigabytes.
- **Neuroscience –** the U.S. based BRAIN Initiative uses big data to map the human brain. By mapping the activity of neurons in the brain, researchers hope to discover fundamental insights into how the mind develops and functions, as well as new ways to address brain trauma and diseases. Researchers plan to build instruments that will monitor the activity of hundreds of thousands and perhaps 1 million neurons, taking 1,000 or more measurements each second. This goal will unleash a torrent of data. A brain observatory that monitors 1 million neurons 1,000 times per second would generate 1 gigabyte of data every second, 4 terabytes each hour, and 100 terabytes per day. Even after compressing the data by a factor of 10, a single advanced brain laboratory would produce 3 petabytes of data annually.
- **Climate sciences –** The NASA Center for Climate Simulation (NCCS) crunches massive amounts of climate and weather information, giving researchers eye-opening visibility into their data — currently around 32 petabytes. Climate and environmental sciences is an excellent proving ground for big data technology as the field has a wide variety and large volume of data which needs to be captured rapidly.
- **Health sciences –** the European Bioinformatics Institute in Hinxton, UK is one of the world's largest biology data repositories and currently stores 20 petabytes of data about genes, proteins and small molecules.

Many scientists are concerned that the data deluge will make it increasingly difficult to find data of relevance and to understand the context of collective data.

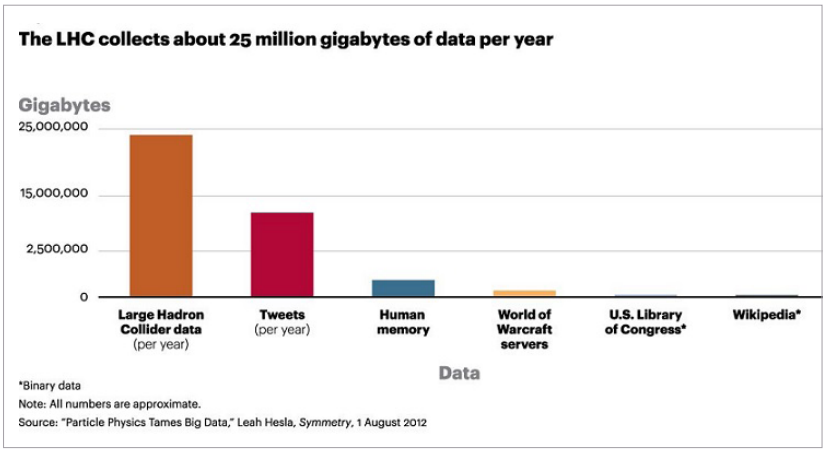
- **Cosmology –** the Square Kilometer Array (SKA) is one of the most ambitious science projects ever undertaken. A consortium of 10 nations, with the involvement of numerous university scientists and industrial companies, plans on setting up a massive radio telescope made up of millions of antennas spread out across vast swaths of southern Africa and Australia. When it's completed in 2024, the array will give astronomers insights into the evolution of the first stars and galaxies after the Big Bang so they can better understand the history of the universe and the nature of matter. Every day, the antennas will gather 14 exabytes of data and store about one petabyte.

Along with the many significant opportunities, data-intensive scientific research also will bring complex challenges. Many scientists are concerned that the data deluge will make it increasingly difficult to find data of relevance and to understand the context of collective data. In addition, the management of data presents increasingly difficult issues. For example, how do international, multidisciplinary and often competitive collaborations of researchers address challenges related to the creation and use of metadata, ontologies and semantics, data curation and still conform to the principles of security, privacy and data integrity? These challenges of a loosely connected community of researchers could be substantial.

There may be distinct challenges with the advent of big data coupling with scientific research, but as with all new technological paradigms, growing pains are de rigueur in anticipation of the benefits enabled for scientific progress. Caveats notwithstanding, there is no doubt that big data is quickly becoming an integral part of scientific research today.

Primary Motivators of Big Data vis-à-vis Scientific Research

With all of the discussion about big data these days, there continues to be frequent reference to the 3 V's that characterize big data: Volume (size of the data set), Velocity (rate at which data is generated), and Variety (types of data collected). These representations of big data are familiar to most researchers, however, they still are useful when assessing the important benefits brought to the table. Most researchers in scientific areas agree that the complexity of the data and its relationships are truly the biggest challenges at all scales and in most applications.



