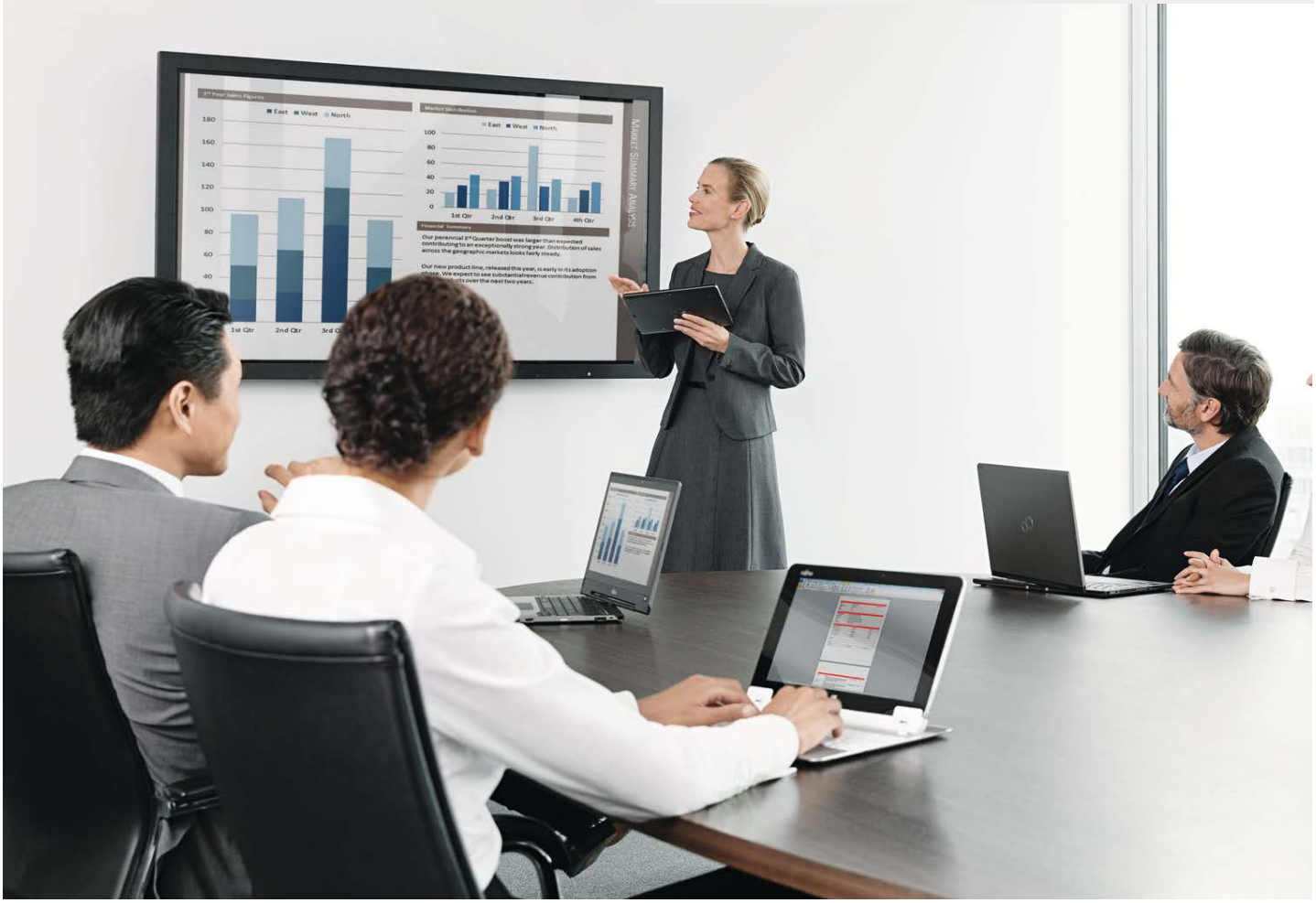
White paper
FUJITSU Integrated System PRIMEFLEX for Hadoop: Genom-Analyse

FUJITSU Integrated System PRIMEFLEX for Hadoop ist eine ganzheitliche und durchgängige Lösung – von der strategischen Beratung über die Entwicklung des Use Case bis hin zu Implementierungsservices – ausgerichtet an spezifischen Kundenanforderungen. In diesem Whitepaper lesen Sie, wie PRIMEFLEX for Hadoop die Genom-Analyse beim Deutschen Krebsforschungszentrum effizienter macht.

