White paper

FUJITSU Integrated System PRIMEFLEX® for Hadoop: Genome Analysis

FUJITSU Integrated System PRIMEFLEX for Hadoop offers end-to-end integration from strategic consulting to use case development and implementation services for specific customer needs. Read how PRIMEFLEX for Hadoop is making an impact on The German Cancer Research Centre for more efficient genome analysis.

Contents

Introduction	2
Efficient Genome Analysis	2
Case study: German Cancer Research Centre	3
Conclusion	4

Introduction

Improving healthcare outcomes is the chief priority of any medical organization and there are fewer more pressing obstacles to this goal than cancer. Cancer, in all its forms, kills more than eight million people globally every year and, although it can sometimes be managed, there is not yet a known cure. This explains why huge amounts of money and manpower are expended on researching the causes and potential treatments of the disease.

Efficient Genome Analysis

Cancer research now largely focuses on genomic data, which involves DNA sequencing, assembly and analysis, in order to identify key markers and variants within our genetic make-up. The human genome, with three billion base pairs, is a classic example of Big Data, as multiple patient samples in a field study with a 1,000-strong cohort can generate petabytes of data. The challenge is to effectively analyze this information quickly and efficiently.

Traditionally, these data sets are explored using specialized computing methods in which notable information is tracked, highlighted and thus the overall volume of data is reduced step by step. However, this approach can take weeks or even months to complete before the next stage in the analysis can proceed. One key problem is that if a certain variant does not initially appear, researchers cannot be sure whether it is not present or has simply been overlooked. This necessitates multiple reductive steps in processing the entire sequence.

"Typically, researchers have to reduce the huge input data by focusing on obvious deviations against a reference genome," explains Dr Fritz Schinkel, Head of Big Data Competence Center and Distinguished Engineer, Fujitsu Technology Solutions. "This is necessary to avoid repeatedly processing the whole raw data but it runs the risk of misinterpretations and restricts the questions one can ask."

PRIMEFLEX for Hadoop

Rather than use a classic High Performance Computing (HPC) set-up, many organizations are now looking at ways in which this data can be stored and processed in parallel for faster compute times. They are also looking at how best to unlock the information so that it can be manipulated by non-programmers in a familiar format in order to take a more human-centric approach to innovation.

"The Big Data approach with plain Hadoop demands a high degree of technical expertise to actually handle the data, meaning it has to be painstakingly prepared so regular researchers can use it," adds Schinkel. "We wanted to create a fresh platform that could not only process the data more efficiently but also present it in a way that would be usable to everyone."

The result is FUJITSU Integrated System PRIMEFLEX for Hadoop, which dramatically reduces the time it takes to gather, process and understand genetic information. Hadoop is the de-facto standard for Big Data and distributed parallel processing, an open source framework primarily for batch operation, written in Java. It is designed to scale up to thousands of nodes and to make data storage and analytics robust against failures.

"Distributed parallel processing provides several advantages. Executing a query or any other data operation by many nodes at the same time increases performance and delivers fast results," adds Schinkel. "You can start small, with just a few servers and then add more servers as they are needed. Basically, your infrastructure will linearly scale up without any limits."

Hadoop is integrated with Fujitsu hardware and Big Data analytics software provided by Datameer® on the front end to enable non-specialist users to easily manipulate huge volumes of data from multiple sources. It simplifies the analysis while removing the need for any programming, coding or scripting thereby opening up the potential of Big Data processing to all areas of research. Datameer is the only end-to-end Big Data analytics application which is purpose-built for Hadoop. It enables the fastest time from raw data to new insights. This reflects Fujitsu's approach to human-centric innovation, which focuses on applying advanced technology to deliver new insights and value from information.

"PRIMEFLEX for Hadoop is a powerful and scalable platform that provides users with a more cost-effective way of creating actionable analytics from Big Data," says Schinkel. "It analyzes large volumes of data to extract and make accessible meaningful research-relevant information, combining the convenience of pre-configured and pre-tested hardware and the economic advantages of open source software plus system support and all-round lifecycle management."

Faster, more accurate results

Essentially, PRIMEFLEX shortens the time it takes to crunch data and provides a more direct route to smarter research. Because the data is processed in place analysis can be completed faster and can use full input data for more accurate results than the classical HPC based approach. This in turn reduces the time it takes to complete projects and speeds up the analysis of genomes.

PRIMEFLEX for Hadoop comes with preinstalled software including RedHat Enterprise OS, Datameer, Cloudera Manager and Cloudera Distribution for Hadoop. The Entry variant is completely installed and configured in the factory and needs only to be connected to the customer network.