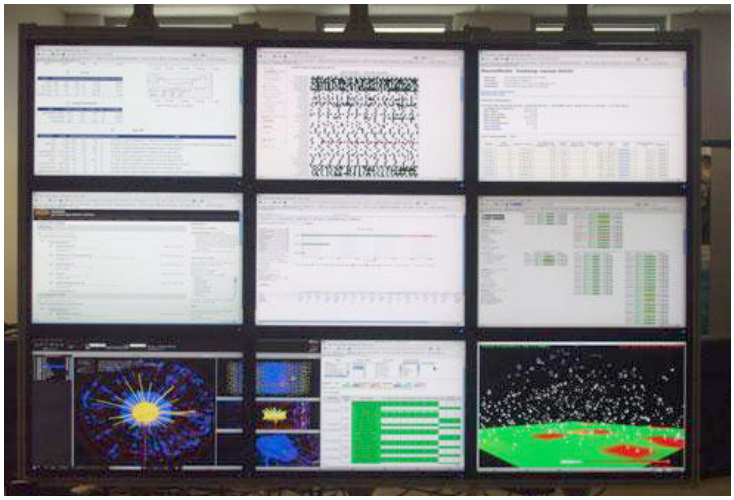
To illustrate how one high-profile, large-scale scientific research project has engaged the primary tenets of big data technology, let's consider the Large Hadron Collider (LHC), the world's premiere particle accelerator located at CERN (Conseil Européen pour la Recherche Nucléaire). CERN is the laboratory, and the LHC is the machine, near Geneva, Switzerland. As the most complex scientific human endeavor ever attempted, LHC is representative of a scientific research project with a strong international collaboration. The goal is to smash protons moving at 99.999999% of the speed of light into each other. The beams of protons collide in four experiment points, each collecting big data class data: Alice, ATLAS, CMS and LHCb.

After a 3 year run beginning in September 2008, the LHC confirmed the discovery of the elusive Higgs boson. The accelerator was shut off on February 14, 2013 and then underwent a 2 year retrofit in order to operate at higher energies - 13 teraelectronvolts (TeV). During 2012, the LHC collected 30 petabytes of data (about 100 terabytes per day during normal operation), for a total of 100 petabytes since inception. As of June 3, 2015 the LHC is now back in full operation.

CERN uses Apache Hadoop for storing the metadata of the experimental LHC data. The data is supported by 3 HBase clusters. The LHC experiment is a good example of big data 3 V's in the following ways:

- Open accessibility of data
- Volume of data
- Data transported in real-time for analysis
- Long term data retention
- Data is semi-structured

An important tool at LHC visualizes the real-time transfers of data in the Hadoop Distributed File System (HDFS) system. Called HadoopViz, this visualization technology shows all packet transfers in the HDFS system as raindrops arcing from one server to the other. The figure below shows a collection of displays in the LHC control center — the Hadoop status page as the screen in the top-right, and the HadoopViz Visualization of Packet Movement as the screen in the bottom-right.



Big Data Technology for Scientific Research

The continued and rapid evolution of big data technology and services has formed a fertile foundation for scientific applications in the past several years. We've seen big data hardware and software solutions promulgate access to analytics and methods of statistical learning like never before. And one thing researchers have learned along the way is that management of computing resources is one of the primary questions to be answered with big data. It is not just a case of determining the scale of resources needed for a project, but also how to configure them, all within the available budget. For example, running a large project on fewer machines might save on hardware costs but will result in a longer project timeline. In some cases, scientific big data are being stored in the cloud instead of on conventional hardware in a research lab. Instead of having to invest in the infrastructure of an on-premises *High Performance Computing* (HPC) cluster to analyze the data, some researchers are using HPC and big data methods in the cloud. The disadvantage of the cloud is that large data sets must be transferred to cloud storage.

In this section we'll examine specific areas of big data technology that scientific researchers are deploying in order to see significant increases in their ability to manage the scientific data deluge.

