White Paper

Big Data in Zeiten von Commercial Forking

Auswege aus dem Lizenz-Dilemma

IONOS

Inhalt

1 Einleitung **3**

2 Die derzeitige Lage: Big Data, Data Science und KI **4**

3 Die Herausforderungen **6**

3.1 Risiken beim Einsatz von Open Source für Big Data 6

3.2 Rückgang in der Nutzung von flexiblen Open-Source-Lizenzen 10

3.3 Flexibilität und Sicherheit im Bereich Big Data und Data Science 12

4 Die Lösung **14**

4.1 Big Data Managed Services 14

4.1.1 Infrastructure-as-Code 15

4.1.2 Open Source in Verbindung mit Infrastructure-as-Code 15

4.1.3 Praxisbeispiel: MARISPACE-X auf Gaia-X 16

4.2 Big Data Frameworks und -Plattformen als Managed Service in der europäischen Cloud 19

4.3 IONOS und Stackable Cloud Data Warehouse 21

5 Fazit **23**

Über IONOS **24**

Impressum **25**

1 Einleitung

Im Big-Data-Bereich kommt besonders häufig Open Source zum Einsatz. Allerdings gilt besonders hier, dass die Lizenzierung von Software als Open Source nicht unbedingt bedeutet, dass eine Software kostenlos ist und ohne Lizenzbedingungen zum Einsatz kommen kann. Hier müssen Unternehmen, Organisationen und auch Behörden darauf achten, welche Bedingungen zum Einsatz einer Software eingehalten werden müssen, und was der tatsächliche Preis ist.

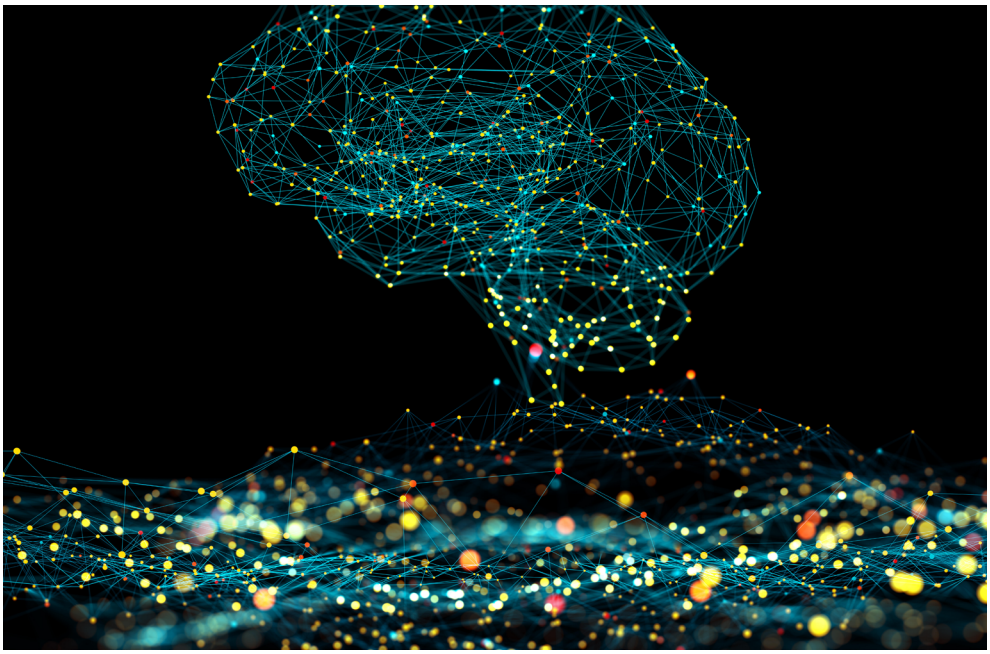
Der Betrieb im eigenen Rechenzentrum, auch On-Premises genannt, ist oft teuer, kompliziert und schwer zu verwalten. Hier kann es sinnvoll sein auf Open-Source-Lösungen aus der Cloud zu setzen. Dabei spielen natürlich Datensicherheit, Datenschutz und auch die Datensouveränität eine wesentliche Rolle. In diesem White Paper wird aufgezeigt, wie sich diese drei Dinge unter einen Hut bringen lassen und gleichzeitig die Big-Data-Analyse maßgeblich erleichtert werden kann.

2 Die derzeitige Lage: Big Data, Data Science und KI

Im Big-Data-Bereich und bei Data-Science-Projekten kommen auch immer mehr Technologien aus den Bereichen künstlicher Intelligenz (KI), Machine Learning (ML) und Deep Learning (DL) zum Einsatz. Hier lassen sich auch große Datenmengen durch KI-Technologien sehr viel effizienter verarbeiten als mit altbekannten Mitteln. Bei diesem Trend kommen häufig mehrere Lösungen zum Einsatz, die erst beim gemeinsamen Betrieb ihr volles Potential erreichen. Die Verbindung dieser Dienste ist nicht nur fachlich, sondern auch lizenzrechtlich eine Herausforderung.

Open Source, KI und Big Data

Big-Data-Projekte sind auf die Speicherung und Analyse von großen Datenmengen spezialisiert, während KI-Lösungen dabei helfen die Daten zu analysieren und zu verarbeiten. In allen beteiligten Bereichen, also Big Data, Data Science und KI, kommen in sehr vielen Fällen Open-Source-Anwendungen zum Einsatz. Allerdings müssen Verantwortliche in Unternehmen, Organisationen und auch bei Behörden einiges beachten.



Die Verwendung von KI-Technologien bei der Analyse von Daten ist einer der wichtigsten Trends, der sich auch in 2022 rasant fortsetzt. Data Science Machine Learning Platforms (DSML) finden eine immer größere Verbreitung und basieren meistens aus mehreren Lösungen, Tools und Frameworks. Bekannte Lösungsanbieter in diesem Bereich sind Dataiku, Databricks oder auch Microsoft.

Grundsätzlich ist der Einsatz von Cloud-Plattformen ideal, da hier nicht nur Lizenzbedingungen wesentlich einfacher umgesetzt werden können, sondern auch die Skalierbarkeit. KI und Big Data benötigen leistungsstarke Hardware und sehr viel

Speicherplatz, der im Netzwerk auch effektiv zur Verfügung stehen muss. Der Einsatz von Cloud-Lösungen ist deshalb schon aus Gründen der Skalierbarkeit ideal. Open-Source-Lösungen und deren optimale Verwendung spielen an dieser Stelle eine wesentliche Rolle.

> *Unternehmen, Organisationen und Behörden wollen ihre Daten in immer größerer Menge, immer schneller und effektiver verarbeiten. Gleichzeitig müssen die Daten sicher gespeichert werden, der Datenschutz muss eingehalten werden und auch die Datensouveränität ist enorm wichtig. Das geht im Grunde genommen nur mit Data Science in Verbindung mit KI-Technologien aus europäischen Rechenzentren. Zu den großen Trends gehören auch KI-spezifische Infrastruktur-Stacks (MLOps und AIOps), Infrastructure-as-a-Service (IaaS) und Platform-as-a-Service (PaaS) sorgen dafür, dass Unternehmen, Behörden und Organisationen die notwendige Hard- und Software schnell und sicher zur Verfügung steht.*

Big Data arbeitet in vielen Bereichen bereits mit Echtzeitanalysen, was die Messlatte für die Infrastruktur im Unternehmen natürlich entsprechend hoch legt. Dazu kommen die ständigen Veränderungen in der Infrastruktur. Neben dem On-Premises-Einsatz, also im lokalen Rechenzentrum, setzen immer mehr Unternehmen auf die Cloud oder hybride Netzwerke. Auch der Einsatz von Multi-Cloud-Infrastrukturen findet eine immer größere Verbreitung.

Modern Data Stacks und Cloud Warehouse

Modern Data Stacks bieten verschiedene Tools und Technologien, mit denen sich Daten analysieren und verarbeiten lassen. Dabei handelt es sich häufig um verschiedene Anwendungen, die miteinander verknüpft sind. Solche Lösungen lassen sich im lokalen Rechenzentrum (On-Premises) genauso betreiben, wie in der Cloud. Allerdings kommt häufig eine Kombination von zahlreichen Tools und Lösungen zum Einsatz, deren lokaler Betrieb kaum Sinn ergibt. Hier mangelt es häufig an notwendigen Grundlagen für Rechenleistung, Speicherkapazität und Manpower, um die Projekte umsetzen zu können.

