DATA SHARING IN CLOUD-ARCHITEKTUREN

13.10.20

Cloud

von Michael Strauch, Jan Jodat

Das Teilen und Verteilen von Daten ist ein wichtiger Bestandteil der Gesamtarchitektur eines Unternehmens. Spätestens im Falle von externer Datenbereitstellung macht die Bewertung einer Cloud-Lösung dabei unbedingt Sinn. Im folgenden Whitepaper befassen sich Jan Jodat und ich uns mit den Möglichkeiten, die durch moderne Cloud-Lösungen geschaffen werden. Dabei zeigen wir insbesondere auf, wie das gemeinsame Arbeiten mit Daten innerhalb eines Unternehmens und über Unternehmensgrenzen hinweg effizient umgesetzt werden kann.

DATA SHARING ALS TREIBER DES UNTERNEHMENSERFOLGS

VORWORT

Der Umgang mit Daten bestimmt einen erheblichen Teil des Arbeitsalltags vieler Mitarbeiter in Unternehmen und dieser steigt kontinuierlich an. Nicht umsonst werden Daten als der Rohstoff des 21. Jahrhunderts gehandelt. Die Analyse und der Erkenntnisgewinn sind hierbei wesentlicher Bestandteil. Dabei werden Anwendungen geschaffen, die eine konsolidierte Sichtweise auf die Vergangenheit und Gegenwart ermöglichen, den Prozess für zukünftige Entscheidungen unterstützen oder diese gar treffen.

Blickt man in einer modernen Applikationslandschaft hinter die Kulissen und analysiert die entstehenden Aufwände, stellt man fest, dass ein Großteil der Bemühungen auf den Transport und die Aufbereitung der benötigten Daten verwendet wird. Ein großer Aufwandstreiber ist insbesondere der Bereich der Datenaufbereitung und -harmonisierung. Häufig bedarf es hierfür komplexer Logik, welche sich nicht ohne tiefe fachliche Kenntnis reproduzieren lässt. Die daraus entstehenden Prozesse sind unverzichtbarer Bestandteil des Datenrückgrats moderner Unternehmen. Gleichzeitig sind sie aber auch Kostentreiber und Fehlerquelle. Umso wichtiger ist es, Redundanzen nach Möglichkeit zu vermeiden, um konsistente, ressourceneffiziente Datengrundlagen zu realisieren.

Genau das Gegenteil passiert in zahlreichen Unternehmen: Verschiedenste Abnehmer benötigen Daten in unterschiedlichen Systemen. Um eine zeitgerechte Disposition zu realisieren, werden häufig standardisierte Integrationsprozesse umgangen und individuelle Aufbereitungsprozesse außerhalb des üblichen Entwicklungsprozesses realisiert. Solche Entwicklungen sind symptomatisch für IT-Landschaften, die auf On-Premises-Architekturen aufbauen. Jede Lösungsarchitektur für sich bildet zweckmäßig einen spezifischen Anwendungsbereich ab. Aber mangels Skalierbarkeit und Flexibilität ist die Architektur auf diesen einen Bereich limitiert. Der Transport von Daten zwischen den unterschiedlichen Architekturen sorgt für zusätzliche Ressourcenbindung und birgt darüber hinaus Komplexität und Fehlerpotenzial.

Das vorliegende Paper befasst sich mit den Möglichkeiten, die durch moderne Cloud-Lösungen geschaffen werden. Dabei wird insbesondere aufgezeigt, wie das gemeinsame Arbeiten mit Daten innerhalb eines Unternehmens und über Unternehmensgrenzen hinweg effizient umgesetzt werden kann.

Das Teilen und Verteilen von Daten ist ein wichtiger Bestandteil der Gesamtarchitektur eines Unternehmens. Die Daten der verschiedenen Geschäftsprozesse müssen zusammengeführt werden, um die effektive Steuerung und Weiterentwicklung des Unternehmens zu ermöglichen. Hinzu kommt die zunehmende Vernetzung der Unternehmen untereinander, die das Teilen von Daten aus verschiedenen Geschäftsprozessen mit Dritten erfordert. In einem weiteren Schritt können Daten und darauf basierende Analysen selbst als Produkt betrachtet werden, das über entsprechende Marktplätze monetarisiert werden kann.