Inhalt



Einleitung

Die Erzielung besserer Behandlungsergebnisse hat höchste Priorität in medizinischen Einrichtungen und Organisationen. Das gilt insbesondere für die Krebsforschung. Die verschiedenen Arten der Krankheit töten weltweit jährlich mehr als acht Millionen Menschen. Auch wenn es manchmal gelingt, die Krankheit einzudämmen, gibt es bis heute keine Heilung. Deshalb werden sehr viele Ressourcen in die Erforschung der Ursachen und in Behandlungsmöglichkeiten investiert.

Effiziente Genom-Analyse

Die Krebsforschung konzentriert sich heute größtenteils auf Genom-Daten und umfasst DNA-Sequenzierung, -Assemblierung sowie –Analyse. Auf diese Weise sollen Schlüssel-Marker und Varianten in unserer Erbanlage identifiziert werden. Das menschliche Genom mit mehr als drei Milliarden Basen-Paaren ist ein klassisches Beispiel von und für Big Data: Mehrere Patientenproben einer Feldstudie mit 1.000 Teilnehmern erzeugen bereits Petabytes an Daten. Die Herausforderung ist, diese Informationen schnell und effizient zu analysieren.

Traditionell werden solche Daten mit speziellen Rechenverfahren ausgewertet, die auffällige Informationen kennzeichnen, nachverfolgen und so Schritt für Schritt das Datenvolumen reduzieren. Bei diesem Vorgehen kann es Wochen oder gar Monate dauern, bis die nächste Stufe der Analyse erfolgen kann. Ein Grundproblem dabei: taucht eine bestimmte Variante anfänglich nicht auf, können die Forscher nicht sicher sein, ob sie überhaupt vorhanden ist, oder aber ob sie einfach „übersehen" wurde. Dies erfordert eine Vielzahl von reduktiven Schritten bei der Verarbeitung der gesamten Sequenz.

„Typischerweise reduzieren Wissenschaftler die Komplexität einer derart großen Datenmenge, indem sie sich auf auffällige Abweichungen von einem Referenz-Genom konzentrieren", erklärt Dr. Fritz Schinkel, Head of Big Data Competence Center und Distinguished Engineer bei Fujitsu Technology Solutions." Das ist notwendig, um eine wiederholte Verarbeitung der gesamten Roh-Daten zu vermeiden. Aber es birgt das Risiko von Fehlinterpretation und schränkt die möglichen Fragestellungen ein."

PRIMEFLEX for Hadoop

Anstelle eines klassischen Setups mit High Performance Computing (HPC) suchen viele Einrichtungen heute nach Möglichkeiten, mit denen sie Daten parallel speichern und verarbeiten können, um so die Verarbeitungszeit zu reduzieren. Und sie benötigen Möglichkeiten, Informationen so einfach zu erfassen, dass auch Nicht-Programmierer damit in einem gängigen Format arbeiten können. Das stellt den Menschen in den Mittelpunkt der Innovation.

„Der Big-Data-Ansatz mit Hadoop erfordert ein hohes Niveau an technischer Expertise, um die Daten tatsächlich bearbeiten zu können. Das heißt, Daten müssen akribisch aufbereitet werden, damit normale Forscher sie nutzen können", sagt Dr. Fritz Schinkel. „Wir wollten eine neue Plattform schaffen, mit der Daten nicht nur effizienter verarbeitet werden können, sondern die auch von jedermann bedient werden kann."

Das Ergebnis ist FUJITSU Integrated System PRIMEFLEX for Hadoop: Damit lässt sich die Zeit für das Sammeln, Verarbeiten und Verstehen von genetischen Informationen signifikant reduzieren. Hadoop ist der De-facto-Standard für Big Data und die verteilte Parallelverarbeitung von Daten. Es ist ein Open-Source-Framework in Java, das sich auf mehr als tausend Knoten skalieren lässt und somit Datenspeicherung und –analyse vor Ausfällen und Fehlern schützt.