Komplexität von Big Data und Data Science

Seit einigen Jahren steigt der Trend zum Einsatz von Big-Data- und Data-Science-Lösungen in der Cloud auch durch den Einsatz der Cloud Warehouses stark an. Bei dieser Technologie erhält das System seine Daten aus verschiedenen Quellen, extrahiert sie, und verarbeitet sie weiter. Dazu gehören auch umfassende Analysen mit KI-Technologien sowie Visualisierung der Informationen, zum Teil auch in Echtzeit. Beispiele dafür sind Amazon Redshift, Snowflake, Google BigQuery und Microsoft Synapse. Es gibt aber zahlreiche weitere Lösungen in diesem Bereich, die Daten in kurzer Zeit extrahieren, laden und transformieren. Dabei kommt zum großen Teil Open Source zum Einsatz, auch bei Lösungen im Bereich Software-as-a-Service (SaaS). Open Source ist in diesem Bereich ein wichtiges Fundament, das bei lokalen Installationen On-Premises eine genauso große Rolle spielt, wie bei SaaS und PaaS.

Neue Tools und Lösungen

Ein weiteres Beispiel dazu ist das DBT (Data Build Tool). Das Open-Source-Tool hilft bei der Transformation von Daten und wird von immer mehr Data Scientists genutzt, um mit SQL-Abfragen Daten von Quellen zu transformieren. Data Engineering spielt in diesem Bereich ebenfalls eine wichtige Rolle. Hier kommen zum Beispiel Tools wie Spark zum Einsatz. Solche Lösungen erfordern Kenntnisse für die Lizenzierung, aber auch umfassende Fachkenntnisse zum Umgang und Betrieb der Lösung. Beim Einsatz in der Cloud ist der Umgang dagegen sehr viel einfacher, und zwar sowohl bezüglich der Lizenzen als auch der tatsächlichen Verwendung und auch beim Datenschutz, der Datensicherheit und natürlich auch der Datensouveränität.

Durch neue Trends, Lösungen und auch Tools findet eine Vermischung der Technologien statt. Dabei spielen Lizenzen genauso eine Rolle wie auch der tatsächliche Betrieb der Lösungen. So wachsen zum Beispiel Data Lakes und Data Warehouses zu Data Lakehouses oder Unified Analytics Warehouses zusammen. Zusätzlich sind hier wieder KI-/ML-Technologien im Einsatz, die dabei helfen mehr oder weniger strukturierte Daten zu transformieren, zu verarbeiten und anschließend zu analysieren. Es findet eine Verschmelzung von Technologien statt, was im Rahmen der Lizenzierung eine genauso große Rolle spielt wie bei der tatsächlichen Umsetzung und Implementierung der Lösungen im Netzwerk oder der Cloud.

Neue Trends, Lösungen und Tools führen zu einer Vermischung der Technologien.

In vielen Fällen ist der Einsatz von Cloud-Lösungen für den Einsatz von Big Data wesentlich besser geeignet als die Verwendung von lokalen installierten Open Source-Lösungen. In der Cloud sind die Lizenzbedingungen tendenziell viel klarer und nebenbei lassen sich Cloud-Lösungen auch noch leichter skalieren und bedienen, wesentlich sicherer betreiben und gleichzeitig schneller und effektiver den Benutzern zur Verfügung stellen.

3 Die Herausforderungen

3.1 Risiken beim Einsatz von Open Source für Big Data

Grundsätzlich ist Open Source ein interessanter Ansatz für die Verbreitung von Software. In den meisten Fällen stehen die Programme zumindest teilweise kostenlos zur Verfügung. Da der Quellcode öffentlich bekannt ist, sind Sicherheitslücken schnell identifiziert und die Community hilft dabei diese zu schließen.

- **Komplexe Lizenzierung kann den Einsatz von Open Source zum Problem machen**
  Im Big-Data-Bereich sind Softwarelösungen in den meisten Fällen sehr viel komplexer aufgebaut. Hier kommt zwar häufig Open Source zum Einsatz, allerdings in vielen Fällen auch im Zusammenspiel mit kommerziellen Produkten. In diesem Fall wird die Lizenzierung komplexer, da herkömmliche Lizenzen und oft komplexe Lizenzbedingungen der Open-Source-Komponenten miteinander vermischt werden. Das muss bei der Planung der korrekten Lizenzierung der Umgebung Berücksichtigung finden. Auch für die Planung der Kosten spielt das eine wesentliche Rolle.

- **Software Forks machen Lizenzierung noch schwerer und oft auch teurer**
  Häufig gibt es bei Anwendungen auch sogenannte Software Forks. Dabei spalten sich neue Projekte aus bereits etablierten Anwendungen ab und werden getrennt davon weiterentwickelt. Das macht den Einsatz natürlich wesentlich komplexer. Kommt dann noch eine neue Lizenzierung dazu, oder sogar die Umstellung ehemaliger Open Source zu einer kommerziellen Lizenzierung, macht das den korrekten, legalen und richtig lizenzierten Einsatz von Open Source sehr schwer.

  Software Forks kommen häufiger bei Open-Source-Software vor, als viele denken und haben bereits eine lange Tradition. LibreOffice spaltete sich von OpenOffice.org ab, das Web-Content-Management-System Joomla entstand aus Mambo. VNC wurde ab 2002 mit der Version 3.3. kostenpflichtig. Dadurch entstanden weitere Implementierungen unter der GPL, die dasselbe Protokoll verwenden.

  Was verbirgt sich hinter GPL?

  Die GNU General Public License ist aus dem GNU Projekt im Jahre 1983 entstanden. Die Allgemeinheit soll freien Zugang zur Software erhalten und gleichzeitig soll das Recht auf Modifikationen bestehen. Die GPL war die erste Lizenz, die Anwendern auch den Zugriff auf den Quelltext einer Software liefert. Die Bedingung ist, dass modifizierte Versionen der Software ebenfalls wieder frei zur Verfügung stehen muss.

  Das letzte Beispiel zeigt, wie schnell es gehen kann, dass aus einer ehemaligen kostenlosen Open-Source-Lösung ein kommerzielles Produkt wird. Durch das Einbinden von Open-Source-Technologien durch Software-Forking in kommerzielle Software, die in vielen Fällen auch teuer verkauft werden soll, wird das Prinzip von Open Source ausgehebelt und für den eigenen Profit genutzt. Anwender haben keinerlei Vorteile mehr, dass Teile der kommerziellen Software auf Open Source aufbauen.

- **Beim Einsatz von Open-Source-Software fehlt oft der Support**
  Der Einsatz von Open-Source-Software birgt einige Risiken, die in diesem White Paper ausführlich aufgezeigt werden. Zusätzlich zu den in den weiteren Abschnitten erwähnten Problemen der Lizenzänderungen gibt es weitere Herausforderungen, die Berücksichtigung finden sollten. Dazu gehört zum Beispiel der oft fehlende Support beim Einsatz der jeweiligen Lösung.

  Open-Source-Software bietet oft keine Garantie und die Wartung ist nicht immer gewährleistet. Außerdem sind Open-Source-Softwareinitiativen nicht immer ausreichend gut finanziert. Das kann darin resultieren, dass ein Projekt entweder eingestellt oder unter einer anderen Lizenz von anderen Verantwortlichen weiter vorangetrieben wird.

Schwachstellen in Open-Source-Software sind öffentlich zugänglich und können dadurch im Prinzip von der Community schnell geschlossen werden. Allerdings haben auch Angreifer Zugang zu den Schwachstellen und können diese für Malware oder Hacker-Angriffe ausnutzen. Neue Funktionen und Fehlerkorrekturen finden nicht bei allen Open-Source-Lösungen ausreichend schnell Einzug. Das hängt natürlich von der unterstützenden Community und den am Source Code mitarbeitenden Entwicklern ab.

- **Open Source-Lizenzbedingungen können sich ändern und damit Usern Probleme bereiten**
  Es gibt aber auch aktuellere Beispiele für Umstellungen durch kommerzielles Forking: Am Beispiel der bekannten Suchengine Elasticsearch wird aufgezeigt, wie auch im Big-Data-Bereich Open-Source-Lösungen sehr schnell zu Problemen bei der Lizenzierung führen können und der Einsatz von Open Source schnell teuer werden kann bzw. rechtliche Probleme aufwirft.

