BI & ANALYTICS IN DER CLOUD

DIE HERAUSFORDERUNGEN VON CLOUD ANALYTICS IN DER PRAXIS

Der „Change", den aktuell viele Unternehmen gehen oder in naher Zukunft gehen werden, um am Markt weiterhin bestehen oder sogar führen zu können, betrifft viele verschiedene Aspekte. Neben organisatorischen Strukturen oder Geschäftsmodellen, gehören vor allem auch die eingesetzten Hard- und Softwarearchitekturen auf den Prüfstand.

Business Intelligence- und Analytics-Systeme stehen hier häufig im Fokus, schließlich geht es auch um starke Veränderungen beim Datenaufkommen, bei Datenarten und den Chancen für die Wertschöpfung aus Daten. Unternehmen erkennen seit geraumer Zeit den Mehrwert, den das Sammeln und das Analysieren von internen und externen Daten bietet.

Allerdings verursachen Anpassungen eines bestehenden On-Prem Data Warehouses (DWH) an stark steigende Datenvolumina in der Regel hohe Lizenz-, Hardware- und Personalkosten. Dazu kommt, dass On-Prem-Systeme in der Regel so konzipiert werden, dass sie auch den rechenintensivsten Anwendungsfall abdecken können. Für eine Vielzahl weiterer Anwendungsfälle sind diese Systeme dann überdimensioniert.

Doch wird Cloud Computing im analytischen Kontext seinem Ruf als häufig angepriesenes Allheilmittel tatsächlich gerecht? In diesem Zusammenhang scheint Cloud Computing Vorteile zu bieten. So versprechen die Anbieter von Cloud-Analytics-Lösungen eine Vielzahl an Vorteilen bei den Bereitstellungs- und Ausführungszeiten sowie eine schnelle und nahezu grenzenlose Skalierbarkeit der Services. Des Weiteren soll die nutzungsbasierte Abrechnung zu erheblichen Einsparungen auf der Kostenseite führen. Dieser Artikel gibt einen Überblick über die Möglichkeiten analytischer Lösungen in der Cloud. Darüber hinaus werden sowohl die Vorteile als auch die Nachteile der Cloud Services kritisch gegenübergestellt. Zu Beginn geht es um die Möglichkeiten und Hindernisse, die mit Analytics in der Cloud in der Praxis verbunden sind. Anschließend werden drei wichtige Architekturkomponenten vorgestellt, auf denen Cloud Services in der Regel basieren und konkrete Services sowie deren Anbieter beispielhaft vorgestellt. Den Abschluss bildet eine Zusammenfassung und ein Resümée.

Die Cloud – Mehr als nur ein Trend

Im jährlich von Gartner veröffentlichten Hype Cycle, der aktuell diskutierte Technologien in fünf Phasen eines Technologie¬lebenszyklus einteilt, wurde Cloud Computing 2008 erstmals als Hype beschrieben. Sechs Jahre später befand sich die Cloud gemäß Gartner im „Tal der Desillusionierung". [1] Das bedeutet keineswegs das Scheitern einer Technologie, sondern lediglich die anfängliche Ernüchterung nach einem Hype und das Antreffen erster Hindernisse. Gartner zufolge sollte Cloud Computing innerhalb der nächsten zwei bis fünf Jahre den Markt breit durchdrungen haben und somit eine am Markt relevante Technologie darstellen. Diese Einschätzung kann mittlerweile bestätigt werden. Die Anzahl der Unternehmen, die cloudbasierte BI-Lösungen nutzen, stieg von 25 % auf 49 % zwischen den Jahren 2016 und 2018. Im gleichen Zeitraum sank die Zahl der Unternehmen, die den Weg in die Cloud nicht gehen wollen von 38 % auf 19 %. [2]

Die Zahlen und Prognosen bestätigen, dass es sich bei BI-Lösungen in der Cloud nicht um einen bald vergessenen Trend handelt, sondern dass Unternehmen Mehrwerte erkennen und Teile ihrer IT-Infrastruktur tatsächlich in die Cloud auslagern. In diesem Kapitel sollen Beweggründe für die Nutzung von Cloud Services im Analytics-Umfeld thematisiert werden. Außerdem wird auf häufig genannte Risiken bei der Auslagerung der IT-Infrastruktur hingewiesen.

Treiber für Analytics in der Cloud

Die Eigenschaften des Cloud Computings bieten auch im BI-Kontext zahlreiche Gründe zur Auslagerung der Infrastruktur. Nachfolgend werden Treiber für Cloud BI näher erläutert.

- **Steigerung der Agilität**
  Besonders im Hinblick auf die fortschreitende Entwicklung zu datengetriebenem Arbeiten ist die Wahl der IT-Systeme für Unternehmen ein entscheidender Faktor für die Steigerung der Agilität. Mit Hilfe von Cloud-Infrastrukturen lassen sich Proof of Concepts (PoC) für moderne Technologien durchführen, ohne die häufig voll ausgelastete interne IT zu sehr oder überhaupt in Anspruch nehmen zu müssen. Des Weiteren kann das Anlegen mehrschichtiger Systeme für die Nutzung verschiedener Umgebungen, etwa für Entwicklung, Test und Produktion, mittels Containertechnologien wie Docker erheblich vereinfacht werden. Die Bereitstellung und Konfiguration von Infrastruktur kann auch mithilfe von Skripten automatisiert werden. Dieser als Infrastructure as Code bekannte Ansatz ermöglicht das Erstellen und Löschen virtualisierter Umgebungen auf Knopfdruck. Die Anbieter haben dieses Potenzial erkannt und stellen den Anwendern beispielhafte Skripte zur Verfügung. Auch die zunehmend hohe Granularität der Komponenten in Microservices-Architekturen, die über Standardschnittstellen gekoppelt werden können, befördert die Agilität. [3]

- **Flexibilität durch elastische Skalierbarkeit**
  Im Bereich der Analytics-Komponenten gilt es häufig, zwischen horizontaler (scale out) und vertikaler (scale up) Skalierung zu unterscheiden. Bei der horizontalen Skalierung werden Recheninstanzen ab- oder dazugeschaltet, während bei der vertikalen Skalierung die Rechenkapazitäten der vorhandenen Instanzen erhöht oder herabgestuft werden. Der entscheidende Vorteil der Cloud-Ressourcen gegenüber On-Prem-Ressourcen ist die Möglichkeit der elastischen Anpassung in der Cloud, das heißt, Ressourcen können schnell oder zum Teil sogar sofort skaliert und an den momentanen Bedarf angepasst werden.