„Die verteilte, parallele Datenverarbeitung schafft viele Vorteile. Denn eine Datenabfrage oder andere Operation zur gleichen Zeit auf mehreren Knoten auszuführen, steigert die Performance erheblich und liefert schnelle Ergebnisse", erklärt Schinkel weiter. „Man kann klein starten, mit nur wenigen Servern, und – wenn nötig – neue hinzufügen. Im Prinzip kann die Infrastruktur unbegrenzt erweitert werden."

Hadoop ist in die Fujitsu Hardware und die Big-Data-Analyse-Software von Datameer® im Front-End bereits integriert. Dadurch können auch Nicht-Spezialisten sehr einfach große Datenmengen aus mehreren Quellen bearbeiten. Die Datenanalyse wird so erheblich vereinfacht. Gleichzeitig entfallen zusätzliches Programmieren oder Kodieren. Das öffnet das Potenzial von Big Data für sämtliche Forschungsrichtungen. Datameer® ist dabei die einzige End-to-End Big-Data-Analyse-Anwendung, die speziell für Hadoop entwickelt wurde. Sie generiert in kürzester Zeit aus Roh-Daten neue Erkenntnisse. Darin zeigt sich der Ansatz der „Human Centric Innovation" von Fujitsu: der Einsatz fortschrittlicher Technologien zur Gewinnung neuer Erkenntnisse und Werte für den Menschen aus Informationen.

"PRIMEFLEX for Hadoop ist eine leistungsstarke und skalierbare Plattform. Anwender können damit bei geringen Kosten umfangreiche Big-Data-Analysen durchführen", so Schinkel. „Es lassen sich große Datenmengen analysieren und daraus aussagekräftige sowie für die Forschung relevante Informationen extrahieren und zugänglich machen. Dabei kombiniert die Lösung den Komfort einer vorkonfigurierten und getesteten Hardware mit den wirtschaftlichen Vorteilen einer Open-Source-Software inklusive System-Support und Rund-um-Lifecycle-Management."

Schneller zu genaueren Ergebnissen

Im Wesentlichen reduziert PRIMEFLEX for Hadoop die Zeit, um Daten zu verarbeiten, und eröffnet so einen direkten Zugang zu intelligenter Forschung. Da die Daten dort verarbeitet werden, wo sie auch gespeichert sind, können Analysen im Gegensatz zu einem klassischen HPC-Ansatz schneller abgeschlossen und dabei sämtliche Roh-Daten für die Gewinnung genauerer Ergebnisse genutzt werden. Das reduziert wiederum Projektlaufzeiten und beschleunigt die Analyse von Erbgut.

PRIMEFLEX for Hadoop wird mit vorinstallierter Software, unter anderem RedHat Enterprise OS, Datameer®, Cloudera Manager und Cloudera Distribution for Hadoop ausgeliefert. Die Einstiegsvariante wird bereits komplett installiert und konfiguriert. Sie muss einfach nur noch vom Anwender mit dem Netzwerk verbunden werden.

„Es ist eine vollständig vorkonfigurierte Standardlösung, die auch von Mitarbeitern ohne tiefgreifende IT-Kenntnisse leicht verstanden wird", erklärt Dr. Fritz Schinkel weiter. „Sie eröffnet Biologen und Medizinern einen einfachen Zugang zu Big Data. Darin steckt das Potenzial, viele Bereiche der genetischen Forschung und Diagnose zu revolutionieren und damit bessere Behandlungsmethoden schneller verfügbar zu machen."

Zudem bietet Fujitsu End-to-End-Services für die Integration und Beratung zu PRIMEFLEX for Hadoop – von der strategischen Beratung über die Entwicklung eines Use Case bis hin zur Implementierung – je nach spezifischer Kundenanforderung. Die Lösung ist ideal für Organisationen, die große Mengen an komplexen Informationen verstehen müssen, wie zum Beispiel medizinische Forschungseinrichtungen.