  Betreiben Unternehmen solche Lösungen selbst, kommen bei allen Herausforderungen zum Betrieb einer lokalen Big-Data-Plattform noch die Themen Lizenzierung, rechtssicherer Betrieb und unter Umständen auch das Austauschen einzelner Komponenten hinzu. Die wenigsten Unternehmen, Organisationen oder Behörden sind dazu in der Lage, auch noch diese Aufgaben selbst zu meistern.

- **Aus einer beliebten Open Source kann schnell kommerzielle Software werden**
  Die bekannte Suchengine Elasticsearch ist zwar Open Source, allerdings hat das Unternehmen Elastic, unter dessen Dach die Software entwickelt wird, die Lizenzbedingungen der Software geändert.[1] Wer in seiner Umgebung die Open-Source-Lösung einsetzt, muss sich mit den rechtlichen und praktischen Verstrickungen auseinandersetzen. Elasticsearch wurde seit 2010 unter der Apache 2-Lizenz als Open Source veröffentlicht. Gleichzeitig baut Elasticsearch auf der Lucene-Bibliothek auf, die ebenfalls als Open Source zur Verfügung steht.

  Im Jahr 2012 hat der Entwickler Shay Banon das Unternehmen Elastic gegründet, das im Jahr 2018 schließlich an die Börse ging. Parallel zur Open Source Elasticsearch hat Elastic Erweiterungen veröffentlicht, die unter eigenen Lizenzbedingungen mit der Bezeichnung „Elastic-Lizenz“ zur Verfügung stehen. Das macht die Komplexität einer korrekten Lizenzierung des Einsatzes von Elastic-search schnell klar.

[1] Banon (2021): ‚Nochmals zum Thema „offen“, Teil II – Fortsetzung unseres Blogs „Doubling down on open“ von vor drei Jahren‘, online verfügbar unter: https://www.elastic.co/de/blog/licensing-change.

- **Open-Source-Lizenzbedingungen sind nicht in Stein gemeißelt**
  Immer mehr Funktionen von Elasticsearch fallen unter diese neue Elastic-Lizenz und sind damit per Definition keine Open Source mehr. Einer der Auslöser: Unternehmen wie Amazon Web Services (AWS) haben damit begonnen Elasticsearch als Cloud-Software-as-a-Service (SaaS) anzubieten. Als Reaktion darauf hat Elastic Komponenten von Elasticsearch unter die Elastic-Lizenz gestellt und es damit nahezu unmöglich gemacht, dass die Software zuverlässig als Open Source zum Einsatz kommen kann.

  AWS hat auch darauf reagiert und Elasticsearch für seine Plattform geforked. Die neue Lösung steht als Open Source zur Verfügung. Allerdings klagt Elastic gegen AWS und auch gegen Softwareanbieter, die Lösungen für die neue Software mit der Bezeichnung Elasticsearch OpenDistro auf AWS anbieten.[2]

- **Neue Lizenzen schränken die Freiheit von Open Source ein**
  Darüber hinaus will Elastic neue Versionen von Elasticsearch unter einer speziellen Lizenz veröffentlichen, die es faktisch unmöglich macht, legal gehostete Versionen von Elastic anzubinden und zu nutzen. Ein weiteres, prominentes Beispiel für eine solche Lizenzpolitik ist MongoDB. Auch hier wurden Lizenzen so geändert, dass das Hosten durch Fremdanbieter kaum mehr möglich ist. Dazu wurde für die Software 2018 die neue Server Side Public License (SPPL) eingeführt, die jetzt auch für Elasticsearch zum Einsatz kommt. Wer auf Produkte unter dieser Lizenz setzt, zum Beispiel MongoDB oder Elasticsearch, muss Code von modifizierten Versionen komplett offenlegen.

  Server Side Public License

  Server Side Public License (SPPL) ist eine Software-Lizenz, die 2018 von MongoDB eingeführt wurde. Die Lizenz ist nicht als Open-Source-Lizenz anerkannt. Nutzen Entwickler eine Software unter der SPPL, müssen sie den Quellcode der eigenen Software ebenfalls veröffentlichen, wenn sie in einem öffentlichen Cloud- oder Webdienst zum Einsatz kommt. Das gilt auch für jeden anderen Quellcode, der mit dem Projekt im Zusammenhang steht. Dazu gehören zum Beispiel auch die Software für das Systemmanagement, Benutzerschnittstellen, Storage Backend oder Backups. Viele Kritiker sehen darin vor allem den Zweck zu verhindern, dass unter der SPPL lizenzierte Software in anderen Diensten genutzt wird.

  Faktisch ist Elasticsearch daher keine Open Source mehr und rechtlich zumindest ein Risiko. Die Open Source Initiative (OSI) hat die neuen Lizenzbedingungen von Elasticsearch nicht akzeptiert, da sie die Freiheiten für Entwickler und auch für Anwender stark einschränken. Kontrolliert ein Unternehmen weitgehende Teile einer Software und deren Lizenzen, können sich die Spielregeln sehr schnell ändern.

---

[2] Förster (2021): ‚Elastic: Amazon ist schuld am Open-Source-Ende', online verfügbar unter: https://www.heise.de/news/Elastic-Amazon-ist-schuld-am-Open-Source-Ende-5030541.html.

Anwender und Anbieter werden dadurch zum Spielball für die Lizenzierung und können das Produkt kaum mehr legal richtig lizenziert einsetzen und schon gar nicht Kunden anbieten. Diese Änderung der Lizenzen von Elasticsearch hat zu einem neuen Fork mit der Bezeichnung „OpenSearch“ geführt, das auf GitHub zur Verfügung steht. Die Entwickler nutzen hier die Apache 2-Lizenz, während Elasticsearch jetzt auf die SPPL setzt.

Elastic hat übrigens Ende 2020 noch angekündigt, dass Elasticsearch dauerhaft unter der Apache 2-Lizenz laufen soll. Das zeigt, wie schnell sich solche Fakten ändern und wie zuverlässig die Lizenzaussagen sind. Elasticsearch ist unter der neuen SPPL-Lizenz keine Open Source mehr. Wer in Zukunft also Dienstleistungen auf Basis von Elasticsearch einsetzt, muss den Einsatz juristisch prüfen lassen. Es kommt dabei zu rechtlichen und letztlich zu massiven Geschäftsrisiken.

- **Neue Software-Forks können schlechter sein als die Quelle**
  Wenn eine Open Source zu kommerzieller Software wird, so wie es bei Elasticsearch faktisch der Fall ist, entwickeln sich häufig Forks, die den Open-Source-Quellcode weiterverwenden. Allerdings ergibt sich hier wiederum die Gefahr, dass die Forks bestimmte Teile des Codes nicht verwenden dürfen und neuen Code komplett unabhängig entwickeln müssen.

  Das geht natürlich auf Dauer auf die Qualität und Kompatibilität des Produktes, da es sich dabei um eine komplett neue Software handelt. Am Beispiel von Elasticsearch ist darüber hinaus zu erwarten, dass sich viele Entwickler von dem Projekt zurückziehen und nicht mehr zu den Open-Source-Komponenten beitragen. Elastic bezeichnet den Code in Elasticsearch auch nicht mehr als Open Source, sondern als Open Code.

3.2 Rückgang in der Nutzung von flexiblen Open-Source-Lizenzen

Die generelle Verwendung von Open-Source-Lizenzen für neue Software geht sukzessive zurück. Bereits 2017 hat Stephen O’Grady von Redmonk festgestellt, dass sich die Verwendungshäufigkeit der GNU General Public License (GPL) von 2010 bis 2017 halbiert hat. Der Trend geht auch bis weit in die 2020er-Jahre weiter.[3]

Open Source ist nicht gleichbedeutend mit kostenlos

Zwar nutzen immer mehr Software-Produkte Apache- und MIT-Lizenzen als Open Source, allerdings bietet nur die GPL maximal offene Verwendung von Quellcode. So fallen unter die GPL beispielsweise auch die Linux-Distributionen Debian, Ubuntu oder Fedora.

[3] O’Grady (2017): ‚The State of Open Source Licensing‘, online verfügbar unter: https://redmonk.com/sogrady/2017/01/13/the-state-of-open-source-licensing/.

