BI Spektrum
tdwi WISSEN
ONLINETHEMENSPECIAL

Advertorial

Big Data in der Cloud versus On-Premises

von Dimitri Gross

Fast zeitgleich wurden Cloud und Big Data in der IT-Welt als neue Technologiestränge aufgenommen. Mit der voranschreitenden Entwicklung dieser beiden Bereiche zeichnen sich aktuell viele Synergien ab. So sprechen wir nicht mehr von Cloud Storage oder SaaS, sondern von Big Data as a Service bzw. aktuell auch von Machine Learning as a Service. Diese Trends bringen viele Vorteile, sind aber auch mit einigen Bedenken bei den Anwendern verbunden. Welche Herausforderungen und Chancen bringt Big Data in der Cloud mit sich? Welche Vor- und Nachteile hat eine Big-Data-Cloud-Plattform? Und ab wann lohnt sich eine On-Premises-Lösung mehr?

Die Cloud als Speicherort

Wer sich mit Big Data beschäftigt, weiß, wie wichtig hier eine effiziente Datenspeicherung ist. Da die Daten in Dateien abgelegt werden, benötigt man für ihre Speicherung ein schnelles und verteiltes Dateisystem. Dieses Dateisystem muss drei zentrale Anforderungen erfüllen:

1. **Unterstützung hoher Datenvolumen:** Die Daten liegen in extrem großen Mengen vor, weil innerhalb eines Prozesses Milliarden von Datensätzen generiert werden können. Diese Mengen zu bewältigen ist eine Herausforderung, mit der heute viele Firmen zu kämpfen haben. Schaut man sich nur die geocodierten GPS-Daten eines Trucks an, wird deutlich,

wie solche Mengen entstehen. Da geht es schnell um einige Terabyte oder Petabyte an Daten. Das Dateisystem muss dafür sehr viele Festplatten unterstützen.
2. **Organisation als verteiltes Dateisystem:** Wenn ein Master mit vielen Workern arbeitet, müssen die Daten jederzeit auch von unterschiedlichen Rechnern aus zugreifbar sein.
3. **Zusammenarbeit mit Big-Data-Software:** Das Dateisystem muss mit der verwendeten Big-Data-Software harmonieren.

Für die On-Premises-Welt scheinen diese Anforderungen kein Problem zu sein. So verwenden Big-Data-Spezialisten zum Beispiel für Apache Hadoop ein Platten-Array, auch JBOD („Just a Bunch Of Disks“) genannt, das mit einem Hadoop Distributed File System (HDFS) ausgestattet ist. Das HDFS erfüllt damit die oben genannten Anforderungen und bildet zeitgleich eine Replikation auf Softwareebene ab, sodass ein physikalisches RAID-System entfällt.

Aber wie sieht es in der Cloud aus? Da die Cloud nicht die Möglichkeit bietet, ein Platten-Array hinzuzufügen, stellen Cloud-Anbieter eigene Storage-Lösungen zur Verfügung. Die Lösungen sind vielfältig:

Lokale Festplatte pro virtueller Rechnerinstanz: Bei manchen Anbietern kann man sich zwischen verschiedenen Speicherplatten entscheiden, beispielsweise zwischen magnetischen Festplatten (HDD) oder Solid State Disks (SSD). Diese werden dann an eine Rechnerinstanz als Platte angehängt. Dieses Vorgehen ist für lokale Daten sinnvoll, wie eine Betriebssysteminstallation, bei der kein verteiltes Dateisystem vorliegt und die deswegen die zentralen Anforderungen an ein Dateisystem (siehe oben) nicht erfüllt. Unabhängig davon ist es natürlich möglich, mehrere Platten lokal anzubinden und mit diesen ein HDFS anzulegen.

Netzwerk-Storage: Der Netzwerk-Storage ist an mehrere virtuelle Rechnerinstanzen angehängt. Er ist vergleichbar mit einem Network Attached Storage (NAS). Netzwerk-Storages sind im Vergleich zu anderen Storage-Varianten relativ teuer. Deswegen spielt dieser Storage im Big-Data-Umfeld auch keine große Rolle.