Spätestens im Falle von externer Datenbereitstellung macht die Bewertung einer Cloud-Lösung unbedingt Sinn, da die großen Cloud Service Provider (CSP) Services zur Verfügung stellen, um Daten auch über die Grenzen des Tenant hinweg zu verteilen. Daten einem Geschäftspartner bereitstellen zu können, erleichtert die Zusammenarbeit und spart mitunter auch Zeit und Geld. Ist diese Bereitstellung der Daten mittels Cloud-Services darüber hinaus noch standardisiert, dann bieten sich neue, bisher nicht gekannte Möglichkeiten der Zusammenarbeit für Unternehmen.

Aber auch für den Einsatz einer Cloud-Architektur innerhalb eines Unternehmens sprechen viele Gründe. So eröffnet die „skalierbare“ Ressourcenverfügbarkeit der CSP Architekturalternativen, die On-Premises nur unzureichend realisiert werden können. Hierbei sollte die generelle Verfügbarkeit aber nicht dazu einladen, eine beliebige Anzahl von individuellen Strukturen zu realisieren, da diese die Wartbarkeit ungemein erschweren. Vielmehr gilt es, zentralisierte Strukturen zu schaffen, die eine gemeinsame Nutzung von Cloud Services ermöglichen, ohne die Realisierung von Anforderungen mit individuellen Services unnötig stark einzuschränken.

DATA SHARING ORIENTIERTE CLOUD-ARCHITEKTUR

Die im Folgenden aufgezeigte Cloud-Architektur nimmt eine funktionsorientierte Trennung der einzelnen Layer vor und weist damit viele Parallelen mit heutigen On-Premises-Architekturen auf. Im Gegensatz zur klassischen Datenverarbeitung und -bereitstellung werden der Storage und Compute Layer in der Cloud-Architektur getrennt voneinander betrachtet. Auch die anderen Layer können in
einer Cloud-Architektur individueller bzw. unabhängiger voneinander gehandhabt werden, als in einer klassischen On-Premises-Architektur. On-Premises werden häufig die Bereiche STORE und COMPUTE fest miteinander verknüpft und der Layer SERVING & ACCESS wird mit dem STORE LAYER gleichgesetzt. Dieses Architekturdesign erschwert eine Skalierung von On-Premises-Architekturen und raubt Flexibilität bei der Anpassung des Systems an aktuelle Anforderungen. Anders in Cloud-Architekturen: Hier wird durch die Verwendung von funktionsorientierten Services die Möglichkeit einer Trennung der einzelnen Layer realisiert.

Als zentrales Bindeglied zwischen den Services der individuellen Anwendungsszenarien bietet sich der STORE Layer an. Im Vergleich zu On-Premises-Architekturen stellt er weniger ein Archiv für Daten im Rohformat dar, vielmehr bildet dieser Bereich einen zentralen Hub für die Gesamtheit aller analyserelevanten Daten.

Wie in der Grafik zu erkennen ist, erfolgt zwischen den Layern STORE und COMPUTE ein bilateraler Austausch. Dennoch ist eine strikte Trennung der Layer zu empfehlen. Einerseits würde der Datenaustausch wieder auf verschiedenen Ebenen erfolgen, was zu Intransparenz führt und die Ablösbarkeit von Services erschwert. Denn bauen unterschiedliche Applikationen gegenseitige Abhängigkeiten auf, dann erschwert dies die Weiterentwicklung und den Betrieb langfristig massiv. Andererseits sind die Kosten für Storage im Vergleich zu denen für Compute-Ressourcen niedrig. Die Konsequenz wären ungenutzte, aber dennoch abgerechnete Compute-Ressourcen.
Darüber hinaus ermöglicht erst diese Aufteilung das Data Sharing rein auf der Basis von Dateninhalten, losgelöst von den Compute Services. Dies sollte auch bei der Realisierung von Lösungen im Auge behalten werden. Orientiert sich das Schneiden der Storage Services am Data Sharing, erleichtert das die Freigabe der Inhalte ungemein.