Colliding Worlds of HPC and Big Data

In order to address the needs of data-centric scientific research projects, the profiles of traditional HPC and big data are merging and becoming closely intertwined. The compute nature of HPC is finding significant benefit from big data analytics and its ability to process high volume, high velocity data sets. The current most effective software platform for big data analytics — Hadoop — has its classic architecture consisting of HDFS and MapReduce, running on commodity cluster nodes, and the HPC environment has a different architecture where compute is distinct from the storage solution. You'd like to leverage your current investment in HPC by doing big data analytics on the architecture you already have. This is where two worlds come together.

One thing researchers have learned along the way is that management of computing resources is one of the primary questions to be answered with big data. It is not just a case of determining the scale of resources needed for a project, but also how to configure them, all within the available budget.

An area of scientific research that's benefiting from HPC and big data merging is genomics. The advancement of this area of research depends on the availability of HPC because of the compute resources needed to process genomic data sets. But these problems also require solving big data issues like analyzing the data, making sense of what's in the data, identifying what patterns emerge from the data, and more. Essentially, every aspect of what genomic researchers do is becoming an opportunity to capture, analyze and use big data. This is the same net effect seen from the perspective of many other scientific disciplines.

Data Sources and Data Integration

Scientific researchers routinely collect extremely large data sets, primarily for computational analysis with an HPC system. These data sets can also be analyzed with big data tools to look for valuable insights with data visualization tools or advanced analysis algorithms. The difference between HPC and big data analytics is primarily that HPC is CPU bound, whereas big data analytical problems are IO bound. As the two environments continue to merge, researchers can apply big data analytics against a primarily HPC data set without moving the data set from the HPC environment to a Hadoop Cluster with HDFS.

The Intel® Enterprise Edition for Lustre software unleashes the performance and scalability of the Lustre for HPC workloads, including big data applications becoming common within today's research labs. Dell built the reference architecture where Lustre is used in place of HDFS.

A paradigm shift in big data analytics relative to scientific research use cases, as well as other use cases where there is value from analyzing HPC data sets, necessitates a new direction away from the common HDFS architecture, and towards using MapReduce on the Lustre parallel file system. The Intel® Enterprise Edition for Lustre software unleashes the performance and scalability of the Lustre for HPC workloads, including big data applications becoming common within today's research labs. Dell built the reference architecture where Lustre is used in place of HDFS. Further, Dell layers analytics on top of this architecture as well.

A key component to connecting the Hadoop and Lustre ecosystems is the Intel Hadoop Adapter for Lustre plug-in (Intel HAL). Intel HAL is bundled with the Intel Enterprise Edition for Lustre software. It allows the users to run MapReduce jobs directly on a Lustre file system. The immediate benefit is that Lustre is able to deliver faster, stable and easily managed storage for the MapReduce applications directly. A potential long term benefit using Lustre as the underlying Hadoop storage would be a higher raw capacity available when compared to HDFS due to the three time replication as well as the performance benefits of running Lustre on InfiniBand connectivity.

Researchers are interested in multiple additional classes of data sets. Many of them may be on premise and consist of multiple types of data. Other data sources may come from publicly available sites, or purchased from one of the many services that provide data sources to labs. Dell's Data Integration Platform as a Service called Boomi could be used for such applications.

Hadoop

The Hadoop distributed computing architecture increasingly is being deployed for scientific applications requiring big data capabilities, specifically managing, collecting and analyzing the data. Dell™ Apache™ Hadoop® solutions for big data provide an open source, end-to-end scalable infrastructure that allows you to:

- Simultaneously store and process large data sets in a distributed environment—across servers and storage—for extensive, structured and unstructured statistical learning and analysis
- Accommodate a wide range of analytic, exploration, query and transformation workloads
- Tailor and deploy validated reference architectures
- Reduce project costs
- Drive important insights from scientific data

Take the complexity out of analyzing research data sets. With Dell's extensive Hadoop-ready library of analytics solutions, you can easily create "what if" scenario dashboards, generate graphs for relationship analysis and innovate over legacy systems. Dell has teamed up with Cloudera and Intel to provide the most comprehensive, easy-to-implement big data solutions on the market for research applications.

Dell's tested and validated Reference Architectures include Dell PowerEdge servers with Intel® Xeon® processors, Dell Networking and Cloudera Enterprise. This broad compatibility can help your research group build robust Hadoop solutions to collect, manage, analyze and store data while leveraging existing tools and resources. The Intel® Xeon® powered Dell | Cloudera solution can give your research group everything it needs to tackle big data challenges including software, hardware, networking and services.