Objekt-Storage: Der Objekt-Storage ist eine preisgünstige Speicher-Variante, die sich großer Beliebtheit erfreut. Er hat typischerweise folgende Eigenschaften:

Der Zugriff auf den Objekt-Storage erfolgt in der Regel nur über eine API. Es gibt keine Möglichkeit, den Objekt-Storage mit einem frei gewählten Dateisystem zu formatieren, sondern Objekte werden über die API gespeichert und gelesen.
Im Objekt-Storage können Objekte, also Binärdaten und somit auch Dateien, abgelegt werden.
Der Objekt-Storage ist nicht an eine Rechnerinstanz gebunden, sondern existiert unabhängig. Dadurch können viele Worker auf den gleichen Objekt-Storage zugreifen.
Der Objekt-Storage ist in der Regel nicht durch eine Maximalgröße beschränkt. Das heißt, es können beliebig viele Daten abgelegt werden. Die Größe einzelner Objekte ist in der Regel allerdings schon beschränkt, oftmals auf fünf Terabyte pro Objekt. Die Anzahl der Objekte hingegen ist unbegrenzt.

Ein Objekt-Storage ist als Datenspeicher also prinzipiell für Big Data geeignet. Problematisch ist allerdings, dass er nicht mit einem entsprechenden Dateisystem versehen werden kann, weil der Zugriff nur über die API des Objekt-Storage erfolgt. Anders als im Hadoop Cluster liegen die Daten nicht auf den Worker-Nodes, sondern müssen beim Zugriff erst zu den Workern transferiert werden. Das kostet Zeit. Eine Mischung mit lokalem HDFS-Speicher für die Speicherung von Zwischenergebnissen bei ETL-Pipelines und S3 für Ausgangsdaten und Endergebnisse kann hier etwas Linderung verschaffen. Mit der Implementierung einer Big-Data-Software könnte dieses Problem gelöst werden: Wird die Software mit entsprechenden Adaptoren ausgestattet, können die gängigen Objekt-Storages der Cloud-Anbieter direkt angesprochen werden. So unterstützt zum Beispiel Apache Hadoop die Objekt-Storages von Amazon S3, Windows Azure Blob Storage, OpenStackSwift und anderen.

Die genannten Hadoop-Distributionen unterscheiden nicht zwischen Cloud- und On-Premises-Lösungen. Die auf dem Markt etablierten Distributionen lassen sich in der Cloud betreiben. Cloud-Provider bieten dafür bereits vorkonfigurierte Images an, die einfache Schnelltests ermöglichen. Möchte man jedoch einen eigenen Hadoop-Cluster von Grund auf selbst in der Cloud installieren, wird ein ähnliches Skillset benötigt wie bei einer On-Premises-Installation.

Daten in die Cloud – mehrere Möglichkeiten

Eine Frage, die immer wieder gestellt wird, ist: „Wie bekomme ich meine Daten in die Cloud?“ Klar ist: Um viele Daten bearbeiten zu können, müssen diese erst einmal zur Verfügung stehen. Zwei Szenarien lassen sich hier unterscheiden:

1. Die Daten werden zyklisch erfasst und über einen Messaging-Mechanismus in die Cloud übertragen.
2. Die Daten liegen in großer Menge on-premises vor.

Im ersten Fall liegt die Lösung auf der Hand: Die zyklisch anfallenden Daten werden über eine geeignete Verbindung (HTTPS, VPN etc.) zu einem Service in der Cloud übertragen, der die Daten in der Cloud speichert. Dies setzt voraus, dass in der Cloud ein entsprechender Service verfügbar ist. Hier kann durch geeignete Konfiguration zum Beispiel ein Kafka Message Broker zum Einsatz kommen. Je nachdem, wie viele Daten anfallen, ist dieser Ansatz über HTTPS oder über einen VPN-Tunnel möglich. Die Übertragungsdauer hängt von der verfügbaren Internetverbindung ab. Wenn der Anwender zum Beispiel eine synchrone 10-Mbit/s-Internetanbindung besitzt, werden für 10 Terabyte Daten 97 Tage benötigt – vorausgesetzt, die Leitungen sind exklusiv und erreichen tatsächlich die vereinbarten 10 Mbit/s. In der Realität wird diese Rate leider oft nicht erreicht. Dieser Ansatz kann daher nur funktionieren, wenn kleinere Datenmengen kontinuierlich über einen längeren Zeitraum gesammelt werden.

