# The Computerworld Honors Program

Honoring those who use Information Technology to benefit society

## Final Copy of Case Study

**Status:**
Laureate

**Year:**
2013

**Organization Name:**
J. Craig Venter Institute (JCVI)

**Organization URL:**
http://www.jcvi.org/

**Project Name:**
Analyzing an Ocean of Data

**Please select the category in which you are submitting your entry:**
Emerging Technology

**Please provide an overview of the nominated project. Describe the problem it was intended to solve, the technology or approach used, how it was innovative and any technical or other challenges that had to be overcome for successful implementation and adoption. (In 300 words or less.)**

JCVI is a world leader in genomic research, best known for the work of Dr. J. Craig Venter and his team in decoding the first draft of the human genome. Today's JCVI's groundbreaking Global Ocean Sampling Expedition (GOS) collects and analyzes microorganisms found in seawater with the aim of helping scientists better understand the evolution of the oceans, microbial biodiversity, climate and environmental changes, and more. Through this important research, more than 60 million genes and thousands of novel protein families have already been discovered. The GOS project involves a massive volume of genomic data that expands in size and scope as more research and analysis is done. While initial analytic data loads may start out relatively small (in the 20-100 gigabyte range), secondary processing and analysis can cause the data to blow up to terabyte levels and beyond. Advances in genetic sequencing technology have also enabled researchers to capture much more genomic data overall. When the Expedition launched in 2003, a data set that included 50,000 sequences was considered a large amount. Today, a typical round of analysis may include 40-50 million sequences. Due to

the growing amount of data, JCVI faced problems efficiently and economically storing, loading and analyzing it all using its existing MySQL database. Analytic query speed began to suffer as the project progressed, and scientists found themselves waiting hours, and sometimes days, for results to come in. The GOS Expedition team addressed this issue though the deployment of a columnar database technology from Infobright specifically designed for high-volume analytics. The solution delivers the right combination of scalability to accommodate data size, speed and affordability, and with it, JCVI can now continue to analyze more data to make new discoveries about the oceans and our world.

**When was this project implemented or last updated? (Please specify month and year.) Has it incorporated new technologies and/or other innovations since its initial deployment? (In 300 words or less.)**

This project was implemented two years ago and last updated in August 2012 with the newer version of Infobright 4.0.6.

**If this is a previously submitted project that has been significantly updated and/or expanded, please describe the nature of the update here. (In 300 words or less.)**

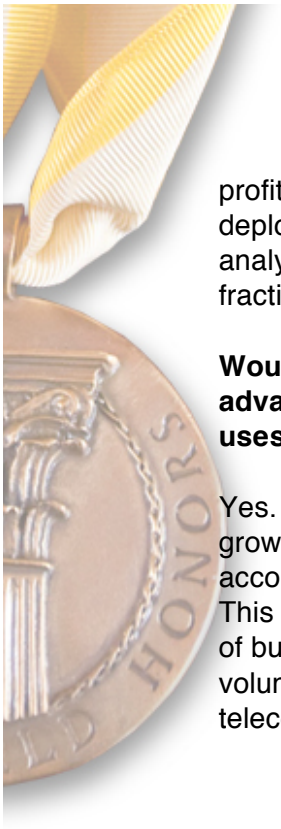
N/A

**Is implementation of the project complete? If no, please describe the project's phases and which phase the project is now in. (In 300 words or less.)**

Yes.

**Please provide at least one example of how the technology project has benefited a specific individual or organization. Feel free to include personal quotes from individuals who have directly benefited from the work. (In 300 words or less.)**

This technology project benefits the JCVI scientists working on the GOS Expedition, the database administrator charged with helping the project team manage their research data, as well as the organization as a whole. With the new database solution in place, queries on genomic data that used to take many minutes (and sometimes hours) to resolve now come back in seconds, resulting in a nearly 10-fold improvement in analytic performance. JCVI has also been able to achieve data compression ratios of 10:1 and in many cases up to 14:1, allowing the organization to speed analysis even more while also cutting down on storage costs. (As an example, one database that was 433 gigabytes in MySQL now takes up only 25 gigabytes of storage in Infobright.) From a maintenance perspective, the deployment has been a lifesaver. JCVI's database administrator used to spend dozens of hours per week tinkering with MySQL queries and creating indexes and partitions to ensure that analysis could be performed. Now, data just needs to be loaded into Infobright, with no customization required, saving IT time and resources. JCVI scientists working on analyzing data from the GOS Expedition can now ask more questions and change analytic parameters on the fly, because they are no longer limited by the bottlenecks that used to be created when queries became too complicated or required too much setup time to perform. More questions ultimately lead to more knowledge, the driving force behind any research endeavor. Finally, as a not-for-




profit research institution, JCVI needs to keep project costs in check. The technology deployed reduces the amount of expensive hardware, storage and resources required to analyze large volumes of data, and the solution itself is extremely affordable, costing a fraction what a more traditional data warehouse would run.

**Would this project be considered an innovation, a best practice or other notable advancement that could be adopted by or tailored for other organizations and uses? If yes, please describe that here. (In 300 words or less.)**

Yes. All types of organizations today are being challenged by unprecedented data growth and, as a result, need to begin rethinking their information strategies to accommodate the storage, processing and analytic challenges brought on by Big Data. This is especially true in the R&D community, but also applies to a range of other types of businesses that need to be able to quickly, simply and affordably manage high volumes of data streaming in from multiple sources. These include power companies, telecommunications service providers, retailers, online service providers and more.