National Center for Genome Resources: Managing a 14-Terabyte Data Store with SAP® Sybase® ASE

NCGR
National Center for Genome Resources

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Company
National Center for Genome Resources

Industry
Life sciences

Products and Services
Sequencing and development of custom bioinformatics resources for research communities

Web Site
www.ncgr.org

SAP® Solutions
SAP Sybase® Adaptive Server® Enterprise

With backing from the National Science Foundation, National Institutes of Health, and other private academic and commercial clients, the National Center for Genome Resources (NCGR) has taken a leadership position in the bioinformatics arena since its establishment in 1994. NCGR uses SAP® Sybase® Adaptive Server® Enterprise (SAP Sybase ASE) to **manage its ever-growing pool of genome information and support the multiple applications** accessing it.

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Supporting the human genome project

The nonprofit National Center for Genome Resources (NCGR) spun off from Los Alamos National Laboratory to support the Human Genome Project. Since 1994, this information-intensive company has relied on SAP Sybase ASE to hold its genetic sequence data. SAP Sybase ASE plays a key role in the success of the company and the pursuit of its mission to improve human health and nutrition. To supply the endless demands of scientists and practitioners, NCGR uses SAP Sybase ASE to manage a total data store of around 14 terabytes (TB), staking its claim among very large databases (VLDBs).

14 terabytes

Total data storage amount

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Using bioinformatics on the building blocks of life

Genomic research looks at differences and similarities between genetic sequences. The human genome and many fully or partially mapped plant and animal DNA sequences are used as baseline data. NCGR's services include bioinformatics – an in-depth analysis of customer-submitted sequences and statistical comparison against existing reference baselines. NCGR also provides sequencing services, analyzing test tube samples of organic material to determine genetic sequences.

NCGR uses SAP Sybase ASE for its data in multiple applications, including the Grindstone Lab Information Management System, which stores mission-critical information in support of the genome sequencing lab, and sequencing data for its Alpheus bioinformatics application. Alpheus data includes two basic sets of information: 7 TB of reference sequences and another 7.5 TB of new data pipelined from the resequencing instruments. The new data is compared against the reference sequences. To maintain manageable table sizes, the data is partitioned by species and sample. NCGR provides these services and information to a variety of clients – from academic organizations to commercial entities, such as agriculture and biotech industries. Neil Miller, principal software engineer and informatics team lead at NCGR, explains, "Many of our clients come to us for sequencing and data analysis assistance. Our industry is currently undergoing what is called the 'second genome revolution' because, in the last year or two, the cost of sequencing has dropped dramatically. However, it still costs more than a small lab might be able to afford, so they pay us to do the sequencing. We also perform complex informatics for many customers."

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Analyzing simple data in daunting quantities

The basics of genome research are very simple: adenine, cytosine, guanine, and thymine (ACGT) are the four bases which form pairs to make up the rungs of the DNA helix. The technology stores these (essentially simple character strings) as a group of 36 to 46 characters that make up a single sequence. The diversity of life on earth is in the number and ordering of these bases.

The human genome has about three billion base pairs, comprising approximately 27,000 individual genes that are grouped into 23 pairs of chromosomes. Species like frogs, with multiple life stages, have even more. While the basic information is very simple, the complexity exists in the quantity and the associated issues of storage, retrieval, annotation, and pattern matching. NCGR has six sequencing instruments that analyze samples. Each instrument produces 50 million new sequences for comparison every three days. This translates into adding 32 GB to SAP Sybase ASE with each run, amounting to over 1/4 TB of data each week. In fact, during the next year, NCGR plans to accommodate loading 60 TB of data.

50 million

New sequences produced in 3 days by 1 instrument

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Using Alpheus – 24x7 software as a service in genome research

A few years ago, a group of scientists approached NCGR to work on a Web-based bioinformatics system. Miller recalls, "They had this idea for sequencing, but there were no available informatics tools to handle what they needed. They asked us to partner on the project. Over the course of six intense weeks, we created the first draft of the system. Initially, we designed the system to service one particular client. Once we realized the value contained in the tool, we began modifying it to broaden its use. The result is our hosted Alpheus system."