- **Kostenverlagerung**
  Die Anbieter der Cloud Analytics Services werben mit erheblichen Kosteneinsparungen gegenüber On-Prem-Systemen. Diese Versprechen sind, insbesondere für einen mittel- bis langfristigen Nutzungszeitraum, mit Vorsicht zu genießen und sollen an dieser Stelle weder belegt noch widerlegt werden. Je nach Unternehmensstrategie oder -größe kann es ein Vorteil sein, wenn hohe Investitionskosten (CAPEX) entfallen und Betriebskosten als bedarfsgerechte Abrechnung (OPEX) verbucht werden (Bild 1).

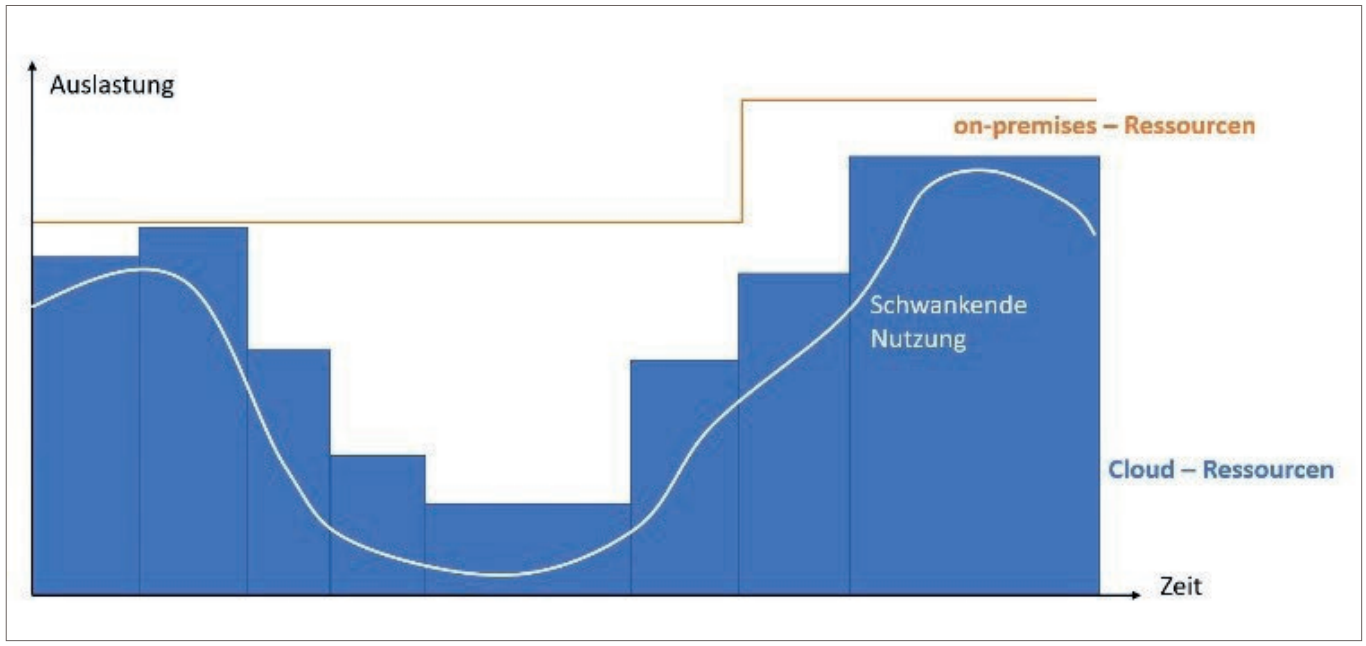

Bild 1: Vergleich der Skalierbarkeit von On-Prem- und Cloud-Ressourcen.

- **Reduzierung administrativer Aufwände**
  Cloud-Anbieter bieten heute schon eine Vielzahl serviceabhängiger Funktionen an wie automatisierte Ressourcenanpassung, automatisches (Security-)Patching und Upgrading, automatische Backups und Verschlüsselung, Ausfallschutz durch die Bereitstellung der Ressourcen auf verschiedenen Rechenzentren, der Einsatz von Algorithmen zur Leistungsoptimierung etc. Die so freiwerdenden Kapazitäten können die Kunden in innovative oder wertschöpfende Tätigkeiten investieren.

- **Kontinuierliche Aktualisierung und Erweiterung der Services**
  Die Tools im Analytics-Bereich werden kontinuierlich weiterentwickelt. Bei vielen Herstellern ist dabei eine „Cloud first"-Strategie erkennbar, das heißt neue Features werden zunächst den Cloud-Nutzern in einer Art Preview zur Verfügung gestellt, ehe sie im nächsten Release allen Nutzern zur Verfügung stehen.

Hindernisse für Analytics in der Cloud

Trotz der zuvor erläuterten vielfältigen Möglichkeiten, die Cloud Computing bietet, sehen viele Unternehmen die Auslagerung ihrer IT-Ressourcen und unternehmensinternen Daten in die Cloud noch skeptisch. Es lohnt sich, sich die befürchteten Risiken einmal genauer ansehen. Vor allem die folgenden Risiken werden hier erfahrungsgemäß genannt:

- **Vendor-Lock-in**
  Gibt man Aufgaben in den Bereichen Administration, Konfiguration sowie IT- bzw. Datensicherheit ab, erhöht dies gleichzeitig die Abhängigkeit vom Anbieter. Je nach Struktur des gewählten Cloud Services kann sich die Migration zu einem alternativen Service schwierig gestalten. Die Möglichkeiten der Extraktion der Daten, ETL-Logiken oder anderer getätigter funktionaler Investments sollten bei der Wahl des Cloud Services ebenfalls berücksichtigt werden, um „böse Überraschungen" zu einem späteren Zeitpunkt zu vermeiden.

- **Supportabhängigkeit**
  Die Abhängigkeit vom Anbieter wird am ehesten ersichtlich, wenn Fehler oder Bugs auftreten. Den Anwendern bleibt dann häufig nichts anderes übrig, als den Fehler beim Anbieter zu melden und ihn mittels detaillierter Fehlerbeschreibung, Systemkonfiguration oder Log-Files bei der Fehlerbehebung zu unterstützen.

