Hadoop* and Bioanalytics: How Intel, Cloudera, and BioDatomics are Teaming Up to Provide Comprehensive Solutions

BIODATOMICS
ACCELERATE YOUR INSIGHT

cloudera®

Disparate Evolutions

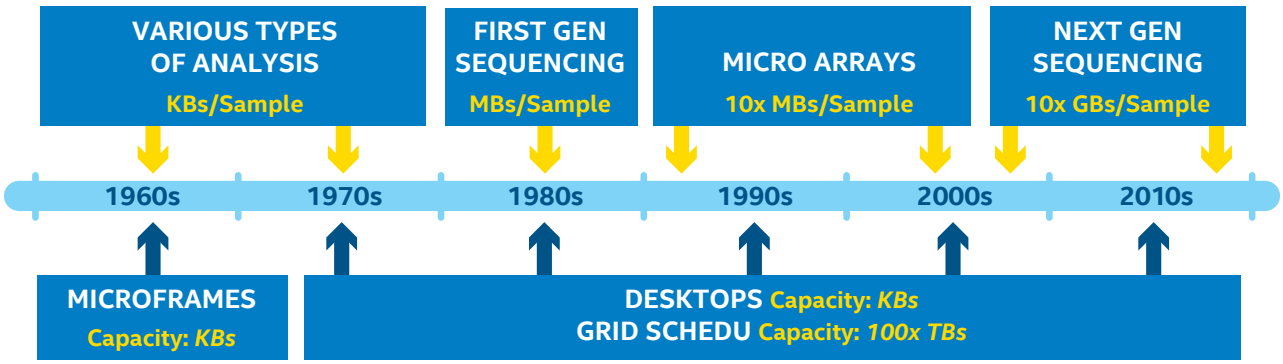
Over the past 50 years, the field of bioanalytics has evolved through several stages—from a diverse variety of analyses in the '60s and '70s, through first generation sequencing in the '80s, micro arrays in the '90s, and finally, from the mid-2000s forward, to Next Generation Sequencing (NGS). Each step has brought improved accuracy, and enhanced possibilities for medical research, drug discovery, and personalized medicine. However, in concert with this evolution there has been a corresponding increase in the size of a single genomic sample. Samples originally measured in kilobytes have given way to megabyte-sized samples in first generation sequencing, and to micro array samples that are tens of megabytes. With NGS, the size of one sample has taken a quantum jump in size to tens of gigabytes.

One might expect the infrastructure supporting analysis of these samples to have evolved commensurately. Oddly, that has not been the case. In the very same timeframe as bioanalytics has taken multiple leaps forward, the supporting infrastructure has barely evolved at all. Instead, since the '70s, it has remained rooted in either desktop systems with a total storage capacity of only a few terabytes, or grid scheduler-based clusters with a maximum capacity of several hundred terabytes. Unfortunately, with single sample sizes now measured in the tens of gigabytes, a suitable infrastructure must be able to handle hundreds of petabytes rather than terabytes.

Grid Scheduler Principles

Grid scheduler clusters consist of disk-less compute servers supported by "farms" of large, external storage servers. Because a piece of data is stored in just one place within the cluster, storage servers are invariably built to enterprise-grade reliability standards.

The Evolution of Bioanalytics



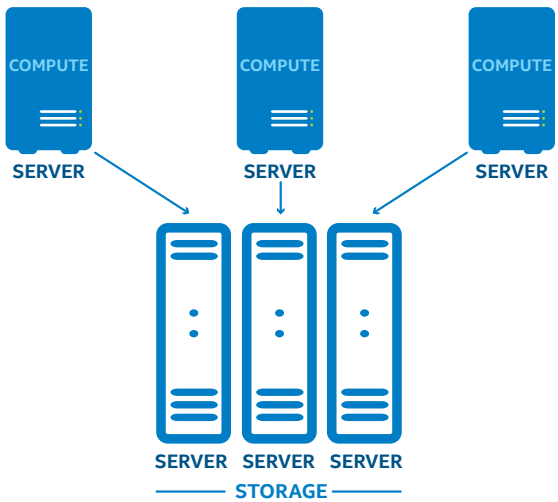
When a task, such as a genomic analysis, is initiated, the grid scheduler moves all relevant data from storage to the centralized compute servers. Therein lies the primary reason why grid-based clusters cannot scale to the levels required by NGS: for large datasets, this all-at-once data transfer is massive and creates a serious bottleneck. The more data, the greater the bottleneck.