Case Study: Deutsches Krebsforschungszentrum nutzt FUJITSU Integrated System PRIMEFLEX for Hadoop für die effiziente Genom-Analyse

In Deutschland wird bei jährlich mehr als 450.000 Menschen eine Krebserkrankung diagnostiziert. Das Deutsche Krebsforschungszentrum (DKFZ) widmet sich als größte biomedizinische Forschungseinrichtung in Deutschland ganz der Krebsforschung. Über 1.200 Wissenschaftler erforschen in mehr als 90 Abteilungen und Arbeitsgruppen wie Krebs entsteht, erfassen Risikofaktoren und suchen nach Strategien, um eine Krebserkrankung zu verhindern.

Eine große Herausforderung dabei sind riesige Datenmengen, die bei der genetischen Analyse von Zellen anfallen, um Auslöser oder Indikatoren für Krebs zu identifizieren. Trotz der Nutzung eines HPC-Clusters kam es beim DKFZ immer wieder zu Engpässen bei der Analyse der Genom-Daten. Das verzögerte den Forschungsfortschritt und führte zu Unzufriedenheit bei den Anwendern.

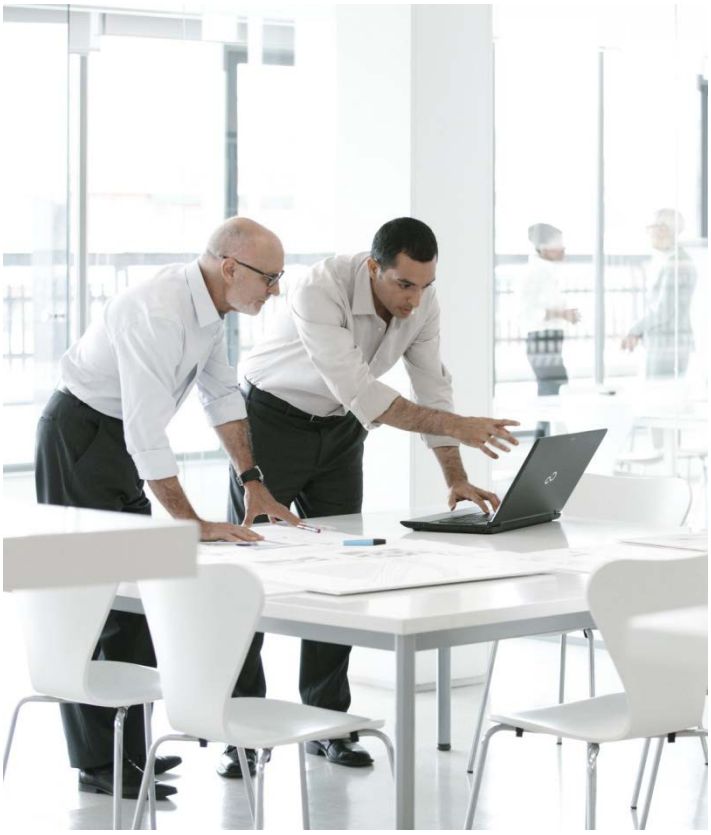
„Wir sind Teil des International Genome Consortiums und einer von weltweit sieben Standorten, die sich der Krebserforschung mittels HPC-Plattformen widmen", erläutert Dr. Matthias Schlesner, Group Leader Computational Oncology am DKFZ. „Wir haben einen standardisierten HPC-Cluster auf Basis der Technologien von Fujitsu, Intel und SAP aufgebaut. Trotzdem hatten wir Probleme mit der Schnittstelle zum File-Server und Schwankungen in der Rechenleistung."

Ein smarter Ansatz für die Datenanalyse

Das DKFZ diskutierte mit seinen Technologiepartnern deshalb Möglichkeiten, um die Leistungsfähigkeit zu verbessern, und erweiterte die zugrundeliegende Hardware. Um diese Bemühungen noch zu unterstützen, setzte es zudem FUJITSU Integrated System PRIMEFLEX for Hadoop ein.