- **Betriebskosten**
  Die häufig als positiv beworbene Verlagerung der Kosten kann sich in gewissen Anwendungsszenarien auch negativ auswirken. So ist der dauerhafte Einsatz hoch performanter Cloud-Infrastrukturen in der Regel sehr kostspielig. Ebenso sind On-Prem-Lösungen dort langfristig monetär überlegen, wo von einer stabilen Nutzungslast und gleichbleibenden Datenvolumina auszugehen ist. Cloud-Anbieter versuchen dem häufig entgegenzuwirken, indem sie großzügige Mengenrabatte auf eine festgelegte Abnahmemenge der Ressourcen über einen gewissen Zeitraum gewähren.

- **Netzwerkkapazitäten und Latenz**
  Besonders im Rahmen der Datenübertragung in Lade- oder Streamingszenarien ist die eigene Netzwerkkapazität mit zu berücksichtigen. Außerdem ist selbst in gut vernetzten Umgebungen stets eine gewisse Latenz vorhanden. Das führt dazu, dass in der Regel nur eine „Near Real-Time"-Datenverarbeitung und -analyse möglich ist.

- **Datenschutz und Sicherheit**
  Die Themen Datenschutz und Sicherheit sowie rechtliche Bedenken sind immer noch Hauptargumente von Unternehmen, die sich gegen die Nutzung von Cloud-Komponenten entscheiden. Generell gilt es natürlich vorab zu klären, ob und wie (personenbezogene) Daten über die eigenen Netzwerkgrenzen hinweg verarbeitet werden dürfen.

Eine pauschale Abwehrhaltung ist heute nicht mehr zeitgemäß: Service Anbieter unterliegen meist viel strengeren Sicherheitsauflagen und haben umfänglichere Sicherheits¬mechanismen implementiert, als das eigene Unternehmen. Ein schönes Beispiel dafür ist ein Fall der Bundespolizei, der Anfang 2019 öffentlich bekannt wurde: Diese speichert Bodycam-Aufnahmen in der AWS Cloud, da derzeit keine den Anforderungen entsprechende staatliche Infrastruktur vorliegt und AWS alle vom Bundesamt für Sicherheit in der Informationstechnik gestellten Sicherheitsstandards erfüllt. [4] Hinsichtlich der Anforderungen an

Datensicherheit und Datenschutz ist allerdings auch ein steigender Aufwand im Rahmen der IT- und Data Governance zu erwarten, wenn IT-Ressourcen und Daten ausgelagert werden. Dies spiegelt auch der vom Business Application Research Center (BARC) veröffentlichte BI Trend Monitor 2019 wider, in dem Data Governance wiederholt auf Platz 4 der Trendthemen im BI- und Analytics-Umfeld landet. [5]

Service- und Bereitstellungsmodelle

Entscheidet sich ein Unternehmen, Cloud Services in seiner Analytics-Architektur zu nutzen oder diese im Rahmen eines Proof of Concepts (PoC) auszuprobieren, sollte es zunächst definieren, welches Cloud-Service-Modell für die eigenen Anwendungsfälle am besten passt. Simultan dazu wählt es die Bereitstellungsart der Cloud-Dienste.

Cloud-Bereitstellungsmodelle

Cloud Services können in unterschiedlicher Weise bereitgestellt werden. Im Wesentlichen sind folgende drei Modelle zu unterscheiden:

- **Private Cloud**
  Hierbei wird die Infrastruktur ausschließlich für eine Organisation betrieben. Betrieb und Organisation der Infrastruktur kann im eigenen Rechenzentrum erfolgen oder an einen Dienstleister ausgelagert werden.

- **Public Cloud**
  Von einer Public Cloud spricht man, wenn ein Anbieter seine Services und Infrastruktur einer breiten Masse von Anwendern zur Verfügung stellt.

- **Hybrid Cloud**
  Eine Hybrid Cloud entsteht, wenn Infrastrukturen aus einer Private Cloud, Public Cloud und dazu eventuell noch On-Prem-Architekturen über Schnittstellen im Verbund genutzt werden. Dabei wird häufig der „Best-of-Breed"-Ansatz verfolgt, um der Architektur das bestmögliche Verhältnis von Individualität und erprobten Standard-Services zu verleihen.

In Bezug auf analytische Systeme ist festzustellen, dass ein Verschieben großer Datenmengen zwischen verschiedenen Infrastrukturen sowohl physikalische als auch wirtschaftliche Herausforderungen birgt. In diesem Zusammenhang wird häufig auch von Data Gravity gesprochen. [6]

Demnach sollten Services und Anwendungen dorthin verlagert werden, wo die Daten entstehen. Um diese Herausforderungen zu bewältigen bieten sich hybride Umgebungen an, die den Unternehmen größtmögliche Variationsmöglichkeiten bieten.

Cloud-Servicemodelle

Cloud Services werden anhand ihrer Servicetiefe unterschieden. Das Servicemodell definiert, welche Zuständigkeiten in der Hand des Anbieters liegen und welche Aufgaben vom Kunden übernommen werden. Zu unterscheiden wären hierbei die drei Servicemodelle:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)

Bild 2 veranschaulicht die Verteilung der Zuständigkeiten je nach Servicemodell. Zu beachten ist, dass durch den steigenden Verantwortungsbereich des Anbieters die Kontrollmöglichkeiten des Anwenders reduziert werden. Dadurch sinken Konfigurations- und Individualisierungsoptionen auf Anwenderseite.
Aus diesem Grund ist es wichtig, die Anwendungsfälle und Nutzergruppen vorab genau zu definieren. Nutzer, die zur Self-Service Analyse auf ein vorgefertigtes Datenmodell zugreifen, benötigen weniger Freiheiten, als es für einen Data-Scientist der Fall ist. Dieser benötigt die Möglichkeit über individuelle Schnittstellen, weitere Quellen anzubinden und diese mittels eigens programmierter Skripte oder Algorithmen aufzubereiten.