Spark

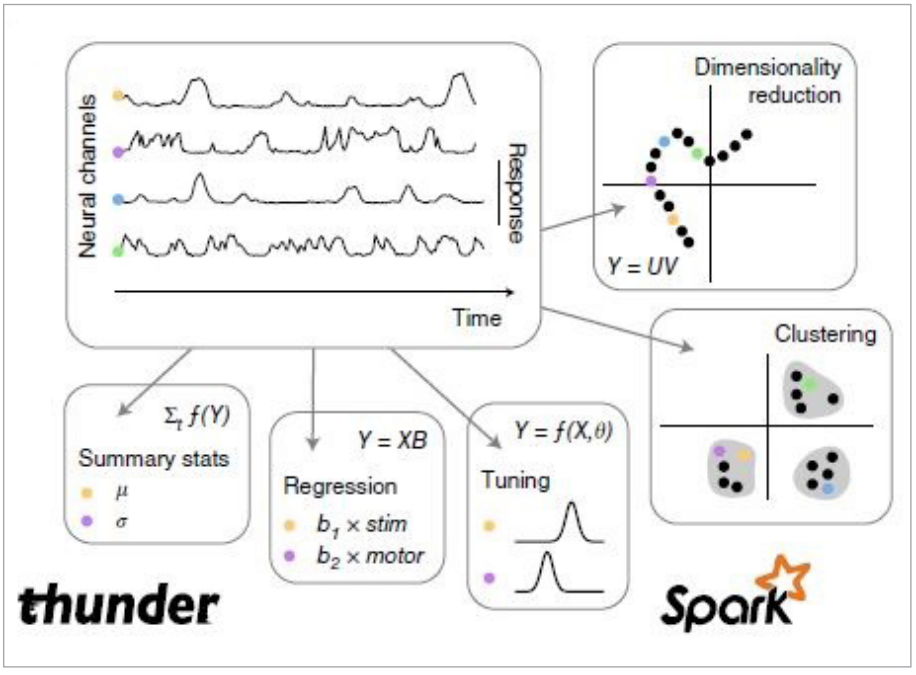
Apache Spark is another distributed processing environment that's gained much interest in the scientific community. Spark is an open-source platform for large-scale distributed computing. MapReduce is a widely adopted programming model that divides a large computation into two steps: a Map step in which data are partitioned and analyzed in parallel, and a Reduce step, in which intermediate results are combined or summarized. Many analyses can be expressed in this model, but systems like Hadoop have key weaknesses in that data must be loaded from disk storage for each operation, which can slow iterative computations (including many machine learning algorithms), and makes interactive, exploratory analysis difficult. Spark extends and generalizes the MapReduce model while addressing this weakness by introducing a primitive for data sharing called a resilient distributed data set (RDD). With Spark, a data set or intermediate result can be cached in the memory across cluster nodes, performing iterative computations faster than with Hadoop MapReduce and allowing for interactive analyses.

One example of a research project taking advantage of Spark is the Howard Hughes Brain Institute. The project's goal is to understand brain function by monitoring and interpreting the activity of large networks of neurons during behavior. An hour of brain imaging for a mouse can yield 50-100 gigabytes of data. The researchers developed a library of analytical tools called *Thunder* which is based on Spark using the Python API along with existing libraries for scientific computing and visualization. The core of Thunder is expressing different neuroscience analyses in the language of RDD operations. Many computations such as summary statistics, regression and clustering can be parallelized using MapReduce.

Statistical Analysis Software

The scientific community is fortunate to have many quality statistical environments and analytics tools with which to use for developing analytical pipelines that connect data to models and then to predictions — SAS, SPSS, Matlab and Statistica. On the open source front, many researchers employ tools like R and Python — each containing vast libraries of statistical functions and machine learning algorithms.

It is important to transform complex and time-consuming manipulation of scientific data sets into a fast and intuitive process. Statistica Big Data Analytics from Dell combines search and analytics in a single, unified environment. Statistica is an advanced content mining and analytics solution that is fully integrated, configurable and cloud-enabled. It deploys in minutes and brings together natural language processing, machine learning, search and advanced visualization.