Steht hingegen schon eine große Datenmenge zur Verfügung und soll diese schnellstmöglich in die Cloud gebracht werden, ist ein Import-Service erforderlich. Dieser Service ermöglicht es dem Anwender, beispielsweise eine oder mehrere Festplatten mit den Daten per Post oder Kurier einzuschicken. Der Cloud-Anbieter baut die Festplatten dann in einem „Übergabebereich“ ein und überspielt die Daten in den privaten Cloud-Bereich des Kunden. Vom Übergabebereich zum privaten Bereich des Kunden geht der Weg nur über das interne Netzwerk des Cloud-Anbieters. Hier kann mit einer Übertragungsrate von einem Gigabit pro Sekunde gerechnet werden. Das bedeutet eine Erhöhung um den Faktor 100 gegenüber obiger Rechnung.

Der Vorteil eines Import-Service liegt auf der Hand: Mehrere Terabyte können ohne Probleme übertragen werden. Das Format und das genaue Vorgehen hängen vom jeweiligen Cloud-Anbieter ab, der einen solchen Service anbietet. Daran schließt sich sofort die Frage an, wie die Daten während des Transports geschützt werden können. Eine Verschlüsselung der Festplatte ist dringend anzuraten! Hierzu muss der Cloud-Anbieter ein Verfahren mit privaten oder öffentlichen Schlüsseln unterstützen: Den privaten Schlüssel besitzt nur der Cloud-Anbieter, den öffentlichen Schlüssel kann der Kunde vom Cloud-Anbieter beziehen. Welches Verfahren ein Cloud-Anbieter unterstützt, muss jeweils angefragt werden. Fakt ist, dass die Übertragung per Festplatte nicht allzu standardisiert ist. So ist es möglich, eigene, auch sehr unterschiedliche Festplatten mit diversen Formatierungen zu verwenden.

Da es für Cloud-Anbieter aufwendig ist, die Datenübertragung mit einer großen Palette zu unterstützen, haben einige Anbieter eine standardisierte Lösung entwickelt, wie beispielsweise Amazon mit Snowball: Auf Anfrage erhält der Anwender eine hochsichere Datenbox, auf die er die Daten übertragen kann. Diese Datenbox ist verschlüsselt und schützt die Daten über mehrfache Redundanzen in ihrem Inneren. Im Lieferumfang der Datenbox befindet sich ein Software Client, mit dessen Hilfe die Daten verschlüsselt auf der Box abgelegt werden können. Die Box selbst ist über das Netzwerk angebunden. Die Kapazität einer solchen Datenbox bewegt sich im Bereich von 100 Terabyte. Wem dies zu wenig erscheint, kann statt einer Snowball ein Snowmobile bestellen, dieser Truck bietet deutlich mehr Speicher. Dies ist aktuell die schnellste und sicherste Methode, um große Datenmengen in die AWS-Cloud zu transferieren.

Big-Data-Vorhaben

Im Vorfeld von Big-Data-Projekten zeichnet sich häufig ein Szenario wie dieses ab: Ein Unternehmen möchte Ideen und Hypothesen zur Einbindung bis dato ungenutzter Datenquellen überprüfen. Dabei stellt sich schnell heraus, dass viele Vorhaben nicht mit Hilfe von relationalen Datenbanken umgesetzt werden können.

In dem Zusammenhang ist aktuell häufig von einem „Analytics Lab“ die Rede und dass bereits mit relativ geringen Mitteln ein schnelles Cluster auf Basis einer freien Hadoop-Distribution aufgebaut werden kann. Doch die Frage, wie groß der Datendurchsatz in den nächsten drei Jahren sein wird, beziehungsweise welche Use-Cases perspektivisch verprobt werden sollen, können die meisten Unternehmen nicht genau beantworten. Und damit ist es ihnen unter Umständen auch nicht möglich, bei der

Hardwarekonfiguration das passende Sizing zu wählen.