Alpheus is NCGR's Sequence Variant Detection Pipeline application. The Web-based, software-as-a-service application imports user data in multiple, popular data formats and matches the imported sequences against NCGR's reference data, looking for notable characteristics and differences. The results are reported both visually and with downloaded data sets. Alpheus is surprisingly easy to use considering the breadth of functionality contained in a Web application. The back end of Alpheus is SAP Sybase ASE, providing massive quantities of data adapting to NCGR's multiple tool sets. Currently, NCGR hosts 49 Alpheus databases, containing data for 22 different organisms and used in a wide variety of studies in the areas of human health and nutrition. Kathy Myers, the senior database administrator at NCGR, describes Alpheus, "We have the ability to drill down into a gene and see what it looks like in a graphical context, right down to the actual nucleotides. Alpheus is a 24x7 system providing access worldwide. One client located and identified a new gene in a plant – a huge scientific breakthrough – using SAP Sybase ASE and the Alpheus tool."

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Innovating bioinformatics applications with SAP Sybase ASE

While most businesses and development organizations refine existing, well-understood areas of knowledge, NCGR has been an early participant in genome information study and continues to innovate in a rapidly evolving field. "This is a brave new world in genetics and we've managed to become one of the largest centers doing this type of work. We have a terrific history in creating the software tools and analysis techniques for genetic data; and in this field, that's actually a fairly new capability," says Myers.

Miller adds, "Our main product was created at a time when nothing existed that enabled scientists and doctors to take advantage of the advances in sequencing technology. Using a relational database is a key component of creating the tools to help people analyze the data. Before that, without these kinds of tools, scientists were left holding this enormous bag of sequence data they didn't know how to make sense of, and they were scrambling to find tools they could use on their own."

The NCGR team has developed many applications featuring SAP Sybase ASE as the central database. The tools are written in Java, PERL, and Ruby on Rails. "Quite simply, it's robust. We trust our data integrity to it," says Miller.

> "SAP Sybase ASE has been remarkably stable. It is an easy database to use and an easy database to understand. . . . The few times when I have needed technical support, it's been terrific. I immediately got to people who knew what they were doing, and they quickly gave me solutions to the few issues I had."
>
> Kathy Myers, Senior Database Administrator, National Center for Genome Resources

Executive overview

Company

Vision

Why SAP

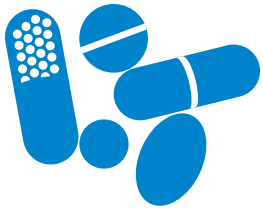
Implementation

Solution

Benefits

Future plans

Applying technology to human health for the greater good



NCGR's scope provides tools for scientists, commercial users, and its own world-class research projects. For example, NCGR is internationally recognized for its work on identifying genetic components of schizophrenia. This type of identification is a boon for early diagnosis and treatment. It reduces guesswork for a disease that untreated, or where treatment is delayed, can be devastating for sufferers and families.

Another study is a public-private multidisciplinary collaboration involving investigators at NCGR, Duke University Medical Center, Henry Ford Hospital, Eli Lilly, Pfizer, Roche Diagnostics, and Metabolon. This study uses advanced bioinformatics technologies to identify specific biomarkers in blood samples to predict outcomes in sepsis and community-acquired pneumonia. Biomarker-based tests enable patient-specific diagnoses and early targeted treatments with higher success rates. Mesothelioma is a cancer linked to asbestos exposure. With NCGR's help, doctors from the International Mesothelioma Program at Brigham and Women's Hospital can characterize genetic mutations found in the tumors. Shedding light on the underlying processes of the disease and the specific genes involved in the mutation is critical to ongoing prevention and treatment efforts.

The predictive nature of genetics is saving lives by identifying individuals who are at risk from what are benign exposures for most people. For example, NCGR recently worked with a company studying adverse reactions to a smallpox vaccine; the reaction appears to have a genetic component. Perfecting this type of test can bring a drug to large groups of people while ensuring the safety of the few who would suffer.

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Insuring the food supply

In the plant world, NCGR's Legume Information System catalogs the genetic similarities and differences of beans. Beans provide a ready source of protein throughout the world and easily grow in depleted soils. The Legume Information System integrates genetic and molecular data from multiple legume species for use in agricultural research. Water molds are a very common plant pathogen, and worldwide they cause crop and ecological damage that are calculated in the hundreds of billions of dollars each year. Water molds impact food crops, ornamentals, forest products, and seafood. NCGR is highly involved in analyzing and annotating gene sequences of various water molds. Indeed, anything living, even entities like viruses that are not necessarily a form of life, are worthy of study and analysis. Using SAP Sybase ASE, the National Center for Genome Resources has become a pillar of this rapid-growth industry.