Grid schedulers are subject to additional limitations that compromise their NGS analysis utility. In NGS, even simple tasks needed for sample preparation (cleaning, aligning, mapping)—let alone executing complex analytics pipelines—are highly time consuming. For datasets of even moderate size, the only practical way to deal with this situation is to employ parallelization. Unfortunately, grid schedulers offer limited opportunities for such parallel processing. Specifically, while grid does permit separate samples to be processed in parallel, it does not offer a means of parallel processing data subsets within an individual sample. For NGS, with its large sample sizes, this leads to exceedingly long execution times.

To accelerate analysis turnaround times, some bioanalytics companies have introduced costly, specialized hardware that can run within a grid. Unfortunately, this hardware is compatible only with custom-developed software. Consequently, utilizing such hardware necessarily restricts scientists to a relatively small, proprietary set of analysis tools. Ideally, a solution to the problem of long bioanalytical turnaround times would employ less expensive hardware and be compatible with the plethora of publicly available, open source bioanalysis resources.

One way to realize that ideal is through the use of Big Data analysis engines. This is a logical approach, since NGS is quite obviously a Big Data application. This class of software is well prepared to handle the massive datasets associated with NGS, to scale as needed to execute complex pipelines, and to manage the necessary levels of parallelization. The most popular, advanced, and comprehensive of these is Hadoop.

Grid Scheduler Cluster Architecture



Hadoop* Principles

Hadoop is a complete, open source infrastructure management ecosystem with roots in the original Google* search engine. In the early 2000s, Google published the spec for the first generation of Hadoop, at which point the open source development community adopted and began evolving the technology. Initially, as with other open source software such as the Linux* operating system, Hadoop was available only as open source, unsupported code. However, as with Linux, Hadoop soon became available as fully supported, enterprise-class software.

Hadoop's architecture differs radically from that of grid schedulers. In Hadoop, compute and storage functions are not separate; rather, they are always co-located. This, in turn, enables data itself to be co-located with processing power. Hadoop's architecture is driven by its fundamental concept, which is to bring processing power to the data—rather than the other way around. When incoming data arrives, it is first segmented, then triplicated, then distributed among nodes. When a compute task is scheduled, the application is moved to every node on which relevant data segments reside, then run on all data in parallel.

This approach has multiple benefits. First, it eliminates data transfer bottlenecks entirely, since data never has to be moved en masse to a centralized point. Further, Hadoop permits parallel processing of data within individual samples. Together, these factors lead to dramatic decreases in execution times.

Hadoop offers additional benefits in terms of scalability. The lack of data transfer bottlenecks plus unbounded parallelization means that the size of a Hadoop cluster is theoretically unlimited. This means that compute capacity is never an issue, nor is the ability to store hundreds of petabytes of data. Indeed, the theoretical scalability has been tested by the likes of Facebook and Yahoo, both of which have Hadoop clusters with well over 10,000 nodes (Yahoo's is 42,000 nodes).

Finally, Hadoop delivers significant cost advantages over traditional HPC. These are derived from several sources. With respect to software licensing, Hadoop—even in enterprise grade, fully supported versions—enjoys low costs since the core software is open source. At the same time, Hadoop-optimized hardware nodes are also less expensive than their grid-oriented counterparts, since clusters are based on the ubiquitous x86 architecture. Healthy competition between suppliers has continuously improved the quality and driven down the cost of this hardware architecture. Innovative companies like Intel have designed Intel® Xeon® processors that are further optimized for Hadoop workloads. This type of platform allows for horizontal scaling when the need arises for additional storage or compute power. Lastly, Hadoop confers major cost savings in the area of storage. Data stored on a Hadoop cluster is triple-replicated by default. This technique delivers both fault tolerance and high throughput. From a cost perspective, this approach also eliminates the need for expensive enterprise storage services, whose license structures were never intended to handle NGS-sized data sets.