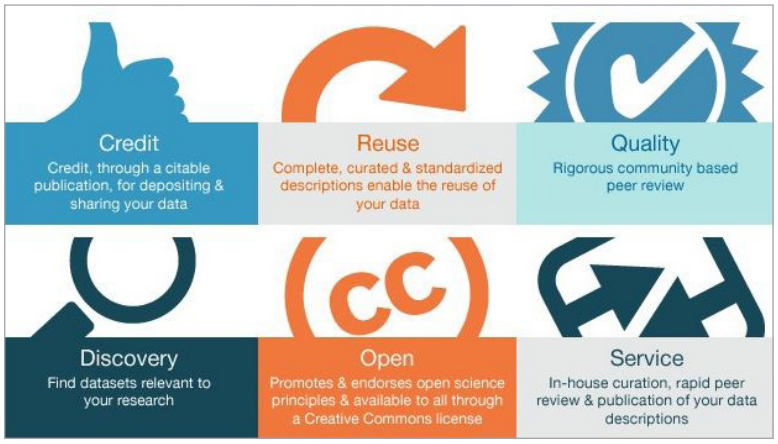
Big Data and Open Science Data

With scientific data sets growing ever larger, researchers are finding that the bottleneck to discovery is no longer a lack of data but an inability to manage, analyze, and share their large data sets. Individual researchers can no longer download and analyze the important data sets in their scientific fields onto their own computers. The goal of the recent trend toward open scientific data is to remove the bottleneck to discovery by providing researchers with access to a variety of key data sets across scientific disciplines and the computing infrastructure to allow scientists to easily manage and share their data and analysis. Big data technologies serve to facilitate these goals by allowing for unparalleled data storage, and analytical capabilities.

More and more these days, research grant proposals require an "open data" element where all data collected by the project is to be made openly available through an easily accessed data store. One of the first, and best known data sharing projects is The Human Genome Project. The sequencing of the human genome was a massive undertaking by many researchers around the world. The results of their efforts have greatly advanced many areas of research in the life sciences and healthcare over the past decade and half, but none of that would have been possible if the genomic sequences had not been widely available. Instead, anyone can freely download human genomic data and use it in conjunction with big data technology. This is what open data is all about.

Here are some key reasons for sharing data and making scientific data open:

- Clearly documents and provides evidence for research in conjunction with published results
- Meet copyright and ethical compliance (i.e. HIPAA)
- Increases the impact of research through data citation
- Preserves data for long-term access and prevents loss of data



- Describes and shares data with others to further new discoveries and research
- Prevent duplication of research
- Accelerates the pace of research
- Promotes reproducibility of research

To promote this level of data sharing, many scientific journals require their authors to make all data underlying their articles openly available from the moment of publication of the article. Opening up research data makes it much easier for other scientists to build upon that work and advance the field.

The open data trend has yielded some interesting initiatives to bring open data to the mainstream of scientific research. A good example is Scientific Data, an open-access, peer-reviewed, online-only publication from Nature Publishing Group, containing descriptions of scientifically valuable datasets, The goal of the publication is to assist researchers publish, discover and reuse research data. *Scientific Data* is open to submissions from a broad range of natural science disciplines, including descriptions of big and small data, from major consortiums and single labs. Scientific Data primarily publishes Data Descriptors, a new type of scientific publication designed to promote an in-depth understanding of research datasets. Data Descriptors combine traditional scientific article content with structured information curated in-house, and are devised to maximize data reuse and enable searching, linking and data-mining.

Case Studies: Big Data and Scientific Research

In order to illustrate how the scientific community is rapidly moving forward with the adoption of the big data technology stack, in this section we'll consider a number of research projects that have benefited from these tools. In addition, these project profiles show how big data is steadily merging with traditional HPC architectures. In each case, significant amounts of data are being collected and analyzed in the pursuit of unparalleled understanding of nature and the universe.

Tulane University

As part of its rebuilding efforts after Hurricane Katrina, Tulane University partnered with Dell and Intel to build a new HPC cluster to enable the analysis of large sets of scientific data. The cluster is essential to power big data analytics in support of scientific research in the life sciences and other fields. For example, the school has numerous oncology research projects that involve statistical analysis of large data sets. Tulane also has researchers studying nanotechnology, the manipulation of matter at the molecular level, involving large amounts of data.