Je nach Use Case und zu erwartender Auslastung muss der Cluster entweder große Rechenlast bewältigen können oder viele IOPs garantieren. Beides richtig zu portionieren, ohne ein klares Bild von den zukünftigen Szenarien zu haben, ist nicht möglich. So wählen viele Unternehmen ein Mittelmaß und erhalten damit meist eine eher suboptimale Konfiguration, mit der Tendenz zur Überprovisionierung – also einer Bereitstellung von zu viel Kapazität, die sie entsprechend viel kostet.

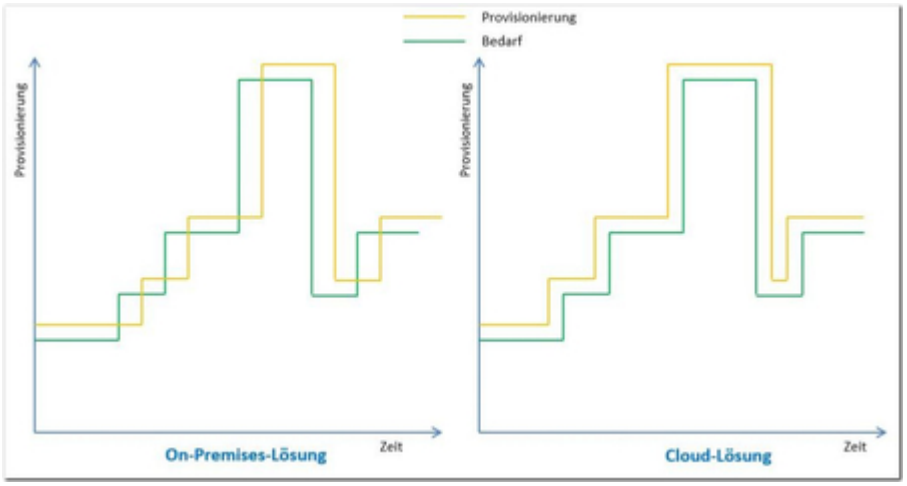

Abb. 1: Provisionierung/Bedarf im zeitlichen Verlauf

Das Problem der richtigen Provisionierung ist nicht neu und begleitet die Infrastrukturseite der IT bereits sehr lange. Eine pragmatische Lösung wäre, zusätzliche Server zu nutzen. Doch diese wären natürlich nicht sofort verfügbar, sondern müssten noch installiert und angebunden werden. Es verginge also wertvolle Zeit. Besser wäre es, in so einem Fall zu überprüfen, wann sich der Kauf einer On-Premises-Plattform tatsächlich lohnt und ob nicht zu Beginn, wenn die Anforderungen noch zu definieren sind, der Betrieb in der Cloud die bessere Wahl wäre, zumal bei einer On-Premises-Installation.

Rechenzentrum in der Cloud oder on-premises?

Wie die bisherigen Ausführungen zeigen, gibt es gute Gründe dafür, Big Data on-premises zu betreiben, aber auch starke Gründe dafür, in die Cloud zu gehen. Entscheidende Faktoren sind hierbei Kosten und Zeit. Im nachfolgenden Beispiel betrachten wir den Kostenfaktor etwas genauer. Ab wann ist es günstiger, in die Cloud zu gehen? Und ab wann ist eine Cloud-Lösung unter Umständen sogar teurer als eine On-Premises-Lösung?