Cloudera and BioDatomics: Marrying Hadoop and Bioanalysis

Despite the clear benefits of Hadoop over grid schedulers—benefits which have proven themselves in numerous vertical industries—bringing the technology to the field of bioanalysis was, until recently, elusive. This is because bioanalytics is a highly specialized field. Bioinformaticians work in building blocks called analysis "tools," which are linked together to form "pipelines" or "workflows" that define analytical tasks. Hadoop has no knowledge of such tools or workflows. Further complicating adoption, all existing tools and workflows were built for desktop or grid-based infrastructures rather than Hadoop.

To solve the problem, BioDatomics worked closely with Cloudera over a two year period to develop BioDT.* BioDT is a bioanalytics platform designed to improve the productivity of biologists and bioinformaticians alike. It does this via a combination of a powerful yet intuitive user interface, and an execution engine that works with Hadoop to parallelize pipelines, NGS data, and popular analysis tools. Importantly, the tools used within BioDT (several hundred are integrated into the platform) need not be optimized for Hadoop in any way; BioDT will emulate a grid scheduler for those tools expecting one.

BioDT offers several additional productivity-enhancement features, including a drag-and-drop interface that obviates the need for command line programming when building pipelines; interactive tabular results that enable results to be filtered, grouped, and searched in real time; real-time collaboration so multiple scientists can confer together on a workflow or results; graphics that enable results to be visualized in a variety of ways, making insights more evident; hundreds of pre-integrated tools, including popular Galaxy and GATK tools, plus an API to easily import new tools; support for workflows that incorporate loops, conditionals and nesting; and version control, which facilitates reproducible research.

BioDT is not the only Hadoop-related innovation in biosciences. Work at institutions such as UC Berkeley, Georgia Tech, and the University of Chicago is creating a growing set of analytics tools optimized to take full advantage of Hadoop. As well, industry standards are moving toward Hadoop; for instance, the Global Alliance for Genomics and Health has specified the Hadoop-optimized Apache Avro as its NGS data format.

The Intel Connection

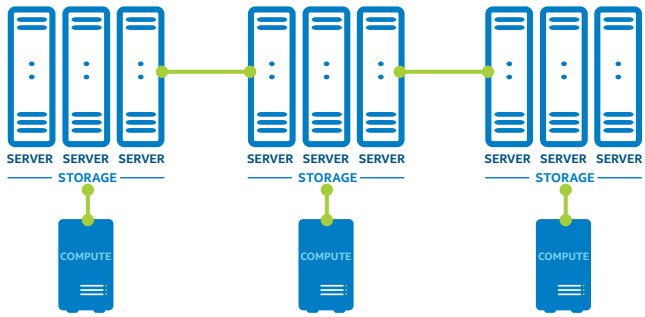
Intel is supporting Hadoop-driven bioanalytics in multiple ways. At a corporate level, Intel and Cloudera have a deep strategic relationship that includes a large investment by Intel in Cloudera, as well as representation on Cloudera's Board of Directors. At a technology level, the two companies are working together to optimize Hadoop software for maximum performance from underlying infrastructure. That infrastructure includes compute, memory (RAM), and storage elements. Intel is also committed to ensuring that Hadoop meets the demands of organizations requiring enterprise-grade Big Data solutions. For example, Intel is working to improve the security strength and performance of encryption through hardware acceleration built into the line of Intel Xeon processors. In addition, Intel is investigating ways to provide more RAM on each node in order to enable Cloudera Impala* and Apache Spark* to execute more in-memory operations.