> "The Cypress cluster, based on Dell technology, is going to bring to us a whole new level of infrastructure that will make our current faculty members want to stay here. It will also help us to attract new faculty members."
>
> *Dr. Nicholas Altiero, Dean of Science and Engineering, Tulane University*

Tulane worked with Dell to design a new HPC cluster dubbed *Cypress*, consisting of 124 Xeon-based Dell PowerEdge C8220X server nodes, connected through the high-density, low-latency Z9500 switch, providing a total computational theoretical peak performance of more than 350 teraflops. Dell also leveraged their relationship with Intel, who in turn leveraged their relationship with leading Hadoop distribution Cloudera – allowing Tulane to do big data analytics using Hadoop in an HPC environment.

Using Cypress enables Tulane to conduct new scientific research in fields such as epigenetics (the study of the mechanisms that regulate gene activity), cytometry (the measurements of the existence of certain subsets of cells within a kind of tissue in the human body), primate research, sports-related concussion research, and the mapping of the human brain.

Arizona State University

ASU worked with Dell to create a powerful HPC cluster that supports big data analytics. As a result, ASU built a holistic *Next Generation Cyber Capability* (NGCC) using Dell and Intel technologies that is able to process structured and unstructured data, as well as support diverse biomedical genomics tools and platforms.

ASU turned to Dell and Intel to expand its HPC cluster. The resulting NGCC delivers 29.98 teraflops of sustained performance for HPC, big data and massively parallel (or transactional) processing with 150 nodes and 2,400 cores. The HPC side of the NGCC includes 100 Dell PowerEdge M620 servers with Intel® Xeon® E52660 processors and 1,360 cores. NGCC's transactional side includes 20 Dell PowerEdge M420 servers, each with Intel Xeon E5-2430 processors.

HPC and Cloudera's Hadoop distribution upon which NGCC is based can handle data sets of more than 300 terabytes of genomic data. In addition, ASU is using the NGCC to understand certain types of cancer by analyzing patients' genetic sequences and mutations.

National Center for Supercomputing Applications

The *National Center for Supercomputing Applications* (NCSA) provides computing, data, networking, and visualization resources and services that help scientists, engineers, and scholars at the University of Illinois at Urbana-Champaign and across the country. The organization manages several supercomputing resources, including the iForge HPC cluster based on Dell and Intel technologies.

One particularly compelling scientific research project that's housed in the NCSA building is the Dark Energy Survey (DES), a survey of the Southern sky aimed at understanding the accelerating expansion rate of the universe. The project is based on the iForge cluster and ingests about 1.5 terabytes daily.

Translational Genomics Research Institute

To advance health through genomic sequencing and personalized medicine, the *Translational Genomics Research Institute* (TGen) requires a robust, scalable high-performance computing environment complimented with powerful big data analytics tools for its Dell | Hadoop platform. TGen optimized its infrastructure by implementing the Dell Statistica analytics software solution and scaling its existing Dell HPC cluster with Dell PowerEdge M1000e blades, Dell PowerEdge M420 blade servers and Intel processors. The increased performance accelerated experimental results, enabling researchers to expand treatments to a larger number of patients.

As gene sequencers increase in speed and capacity, TGen scaled its HPC cluster to 96 nodes. This was done with cutting-edge PowerEdge servers that featured Intel® Xeon® processors that achieved 19 teraflops of processing. The cluster supports 1 million CPU hours per month and 100% year-to-year data growth. To manage this level of big data, TGen scaled its existing Terascala storage so it can hold 1 petabyte.

Summary

The explosion of big data is transforming how scientists conduct research. Grants and research programs are geared at improving the core technologies around managing and processing big data sets, and speeding up scientific research with big data. The emergent field of data science is changing the direction and speed of scientific research by letting people fine-tune their inquiries by tapping into giant data sets. Scientists have been using data for a long time. What's new is that the scale of the data is overwhelming, which can be an infrastructure challenge. Researchers now need to be able to tame large data sets with new big data software tools and HPC to make rapid advances in their fields.