CASE STUDY
Intel® Distribution for Apache Hadoop* Software
Intel® Xeon® Processor E5 Family
Biotechnology/Life Sciences
Big Data

NextBio Powers Genomic Data Analytics Breakthroughs with Intel® Technologies

Intel® Distribution for Apache Hadoop* software and Intel® Xeon® processor E5 family deliver scale and performance for life sciences computing

NEXTBIO

"We got an 85 to 90 percent increase in total throughput on the Intel processor-based cluster, which is an impressive result. The Intel-based cluster had 59 percent fewer nodes but 70 percent more cores, so it gives us a tremendous increase in density and horsepower."

– Richard Theige, Senior Director of Operations, NextBio

Can big data cure cancer—or at least enable personalized treatments that improve results and minimize side effects? Advances in analyzing molecular, genomic, and clinical data are poised to drive breakthroughs in understanding, treating, and curing cancer and other diseases. But progress requires ongoing improvements in scale, performance, and cost-effectiveness to analyze vast data sets and keep pace with the rising output of DNA sequencing machines and other data sources. NextBio's big data platform is incorporating the Intel® Distribution for Apache Hadoop software and Intel® Xeon® processors into its software-as-a-service (SaaS) cloud offerings to help meet those needs.

CHALLENGES

- **Enable progress in genomics-based research.** NextBio's customers use its sophisticated analytics engine and curated data sources in critical research and clinical studies.
- **Optimize the data center.** NextBio needs an optimized data center to handle the performance and growth needs of its customers' data-intensive analytics.

SOLUTIONS

- **Intel® Distribution for Apache Hadoop software.** Intel Distribution for Apache Hadoop software provides the advantages of open standards and is optimized for great performance and throughput on Intel Xeon processor-based infrastructure. Intel® Manager for Apache Hadoop software offers robust tools to streamline setup, management, security, and troubleshooting for Hadoop clusters.
- **Intel Xeon processors.** The Intel Xeon processor E5 family delivers balanced computing and I/O performance for big data analytics.

TECHNOLOGY RESULTS

- **Performance.** With highly tuned hardware and software, NextBio accelerates the processing of massive data sets and gains a sixfold data reduction using native Hadoop compression.
- **Density.** NextBio saw an 85 to 90 percent increase in total throughput on an Intel Xeon processor E5-2680-based cluster with 59 percent fewer nodes but 70 percent more cores than its previous-generation cluster.
- **Scale and reliability.** With distributed data and computation, NextBio can partition its massive data sets into manageable sizes and avoid creating a single point of failure.

BUSINESS VALUE

- **Faster return on investment (ROI) for customers.** NextBio customers can take advantage of falling prices for genomics analysis and rising volumes of genomic data, helping researchers make faster progress and pharmaceutical companies speed their drug development and approval processes.
- **Scalable growth, operational savings.** NextBio reduces costs for server, storage, network infrastructure, and power. Its optimized data center can support more customers and larger data sets, positioning NextBio to drive the next generation of genomics-based innovation.

Intel® Distribution for Apache Hadoop* software focuses on delivering a total solution optimized across software, server, storage, and network infrastructure

"There was no way we could scale in a cost-effective manner without the use of Apache Hadoop software on commodity hardware. That's the bottom line."

– Satnam Alag, Chief Technology Officer and Vice President of Engineering, NextBio

Big Data, Revolutionary Science

Modern medicine is undergoing a transformation, powered by rapid reductions in the cost of genome sequencing. Soon, your DNA profile may become as much a part of your medical record as your latest electrocardiogram or CT scan, and your medical team may use it to develop personalized treatment plans tailored to your unique data profile.

Founded in 2004, NextBio is an innovator in big data analytics for life sciences research and translational medicine—the discipline that works to convert basic research findings into clinical breakthroughs. "Genomics is seeing a revolution in terms of the volume, velocity, and price of the molecular data being generated," says Satnam Alag, chief technology officer and vice president of engineering for NextBio. "NextBio is at the intersection of big data and genomics. By using big data technologies and secure cloud-based infrastructure, we can compare one genomic data set against billions of other biological data points. We're now able to meaningfully interpret a person's molecular data in a very short time, and that has phenomenal implications for research and patient care. NextBio's big data platform can mine large numbers of these big data sets in aggregate to help researchers discover new biomarkers and drug targets."

