Solution Brief
Intel® Xeon® Processor E5 v2 Family
Big Data
Life Sciences/Biotechnology

Improved genome sequencing with big data solution

Whiteklay enhances performance of its BioDek* distribution system and speeds up genome next-generation sequencing (NGS) with Intel® Xeon® processor E5 v2 family and CDH*, a distribution of Apache Hadoop* from Cloudera

"The Intel® Xeon® processor E5-2680 v2 significantly reduced the processing time for large data sets, resulting in lower operating costs to a degree not previously possible. Combined with the power of CDH*, a distribution of Apache Hadoop* from Cloudera, BioDek* performance grew by 30 percent, bringing down the time to analyze NGS data by nearly 50 percent."

– Amit K Parija
General Manager
Consulting and Data Strategy
Whiteklay

Whiteklay began as a data management company but has evolved into an independent software vendor focused on the three major verticals of big data, high-performance computing, and cloud OpenStack*. With offices in Singapore and Pune, India, this Intel Solutions Integrator deploys Intel® technology and innovation to optimize its BioDek* distributed system, which can harness the power of CDH*, a distribution of Apache Hadoop* from Cloudera, to help scientists and bioinformaticians analyze the massive amounts of data generated from NGS. By using CDH to improve the quality and speed of sequence data analysis, Whiteklay's BioDek environment enables researchers to pursue the study of genomics at an unprecedented level.

Challenge

- **Reduce data conversion time.** Substantially reduce the time to convert FASTQ and FASTA genome formats, file formats which are commonly used for sequencing data, to sequence alignment map (SAM) files, and remove duplicate reads during post-processing to help scientists speed up preparation of NGS data for further analysis and reduce the processing time for large data sets.

Solution

- **Utilize Intel® Xeon® processor E5-2680 v2-based servers and CDH.** Upgrade BioDek from its initial testing platform using Apache Hadoop 1.x and non-Intel® architecture to CDH running on Intel Xeon processor E5-2680 v2 to speed up performance in storing NGS data and achieve maximum energy efficiency.

Technology Result

- **Improved performance and productivity.** Using the Intel Xeon processor E5-2680 v2 and CDH resulted in a 30 percent increase in BioDek's performance while cutting the distribution system's time to analyze massive data generated from NGS by almost half[1].

Business Value

- **More seamless and efficient NGS analysis.** By improving the speed and quality of data analysis, scientists and researchers can accomplish more tasks quickly and easily, exploring uncharted areas of genomics previously limited by computing throughput.

DNA sequencing has significantly impacted healthcare and biotechnology by greatly accelerating medical and biological research. It has even become indispensable in areas like forensic science. Modern DNA sequencing technology has made it possible to use DNA sequencing to advance its potential benefits in the fields of diagnostics and early detection of genetic predisposition to disease to improve agriculture and livestock breeding and processing. The possibilities seem endless, hampered only by limitations in computing throughput, speed, scalability, and resolution. This changed with the advent of NGS.

NGS has enabled scientists and researchers to maximize the potential of data sequencing. This is accomplished by searching, and comparing billions of DNA fragments. As a result of this processing, however, NGS produces massive amounts of data that pose challenges in terms of storage, analysis, management, and sharing of data. Without addressing these new NGS issues, researchers would face a new infrastructure bottleneck.

Whiteklay looked for a solution to address these issues and improve the quality and speed of NGS data handling. It developed the BioDek distributed system, which combines the post-processing of DNA reads after sequencing to prepare the NGS data for further analysis and the post-processing steps of sequence alignment. BioDek addressed the challenge in two steps: first, FASTQ and FASTA genome formats were converted to a sequence alignment map (SAM) file, which is a fundamental post-

Intel® Xeon® processor E5 v2 family and big data solution speed up genome next-generation sequencing

processing step in nearly all applications of deep sequencing technologies; and second, BioDek improved the quality and reduced the amount of stored data by removing duplicate reads.

The BioDek environment covered such tasks as annotating sequence data, browsing annotations mapped to a reference genome, and comparing and analyzing genomic sequences. Whiteklay initially tested BioDek using Apache Hadoop 1.x on a different server architecture, but subsequently deployed the CDH running on servers based on the Intel Xeon processor E5-2680 v2. This migration to CDH running on Intel architecture increased BioDek's performance by managing the server resources on an optimum note, speeding up the whole analysis process of converting FASTQ or FASTA genome format to SAM. This resulted in a 30-percent performance improvement that cut in the time to analyze NGS data by almost 50 percent[1]. The resulting increase in productivity lets scientists and researchers accomplish more sequencing in less time, cutting operating costs and increasing ROI by allowing for more sequences to be processed on the same infrastructure.

Addressing massive data issues in NGS

Currently, sequence data are in FASTQ files generated from Illumina* using the BWA* algorithms for post-processing and sequence alignment. Under the BioDek environment, the distributed approach would combine BWA with duplicate read detection and remove and convert the files to SAM.

BioDek allows large data to be split into smaller chunks for independent processing using MapReduce, which has become a significant process for its scalability in processing large sequencing datasets.

However, since MapReduce requires technical expertise in Apache Hadoop and Java*, it was still too technical for many bioinformaticians. To address this issue, BioDek uses the Apache Pig* engine to automatically parallelize and distribute data processing tasks.

By combining CDH and servers based on Intel architecture, BioDek significantly improved performance by using varying input and cluster sizes, with each node equipped with Intel Xeon processor E5-2680 v2-based servers at 3.5 GHz, with 16GB of RAM and two 250GB SATA* disks.

Working with one of the industry's best

While the Intel Xeon processor E5 v2 family and CDH give a performance boost to BioDek, Whiteklay wanted to further enhance the operational efficiency of its data centers with Intel® Node Manager.

Intel Node Manager enabled the resource monitoring of CPU, memory, and network usage, swapping, HDFS capacity, HBase compaction storm, disk capacity, disk I/O utilization, frequent Java virtual machine (JVM) garbage collection (GC), and MapReduce job failure statistics. This has helped Whiteklay system administrators optimize server utilization and distinguish high-level events from job failures by combining multiple metrics and then triggering alerts with specified thresholds.

The next step for Intel technology-supported BioDek

Using the high-speed, production-ready CDH, BioDek has enabled simple and scalable conversion, manipulation, and analysis of genomic data. With the performance enhancement enabled by CDH and Intel Xeon processor E5-2680 v2, Whiteklay is now looking into building a package that will easily load, extract, and analyze NGS data in a way that even bioinformaticians without any working knowledge of programming can easily use.

Committed to developing open-source software, Whiteklay is looking to integrate the BioDek solution with OpenStack to also make its solution available on cloud. Whiteklay plans to do this by working closely with Intel, which continues to provide significant contributions to the OpenStack community in its ongoing commitment to open source.

Find the solution that's right for your organization. Contact your Intel representative, visit Intel's **Business Success Stories for IT Managers,** and check out IT Center, Intel's resource for the **IT industry.**

- **With the BioDek* project, Whiteklay aims to inculcate the importance of having the necessary hardware and software in the correct specifications before building the Apache Hadoop* cluster and configuring the memory requirement correctly for MapReduce* JobTracker*, task tracker, name node, secondary name node, HBase* region server, HBase master, Hive*, and the clients.**
- **Whiteklay recommends exploring the CDH* within the BioDek environment to boost performance, especially since the human genome file fits the MapReduce architecture, which has the scalability, speed, and throughput to handle the large volume of data involved in genome sequencing in many laboratories around the world.**

1014/EDH/PMG/XX/PDF 331307-001