DATA LAKE FÜR DIE MEDIZINISCHE FORSCHUNG

Berlin Institute of Health (BIH) und Charité optimieren Daten-Speicherung und -Analyse

AUF EINEN BLICK

AUFGABE

Data-Lake-Plattform zur zentralen und sicheren Daten-Speicherung und -Auswertung

SYSTEME UND SOFTWARE

HARDWARE:

> 25 HPE Server: DL360 + DL380 im Durchschnitt 12 Cores mit 384 GB RAM und 32 TB HDD, 25 GB Ethernet

SOFTWARE:

> Cloudera HDP 3.1.4 Distribution

> Aus der Distribution eingesetzte Software:

- HDFS, YARN, Ambari, Grafana, Apache Ranger, Apache Knox, Apache Superset, Apache Hive, Apache Oozie, Apache Zoo-keeper, Apache Spark, Apache Zeppelin, Tez, Apache Nifi

> Doku-Wiki

> DevOps Tools: GitLab für CI/CD und Ansible

VORTEILE

> Sichere und zentrale Daten-Speicherung

> Zentrales Berechtigungskonzept

> Einfachere und schnellere Daten-Nutzung

BIH & CHARITÉ

Das Berlin Institute of Health (BIH) ist eine Wissenschaftseinrichtung für Translation und Präzisionsmedizin und widmet sich neuen Ansätzen für bessere Prognosen und neuartigen Therapien bei fortschreitenden Krankheiten und ungelösten Gesundheitsproblemen, um Menschen Lebensqualität zurückzugeben oder sie zu erhalten. Die Charité – Universitäts-medizin Berlin ist eine der BIH-Gründungsinstitutionen und dort arbeiten jeden Tag rund 4.255 Wissenschaftlerinnen, Wissenschaftler, Ärztinnen und Ärzte in über 1.000 Projekten, Arbeitsgruppen und Kooperationen daran, zukunftsweisende Entwicklungen auf dem Gebiet der Medizin bei höchsten Anforderungen an Qualität und Nachhaltigkeit voran-zubringen.

HERAUSFORDERUNG DATENKONSOLIDIERUNG

Das BIH und der Bereich Forschung & Lehre der Charité hatten das Ziel, eine zentrale Data-Lake-Plattform zu schaffen. Damit sollten strukturierte und unstrukturierte Forschungs- sowie klinische Daten zentral mit einem einheitlichen Sicherheits- und Berechtigungskon-zept gespeichert werden. Außerdem sollten so Rechenressourcen und ein einheitliches Toolset zur explorativen Auswertung der gesammelten Forschungsdaten zur Verfügung gestellt werden. Dadurch sollten nicht nur die zentrale und übergreifende Verwendung der Forschungsdaten ermöglicht, sondern auch Kosten und Zeitaufwand reduziert werden.

LÖSUNG: DATA LAKE MIT CLOUDERA

Die SVA-Experten konnten bei der Ausarbeitung einer Lösung nicht nur auf langjährige praktische Erfahrung mit dem Aufbau von Data-Lake-Plattformen und deren Tools setzen, sondern vor allem auch auf den engen Austausch mit dem Kunden. Die Basis bilden zu-nächst flexible und leistungsstarke HPE ProLiant Server. Sowohl DL360 als auch DL380 Systeme kommen hier zum Einsatz – skalierbar und sicher.

GESAMTLÖSUNG AUS EINER HAND

Die erfolgreiche Partnerschaft der SVA mit Cloudera führte zum Einsatz des kosteneffizienten Open Source Frameworks Cloudera Hortonworks Data Platform. Für die Plattform sprachen Vorteile wie agile Implementierungszeiten bei geringeren Gesamtbetriebskosten und unternehmensweite Zugriffskontrolle und Metadaten für Sicherheit und Governance. Ein SVA-Team aus den Bereichen Data Engineering, Big Data Architecture und Data Science konnte die Konzeption und Installation von Hardware und Plattform sowie die Prozesse für die Etablierung des weiteren Workflows aus einer Hand liefern.

Wichtige Punkte in diesem Projekt waren neben der Absicherung des Clusters mittels Kerberos & In-Flight-Encryption, die Anbindung der einzelnen Clusterkomponenten an das zentrale AD (LDAP) sowie die Entwicklung von Rollen- und Rechtekonzepten und deren durchgängige Implementierung. Außerdem wurden ein Data-Governance-Konzept für die Plattform erstellt, Quellsysteme über verschiedene Schnittstellen (u. a. Apache Nifi) angebunden und Datapipelines zur Aufbereitung der Quelldaten (u. a. mit Apache Spark und Kafka) implementiert.

PROJEKTUMFELD

Der Data Lake ist als ein maßgeblicher Teil der Health Data Plattform (HDP) ein wesentlicher und integraler Bestandteil der Digitalisierungsstrategie der Charité. Die HDP beherbergt auch das Datenintegrationszentrum, welches die Daten für die Medizininformatik-Initiative (MII) und das Konsortium HiGHmed aufbereitet. Im Rahmen der MII und HiGHmed ist Interoperabilität ein zentraler Aspekt – strukturierte Daten werden in den Data Lake mittels openEHR und FHIR transferiert. Im Zusammenhang mit der SARS-CoV2-Pandemie konnte auch mit Hilfe der HDP und des Data Lakes mittels eines POC innerhalb von ein paar Tagen ein Transfer von Behandlungsdaten in zentrale MII-Komponenten beispielhaft durchgeführt werden. Auf den Erkenntnissen des POC basierend wird derzeit auch die COVID-19-Infrastruktur der MII für die deutsche Universitätsmedizin umgesetzt.

AGILES UND ERFOLGREICHES PROJEKT

Die Implementierung der Data-Lake-Plattform ist ein anhaltender Prozess mit immer neuen und sich verändernden Anforderungen und wird über die Laufzeit der Plattform nicht abgeschlossen sein. Daher wurde gemeinsam schnell die Einführung von SCRUM als agile Methodik beschlossen und DevOps Tools wie GitLab und Ansible eingesetzt. Trotz oder wegen dieser Art Kulturwandel – mit neuen Rollen in der IT und deren Ausgestaltung (Data Architect, Data Engineer, Data Scientist, DataOps, DevOps) – läuft das Projekt äußerst zufriedenstellend. Die nun einfachere und schnellere Nutzung von Forschungsdaten und Gesundheitsdaten führt schon jetzt zu einer stärkeren Vernetzung der Forschungseinrichtungen und bringt neue Impulse für den Bereich digitale Medizin.

KONTAKT
SVA System Vertrieb Alexander GmbH
Borsigstraße 14
65205 Wiesbaden
Tel. +49 6122 536-0
Fax +49 6122 536-399
mail@sva.de
www.sva.de