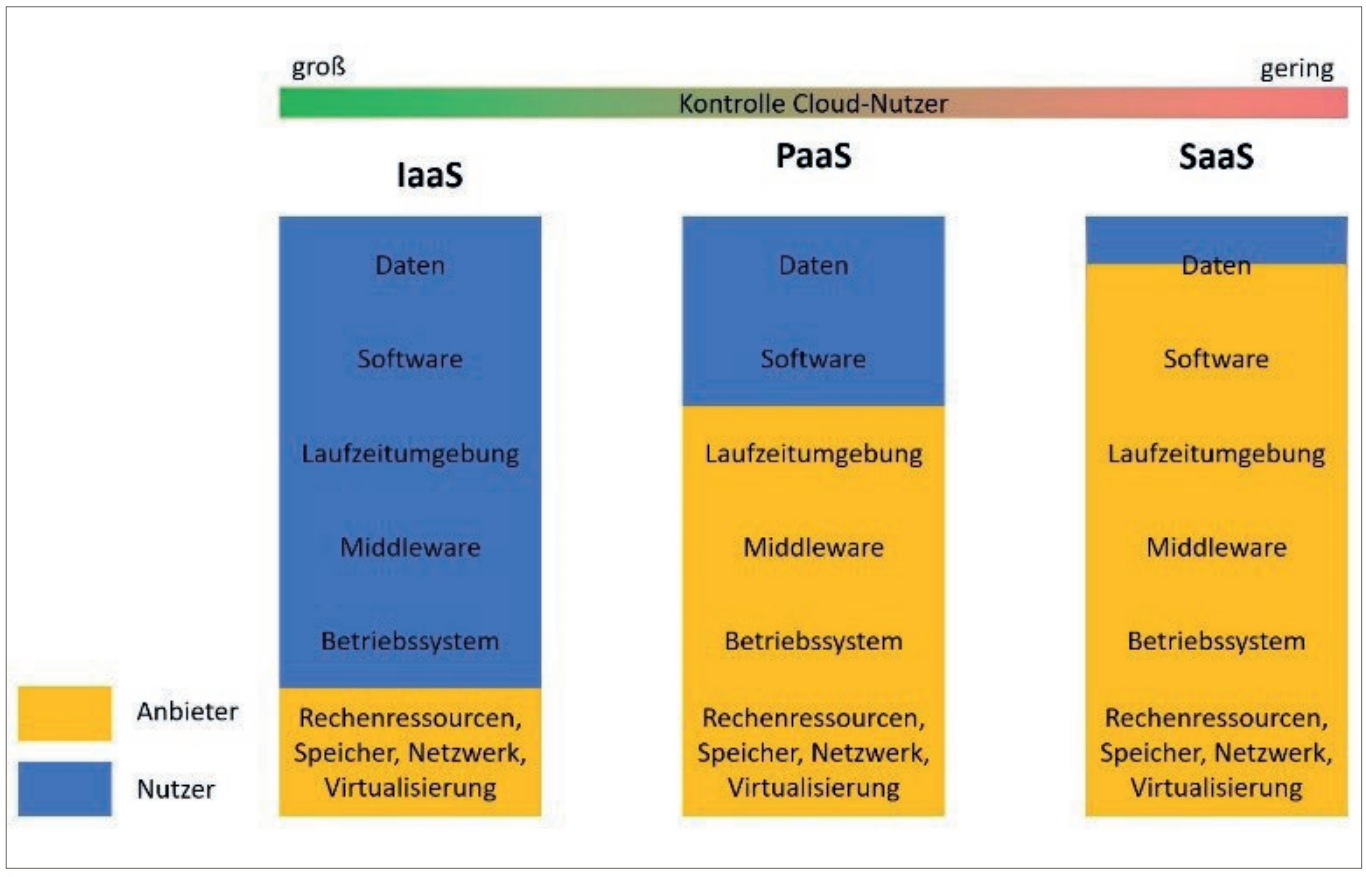

Bild 2: Vergleich der Cloud-Service-modelle.

Architekturkomponenten & Anbieter

Mittlerweile ist die Umsetzung einer ganzheitlichen analytischen Plattform mittels Cloud-Komponenten möglich. Welche Services stehen am Markt zur Verfügung? Im Rahmen des Whitepapers kann natürlich nur eine kleine Auswahl der verfügbaren Services betrachtet werden. Die Auswahl beschränkt sich auf Services, die zur Umsetzung der Komponenten ETL, DWH und Reporting einer Analytics-Architektur dienen und als PaaS oder SaaS bereitgestellt werden.

Aus jedem Bereich haben wir zwei Cloud Services ausgewählt, die jeweils signifikante Unterschiede in Funktionsweise und -umfang sowie Service- oder Preismodell aufweisen. Bild 5 zeigt die nachfolgend vorgestellten Cloud Services und je ein bzw. zwei Charakteristiken, die die Services unterscheiden.
Bei der Auswahl der Services wurde zudem darauf geachtet, dass diese auch gemeinsam in einer Architektur implementiert werden können und sich somit ein schlüssiges Gesamtbild ergibt. Selbstverständlich könnten auch weitere, hier nicht aufgeführte Cloud Services, mit den genannten Services innerhalb einer Architektur genutzt werden.Darüber hinaus gibt es am Markt eine Vielzahl weiterer Services, mit denen auch weitere Komponenten einer analytischen Architektur umgesetzt werden können. Dazu zählen beispielsweise Tools zum Monitoring, zur Erstellung eines Metadatenkatalogs oder zum Einsatz in Big Data- und Streaming-Szenarien.

ETL as a Service

Die Grundlage eines analytischen Systems ist bekanntlich eine gut aufbereitete Datenbasis. Dafür werden Daten aus unterschiedlichen Quellsystemen homogenisiert, bereinigt und angereichert. Für die Umsetzung braucht es eigens geschriebene (SQL-)Mappings oder ein Tool zur Realisierung des ETL-Prozesses.

ETL steht für Extract, Transform, Load. Entsprechend unterstützt ein solches Tool bei der Erstellung, Dokumentation, Wartbarkeit und Standardisierung komplexer Transformationen, indem es den Prozess durch die modellbasierte Umsetzung der Transformationslogiken effizienter gestaltet und diesen Prozess auch für Personen mit wenig bis keiner Programmiererfahrung verständlich visualisiert.

Aufgrund gestiegener Erwartungen an die Datenanalyse sowie einer vermehrten Heterogenisierung der Quellsysteme bei insgesamt beschleunigtem Datenaufkommen, unter anderem durch die Verwendung von Sensordaten oder Kundentracking im Browser, sind auch die Anforderungen an die Datenintegration größer geworden. Daher kann es in Anwendungsfällen mit schwankenden Ressourcen-anforderungen Sinn ergeben, dass auch die Kapazitäten für den ETL-Prozess skalierbar und flexibel an neue Gegebenheiten anpassbar sind.

Am Markt existieren eine Vielzahl solcher Services. So bieten zum Beispiel Informatica, Oracle und Talend ihre aus der On-Prem-Welt bekannten ETL-Tools jetzt auch als nutzungsbasierten Cloud-Service an. Microsoft bietet mit der Azure Data Factory ebenfalls einen vollverwalteten und für die Cloud optimierten ETL Service an, der es zusätzlich ermöglicht, bereits On-Prem erstellte SQL Server Integration Services (SSIS) Packages auch weiterhin in der Cloud zu nutzen.