„Nachdem wir die Plattform für die lokale Architektur optimiert hatten, haben wir auf Hadoop migriert. Wir wollten herausfinden, ob wir Engpässe so überwinden können. Normalerweise suchen wir bei unseren Untersuchungen nach relevanten genetischen Merkmalen bei einer Patientenkohorte. Wenn wir solche nicht finden, können wir jedoch nicht genau sagen, ob sie tatsächlich nicht vorhanden sind, oder ob wir sie nur nicht sehen können. Das heißt: wir müssen dann zusätzliche Rechenschritte unternehmen und die gesamten Rohdaten erneut untersuchen. Mit Hadoop können wir auch große Datenmengen parallel verarbeiten. So können wir Reduktions-Schritte überspringen und ohne Performanceeinbußen weiterarbeiten."

Die FUJITSU PRIMERGY CX400 Server-Lösung ermöglicht einen neuen, Daten-orientierten Analyseansatz. Sie nutzt dazu die End-to-End-Analysesoftware von Datameer® und kann dabei strukturierte und unstrukturierte Daten gleichermaßen einbeziehen. Dadurch können auch Nicht-Programmierer und Business-Anwender Daten einfach nutzen und bearbeiten – intuitiv in einer Tabelle, die die Ausgangsdaten und die Bearbeitung aufzeigt. Heute nutzt ein Forscherteam FUJITSU PRIMEFLEX for Hadoop, um größere Genom-Kohorten zu analysieren und dadurch besser zu verstehen, wie der Krebs agiert und wie er aufgehalten werden kann.



Das Genom entschlüsseln

Mit dem neuen System konnte das DKFZ die Datenanalyse vereinfachen, den Analyseprozess beschleunigen und effizienter genetische Varianten bei einem Patienten identifizieren. Und dabei alle möglichen Genom-Positionen untersuchen, ohne Daten zu transferieren. Der Ansatz der „Human-centric Innovation" trägt so zu einem besseren Verständnis der Erkrankung und zu besseren Behandlungsergebnissen für die Patienten bei.

„FUJITUS PRIMEFLEX for Hadoop macht alles wesentlich einfacher und ermöglicht uns schnelle und unabhängige Analysen von mehr als 900.000 bekannten und klinisch relevanten genetischen Merkmalen", führt Schlesner fort. „Mit Hadoop erfolgt die Verarbeitung wesentlich schneller als mit einem klassischen HPC-Setup und ohne Netzwerk-Engpässe. Die parallele Datenverarbeitung kann die Analyse eines einzelnen Genoms um den Faktor Vier oder mehr beschleunigen."

Das DKFZ profitiert zudem von der Zuverlässigkeit der Fujitsu Hardware, die seit Beginn der Pilotphase keinen einzigen Ausfall hatte. So kann sich die Forschungseinrichtung ihrer Kernkompetenz, der Krebsforschung, widmen.

Daten-Analyse der Zukunft

Das DKFZ unternimmt regelmäßig Performance-Tests, um Fujitsu for Hadoop weiter zu justieren. „Unsere Erfahrungen mit der neuen Plattform sind sehr gut. Wir konnten unser genetisches Fachwissen perfekt mit der technischen Expertise von Fujitsu kombinieren," resümiert Schlesner. „Nun können wir Routine-Analysen für die Entwicklung neuer Forschungsmethoden nutzen."

Zusammenfassung

FUJITSU Integrated System PRIMEFLEX für Hadoop ist eine leistungsstarke und skalierbare Plattform für die schnelle Analyse großer Datenmengen. Sie kombiniert vorkonfigurierte und vorab getestete Hardware, die auf Industrie-Standard-Komponenten basiert, mit Open-Source-Software von Cloudera und Big-Data-Analysesoftware von Datameer®. Das macht PRIMEFLEX for Hadoop zu einer der kosteneffizientesten, genauesten und schnellsten Lösung für Forschungseinrichtungen, die damit Daten bestmöglich für ihre Patienten nutzen können.

Kontakt
FUJITSU
Fujitsu Technology Solutions GmbH
Mies-van-der-Rohe-Strasse 8, 80807 München, Deutschland
Telefon: 00800 37210000*
E-Mail: cic@ts.fujitsu.com
Website: http://de.fujitsu.com
05-2016
*verfügbar und kostenfrei aus allen Netzen in D/A/CH