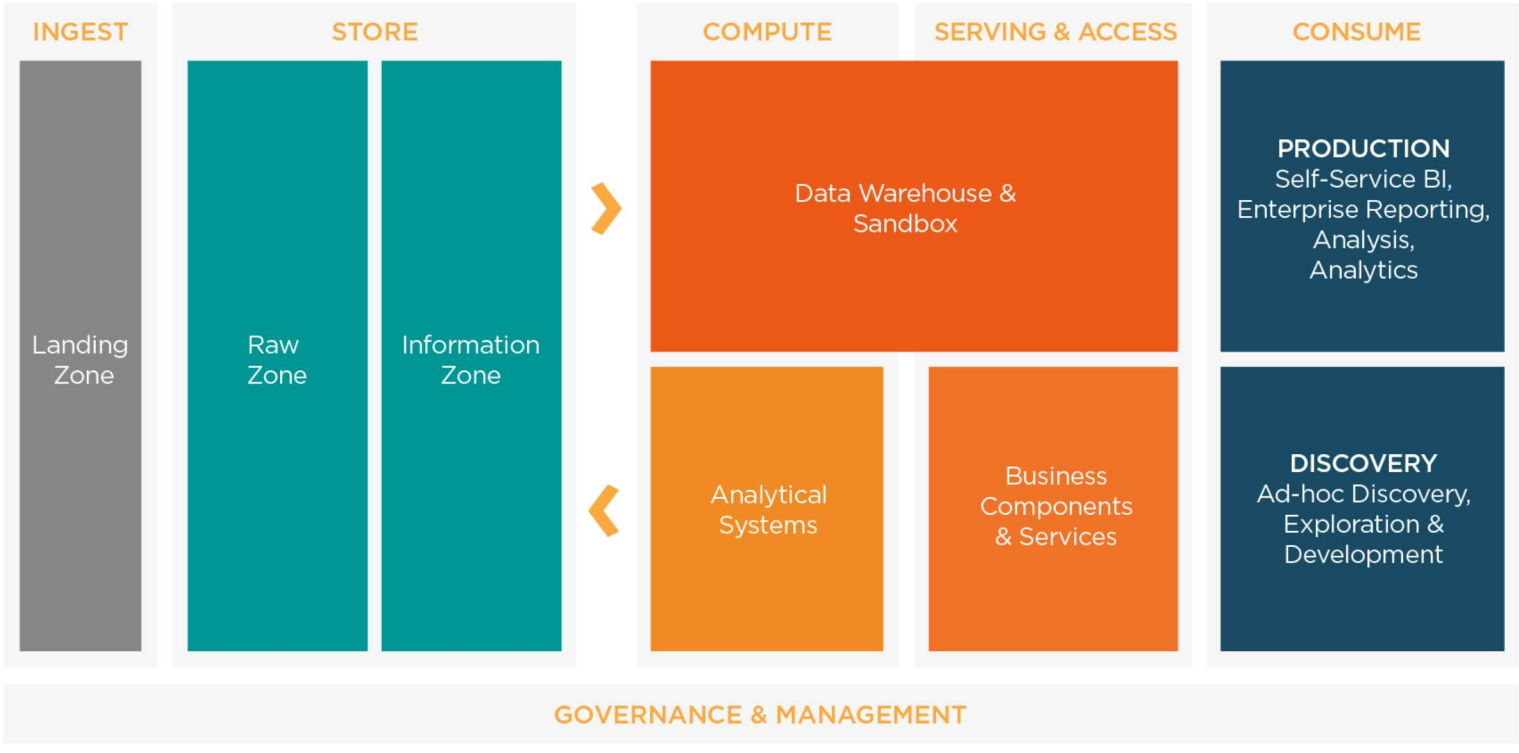

Abb.: Cloud-Architektur mit funktionsorientierten Schichten

Von einer schnelleren und detaillierteren Unterstützung bei der unternehmerischen

Entscheidungsfindung bis hin zu Potenzialen für Innovationen können sich erhebliche geschäftliche Vorteile gegenüber Wettbewerbern ergeben, die Data Sharing nicht als Option für ihr Unternehmen in Betracht ziehen.