Wenn ein Projekt den Code eines anderen GPL-Projektes nutzt, dann muss auch dieses Projekt unter der GPL laufen. GPL-Produkte dürfen kommerzialisiert werden, allerdings ist die GPL sehr starr, wenn es darum geht Software kommerziell zur Verfügung zu stellen. Das ist einer der Gründe, warum sich viele Entwickler von der GPL zurückziehen und auf andere Lizenzierungen setzen. Wer Open Source im lokalen Netzwerk betreibt, muss sich wohl oder übel damit auseinandersetzen. Wer aber Open Source in der Cloud nutzt, delegiert auch diese Aufgaben zum PaaS- oder IaaS-Anbieter.

Immer mehr Unternehmen denken laut darüber nach, mit Open Source Geld zu verdienen. Neben reinen Dienstleistungen und Distributionen wollen Unternehmen Lizenzen nutzen, die es ermöglichen bei entsprechender Verbreitung der Software flexibler zu reagieren. Der Kompromiss zwischen der Bereitstellung kostenloser Open Source und der Möglichkeit Geld zu verdienen ist für viele Unternehmen schwierig.

GPL versus Apache 2, MPL und MIT

Die GPL setzt faktisch die umfassende, kostenlose Nutzung von Open Source voraus. Das ist in weiten Teilen der Anwenderschaft, aber auch bei Entscheidern in Unternehmen das Grundverständnis von Open Source. Allerdings müssen Entwickler und Unternehmen, die eine Software bereitstellen, in der Lage sein den Aufwand zu monetarisieren. Hier sind Apache 2- und MIT-Lizenzen wesentlich flexibler, genauso wie die MLP-Lizenz. Im Big-Data-Bereich kommt häufig auch noch die Mozilla Public License (MPL) zum Einsatz. Hier ist es erlaubt, auch bei der Verwendung anderer MPL-Projekte im eigenen Projekt, andere Lizenzarten zu verwenden. Wer eine neue Big-Data-Lösung entwickelt und dabei auf MPL-Projekte zurückgreift, kann sein Projekt problemlos unter der Apache 2- oder MIT-Lizenz veröffentlichen. Diese Lizenzen sind daher sehr viel flexibler. Allerdings muss der MPL-lizenzierte Code dabei vom restlichen Code klar abgetrennt sein.

Die Apache 2-Lizenz wird durch die Apache Software Foundation (ASF) veröffentlicht. Bei der Verwendung dieser Lizenz dürfen Entwickler beliebig andere Lizenzen nutzen, müssen die Lizenzen der verwendeten Produkte aber erwähnen und alle Änderungen dokumentieren. Der Code darf auch in Closed-Code-Szenarien genutzt und kommerziell verwendet werden.

In den letzten Jahren hat sich die Einstellung vieler Entwickler verändert. Es sollen weiterhin Lösungen auch als Open Source zur Verfügung stehen, es muss aber auch möglich sein damit Geld zu verdienen. Es ist daher zu erwarten, dass auch in Zukunft immer mehr Lösungen auf Basis von Apache 2 und MIT veröffentlicht werden. Prominentes Beispiel für die Apache-Lizenz ist das Big-Data-Framework Apache Hadoop.

Die MIT-Lizenz des Massachusetts Institute of Technology gibt es seit den 1980er-Jahren. Die Lizenz gehört zu den einfachsten Lizenzen am Markt und auch zu den flexibelsten. Daher kommt sie auch bei vielen Big-Data- und Data-Science-Projekten zum Einsatz. Mit der Lizenz gibt es nahezu keine Einschränkungen, was die Verwendung von Open-Source-Komponenten betrifft. Es ist nur notwendig, das originale Copyright und die Lizenzbedingungen der verwendeten Software einzubinden. Autoren werden komplett von der Haftung entbunden.

GPL wird in Zukunft vermutlich vor allem für freie Software genutzt werden. Unternehmen, Organisationen und Behörden, die auf Open Source setzen, sollten also auch rechtzeitig planen, wie ein Wechsel der Lizenzen oder das Anpassen der Lizenzbedingungen kompensiert werden kann. Der Grundsatz, dass Open Source für alle Zeiten auch mit kostenlos gleichgesetzt werden kann, gilt schon lange nicht mehr.

Fazit

Lizenzänderungen können ein massives, geschäftliches Risiko darstellen, wenn Unternehmen, Organisationen oder Behörden eine Software für grundlegende Dienste einsetzen. Denn es ist durchaus möglich, dass bei Änderungen ein legaler Einsatz überhaupt nicht mehr möglich ist und umfassende Änderungen in der lokalen Infrastruktur notwendig sind.

3.3 Flexibilität und Sicherheit im Bereich Big Data und Data Science

Im Bereich von Big Data und Data Science kommt Open Source besonders häufig zum Einsatz. Das liegt unter anderem auch daran, dass große Unternehmen wie Netflix, Twitter, Facebook, Microsoft oder auch Google Lösungen entwickeln, die sie zunächst intern zur Verarbeitung ihrer enormen Datenmengen nutzen. Sobald das Tool stabil genug ist, wird es als Open Source der Allgemeinheit zur Verfügung gestellt. Für das Unternehmen hat das den Vorteil, dass die Lösung von einer großen Community ständig weiterentwickelt und verbessert wird. Die Herausforderung für Unternehmen, die diese Lösung einsetzen wollen, liegt darin, die passende zu finden, diese richtig zu lizenzieren und gleichzeitig dafür zu sorgen, dass die Komponenten optimal zusammenarbeiten.

Zugang zu Open Source Tools am Beispiel von Cloudera und Hortonworks

Spätestens seit 2019 hat die bisher beschriebene Lizenzproblematik auch im Big-Data-Markt einen neuen Höhepunkt erreicht. Durch die Fusion von Cloudera und Hortonworks wurden nahezu alle Lösungen von Cloudera kostenpflichtig. Wer auf die Lösungen des Anbieters zurückgreift, muss seither über Subscriptions regelmäßig den Einsatz der Software bezahlen.

[4] Cloudera (2021): ‚Cloudera pricing & licensing updates', online verfügbar unter: https://www.cloudera.com/products/pricing/pricing-update.html.

Dazu kommen Preiserhöhungen für bereits kostenpflichtige Produkte und umfassende Änderungen bei der Verwendung von Cloud-Lösungen. Seit dem Frühjahr 2021 bemerken alle Cloudera-Kunden schmerzhaft diese Änderungen der Lizenzpolitik.[4] Das zeigt mehr als deutlich, dass sich Unternehmen, Organisationen und Behörden immer auch mit Alternativen beschäftigen müssen, um im Bedarfsfall vorbereitet zu sein, wenn sich Änderungen ergeben.

Cloudera und Hortonworks bieten daher Distributionen oder auch Sammlungen von verschiedenen Tools, die als Gesamtlösung zur Verfügung stehen. Zwar sind einzelne Tools weiterhin Open Source, die komplette Gesamtlösung wird aber kostenpflichtig. Die kompletten Distributionen sind kommerzialisiert.

Das White Paper hat bisher aufgezeigt, welcher ungeheuren Dynamik der Open-Source-Markt bei Big Data und Data Science unterliegt. Es gibt bereits zahlreiche Technologien, Einsatzgebiete, Lösungen und Tools. Ständig erscheinen neue Herangehensweisen, Anwendungen dazu und viele dieser Lösungen arbeiten eng zusammen, fusionieren oder ergänzen sich. Es ist daher wichtig, dass sich Unternehmen, Organisationen und Behörden keiner starren Infrastruktur unterwerfen, sondern flexibel und zügig reagieren können. Dazu kommt die Notwendigkeit immer auf dem neuesten Stand zu sein, Software schnell aktualisieren zu können und gleichzeitig auch fortlaufend die Umgebung zu optimieren.

Flexibilität und Sicherheit spielen für Big Data und Data Science eine wesentliche Rolle.

Viele Unternehmen gehen in diesem Bereich auch den Infrastructure-as-Code-Ansatz. Dabei wird die maximale Flexibilität für Big-Data-Umgebungen erreicht. Alle Lösungen innerhalb der Big-Data-Infrastruktur lassen sich dadurch als Code darstellen, der durch Versionierung und dem CI/CD-Ansatz auch jederzeit erweiterbar und schnell anpassbar ist. Im eigenen Rechenzentrum lassen sich solche Ansätze aber kaum verfolgen, da hier häufig die Zeit fehlt und auch das notwendige Fachwissen kaum vorhanden ist.