"It's an off-the-shelf solution that comes fully pre-configured so that it makes sense even to non-technical staff," continues Schinkel. "Bringing Big Data to biologists or physicians has the potential to revolutionize exciting new fields in genetic research and diagnosis, which could mean more effective treatments become available more quickly."

Fujitsu also offers end-to-end integration and consulting services for PRIMEFLEX for Hadoop, from strategic consulting to use case development and implementation services for specific customer needs. This makes the solution ideal for any organization that needs to understand large volumes of complex information, such as medical research organizations.

Case study:
The German Cancer Research Centre deploys FUJITSU Integrated System PRIMEFLEX for Hadoop for more efficient genome analysis

More than 450,000 people are diagnosed with cancer each year in Germany. The German Cancer Research Centre (DKFZ) is the largest biomedical research institute in Germany and is devoted to understanding the disease. In over 90 divisions and research groups, more than 1,200 scientists are investigating the mechanisms of cancer, identifying risk factors and trying to find strategies to prevent people from getting cancer.

The challenge is the amount of data involved in analyzing cells at a genetic level in order to identify common triggers or indicators relevant to cancer. Even using a HPC cluster, the DKFZ was still experiencing bottlenecks relating to genome data. This holds up progress and creates frustration among users.

"We are part of the International Genome Consortium, one of seven global sites dedicated to cancer research on HPC platforms," explains Dr Matthias Schlesner, Group Leader, Computational Oncology, DKFZ. "We had established a standardized HPC cluster based on Fujitsu, Intel and SAP technology however we had problems interfacing with the file server and experienced peaks and troughs in processing capability."

A smarter approach to data analysis

The DKFZ discussed proposals to improve performance with its technology partners and upgraded the underlying hardware. It also deployed FUJITSU Integrated System PRIMEFLEX for Hadoop in order to support this endeavour.

"Once we had optimized the platform for the local architecture, we migrated to Hadoop to evaluate whether it could overcome the bottlenecks. Typically, we look for variant positions within patient cohorts but if no variant presents we can't know whether it isn't there or we simply can't see it," says Schlesner. "That meant we usually have to perform additional processing steps looking at complete input data. With Hadoop, we can run parallel on large datasets, enabling us to skip the reduction step and work without performance degradation."

This Fujitsu PRIMERGY CX400 server solution enables a very different, data-orientated approach to analysis, using Datameer end-to-end analytics software to integrate structured and unstructured data. This makes it possible for non-programmers and business users to handle and manipulate data more easily and intuitively on a spreadsheet presenting the Big Data input and manipulation. Now, a team of researchers uses Fujitsu PRIMEFLEX for Hadoop to analyze bigger cohorts of genomes in order to better understand how cancer operates and how it can be stopped.

Unlocking the genome

By simplifying data analysis, the new system has speeded up the process and enabled the DKFZ to more easily identify genetic variants in each patient, looking at all possible positions on the genome without the need to transport data. The ultimate hope is that this human-centric approach to innovation will lead to an improved understanding of the disease and better patient outcomes.

"Fujitsu PRIMEFLEX for Hadoop makes everything more convenient and gives us the ability to perform quick and unbiased analysis of up to 900,000 known, clinically important genetic positions," continues Schlesner. "This is processed much faster on Hadoop than on HPC without facing any bottlenecks with network. The parallelization approach can even speed up the analysis of a single genome by a factor of four or more."

The DKFZ has also been impressed by the reliability of the Fujitsu hardware which has not experienced any downtime since the pilot began. This gives the organization peace of mind and allows it to focus on the vital business of cancer research.

A path to future analytics

The DKFZ is continuing to undertake performance tests to fine-tune how Fujitsu for Hadoop operates, with the results due to be published at a forthcoming summit. It is hoped that other medical research bodies will also see the benefits that the platform offers for Big Data analysis.

"Our experience with the new platform has been good and combining our genomic expertize with Fujitsu's technical skill has worked out well," concludes Schlesner. "We are hoping that carrying out routine analysis will enable us to develop new research techniques."

Conclusion

FUJITSU Integrated System PRIMEFLEX for Hadoop is a powerful and scalable platform that can analyze large volumes of genetic data at high speeds. It combines pre-configured and pre-tested industry-standard hardware with open source software provided by Cloudera and Big Data Analytics from Datameer, making it the most cost-effective, accurate and fast way for research companies to best explore their data for the most effective patient outcomes.

In collaboration with

Contact
FUJITSU
E-Mail: cic@ts.fujitsu.com
Website: www.fujitsu.com
2016-04-12