Zusätzlich existieren von diesen und anderen Anbietern komplette Datenintegrationssuiten. Diese Suiten bündeln mehrere Cloud Services mit Funktionen für Integration, Monitoring, Metadatenmanagement und weiteren innerhalb einer Plattform.

Im Folgenden werden wir uns die Cloud-nativen ETL Services AWS Glue und Matillion einmal näher ansehen. Sie eignen sich als Beispiele, weil die Services deutlich anhand ihrer Funktionen und Bedienung voneinander abgegrenzt werden können. Ferner bieten beide Services Möglichkeiten zur Datenintegration DWH Services.

AWS Glue

Bei Glue handelt es sich um einen serverlosen, vollverwalteten und Cloud-optimierten ETL Service von AWS, der auf einer skalierbaren Apache Spark Umgebung aufbaut. Vor Erstellung der ETL Jobs analysieren sogenannte Crawler die Quell- und Zieldatenspeicher, um anhand der identifizierten Metadaten die Schemata zu bestimmen. Die Metadaten werden daraufhin in einem Datenkatalog gespeichert.
Anschließend können ETL Jobs für diese Quellen und Ziele erstellt werden. Dazu wird mit dem Verweis auf die Quelle und dem Ziel Code für die Apache Spark Umgebung erzeugt, der die Teilprozesse Extract und Load abdeckt. Jegliche notwendige Transformationslogik muss per Code in Scala oder Python ergänzt werden. Eine GUI, die die Erstellung von ETL Jobs mittels Drag & Drop von verschiedenen Operatoren erleichtert wie in anderen Tools, gibt es in AWS Glue nicht. Die erstellten Jobs können bedarfs-, zeit- oder ereignisgesteuert gestartet werden.

Generell eignet sich Glue besonders gut, wenn Daten bereits in AWS Umgebungen oder in Services wie S3 oder RDS lagern und in andere AWS-Speicher, wie z. B. Redshift, geladen werden sollen. Es ist aber auch möglich, über JDBC-Konnektoren andere Datenquellen anzubinden.

Matillion

Matillion ist ein vom gleichnamigen Hersteller bereitgestellter ETL Service, der speziell für die Beladung von Amazon Redshift, Google Big Query und Snowflake entwickelt wurde. Dabei definiert der Zieldatenspeicher, welche Produktversion zu wählen ist und welche Quellen unterstützt werden. Die Produktversion definiert auch in welcher Cloud-Infrastruktur der Service gehostet wird, zum Beispiel Google Cloud, AWS oder Azure.

Im Gegensatz zu AWS Glue erstellt der Nutzer bei Matillion die ETL-Prozesse in einer webbasierten grafischen Benutzeroberfläche per Drag & Drop. Neben einer Vielzahl an Operatoren ist es möglich, Skripte wie SQL oder Python als Komponente in den ETL-Prozess zu integrieren, falls keiner der Operatoren die gewünschte Transformationslogik erfüllen kann. Bei der Ausführung des ETL Jobs extrahiert Matillion die Quelldaten und lädt diese in die Zielumgebung. Zusätzlich werden die vom Anwender erstellten Transformationslogiken an die Zieldatenbank übermittelt und dort ausgeführt.

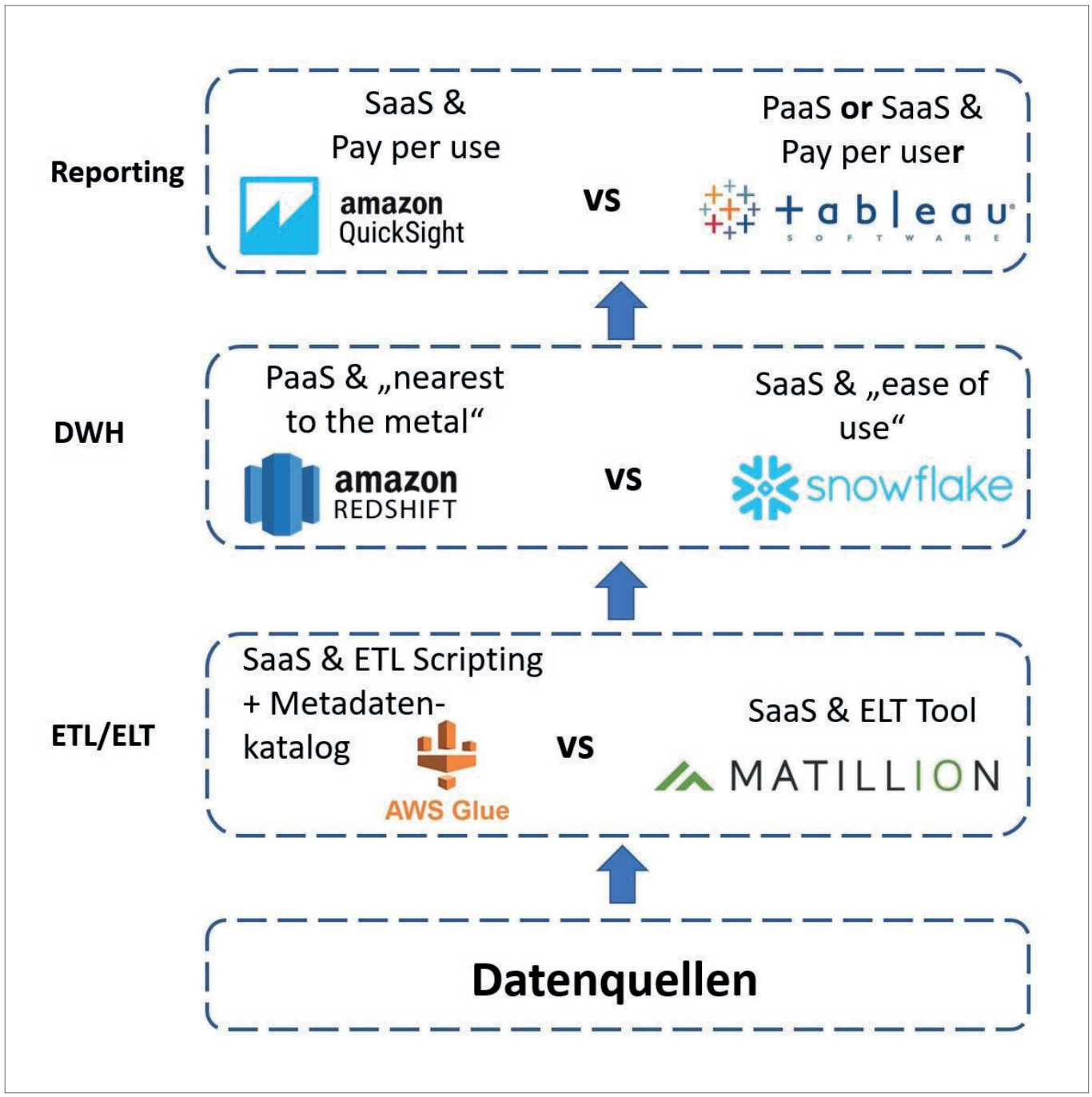

Bild 3: Differenzierung AWS Glue & Matillion.