Die Entscheidung hierfür kann nicht allgemeingültig getroffen werden, sondern ist von den zu betrachtenden Use-Cases abhängig. Zu berücksichtigende Parameter sind die Auslastung des Systems („Läuft die Berechnung ständig oder nur ab und zu?“) und die Kapazität des Systems („Habe ich immer die gleichen Anforderungen an die Rechenkapazität, also an CPU und RAM?“, „Oder sind diese Anforderungen von Use-Case zu Use-Case unterschiedlich?“)

Warum ist der genaue Blick auf diese Parameter für einen Vergleich so relevant? Die Antwort ist offensichtlich: Bei einer On-Premises-Lösung fallen Investitions- und Betriebskosten für die Anschaffung der Hardware, für Strom, Kühlung, Infrastruktur, Betriebsmannschaft und so weiter an. Bei einer Cloud-Lösung hingegen werden nur Betriebskosten fällig, die nach Nutzung abgerechnet werden.

Wenn wir das Thema ganz einfach angehen, würden wir die Kosten für eine On-Premises-Hardware-Beschaffung inklusive Gemeinkostenumlagen für Rechenzentrum und entsprechendes Personal mit denen für eine identische Ausstattung in der Cloud vergleichen, um festzustellen, dass diese in der Cloud höher ausfallen. Das liegt daran, dass wir die in der Cloud typische Abrechnung nach Nutzung nicht berücksichtigt haben: In den seltensten Fällen nutzen wir ein System 24 Stunden pro Tag und 365 Tage im Jahr. Typischer ist, dass wir eine Berechnung durchführen, bei der die Systeme zum Beispiel fünf Tage mit Volllast laufen, an anderen Tagen jedoch unter Umständen mit Leerlauf oder mit geringer Last.

Bei einer On-Premises-Installation ist normalerweise nicht ausschlaggebend, ob diese bezahlt und verfügbar ist. Die Kosten sind unabhängig von der Nutzung entstanden. In der Cloud sieht das anders aus: Wenn die Nutzung unterbleibt, fallen allenfalls sehr geringe Kosten für die Bereithaltung der Ressourcen an. Im Übrigen entsteht Aufwand bei einem reinen Pay-per-use-Modell nur für die Minuten und Stunden, die wirklich verwendet werden. Die Auslastung einer Anlage und damit die zu erwartende Nutzung ist also ein Faktor, der die Kosten mitbestimmt.

Ein weiterer Faktor ist die Kapazität, die eine Anlage haben muss. Wenn ein Anwenderunternehmen Hardware beschafft, muss dies im Normalfall so erfolgen, dass die Anzahl von CPUs und RAM für alle Use-Cases ausreichend sind. Deshalb wird typischerweise eine Anlage beschafft, die stärker ist, als der aktuelle Bedarf dies vorgibt, damit noch genügend Reserve vorhanden ist. Diese Reserve stellt häufig gebundenes Kapital dar und belastet damit die Gesamtinvestition. In der Cloud kann die Kapazität hingegen dynamisch angepasst werden: Wenn mehr oder weniger Rechenleistung benötigt wird, kann horizontal skaliert werden, indem neue Worker ergänzt bzw. abgeschaltet werden. Oder die Skalierung erfolgt vertikal, indem die Anzahl von CPUs und RAM pro Worker erhöht bzw. verringert wird, und das schlägt sich auf den Preis nieder. Je näher die gewählte Kapazität an der tatsächlich benötigten Kapazität liegt, desto günstiger kann die jeweilige Leistung erworben werden.

Deutlich wird der Unterschied im konkreten Anwendungsfall. Nehmen wir folgendes Szenario an: Wir benötigen eine schnell skalierende analytische Plattform auf Basis von Hadoop, die wir in der Cloud betreiben möchten. Beim Cloud-Anbieter AWS hätten wir beispielsweise zwei Alternativen:

1. Betrieb einer etablierten Hadoop-Distribution und Provisionierung passender Instanzen
2. Betrieb eines Hadoop-Clusters auf Basis von Amazon Elastic Map Reduce (EMR)