Michael Strauch, Lead Consultant, INFOMOTION GmbH

DATA SHARING AM BEISPIEL DER INFOMOTION TAILORED DATA PLATFORM ON AZURE

Eine konkrete Ausgestaltung der zuvor beschriebenen logischen Architektur findet sich in der INFOMOTION Tailored Data Platform wieder. Hier am Beispiel von Microsoft Azure als CSP. Abhängig vom CSP ändern sich die verwendeten Services, der Grundgedanke hinter der Architektur bleibt jedoch identisch. Spezifische Services für einzelne Applikationen werden nach Bedarf provisioniert und entsprechend der Last individuell skaliert. Die Verwendung eines gemeinsamen Store Layers durch alle Applikationen schafft hier das Datenrückgrad des Unternehmens. Auf dieser Ebene werden sowohl Rohdaten als auch bereits verarbeitete Daten abgelegt und stehen in Echtzeit anderen Applikationen zur Verfügung. Der Freigabeprozess erfolgt auf Ebene der individuellen Speicherressourcen/-container.

Über dieses Szenario können sämtliche Anforderungen innerhalb des gleichen Cloud-Abonnements abgedeckt werden. Auch eine Verteilung der Daten hin zu anderen Azure Tenants ist realisierbar über den Einsatz eines separaten Service, der generalistisch im Store Bereich verwaltet wird. Hierbei gibt es durch Azure Data Share sowohl die Option für direkte Zugriffe durch Dritte auf die Daten als auch die Alternative, regelmäßige Replikationen in die Cloud des Dritten zu realisieren.

Sobald es um die Überlegung geht, auch die Konnektivität zu anderen CSP herzustellen, stößt man mit dem Einsatz nativer Möglichkeiten der CSP an Grenzen – insbesondere, wenn es darum geht, Daten nicht redundant zu halten.
Um hier eine Alternative zu schaffen, wird im folgenden Abschnitt das Third Party Tool Snowflake vorgestellt. Mit diesem kann eine CSP-übergreifende Datenplattform realisiert werden, die diverse Möglichkeiten zum externen Data Sharing bietet.

Abb.: INFOMOTION Tailored Data Platform am Beispiel von Microsoft Azure

DATA SHARING EMPOWERMENT DURCH EINSATZ VON SNOWFLAKE ALS ZENTRALE DATENPLATTFORM

Snowflake ist eine Cloud-basierte Datenplattform, die auf den Bereich des Data-Warehousing spezialisiert ist. Sie bietet neben beliebig horizontal und vertikal skalierbaren Rechenressourcen eine volle SQL Unterstützung, was insbesondere Usern ohne Programmierkenntnisse entgegenkommt und die Einstiegshürde in eigenständige Datenanalysen verringert.
Mit Hilfe von Snowflake können sowohl strukturierte als auch semi-strukturierte Datenformate gespeichert und verarbeitet werden, die einen großen Anteil der analyserelevanten Daten in Unternehmen ausmachen.

Zum Erstellungszeitpunkt des Papers ist Snowflake auf den drei großen CSP (Azure, AWS, GCP) verfügbar und kann direkt mit deren Flat Storage Services kommunizieren, was eine direkte Integration von Data Lakes ermöglicht. Kombiniert mit ODBC/JDBC Unterstützung sowie Konnektoren zu Spark, Python und Kafka ist bereits heute eine breite Konnektivität sichergestellt und qualifiziert Snowflake als interoperable Datenplattform. Mit den inline Funktionen von Snowflake Stage, Stream und Task lässt sich innerhalb des CSPs eine vollautomatisierte Verarbeitungsstrecke erstellen, die einer Pipeline nahekommt.