Schlussendlich spielt natürlich auch der Preis eine Rolle. Wie das Beispiel Cloudera/Hortonworks zeigt, sind Lizenzmodelle ein wichtiger Faktor bei der Planung einer Data-Science-Umgebung. Modularität ist genauso wichtig. Funktionen innerhalb der Infrastruktur müssen schnell ersetzt werden können und dabei auch skalierbar bleiben. Hier spielt auch der Zusammenhang mit Skalierbarkeit und Preisentwicklung eine wichtige Rolle.

Sicherheit ist für Big Data und Big Data Science enorm wichtig. Der sichere Umgang mit Daten ist enorm wichtig, nicht nur wegen der DSGVO. Das BSI warnt in seinem Lagebericht zur IT-Sicherheit in Deutschland 2020 vor einer stark anwachsenden Zahl von Angriffen auf die IT-Infrastruktur von Unternehmen und Organisationen. Allein die Anzahl neuer Schadprogramm-Varianten hat 2020 um rund 117,4 Millionen zugenommen. Das sind knapp 320.000 neue Malware-Programme pro Tag. Im Vergleich zum Vorjahr beträgt die Steigerung noch einmal über ein Drittel.

Nach einer repräsentativen Studie des Bitcom aus 2021[5] entstand der deutschen Wirtschaft, quer durch alle Branchen, im Jahr 2020 ein Schaden von 220 Milliarden Euro durch kriminelle Cyber-Attacken. Das ist das Doppelte des Schadens aus dem Vorjahr.

Aus den Angriffen resultierten zahlreiche Sicherheitsvorfälle verbunden mit erfolgreichen Verschlüsselungsvorgängen von Daten durch Ransomware. In der Folge kam es bei den betroffenen Unternehmen zu Erpressung oder begleitender Computersabotage inklusive Unbrauchbarmachungen von IT-Systemen. Allein dieser Bereich hat sich 2020 im Vergleich mit den Jahren 2018/2019 vervierfacht. Von den befragten Unternehmen sehen fast 10 Prozent ihre Existenz bedroht. DDoS-Angriffe sind seit Jahren bekannt und steigen jährlich an, zuletzt um 10 Prozent von 2019 auf 2020. Im August 2021 hat Microsoft den bisher größten DDoS-Angriff von 70.000 Rechnern auf seine Azure-Cloud abgewehrt.

Mit der Komplexität der Umgebungen steigt auch die Komplexität der Sicherheitsinfrastrukturen, die notwendig ist, um die Daten zu sichern. Hier sollten sich Unternehmen einen Partner ins Boot holen, der die Infrastruktur leistungsstark, flexibel und gleichzeitig sicher zur Verfügung stellen kann.

4 Die Lösung

4.1 Big Data Managed Services

Open-Source-Komponenten für den Aufbau und Betrieb skalierbarer Data- und Streaming-Infrastrukturen können bei optimaler Planung der richtige Weg sein, um Big Data im Unternehmen zu betreiben. Dazu gehören vor allem die Komponenten:

- Moderne Data Warehouses und Data Lakehouses
- Event Streaming
- Machine Learning und künstliche Intelligenz

Wichtig ist an dieser Stelle die Modularität im Auge zu behalten und gleichzeitig für Alternativen zu sorgen. Es ist generell nie eine gute Idee, sich auf eine einzige Softwarekomponente zu konzentrieren. Die Beispiele in diesem White Paper zeigen, dass die Konsequenzen teuer, riskant und schlussendlich auch gefährlich für den wirtschaftlichen Bestand des Unternehmens sein können.

[5] bitkom (2021): ‚Angriffsziel deutsche Wirtschaft: mehr als 220 Milliarden Euro Schaden pro Jahr', online verfügbar unter: https://www.bitkom.org/Presse/Presseinformation/Angriffsziel-deutsche-Wirtschaft-mehr-als-220-Milliarden-Euro-Schaden-pro-Jahr.

Mit diesem Ansatz erreichen Unternehmen die maximale Flexibilität und gleichzeitig den größten Datenschutz. Auch die Komplexität der Konfiguration und des Betriebs ist deutlich reduziert. Natürlich profitieren davon auch die Wartung und Aktualisierung der kompletten Umgebung. Das spielt in den nächsten Jahren weiterhin eine wichtige Rolle. Unternehmen und Organisationen müssen sich in einem solchen Szenario nicht um Sicherheitsupdates kümmern. Die Tools arbeiten optimal zusammen, der Anbieter achtet auch bei Updates darauf, dass das so bleibt.

4.1.1 Infrastructure-as-Code

Big Data Managed Services ermöglicht den Betrieb von Big-Data-Infrastrukturen mit einer Nutzerzentrierung und aufwandsparender Bereitstellung. Hier kommt wieder der bereits erwähnte Infrastructure-as-Code-Ansatz zum Tragen. Das bietet einige Vorteile bei der Implementierung neuer Funktionen, Updates und auch für das Ersetzen von Komponenten, wenn sich Lizenzbedingungen oder andere Dinge ändern. Bei diesem Infrastructure-as-Code-Ansatz kommen folgende Bereiche zum Tragen:

- Automatisierte Provisionierung
- Konfiguration
- Monitoring
- Aktualisierung
- Wartung

4.1.2 Open Source in Verbindung mit Infrastructure-as-Code

Dabei kommen gängige Lösungen für Big Data und Data Science auf Basis von Open Source zum Einsatz. Dabei lassen sich gängige Software-Lösungen für Big Data in der IT des Unternehmens einbinden. Hier spielt es zunächst keine Rolle, ob die Lösungen im lokalen Rechenzentrum (On-Premises) oder in der Cloud zum Einsatz kommen. Auch hier sind wieder hybride Netzwerke und auch Multi-Cloud-Infrastrukturen problemlos denkbar. Ideal sind an dieser Stelle bereits vorkonfigurierte Distributionen aus den Bereichen:

- Big Data
- Stream Processing
- Business Intelligence
- Machine Learning
- Künstliche Intelligenz

Wie bereits ausgeführt, sind hier auch ausdrücklich Kombinationen dieser Technologien sinnvoll einsetzbar. Dazu trägt auch der modulare Ansatz bei, der es jederzeit ermöglicht, zusätzliche Module einzubinden oder diese zu ersetzen. Abgerundet wird eine solche Lösung als Managed Service mit dem Extra an Sicherheit durch professionelle Absicherung der Big Data unterliegenden Systeme.

4.1.3 Praxisbeispiel: MARISPACE-X auf Gaia-X

Bei der Verarbeitung von Daten für Big Data und Data Science spielt natürlich die Datensouveränität eine wesentliche Rolle. IONOS Cloud bringt gemeinsam mit seinem Partner Stackable eine Cloud-Infrastruktur sowie Services für die Zusammenführung und Analyse von Daten in das europäische Gaia-X-Projekt ein. Gaia-X ist ein Projekt, das auch von den Regierungen in Deutschland, Frankreich und weiterer EU-Staaten vorangetrieben wird. Es geht um eine von anderen Regionen unabhängige Multi-Cloud-Architektur, die umfassende Datensouveränität durch harmonisierte Datenaustauschräume für europäische Kunden bietet.

Datensouveränität durch eigene Cloud-Infrastruktur in Europa mit Gaia-X

Gaia-X will eine Dateninfrastruktur für Europa aufbauen, die unabhängig von anderen Regionen der Welt ist. Das Projekt wurde beim Digital-Gipfel 2019 durch das Bundeswirtschaftsministerium vorgestellt. Vorbild für Gaia-X ist das europäische Airbus-Konsortium. Am Projekt arbeiten unter anderem IONOS, die deutsche Telekom, Bosch, das Fraunhofer-Institut, SAP und Siemens mit. Die Software, die für Gaia-X zum Einsatz kommt, basiert zum größten Teil auf Open Source.

KI spielt für immer mehr europäische Unternehmen eine wichtige Rolle. Hier wird schnell klar, dass die europäischen Unternehmen, Behörden und Organisationen nicht unbedingt amerikanischen Konzernen heikle Daten in diesem Bereich anvertrauen wollen. Genau hier setzt Gaia-X an.