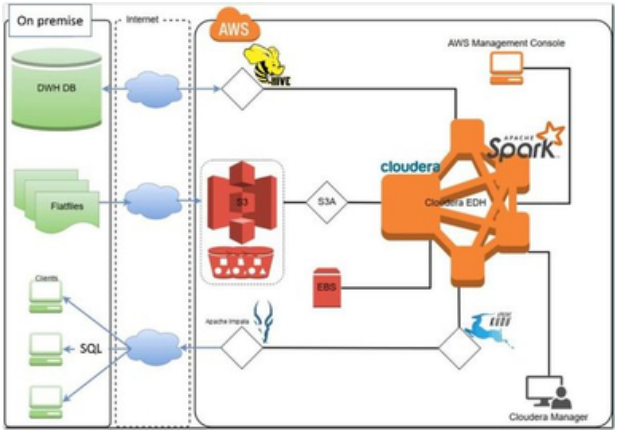

Abb. 2: Cloudera EDH auf AWS

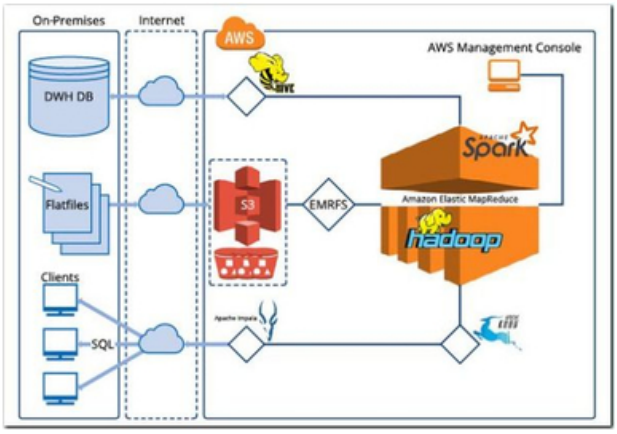

Abb. 3: Amazon Elastic Map Reduce (EMR)

Wie wir sehen, gibt es bis auf den Connector zum S3 Objekt-Storage keinen Unterschied im Tooling. Der eigentliche Unterschied zeigt sich im Preis. Beim Betrieb eines Hadoop-Clusters auf Basis von EMR muss lediglich der Cluster bezahlt werden, auf dem die Umgebung läuft. Wählen wir Cloudera oder eine andere Hadoop-Distribution, kommen noch weitere zu berücksichtigende Lizenzkosten hinzu.

Doch warum nutzen Unternehmen Konstrukte wie eine Hadoop-Distribution + AWS, wenn sie eine Lösung theoretisch günstiger mit EMR betreiben könnten? Hier kommen Faktoren wie Metadatenmanagement, Auditing, Perimeter-Security oder Authentication hinzu, die in Enterprise-Ready-Distributionen gut gelöst sind und auch GDPR-konform quasi „out of the box“ betrieben werden können. Bei anderen Lösungen muss dies noch zusätzlich berücksichtigt werden.

Gibt es noch weitere Gründe für die Cloud?

Die Frage ist eindeutig mit Ja zu beantworten. Unser Rechenbeispiel hat gezeigt, dass sich für gewisse Anwendungsfälle die Cloud mehr lohnt als eine On-Premises-Installation. Umgekehrt gibt es Use-Cases, die eindeutig für die Verwendung einer On-Premises-Lösung sprechen.

Ein wichtiger Punkt, der für die Cloud spricht, sind ihre Vorteile für innovative Geschäftsideen: Die Cloud bietet alle nutzbaren

Ressourcen virtuell an. Diese können über APIs angelegt bzw. gestartet werden. Selbstverständlich sind Stoppen und Löschen ebenso möglich. Die Cloud-Anbieter nutzen diesen Vorteil und haben eine deklarative Beschreibung der virtuellen Umgebung entwickelt, mit deren Hilfe ganze Rechenzentren automatisiert virtuell aufgebaut und gestartet werden können. Dieser Ansatz nennt sich Infrastructure as Code. Mit Infrastructure as Code ist es möglich, eine komplette Big-Data-Plattform in Skripten zu formulieren und nach Bedarf zu starten oder zu löschen.