Durch die Trennung von Storage und Compute in Snowflake kann sich der Service nahtlos in die vorgestellte Architektur einfügen und bereits etablierte Services unterstützen oder diese anforderungsgerecht ablösen. Dabei ist Snowflake nicht nur für einen Greenfield-Ansatz empfehlenswert, sondern kann auch in bestehenden Cloud-Architekturen einen Mehrwert darstellen.

In untenstehender Abbildung wird demonstriert, in welche Layer Snowflake einzugliedern ist. Der STORE Layer profitiert von den Integrationsalternativen und ermöglicht einen direkten Durchgriff auf den Flat Storage. Hier ist die Entscheidung zu fällen, in welchem der Services die Daten persistent gehalten werden oder ob eine redundante Datenhaltung mit automatisierter Aktualisierung von Vorteil ist.

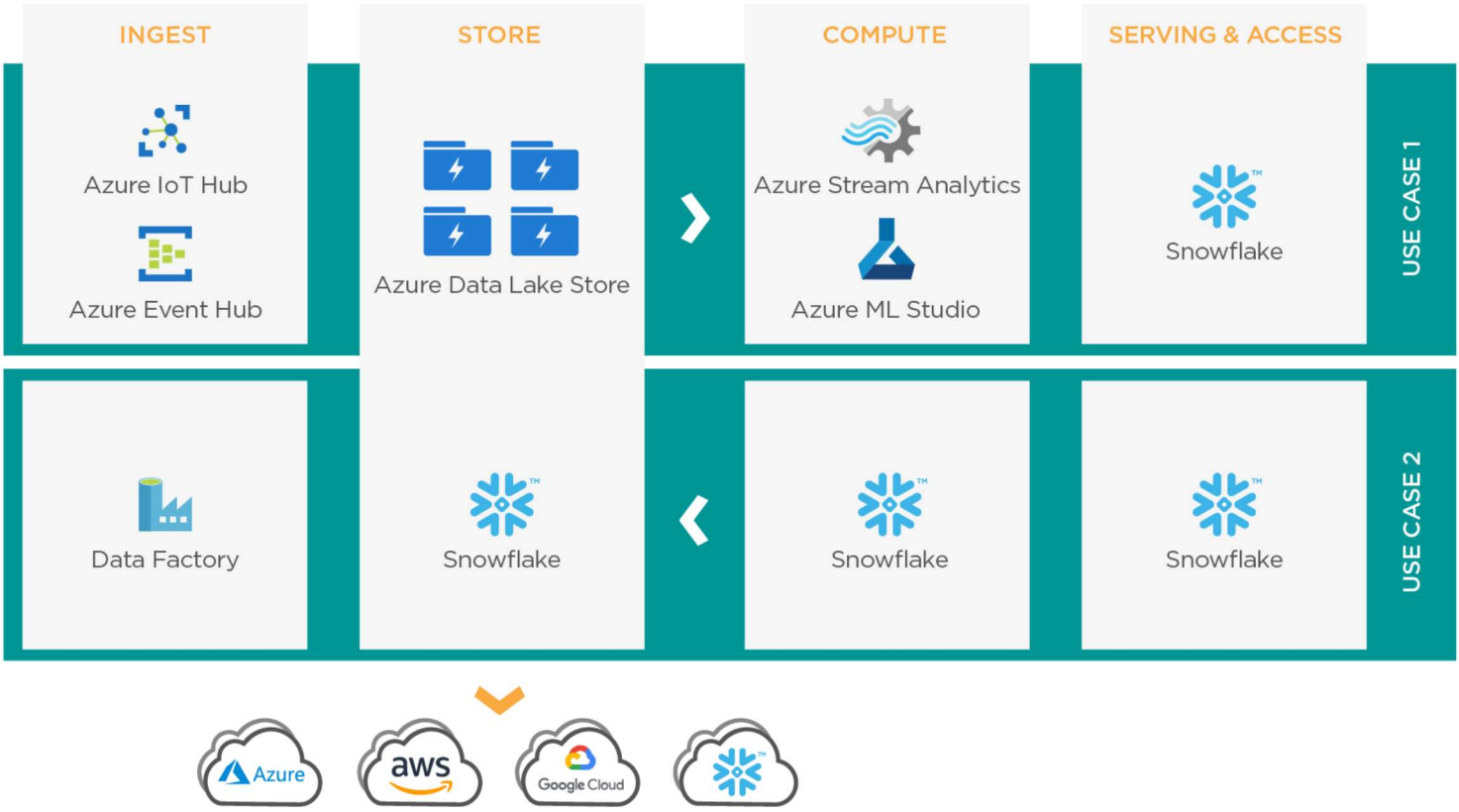

Abb.: INFOMOTION Tailored Data Platform erweitert um Snowflake

Der COMPUTE Layer profitiert von der horizontalen und vertikalen Skalierbarkeit der Snowflake Compute-Ressourcen. Horizontal skalieren bedeutet, dass der Compute-Layer – im Snowflake-Kontext „Warehouse“ genannt – automatisch um mehrere parallel laufende Cluster erweitert wird. Die Cluster werden von Snowflake eigenständig gestartet und heruntergefahren. Bei einer vertikalen Skalierung wird die Anzahl der genutzten Server erweitert und kann von 1 bis 128 Servern beliebig gewählt werden. Im Vergleich zu zahlreichen anderen Services werden hier nur die Kosten für tatsächlich provisionierte Ressourcen berechnet und nicht für die generelle Verfügbarkeit. Die Provisionierung selbst geschieht zur Laufzeit der Kalkulationen ohne signifikante Laufzeitverlängerungen und auch die Deprovisionierung erfolgt automatisiert bei Inaktivität.

Im SERVING & ACCESS Layer kann Snowflake wie eine normale Datenbank angesprochen werden und sich mit allen ODBC/JDBC fähigen Abnehmersystemen verbinden.

Durch die zahlreichen Konnektivitätsalternativen kann Snowflake verwendet werden, um eine Brücke zwischen den verfügbaren CSP zu schlagen. In dem untenstehenden Beispiel ist eine Verbindung der bereits vorgestellten Layer über CSP Grenzen hinweg dargestellt.