MARISPACE-X bietet einen Datenschatz, der Leben retten kann

MARISPACE-X von IONOS Cloud hat die Sammlung von maritimen Daten im Fokus. Viele europäische Unternehmen und auch Organisationen sammeln Unmengen von Geoinformationen aus den Weltmeeren und verarbeiten diese jeweils für einen speziellen Zweck. Allerdings wurden bisher die gesammelten Daten der verschiedenen Unternehmen nicht untereinander geteilt, um eine effektive Nutzung zu ermöglichen, die weit über die Verwendung eines einzelnen Einsatzgebietes hinausgeht. Drohnen, Satelliten, Sensoren und eine Vielzahl von Messinformationen laufen parallel zueinander, sammeln Daten und verarbeiten diese auch.

Würden diese Daten aber in einem sicheren Datenraum zusammenfließen, ergäbe sich eine der umfassendsten Datenquellen für die maritime Wirtschaft mit ungeheuren Datenmengen. Genau diesen Ansatz verfolgt MARISPACE-X.

Dieser Datenschatz ist für europäische Unternehmen für zahlreiche Anwendungsfälle und Business Cases ungeheuer interessant. Ein Beispiel dafür ist die Suche nach und Bergung von Munition aus den Weltmeeren. Die Munition stellt eine große Gefahr für die Schifffahrt dar, da Granaten regelmäßig explodieren. Dazu kommen austretende Giftstoffe der Munition, welche auch das maritime Leben in Gefahr bringen. Alleine in der Ostsee sind noch 1,6 Millionen Tonnen Munition versenkt, in der Nordsee sollen es weitere 1,3 Millionen Tonnen sein.

Big Data und KI helfen bei der Suche nach Munition in den Weltmeeren

Die north.io GmbH aus Kiel arbeitet an MARISPACE-X entscheidend mit und stellt eine KI-gestützte Software bereit, die historische Dokumente nach Munitionsfundstellen untersucht. Dazu kommt das Errechnen von Munitionsrisiken für verschiedene Regionen und Big-Data-Analyse zur Munitionsräumung. Solche Anwendungen benötigen riesige Datenmengen und enorme Rechenpower.

Die Herausforderungen gehen weit darüber hinaus, was Unternehmen oder Organisationen in eigenen Rechenzentren bezahlbar zur Verfügung stellen können. Das ist auch der Grund, warum north.io auf die IONOS Cloud setzt. Zur Analyse setzt das Unternehmen parallel noch auf Massendaten anderer Partnerunternehmen und deren Analyse durch weitere Partner im Verbund von MARISPACE-X. Ein Beispiel dafür ist hier Stackable und seine freie und – anders als in diesem White Paper bereits beschriebenen Softwarekomponenten – wirklich offene Distribution bestens eingeführter Open-Source-Projekte für moderne Datenplattformen.

Gemeinsame Storage-Plattform schafft neue Business-Modelle

Rainer Sträter, Head of Global Platform Hosting bei IONOS dazu: „Mit MARISPACE-X wollen alle Partner einen gemeinsamen virtuellen Topf für alle verfügbaren maritimen Daten schaffen und dadurch auch neue Business-Modelle erschließen“. Damit das möglich wird, stellt IONOS als Konsortialführer die Cloud-Infrastruktur für MARISPACE-X zur Verfügung. Die Daten werden dazu auf einer gemeinsamen Storage-Plattform zusammengeführt und analysiert. Zum Einsatz kommt dabei Open Source.

Damit die Daten effektiv analysiert werden können, ist auch entsprechende Compute-Performance Bestandteil des Projektes. In Zusammenarbeit mit dieser Rechenpower, der gemeinsamen Storage-Plattform und des riesigen Datenvolumens, können Analyse-Anwendungen exakt ermitteln, wo Gefahrenstoffe lagern und welcher Aufwand für die Bergung notwendig ist.

Klimaschutz, Offshore Windenergie und Internet of Underwater Things mit Big Data

Weitere Einsatzgebiete von MARISPACE-X in diesem Bereich sind das Kalkulieren von Standorten für Windparks oder die optimale Verlegung von Seekabeln. Die Nutzung geht daher weit über das Finden von Munition heraus. Bereits seit Mitte 2021 stehen auch Offshore Windenergie, Internet of Underwater Things (IoUT) und biologischer Klimaschutz im Fokus des Projektes.

Satellitenbilder und Unterwassersensoren helfen bei der Suche nach Algenflächen und Seegraswiesen, die bei der Speicherung von Kohlendioxid eine wesentliche Rolle spielen. MARISPACE-X hilft bei der Analyse bestehender Flächen und kann gleichzeitig neue Anbauflächen durch die Analyse finden. Dabei sind auch Bodenbeschaffenheit und Meeresströmungen Bestandteil der umfassenden Berechnungen.

Gemeinsamer Datenaustausch mit International Data Spaces

Bei MARISPACE-X steht der gemeinsame Datenaustausch zwischen den beteiligten Partnern im Fokus. Dazu stehen auch Konnektoren zur Verfügung, die Datenquellen und die Verarbeitungssysteme miteinander verbinden. Der hier verwendete Standard mit der Bezeichnung International Data Space Association (IDSA) ist schnell und sicher. Der Standard unterliegt den europäischen Richtlinien und Gesetzen zum Datenschutz und Datensicherheit. Solche Richtlinien werden nur von europäischen Unternehmen umfassend eingehalten, da auch die gesammelten Daten nur in Rechenzentren der europäischen Union gespeichert werden.

IONOS Cloud stellt die dazugehörige Rechenpower und den Storage zur Verfügung. Gleichzeitig sorgt IONOS Cloud für die notwendige Kompression der Daten. Zum Konsortium gehören neben north.io und IONOS auch zahlreiche andere europäische Unternehmen, die Daten zur Verfügung stellen und gemeinsame Berechnungen und Analysen mit MARISPACE-X durchführen. Mitglieder sind unter anderem die Unternehmen TrueOcean GmbH, Stackable GmbH, MacArtney Germany GmbH, Siemens Gamesa Renewable Energy, Quality Positioning Services B.V., WINDEA, Offshore GmbH & Co. KG, OffCon24 und Wallaby Boats.

Auch wissenschaftliche Einrichtungen, öffentliche Institutionen und Verbände arbeiten an dem Projekt mit. Dazu gehören das Fraunhofer-Institut für Graphische Datenverarbeitung IGD, GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel – AG Deep Sea Monitoring, Universität Rostock, Christian-Albrechts-Universität zu Kiel, das Maritime Cluster Norddeutschland, die Gesellschaft für Maritime Technik e.V., TransMarTech Schleswig-Holstein, IHK Schleswig-Holstein, Labs Network Industrie 4.0 e.V. sowie das Ministerium für Energie, Landwirtschaft, Umwelt, Natur und Digitalisierung Schleswig-Holstein und die Ocean Data Alliance.

4.2 Big Data Frameworks und -Plattformen als Managed Service in der europäischen Cloud

Die zuvor erwähnten Möglichkeiten sind nur einige Beispiele für die Funktionen, die Big Data in Zusammenarbeit mit KI bietet. Für die Analyse kommen zum großen Teil Open-Source-Lösungen zum Einsatz, die in einer Managed Cloud zusammengeführt, verwaltet, aktualisiert und gewartet werden.

Apache Kafka – Managed Big Data aus der europäischen Cloud

Apache Kafka gehört zu den bekanntesten Open-Source-Lösungen im Bereich Big Data. Die ursprünglich von LinkedIn entwickelte Lösung ist in der Lage, große Datenmengen aus verschiedenen Quellen zu importieren und weiterzuverarbeiten. Apache Kafka besticht vor allem durch den sehr hohen Datendurchsatz und kann Realtime Events, Logs und andere Daten nutzen. Dabei ist die Software stark im Event Streaming. Diese Events lassen sich für die Weiterverarbeitung in Dateien ablegen. Hadoop dagegen ist stark im Verarbeiten von riesigen Datenmengen in Dateien, die wiederum von Kafka erstellt werden können.

Kafka verbindet Arbeitsspeicher, Cache von Speichersystemen und die Speicherverwaltung des lokalen Betriebssystems miteinander. Das ermöglicht effiziente Verteilung der Rechen- und Speicheraufgaben. Wenn die zu Grunde liegende Infrastruktur diese Daten schnell zur Verfügung stellen kann, wie zum Beispiel in der IONOS Cloud, kann Kafka wahre Wunder vollbringen.