NextBio has developed sophisticated SaaS platforms that help life science researchers extract insights from a variety of public and proprietary genomic databases. The company curates multi-terabyte data sets of genomic and other data to ensure the data's quality, and provides advanced capabilities for search, correlation, analysis, and collaboration. Its customers include pharmaceutical leaders as well as a range of life sciences research centers.

With NextBio's easy-to-use interface and application programming interfaces (APIs), customers often combine the curated public data with their own private databases to gain new insights. For example, drug companies might search NextBio's billions of precomputed correlations and combine them with their own privacy-protected patient records, looking for biomarkers that suggest which patients might benefit from a particular treatment or which might experience side effects.

'Really Big, Big Data'

NextBio deals with stunningly large data volumes, much of it semi-structured. Each human genome has approximately 4 million variants, and a person's full genome, with its 3.2 billion base pairs, can take as much as 10 million rows to represent in a traditional database. "When you think of the billions of people in the world, you're looking at really big, big data," says Richard Theige, senior director of operations at NextBio. "It would be unmanageable without the distributed computing that we get with Apache Hadoop software."

The cost of genome sequencing has been dropping so dramatically that many in the industry say that a price tag of USD 1,000 to sequence a person's genome is well within reach. That drop in price will only accelerate the growth of genomic data.

Eyeing these trends and understanding the deep value such data can provide, NextBio in 2008 became an early adopter of the Apache Hadoop software framework. The company pairs the open source distributed data framework with scalable 1U and 2U Intel® architecture-based servers and modular, standards-based storage.

"There was no way we could scale in a cost-effective manner without the use of Apache Hadoop on commodity hardware," Alag says. "That's the bottom line. You write your algorithm once, and if you get it right, you can scale very easily by adding more nodes to the Hadoop cluster."

For many computing scenarios, NextBio's technologists partition the data space, implement their algorithms in the Hadoop MapReduce* framework for distributed computing, and use the HBase* database and Hadoop Distributed File System* (HDFS*) for fault-tolerant management of terascale and petascale files. Applications can dynamically partition HBase as data size grows. The company also uses sharded MySQL* databases and search indexes where appropriate.

"We're aggregating the world's genomic data, so we have a critical need to scale both storage and computation," says Alag. "We use the Hadoop stack to help us achieve that. For example, when we bring a new data set into the system, we compare it to all the known data sets and assign a score that represents how close the data sets are. We do all that computation in Hadoop. To take a clinical example, you might select two subsets of patients and search across various genes to find the biomarkers or other genomic information where the two cohorts differ—for example, which patients had side effects and which ones didn't. All this happens with Hadoop technology. When we bring in a full genome for a particular patient and want to generate a knowledge base of biomarkers to see what can be determined about that patient, we run a MapReduce job to generate that report."

Moving to Intel Distribution

As Intel was developing its Apache Hadoop distribution, NextBio was comparing multiple distributions. "From a per-node license cost perspective, some didn't scale well," Theige recalls. "Another distribution was proprietary in nature, making it difficult to troubleshoot. Another distribution was not mature in its support staffing. With Intel, we had the opportunity to work closely with a strategic collaborator to tune our code and enable their engineers to tune Intel® technology to provide performance improvements for genomic applications and to ease administration tasks. Intel's adherence and commitment to open source standards were also key for us."

NextBio has entered into the production phase with Intel Distribution, a comprehensive solution that includes the full distribution from the Apache Hadoop open source project, MapReduce, HDFS, and related components such as the HBase data warehouse infrastructure and the Hive* and Pig* data languages. Intel Distribution adds extensions to HBase and Hive that improve tasks such as real-time distributive query, automated machine-learning tuning controls, cross-data center table replication, and access control security.

Intel Distribution also provides a central management console, Intel® Manager, to secure, administer, monitor, and configure the cluster efficiently. Solution elements are pre-integrated to simplify management and deployment and enable faster time to market. NextBio is using Intel Manager for Apache Hadoop software to set up and configure its Hadoop clusters.