Diese Freiheit kombiniert mit dem „Pay as you use"-Ansatz fördert Innovationen. Einer Firma ist es jetzt möglich, in wenigen Minuten eine Big-Data-Plattform aufzubauen, einen Proof of Concept zu fahren und anschließend alles wieder abzureißen. Für einen solchen Proof of Concept sind die Kosten gering und das Risiko eines wirtschaftlichen Schadens aufgrund von Investitionen in On-Premises-Rechenzentren kann komplett ausgeschlossen werden. Somit ist es möglich, sehr schnell ohne große Evaluierungsprozesse neue Ideen zu testen und zu verwerfen oder weiterzuführen und auszubauen. Frei nach dem Motto „fail early" werden Konzeptionsfehler somit sehr früh erkannt, und es kann gegengesteuert werden, bevor ein hoher Schaden entsteht. Genau diese Freiheit ist eine Keimzelle für mehr Innovation im Unternehmen. Das zeigt sich insbesondere bei Big-Data-Plattformen, da die Kosten für On-Premises-Lösungen sehr hoch sind und somit einen Trial-and-Error-Ansatz eher verhindern.

Natürlich ist es aufwendig, eine komplette Big-Data-Plattform als Skript zu entwickeln. Aber auch hier gibt es Hilfe: Die Cloud-Anbieter wollen ihren Kunden die Vorteile von Infrastructure as Code schmackhaft machen und bieten deswegen schon vorgefertigte Skripte zum Download an, die viele Standardprobleme auf unterschiedlichen Gebieten lösen. Diese „Templates" können einfach an eigene Bedürfnisse angepasst und kombiniert werden. Sie beschleunigen damit die Erstellung individueller Skripte.

Ein weiterer Punkt, der für die Cloud spricht, ist, dass die Cloud-Anbieter sehr an einer stetigen Verbesserung interessiert sind. Die Konkurrenz und der Wettkampf um die Gunst der Kunden sind groß. Die Anbieter versuchen ihr Angebot also besonders attraktiv zu gestalten, indem sie erstens Kosten senken und zweitens hochwertige Services anbieten. Mit diesen Services vereinfacht sich die Nutzung der Cloud. So muss eine Big-Data-Plattform zum Beispiel nicht mehr selbst installiert und gewartet werden, sondern der Cloud-Anbieter stellt diese als Service bereit und betreibt sie für den Anwender. Der Trend zu immer höherwertigen Services prägt schon seit längerem das Vorgehen der Cloud-Anbieter. Daher empfiehlt es sich, die Angebote der Cloud-Anbieter individuell zu prüfen und zu entscheiden, ob es eventuell schon einen Service gibt, der ein bestimmtes Problem zu einem hohen Prozentsatz löst.

Fazit

Unsere einfachen Rechenbeispiele machen eins klar: Die Preis-Leistungs-Abwägung zwischen On-Premises- und Cloud-Lösungen hängt von vielen Faktoren ab. Eine deutliche Tendenz lässt sich aber in jedem Fall erkennen: Je klarer die Use-Cases sind, desto genauer kann die Wirtschaftlichkeitsrechnung erfolgen. Schon eine grobe Schätzung wie in unserem Beispiel zeigt, dass sich der Weg in die Cloud in vielen Fällen lohnen kann.

Doch was, wenn die Use-Cases kaum bekannt sind? In dem Fall empfiehlt es sich, auf die Cloud zu setzen. Dort können Systeme und Umgebungen zu vertretbaren Kosten schnell und spontan aufgesetzt und ebenso ohne großen Verlust wieder weggeworfen werden.

Dimitri Gross

verfügt über zwölf Jahre IT-Erfahrung und arbeitet als Managing Consultant Solutions bei der OPITZ CONSULTING Deutschland GmbH. Im Competence Center Big Data leitet er den Bereich Big-Data-Architektur, außerdem beschäftigt er sich mit Werkzeugauswahl, Lösungsdesign und Aufbauorganisation in Big-Data-Projekten. Darüber hinaus ist er als Dozent an der Hochschule für Ökonomie & Management München tätig, tritt als Sprecher auf TDWI- und anderen Konferenzen auf und ist Autor mehrerer Fachpublikationen im Bereich Big Data und Künstliche Intelligenz.

E-Mail: Dimitri.Gross(at)opitz-consulting.com

Bildnachweise:

OPITZ CONSULTING Deutschland GmbH