Genau genommen handelt es sich bei Matillion also um ein ETL-Tool, das die Daten in den Zieldatenspeicher lädt, bevor die Daten transformiert werden. Dies ist besonders im Hinblick auf die Kosten zu beachten. Aufgrund dieser Architektur fallen zusätzlich zu den Kosten für den Service von Matillion noch Kosten für die Rechenressourcen auf der Zieldatenbank an.

Data Warehouse as a Service

Das grundlegende Konzept eines DWHs, das eine themenbezogene, integrierte, konsistente und zeitraumbezogene Speicherung von Daten vorsieht, kann in der Cloud unverändert bestehen bleiben. [7] Zusätzlich bieten einige Anbieter die Möglichkeit, auch nicht relational strukturierte Daten zu speichern und zu analysieren. Diese Funktionalität ersetzt im Normalfall aber nicht den Data Lake, sondern dient dazu, das DWH als konsolidierten Single-Point-of-Truth beizubehalten. Darüber hinaus bringen die Cloud Services im PaaS- oder SaaS-Modell weitere Vorteile wie automatisierte Backups, automatisierte Patches und die simple Replikation der Daten zum Anlegen verschiedener Umgebungen mit sich.

Beispielhaft werden nachfolgend mit Amazon Redshift und Snowflake zwei sehr verbreitete Cloud-DWH-Anbieter vorgestellt. Zusätzlich zur Relevanz auf dem Markt eignen sich die beiden gut zur Demonstration von unterschiedlichen Service-Arten im DWH-Bereich.

Dazu erläutern wir die jeweiligen Architekturen und stellen kurz die relevanten Charakteristiken vor.

Amazon Redshift

Amazon bietet in seiner marktführenden Cloud-Computing-Plattform AWS seit Februar 2013 mit Redshift einen Datenbank-Service an, der für die Entwicklung eines DWHs durch die spaltenorientierte Speicherung optimiert ist. Zusätzlich kann der Nutzer durch die Funktion Redshift Spectrum nicht relational strukturierte Daten abfragen, ohne diese per ETL vorab ins DWH geladen zu haben. Dafür werden die Daten im Objektspeicher-Service Amazon S3 gelagert.

Redshift baut auf der Massive-Parallel-Processing-Cluster-Architektur (MPP) von ParAccel auf, einem ehemaligen Hersteller analytischer DBMS. Gemäß dem Shared-Nothing-Prinzip, enthält dabei jeder Knoten (Compute Node) im Cluster seine eigene CPU sowie Arbeits- und Massenspeicher. Der Arbeits- und Massenspeicher und somit auch die Daten werden auf verschiedene Slices verteilt (vgl. Bild 5). Eine sinnvolle Verteilung der Daten ist für optimale Abfrageergebnisse essenziell und wird durch den Nutzer definiert. Der in der Praxis relevanteste Ansatz in der DWH-Entwicklung ist die Verteilung anhand eines Schlüsselattributs.

Grundsätzlich bietet Redshift die Wahl, das Cluster im Single-Node- oder Multi-Node-Modus zu betreiben. Im Multi-Node-Betrieb existiert neben der gewünschten Anzahl an Compute Nodes ein weiterer Knoten, der sogenannte Leader Node. Dieser koordiniert die Rechenknoten und übernimmt über Schnittstellen die Kommunikation zu anderen Applikationen. Die Anzahl der benötigten Knoten richtet sich dabei einerseits nach Datenvolumen sowie andererseits nach Abfragelast. Im Single-Node-Modus werden die Aufgaben des Leader Nodes von einem einzigen Compute Node im Cluster übernommen.

Die Kopplung von Rechen- und Speicherressourcen in der Architektur verhindert eine separate Skalierung. Das Hinzufügen oder Entfernen von Compute Nodes zum Cluster ermöglicht eine horizontale Skalierung. Außerdem kann durch die Anpassung des Knotentyps der Compute Nodes vertikal skaliert werden. Allerdings geht jede Art der Skalierung mit einer Downtime daher, da die Daten auf den Knoten neu

Jan-Hendrik Groth
Developer Business Intelligence
& Analytics
OPITZ CONSULTING
Deutschland GmbH
www.opitz-consulting.com

verteilt werden müssen. Durch die Kopplung von Speicher- und Rechenressourcen ist es zudem nicht möglich, das Cluster per Knopfdruck oder während Ruhephasen automatisch zu pausieren, um Kosten zu sparen.

Snowflake

Snowflake wird seit 2014 vom gleichnamigen Hersteller bereitgestellt. Seine Verbreitung ist in den letzten Jahren so gestiegen, dass es im Jahr 2019 bereits als Leader in Gartners „Magic Quadrant for Data Management Solutions for Analytics" genannt wurde. Den steilen Aufstieg verdankt Snowflake sicherlich unter anderem seiner Benutzerfreundlichkeit sowie der stetigen Weiterentwicklung seiner Funktionen und der Erweiterung des Ökosystems durch Anbindung diverser ETL- und Analytics-Tools.

Snowflake betreibt im Gegensatz zu seinen Wettbewerbern Amazon, Microsoft, Google, SAP oder Oracle keine eigene Cloud-Infrastruktur. Der Service wird wahlweise auf AWS, Azure oder Google Cloud Ressourcen betrieben. Zudem weist Snowflake im Vergleich zu anderen Anbietern keine RDBMS-On-Prem-Vergangenheit auf. Der Service wurde also von Beginn an für die Cloud-Nutzung konzipiert. Dies zeigt sich in mehreren Facetten seiner Drei-Schichten-Architektur:

Die Speicherschicht wird mit Objektspeichern realisiert, während Rechenschicht und Serviceschicht über Recheninstanzen der jeweiligen Cloud-Infrastruktur realisiert werden. Der Anwender kann in Snowflake mehrere unabhängige Rechencluster, sogenannte Virtual Warehouses für unterschiedliche Aufgaben oder Nutzergruppen anlegen. Dabei besteht jedes Virtual Warehouse aus einem oder mehreren Clustern von Recheninstanzen. Ein Beispiel:

Für die BI-Abteilung wird ein Virtual Warehouse der Größe S betrieben. Das besteht aus zwei Servern bzw. Recheneinheiten pro Cluster. Eine genaue Spezifikation darüber, wie viele CPUs, wie viel RAM etc. sich hinter den Servern verbergen, wird nicht offengelegt. Aufgrund häufig parallellaufender Abfragen skaliert das Warehouse zwischen ein bis maximal drei Clustern automatisch. Demnach werden bei höchster Auslastung drei Cluster mit je zwei Servern für die BI-Abteilung betrieben.