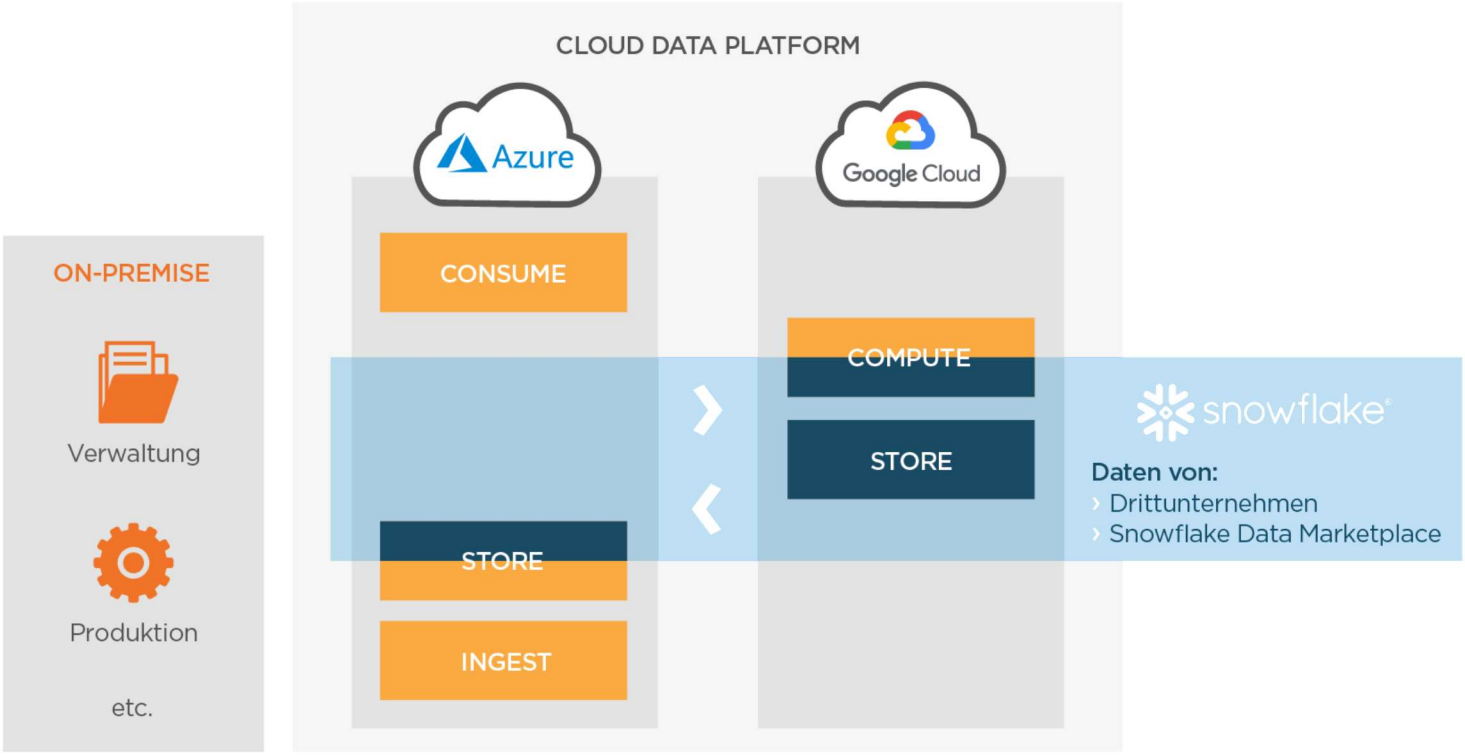

Abb.: Data-Sharing mit Snowflake und unterschiedlichen CSP im Zusammenspiel

So ist es vorstellbar, dass Azure als primärer CSP verwendet wird, aber einige Anwendungsfälle leichter bei einem anderen CSP realisiert werden können. Gründe hierfür können Unterschiede im Service Angebot sein oder schlicht die Verfügbarkeit von Entwicklungsressourcen mit Erfahrung bei einem anderen CSP. Zu entscheiden ist hierbei, ob man einen einzigen Snowflake Tenant verwenden möchte oder ob sich mehrere, unterschiedlich gehostete Tenants anbieten. Diese Entscheidung ergibt sich maßgeblich aus der jeweiligen Verwendung der Infrastruktur auf den unterschiedlichen CSPs.

Letztere Möglichkeiten zum Data Sharing, also die Verbindung mehrerer Snowflake Tenants, kann auch über die Unternehmensgrenzen hinweg realisiert werden. So ist es denkbar, dass sich Unternehmen einzelne Datensegmente gegenseitig freigeben und somit konsolidierte Datenanalysen – über Unternehmensgrenzen hinweg – möglich machen.

Auch ohne ein Snowflake Tenant von Dritten gibt es die Möglichkeit, diesen Zugriff auf Daten zu ermöglichen. Sogenannten Reader Accounts können feingranular mit Berechtigungen ausgestattet werden, sodass man zu jedem Zeitpunkt Einfluss auf die Verwendung von Daten und Rechenressourcen hat. Als neues Feature im Snowflake Repertoire präsentiert sich der Data Marketplace. Hierüber können Datenpools auf einem allgemein zugänglichen Marktplatz angeboten und erstanden werden. In dem Marketplace werden Finanzdaten, Wetterdaten, Konjunkturdaten wie auch Gesundheitsdaten zur Verfügung gestellt. Außerdem kann anhand von Stichwörtern kategorisiert werden, ob es sich um Standortdaten handelt und in welchem Tonus die Daten aktualisiert werden.