"Doing resequencing, we're learning more about human genetics, and SAP Sybase ASE has been a big part of that for us."

Neil Miller, Principal Software Engineer and Informatics Team Lead, National Center for Genome Resources

Executive overview

Company

Vision

Why SAP

Implementation

Solution

Benefits

Future plans

Looking ahead

Over the lifecycle of SAP Sybase ASE at NCGR, its use has evolved. "SAP Sybase ASE will continue to be our solution for the individual database projects that will be generating data in the hundreds of gigabyte range," says Myers. With regard to data sizes, Myers notes, "This whole area of research becomes a problem of statistics. How many healthy people will we need to sequence to come up with a true normal – a true genetic sequence to compare against? Once we establish a range for normal, then how many sets of patients will we need to compare against to have a statistically significant number? The numbers and amounts become enormous; for a database person this is a very exciting time."

With the second genome revolution, data sizes are exploding; and at nearly 15 terabytes, with a projected bloom in data store size on the near horizon, NCGR is looking to move into the next level of very large storage systems. A prime candidate is SAP Sybase IQ, a column-based analytics server. With a simple migration path from SAP Sybase ASE, many customers use a distributed data management architecture with a mix of SAP Sybase ASE and SAP Sybase IQ. As NCGR stores more reference sequences to use as baselines for informatics analysis, the needs are changing from a highly transactional based approach into one of a large data warehouse.

"SAP Sybase ASE is really the cornerstone of our research. It performs remarkably well, given the exceptional amount of data we are throwing at it right now."

Kathy Myers, Senior Database Administrator, National Center for Genome Resources

Executive overview

Company

Vision

Why SAP

Implementation

Solution

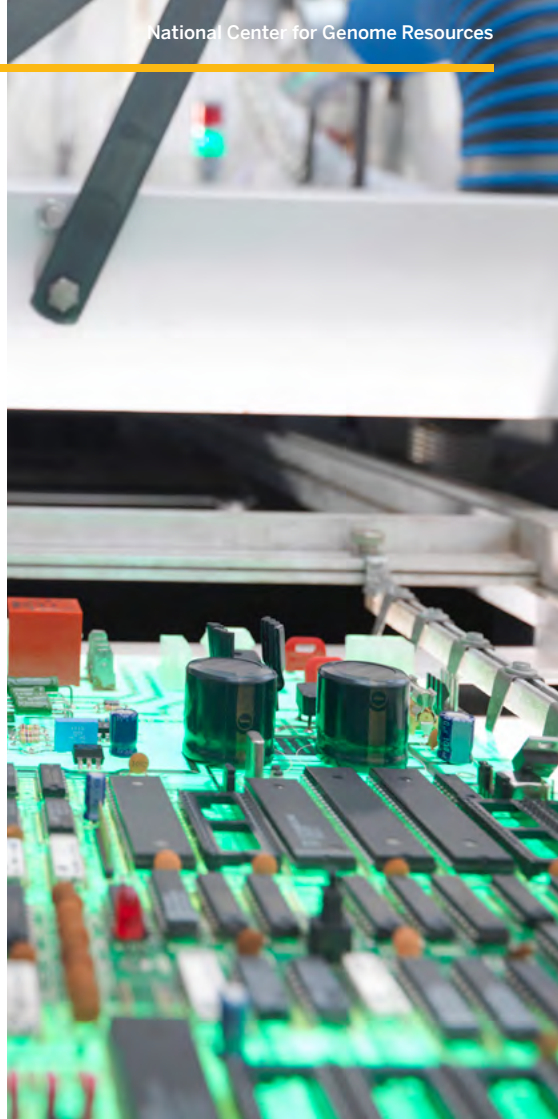
Benefits

Future plans

Hitting the target

The National Center for Genome Resources has a concrete and powerful mission: to improve human health and nutrition by genome sequencing and analysis. NCGR plays an important role in genome research by furthering the state of the art in genetic research. It could be said that genetic research has reached the end of the beginning. The core elements are in place, and with increasingly affordable sequencing costs, what comes next is an explosion of baseline sequence libraries and a surge in applied genome techniques.

The National Center for Genome Resources is a non-profit organization that is literally paying dividends to future generations.



CMP21221 (12/08)

Copyright/Trademark | Privacy | Impressum