Beim Anlegen bestimmt der Anwender die Größe des Warehouses, die Anzahl der Cluster und die Einstellung zur automatischen Skalierung sowie zur Abschaltung bzw. automatischer Wiederaufnahme des Betriebs. Die Größe der Warehouses rangiert von XS bis 4XL und bezieht sich auf die vertikale Skalierung. Die Größe beeinflusst die Ausführungszeit je Abfrage. Die Anzahl der Cluster rangiert von 1 bis 10, ermöglicht eine horizontale Skalierung und nimmt so Einfluss auf die Parallelisierung mehrerer Abfragen innerhalb eines Warehouses. Die Trennung der Speicher- und Rechenressourcen und die damit ermöglichte automatische Pausierung der Rechenressourcen in Ruhephasen können zur Kostenreduktion beitragen.

Wie bei anderen Tools übernimmt auch bei Snowflake die Serviceschicht die Kommunikation zu externen Anwendungen. Außerdem sind die Recheneinheiten der Serviceschicht zuständig für Datenmanagement und -verteilung, Transaktionsmanagement und -optimierung, Security, Metadatenverwaltung sowie die Snowflake eigene Data-Sharing-Funktionalität.

Darüber hinaus nutzt der Service Algorithmen zum Clustern und Verteilen der Daten, um diese möglichst komprimiert in sogenannte Mikropartitionen abzuspeichern. Dem Anwender wird somit der Aufwand der Partitionierung, Indizierung und Festlegung eines Verteilungsstils abgenommen. Darauf aufbauend nutzt der Service Metadaten der gespeicherten Daten und Mikropartitionen für die Optimierung von Abfragen. Betrachtet man Bild 2 ließe sich Amazon Redshift dem PaaS-Modell und Snowflake dem SaaS-Modell

zuordnen. Snowflake bietet, neben der Datenbank, eine Web-Oberfläche zur Interaktion mit der Datenbank an und befreit den Anwender größtenteils von Aufgaben, die Performanceoptimierung, Ressourcen-Skalierung und Datenverteilung betreffen. In Redshift hingegen ist die Datenbank nur über Third-Party Clients (SQL) oder Tools abfragbar, Datenverteilung, Performanceoptimierung sowie Ressourcenskalierung liegen beim Anwender.

Zusammengefasst kann man sagen, dass Redshift eher eine individualisierbare Datenbank ist, die bezüglich ihrer Nutzung einer On-Prem-Datenbank nahekommt. Snowflake hingegen verfolgt verstärkt einen „Ease of use"- Ansatz.

Reporting as a Service

Die Funktionen und die Bedienung von On-Prem Reporting Tools gleichen, aufgrund ihrer weborientierten Oberfläche, häufig ihren Pendants aus der Cloud. Besonders naheliegend erscheint die Nutzung eines Cloud Services für die Erstellung von Reports, Dashboards und Self-Service-BI-Anwendungen, wenn die auszuwertenden Daten bereits in der Cloud lagern.
Nachfolgend sollen die Reporting Services Tableau und Amazon Quicksight kurz vorgestellt und voneinander abgegrenzt werden. Auch hierbei handelt es sich lediglich um einen Bruchteil der am Markt vorhanden Cloud Services.

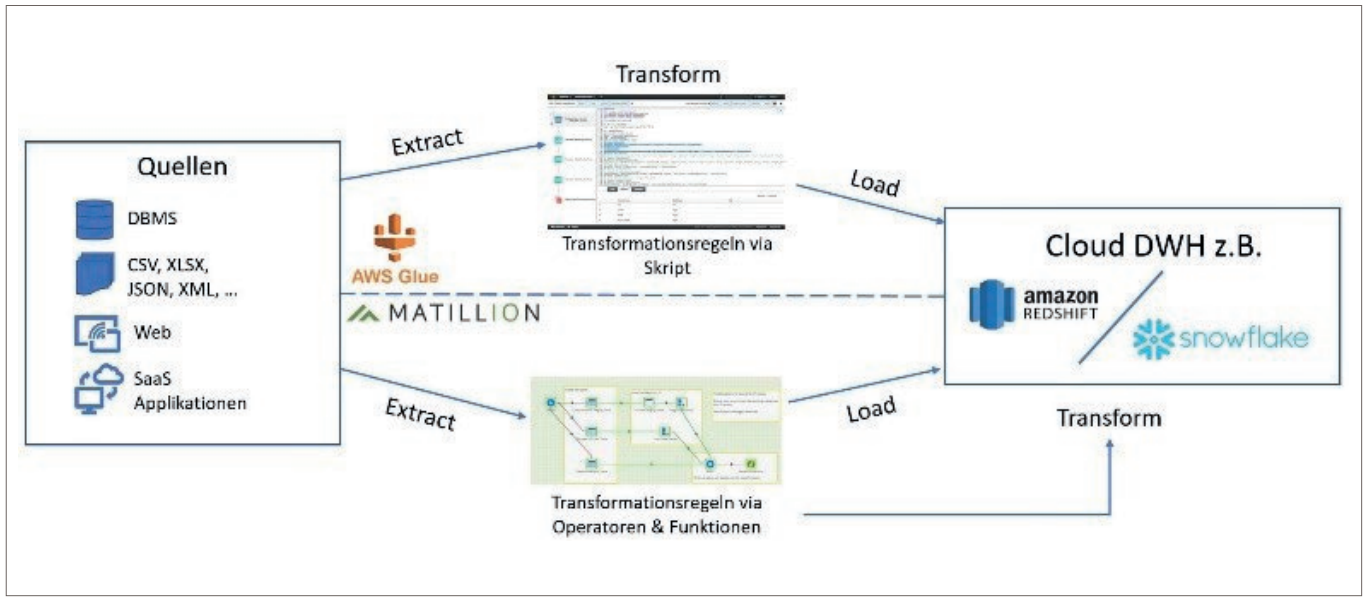

Bild 4: Auswahl BI Cloud Services.