FAZIT

Data Sharing in der Cloud soll standardisierte Schnittstellen für den unternehmensinternen und unternehmensübergreifenden Datenaustausch bieten. Diese Standards reduzieren den Aufwand bei der Konzeption sowie der Implementierung und vereinheitlichen den Betrieb von Applikationen. Snowflake bietet aktuell eine innovative Lösung für das Data Sharing an. Als Cloud Data Platform haben sie eine Lösung entwickelt, mit der es ermöglicht wird, Daten Tenant- und CSP-unabhängig zu teilen. Provider-übergreifendes Data Sharing stellt aktuell eine Innovation dar, die ohne ein Third Party Tool nur schwer realisiert werden kann. Da sich das Angebot an Cloud Services rasant fortentwickelt, sind neue Alternativen fortlaufend zu validieren, um zukünftige Potenziale noch besser auszuschöpfen.

Dieser Artikel hat ausschließlich die technischen Aspekte des Data Sharing betrachtet. Daneben müssen auch organisatorische

Herausforderungen bewältigt werden: Heute ist es nicht selbstverständlich, Daten mit Dritten auszutauschen. Die Bereitschaft zum Teilen von Unternehmensdaten, das dafür notwendige Vertrauen zum Geschäftspartner sowie eventuelle Risiken und nicht zuletzt rechtliche Fragen müssen geklärt werden, bevor eine Umsetzung beginnen kann. Die Vermarktung und Monetarisierung der eigenen Daten ist ein umso größerer Schritt. Allgemein muss ein Umdenken innerhalb der Unternehmen stattfinden. Es gilt Prozesse und Rollen einzurichten, die sich explizit mit dem Data Sharing befassen.
Ist die organisatorische Hürde zusätzlich zu der technischen überwunden, bieten sich Unternehmen eine Vielzahl neuer Möglichkeiten, um existierende Daten schneller und zielgerichteter zu beziehen und zu verteilen. Es können sich aber auch komplett neue Datenkonstellationen ergeben, die neue Erkenntnisgewinne fördern. Von einer schnelleren und detaillierteren Unterstützung bei der unternehmerischen Entscheidungsfindung bis hin zu Potenzialen für Innovationen können sich erhebliche geschäftliche Vorteile gegenüber Wettbewerbern ergeben, die Data Sharing nicht als Option für ihr Unternehmen in Betracht ziehen.

AUTOREN

MICHAEL STRAUCH

Lead Consultant

Michael Strauch studierte Wirtschaftsinformatik an der Hochschule Fulda. Nach seinem Berufseinstieg 2011 bei INFOMOTION betreute er KAGs bei dem Betrieb und der Weiterentwicklung um regulatorische Anforderungen des BI-Systems. Neben dem Beruf machte er seinen Abschluss zum Master of Science in Wirtschaftspsychologie an der FOM Hochschule für Ökonomie und Management. Aktuell ist Michael Strauch als Architekt für die Konzeption von Anforderungen sowie der technischen Projektleitung von mehreren Teams im Retail Bereich verantwortlich.

E-MAIL SENDEN

JAN JODAT

Senior Consultant

Jan Jodat studierte Wirtschaftsingenieurwesen und Unternehmensführung an der Technischen Hochschule Mittelhessen. Seit seinem Berufseinstieg 2015 bei INFOMOTION betreute Herr Jodat branchenübergreifend Kunden bei der Konzeption und Implementierung ihrer BI-Lösungen. Aktuell ist Jan Jodat als Entwickler, Architekt und technischer Projektleiter in mehreren Projekten für einen Kunden in der Retail Branche tätig.

E-MAIL SENDEN