Damit die Komponenten miteinander interagieren können und dabei leistungsstark zur Verfügung stehen, wird wiederum auf Apache Zookeeper gesetzt. Dabei handelt es sich ebenfalls um Open Source. Zookeeper ist ein zentraler Dienst für die Pflege von Konfigurationsinformationen, die Benennung von Objekten und die verteilte Synchronisation von Gruppendiensten. Die Lösung soll vor allem Wildwuchs in der Infrastruktur verhindern und dabei helfen, eine einheitliche Konfiguration zu erreichen. Das ist ideal für eine stabile und effektive Cloud-Infrastruktur.

Kafka kann auch dazu verwendet werden, große Datenmengen zu verarbeiten und direkt in das Hadoop-System zu senden. Apache Kafka arbeitet auch mit Apache Storm zusammen. Apache Storm kann Daten aus anderen Streams auslesen, bevor diese persistent gespeichert und abgeschlossen werden, auch aus Apache Kafka.

Big Data in neuen Dimensionen mit Apache Spark und NiFi

In großen Hadoop- oder Big-Data-Umgebungen reichen die Standardmöglichkeiten und Abfragen häufig nicht aus, um effizient Daten analysieren zu können. Das Apache-Projekt Spark hat sich diesem Problem angenommen und bietet eine effiziente Echtzeitanalyse von Daten in Hadoop-Clustern.

Das Apache-Projekt Spark bietet eine effiziente Echtzeitanalyse von Daten in Hadoop-Clustern.

Im Daytona Gray Sort Benchmark siegte Spark in der 100-TByte-Klasse mit einem neuen Weltrekord. Der alte Weltrekord lag bei 72 Minuten und wurde von einem Hadoop-MapReduce-Cluster aufgestellt. Spark hat den alten Rekord mit 23 Minuten geschlagen und das mit einem Zehntel der Rechenkraft. Es wird schnell klar, dass Spark in Bereiche der Big-Data-Verarbeitung vordringen kann, die für Hadoop nicht möglich sind.

Apache Spark erweitert die Möglichkeit von Hadoop-Clustern um Echtzeitabfragen. Dazu bietet das Framework In-Memory-Technologien, kann also Abfragen und Daten direkt im Arbeitsspeicher der Clusterknoten speichern. Da die Abfragen sich auch parallel auf mehrere Knoten verteilen lassen, steigt die Leistung enorm. Für den Einsatz ist ebenfalls leistungsstarke Hardware notwendig, die über die IONOS Cloud bei Bedarf zur Verfügung gestellt werden kann.

Apache Spark soll MapReduce in Hadoop ablösen und bietet eine extrem schnellere Abfragegeschwindigkeit von Daten. Die Entwickler selbst sprechen von einer hundertfachen Geschwindigkeit. Das Framework wird bereits von großen Unternehmen eingesetzt, die eine große Datenmenge verarbeiten müssen. Prominente Beispiele sind NASA, Intel und IBM. Der Online-Musikdienst Spotify optimiert seine Wiedergabelisten ebenfalls mit Spark und auch in der IONOS Cloud, zum Beispiel beim Projekt MARISPACE-X, kommt Spark zum Einsatz.

In diesem Zusammenhang spielt auch Apache NiFi eine wesentliche Rolle. Dabei handelt es sich um eine weitere Open-Source-Lösung auf Basis der Apache-Lizenz. NiFi hat die Aufgabe, den Datenfluss zwischen Systemen zu automatisieren. Im Zusammenspiel mit Spark, NiFi und Kafka lassen sich ungeheure Datenmengen verarbeiten, transformieren und auch analysieren. Geschieht das auf einer gemeinsamen Plattform, welche für die Zusammenarbeit optimiert ist, und die über entsprechende Rechen- und Storagepower verfügt, ergeben sich ungeheure Vorteile bei der Geschwindigkeit der Analysen.

Apache Druid: Datenspeicher mit geringen Latenzen

Apache Druid ist ein Open-Source-Analysespeicher, der Business-Intelligence-Abfragen von Ereignisdaten mit geringer Latenz ermöglicht. Echtzeitzugriffe sind genauso möglich wie eine schnelle Datenaggreation. Die Open-Source-Lösung kann als Alternative zu herkömmlichen Data Warehouses zum Einsatz kommen. Druid hat seinen Schwerpunkt im Bereich von Ereignisdaten und Zeitreihen, kann aber auch andere Daten analysieren.

Wenn eine hohe Leistung bei der Analyse notwendig ist, kann der Betrieb von Druid sinnvoll sein.

In vielen Umgebungen wird Druid auch parallel zu Data Warehouses eingesetzt. Wo immer Echtzeit-Analyse, interaktive Oberflächen mit aktualisierenden Daten oder hochmoderne Abfrage-Apps im Einsatz ist, kann es sinnvoll sein, diese Daten aus Apache Druid zu beziehen. In einem solchen Szenario kann Druid wiederum seine Daten aus dem Data Warehouse erhalten, aufbereiten und zur Verfügung stellen. Das Data Warehouse kann dann parallel Berichte und Archivdaten liefern.

Druid unterstützt zahlreiche Datenquellen, auch in der Cloud. Beispiele dafür sind Amazon S3, Apache Kafka, Azure Event Hub, Amazon Kinesis, Google Cloud Storage, Azure Data Lake, lokale Daten und natürlich auch Daten aus Datenbanken sowie lokal gespeicherte Daten. Nach der Anbindung der Quelle liest Apache Druid diese ein und indexiert die Daten. Auch Hadoop wird von Druid unterstützt.

Apache Druid kann mit Apache Hive und Apache Ambari integriert werden. Auch die Erstellung von OLAP-Cubes mit SQL oder das Abrufen bereits vorhandener Druid-Cubes ist im Rahmen der Echtzeitanalyse möglich.

Bei optimaler Konfiguration kann Open Source bei der Analyse für wahre Wunder sorgen

Die Beispiele in diesem Abschnitt des White Papers zeigen, dass es zahlreiche Open-Source-Lösungen gibt, die Unmengen an Daten verarbeiten und analysieren können. Arbeiten Tools wie Kafka, Spark, NiFi, Druid und Hadoop zusammen, können Unternehmen, Organisationen und auch Behörden nahezu jede Datenmenge schnell und sicher analysieren. Allerdings gehören dazu enorme Rechenpower, Storage-Kapazitäten und ungeheures Fachwissen. Stackable, ein Start-up aus Deutschland mit einer Data Platform, die nur Software mit nicht restriktiver Open-Source-Lizenzierung einsetzt, bringt Licht in den Lizenzdschungel und ermöglicht vertrauensvolles Big Data in einem durch und durch transparenten Umfeld.

Um die Lösung zusammenarbeiten zu lassen, und erst dann ergibt sich ein optimaler Nutzen, müssen Hard- und Software zusammenspielen und optimal konfiguriert sein. Dazu kommen regelmäßige Aktualisierungen und Wartungen. Ein solches Umfeld lässt sich im Grunde genommen nur in der Cloud schaffen, idealerweise durch einen Partner der zuverlässig ist und auf dem Boden der europäischen Gesetzgebung steht.

Beim Betrieb von Big Data-Analysen in europäischen Clouds, wie zum Beispiel IONOS Cloud, wird großes Augenmerk darauf gelegt, dass die Datensouveränität gewahrt bleibt. Gleichzeitig sorgt der Anbieter dafür, dass die Lösungen immer optimal zusammenarbeiten, aktualisiert sind, und die verwendete Hardware optimal mit der Software zusammenarbeitet. Durch die modulare Struktur ist darüber hinaus auch sichergestellt, dass sich einzelne Komponenten austauschen lassen, während die Gesamtheit der Lösung bewahrt bleibt.

4.3 IONOS und Stackable Cloud Data Warehouse

Das Stackable Cloud Data Warehouse stellt eine moderne BI-Lösung zur Verfügung, das aus leistungsstarken Datenverarbeitungs- und Verteilungspipelines besteht. Auch die Datenspeicherung ist stark skalierbar. Zum Einsatz kommt dazu S3 Object Storage.

Das Rechenzentrum von IONOS Cloud befindet sich in Europa und unterliegt daher der DSGVO.