Performance

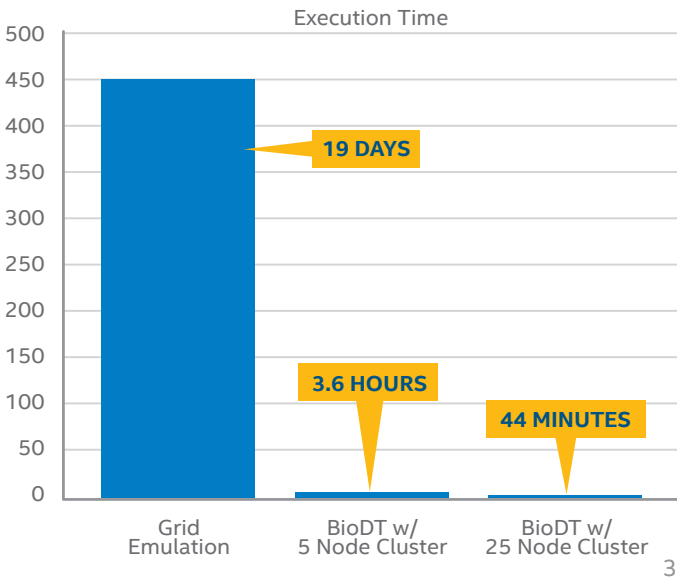
As noted below, because Hadoop segments and distributes data across a theoretically infinite number of compute nodes, the framework has the potential for virtually unlimited scalability and a speed increase of multiple orders of magnitude compared to legacy approaches. These developments make it possible to dramatically accelerate the analytical processing of NGS data. The execution speed of NGS data analysis directly impacts the speed at which research findings become available. Currently, analysis of a large dataset via a complex workflow can take days—in some cases weeks. This represents a significant impediment to research productivity. The goal of BioDatomics, Cloudera, and Intel is to reduce these protracted analysis times by optimizing bioinformatics algorithms for Hadoop and executing them on Hadoop-compatible software and a Hadoop-optimized cluster.

To illustrate the degree of acceleration this approach achieves, BioDatomics compared the time to perform an analysis using

Hadoop* Cluster Architecture



BioDT Performance



Hadoop and non-Hadoop techniques. The genomic file was not large by life sciences standards, just 50 GB. However, the analysis was typically complex. In this case, the DNA samples within the file were first stripped of human DNA. Then, using a popular analysis algorithm called BLAST, the resulting non-human DNA was compared against an extensive database of viral, bacterial, and other organisms. The purpose was to find matches, if any.

To emulate a grid environment, BioDatomics ran the analysis using a non-Hadoop optimized, non-parallelized version of BLAST[1]. As the figure above illustrates, the non-Hadoop analysis took 19 days. Although this sounds like a long time, it is not at all atypical for genetic analysis. In stark contrast, using a Hadoop-optimized version of BLAST running on a modest 5-node Intel®-based cluster resulted in an analysis time of just 3.6 hours—a 127x improvement. Further, thanks to the nearly linear scalability of Hadoop clusters, the same analysis on a 25-node cluster completed in just 44 minutes.

Synergistic Offerings

One of the ways in which the three companies are working together to deliver powerful yet accessible solutions is through BioDT SaaS, a cloud-based service that enables small organizations and individual researchers or clinicians to take advantage of the combined power of Intel, Cloudera, and BioDT—even if they have no Hadoop infrastructure whatsoever. As with most cloud-based services, users pay only for the resources they use. In this service, processing power is delivered by Intel, Cloudera provides the Hadoop management ecosystem, and BioDT supplies the bioanalytics application. Further, the vast majority of the open source analysis tools integrated into the BioDT platform are optimized for Intel® processors.

The three companies have also teamed to create a BioDT Dedicated SaaS service. Though also a cloud-based offering, this service combines Cloudera/BioDT software with an Intel-powered cluster dedicated to one customer. Dedicated SaaS is optimal for larger organizations seeking a powerful, secure analytics platform that nonetheless requires no on-site infrastructure. Furthermore, organizations avoid the up-front cost of acquiring such a system; rather, the system is billed on a monthly basis for as long or as short as it is needed.

Summary

In summary, genomic bioanalytics has been evolving more rapidly than the supporting infrastructure. However, with Hadoop, infrastructure has caught up to modern bioanalytics requirements. Hadoop offers significant speed and scalability over traditional grid scheduler clusters. Consequently, as can be seen from a ramp-up in commercial pilot tests, standards activity, and work taking place at academic institutions, the industry is moving inexorably to Hadoop.

But Hadoop alone does not solve a single bioanalysis research problem or provide usable patient insight. BioDT from BioDatomics harnesses the power of Cloudera Hadoop to bring Hadoop's Big Data benefits to the world of bioinformatics. Intel complements these elements with processors that have been optimized to extract maximum performance and security from Hadoop installations.

1. Number of cores/nodes: 6 physical (12 virtual) Type of Intel processor: Intel® Xeon® processor E5-1650 v2 RAM/node: 128 GB.

 0315/DW/HBD/PDF 332172-001US