LESSONS LEARNED

Looking to deploy Apache Hadoop? Satnam Alag and Richard Theige offer these suggestions:

- **Match the tools to the business case.** The Apache Hadoop framework is well suited to large, unstructured data sets that are outgrowing traditional approaches, particularly if you have computations that also need to scale.
- **Take advantage of open standards.** Adhering to open standards helps you control costs and benefit from the rapid innovation of the open source community. Stay current with the open source distribution so you're taking advantage of ongoing improvements.
- **Pay attention to performance.** Speedups in Hadoop processing not only provide results faster, but also reduce the need for server, storage, and network capacity. Choose scalable hardware that provides high compute and I/O performance.
- **Get the tools and support you need.** A powerful tool set can help you deploy Apache Hadoop software quickly and manage the environment effectively. If you need 24/7 global support, make sure the vendor provides it.
- **Consider 10 Gigabit Ethernet networking.** You'll need a robust network for storage backups and communication across high-density server racks. Maximize throughput by isolating your Apache Hadoop software clusters from the rest of your network.

"The performance on Intel's current platforms is so good that the compression and decompression is almost invisible. It's a huge performance win by reducing disk and network I/O, and it means we will always use compression. It will save us on storage, power, and ultimately disk-drive repairs."

– Richard Theige, Senior Director of Operations, NextBio

NextBio is also collaborating with Intel engineers to enhance the Apache Hadoop stack for use in genomic data analysis. The two companies plan to contribute their enhancements to the open source community.

Performance that Cuts Costs Across the Infrastructure

Intel's software teams have optimized the open source Apache Hadoop stack to take full advantage of the Intel Xeon processor E5 family and instruction sets such as SSE4.2 for outstanding performance. Since Hadoop code is highly distributed, coding efficiencies are multiplied across the infrastructure, improving performance, energy consumption, and capacity requirements for servers and storage controllers. The optimizations in the Intel Distribution, along with the Intel Xeon processor E5 family's high CPU and I/O performance, also reduce network impacts.

NextBio saw an example of the Intel solution's optimized performance when it used HDFS and MapReduce on two clusters. One was based on the Intel Xeon processor E5-2680 at 2.7 GHz with 8 cores and 48 GB of memory; the other was based on quad-core, non-Intel processors with 8 GB of memory. The program placed static data in the distributed Hadoop caches, replicated the dynamic data across 90 to 100 percent of the available nodes, and ran CPU-intensive calculations on the data.

On the non-Intel processor-based cluster, the program spent 60 percent of its time performing I/O tasks and just 40 percent on actual calculations. On the Intel Xeon processor-based cluster, the program spent just 25 percent of its time on I/O, leaving 75 percent of its total time in computation. More importantly, the total time the Intel Xeon processors needed to perform the calculations dropped by about 50 percent—an indicator of the higher performance and greater memory capacity provided by the Intel processors, according to Theige.

"Intel has grown the whole pipeline," Theige says. "We got an 85 to 90 percent increase in total throughput on the Intel Xeon processor E5 family, which is an impressive result. The Intel processor-based cluster had 59 percent fewer nodes but 70 percent more cores, so it gives us a tremendous increase in density and horsepower."

The NextBio team also saw a sixfold data reduction using native Hadoop compression on its Intel Xeon processor-based cluster. "The performance on Intel's current platforms is so good that the compression and decompression is almost invisible," Theige says. "It's a huge performance win by reducing disk and network I/O, and it means we will always use compression. It will save us on storage, power, and ultimately disk-drive repairs."

Life-Changing Advances

Looking forward, the leaders of NextBio are excited to be on the forefront of a new era of personalized medicine. "Our hope is that we will be the Google of DNA," Theige says. "Look at how much information you find today by using Google. We're enabling some incredibly sophisticated search capabilities with molecular data, which takes things to a whole other level. DNA sequencing and personalized medicine are going to become pervasive just like CT scans, MRIs, and blood tests are. They will become commonplace, and they will be life-changing."

Find a solution that is right for your organization. Contact your Intel representative, visit **Business Success Stories for IT Managers**, or explore the **Intel IT Center**.

Watch a video interview with NextBio's Satnam Alag: **www.intel.com/content/www/us/en/big-data/big-data-genome-data-analysis-video.html**

To learn more about big data analytics, visit the IT Center at **intel.com/bigdata**

Download Intel® Distribution for Apache Hadoop* software at **hadoop.intel.com**

0213/LJ/TDA/XX/PDF 328067-002US