Durch schnelleund übergreifende SQL-Abfragen und OLAP-basierte Business-Analysen lassen sich Charts und Dashboards mit relevanten Daten füllen, die auf Basis verschiedener Open-Source-Lösungen erstellt werden. Dazu kommen vor allem HBase und die Hadoop-Plattform zum Einsatz, zusammen mit Kafka, Spark, Superset, Atlas und vielen anderen Open-Source-Tools, die Stackable zu einem Ökosystem für Big Data Managed Services verbindet. Die Entwickler erweitern, aktualisieren und verbessern die Open-Source-Tools in Zusammenarbeit mit der Open-Source-Community dauerhaft.

Das Rechenzentrum für die cloudbasierte Lösung befindet sich in Europa und unterliegt daher der DSGVO. Wichtig ist hierbei auch, dass auch der Provider der Cloud dahinter ein europäisches Unternehmen ist und somit maximal mehr Sicherheit vor dem US CLOUD Act bietet. Dazu kommt eine Zertifizierung nach ISO 27001. Der Support für die Lösung kommt aus Deutschland, genauso wie das Consulting zum Aufbau eines Cloud Data Warehouses. Daher profitieren Kunden von kurzen Wegen, wenn es darum geht, schnell Unterstützung bei Aufbau, Betrieb oder Ausbau zu erhalten.

Mit der intuitiven Benutzeroberfläche Data Center Designer (DCD) erstellen Unternehmen und Organisationen mit Leichtigkeit ihr virtuelles Rechenzentrum. Basis der Stackable Big-Data-Lösungen ist die Container-Orchestrierung Kubernetes. Dazu stellt Stackable auch vorgefertigte Konfigurationen zur Verfügung. Die Einrichtung ist wesentlich einfacher als vergleichbare Lösungen von US-amerikanischen Anbietern; ferner grenzt sich das Cloud Data Warehouse durch besonders qualifizierten Support in lokaler Sprache und ohne Aufpreis von ähnlichen Angeboten ab.

Durch die Verwendung von Vanilla Kubernetes profitieren Kunden gleichzeitig von maximaler Flexibilität, da diese ermöglicht, die ganze Umgebung auch in ein anderes Rechenzentrum umziehen zu können. Kubernetes hat sich zu einem marktgängigen Standard entwickelt. Beim Einsatz von Stackable Cloud Data Warehouse besteht daher zu keiner Zeit die Gefahr eines Vendor Lock-ins. Die Lösung ist stark modular aufgebaut und setzt als Basis auf bekannte und beliebte Open-Source-Tools. Ein Wechsel der Cloud-Plattform ist jederzeit möglich, genauso wie ein Umzug in das eigene Rechenzentrum. Stackable Cloud Data Warehouse bietet eine hohe Transparenz und steht als Open-Source-Lösung zur Verfügung. Dabei setzt das Unternehmen auf die originale Form der eingesetzten Tools, deren Quellcode jederzeit einseh- und damit nachvollziehbar ist.

Unternehmen und Organisationen haben dadurch die Möglichkeit, ein riesiges Cloud Data Warehouse aufzubauen, ohne dabei internes Wissen für den Aufbau eines Clusters zur Verfügung stellen zu müssen. Es gibt derzeit und auch in absehbarer Zukunft wenige Experten, die in der Lage sind, ein solches umfassendes Ökosystem aufzubauen. Hier profitieren Unternehmen und Organisationen vom Know-how der Experten bei Stackable. Stackable bündelt seine Kompetenz mit denen von IONOS Cloud und b.intelligent, einem ausgewiesenen Spezialisten in Consulting-Fragen zu Big Data. Wer im Unternehmen über genügend Wissen verfügt, kann aber an das Stackable Cloud Data Warehouse über Software Container jederzeit eigene Dienste anbinden. Schließlich ist das System vollständig transparent dank Open-Source-Komponenten.

Da das System Kubernetes nutzt, ist die Integration weiterer Container und Tools jederzeit möglich. Dabei können auch hybride Clouds und Multi-Cloud-Umgebungen zum Einsatz kommen. Wer auf gängige Cloud-Umgebungen setzt, kann auch Authentifizierungen auf Basis von Azure Active Directory und Active Directory umsetzen, aber auch zahlreiche andere Systeme nutzen. Das System basiert auch hier auf Open Source, sodass eine flexible Erweiterung schnell möglich ist. Dabei unterstützen auch die Experten von Stackable bei der Umsetzung. Zudem stellt Stackable sicher, dass durch Commercial Forking keine Lizenzprobleme beim Einsatz der Open-Source-Softwarekomponenten auftreten.

5 Fazit

In einer Managed Cloud sind alle diese Herausforderungen schlicht und ergreifend der einzige Weg, um den Herangehensweisen von umfassenden Datenanalysen in den nächsten Jahren gerecht zu werden. Davon können Unternehmen, Organisationen und Behörden aller Größenordnungen profitieren, da sie keine eigene Hard- und Software bereitstellen und keine Cluster betreiben müssen. Durch das Buchen der richtigen Cloud-Infrastruktur aus europäischen Rechenzentren bleiben gleichzeitig Datenschutz, Datensicherheit und auch Datensouveränität erhalten.

Managed Services stellen zudem sicher, dass die Dienste und damit auch die Daten maximal sicher gespeichert und verarbeitet werden. Außerdem ist die Einrichtung sehr viel einfacher als beim Betrieb eines eigenen Clusters im lokalen Rechenzentrum. Und schlussendlich sind solche Lösungen leichter zu bedienen, einfacher zu steuern, sehr viel besser zu skalieren und auch sehr viel günstiger als das Einsetzen von Software im eigenen Rechenzentrum.

Darüber hinaus erleichtern Managed Services insbesondere bei noch neuen Technologien wie Big Data und KI den Einstieg für den Nutzer, da viel weniger IT-Administrationsaufwand anfällt. Werden wirklich quelloffene Open-Source-Software-Module in einer Distribution gebündelt und transparent bereitgestellt, profitiert der Anwender deutlich mehr von den Agilitäts- und Sicherheitsvorteilen Community-betriebener Software.

Über IONOS

IONOS ist mit mehr als acht Millionen Kundenverträgen der führende europäische Anbieter von Cloud-Infrastruktur, Cloud-Services und Hosting-Dienstleistungen. Das Produktportfolio bietet alles, was Unternehmen benötigen, um in der Cloud erfolgreich zu sein: von Domains über klassische Websites und Do-It-Yourself-Lösungen, Online-Marketing-Tools bis hin zu vollwertigen Servern und einer IaaS-Lösung. Das Angebot richtet sich an Freiberufler, Gewerbetreibende und Konsumenten sowie an Unternehmenskunden mit komplexen IT-Anforderungen.

IONOS Cloud ist die europäische Cloud-Alternative und Teil von IONOS. Unser Produktportfolio umfasst mit der Cloud Compute Engine eine IaaS Compute Engine mit eigenem Code Stack für Virtualisierung, Managed Kubernetes für Container-Anwendungen, eine Private Cloud powered by VMware sowie S3 Object Storage. Mit unserem Angebot bieten wir etablierten mittelständischen und großen Unternehmen, regulierten Industrien, der Digitalwirtschaft und dem öffentlichen Sektor alle notwendigen Dienste und Services um in und mit der Cloud erfolgreich zu sein.

IONOS entstand 2018 aus dem Zusammenschluss von 1&1 Internet und dem Berliner IaaS-Anbieter ProfitBricks. IONOS ist Teil der börsennotierten United Internet AG (ISIN DE0005089031). Zur IONOS Markenfamilie gehören STRATO, Arsys, Fasthosts, home.pl, InterNetX, SEDO, United Domains und World4You.

Impressum

IONOS SE
Berlin Office
Revaler Straße 30
10245 Berlin, Germany

IONOS Cloud Kontakt

Telefon +49 30 57700 840
Telefax +49 30 57700 8598
E-Mail produkt@cloud.ionos.de
Website https://cloud.ionos.de

Vorstand

Hüseyin Dogan, Dr. Martin Endreß, Claudia Frese, Hans-Henning Kettler, Arthur Mai, Britta Schmitt, Achim Weiß

Aufsichtsratsvorsitzender

Markus Kadelke

Handelsregister

IONOS SE: Amtsgericht Montabaur / HRB 24498

Umsatzsteuer-IdentNummer

IONOS SE: DE815563912

Copyright

IONOS