Tableau

Tableau ist neben Microsoft, Qlik und ThoughtSpot einer der Leader in Gartners jährlicher Marktanalyse der Analytics & BI Plattformen. Neben seiner On-Prem Desktop-Anwendung und der Server-Version zum kollaborativen Arbeiten, bietet Tableau ebenfalls Cloud-Bereitstellungen an. Der Nutzer kann dabei zwischen Tableau Online als vollverwaltetem SaaS-Angebot oder dem Deployment der Server Variante in der AWS, Azure oder Google Cloud Plattform wählen. Das Angebot bietet dem Anwender die Möglichkeit, sein präferiertes Bereitstellungmodell frei zu wählen. Bei beiden Varianten wird monatlich pro Nutzer und jeweiliger Nutzerlizenz abgerechnet (Creator, Explorer oder Viewer), wobei die Gebühren für Tableau Online höher sind.

Mit Tableau können Nutzer aussagekräftigen Visualisierungen und Dashboards einfach und effizient erstellen. Das Tool besticht dabei besonders durch eine große Zahl anbindbarer Datenquellen und

vielfältige Visualisierungsmöglichkeiten der Daten. Dazu bietet Tableau fortgeschrittene Möglichkeiten der Datenexploration und mit der Storytelling-Funktion lassen sich Datenpräsentationen interaktiv und erkenntnisbringend aufbereiten.

Durch die im Sommer 2019 bekanntgewordene Übernahme durch den CRM-Softwarekonzern Salesforce, stellt sich die Frage, inwieweit sich das Angebot Tableaus zukünftig ändern wird. Tableau soll weiterhin als eigenständiges Unternehmen arbeiten. Denkbar ist, dass Tableau durch Funktionen aus anderen Sparten von Salesforce ergänzt wird und umgekehrt. Besonders nahe liegen dabei Ergänzungen mit der in Salesforce befindlichen analytischen Plattform Einstein.

Amazon Quicksight

Mit Quicksight stellt Amazon seit 2016 ein vollverwaltetes und serverloses BI-Tool in AWS bereit. Mit diesem Service soll die Erstellung von Visualisierungen und interaktiven Dashboards einsteigerfreundlich sowie die Analyse der Daten sehr performant möglich sein. Darüber hinaus grenzt sich Amazon durch die „Pay per use"-Preispolitik von Wettbewerbern ab und wirbt mit großen Kostenvorteilen für den Kunden.

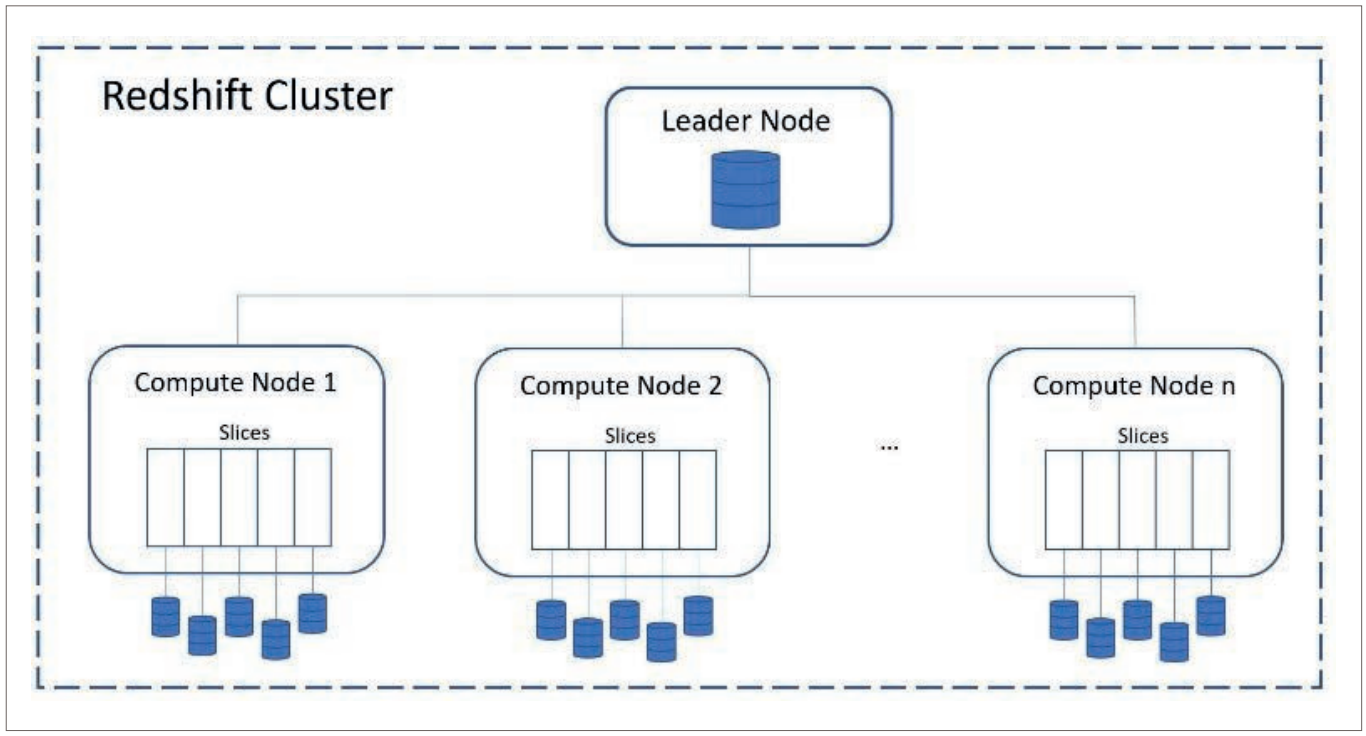

Bild 5: Architektur Amazon Redshift.

Die Daten in Quicksight kommen zum Großteil aus den aktuell verfügbaren Datenquellen von Amazon. Zusätzlich können weitere relationale Datenbanken wie MySQL, Microsoft SQL Server und Snowflake angebunden werden. Obendrein werden Dateiformate wie CSV, JSON und XLSX (Excel) sowie SaaS-Quellen wie Salesforce und Twitter unterstützt.

Das Kernstück von Amazon Quicksight bildet die für die Cloud konzipierte In-Memory Engine „SPICE". Die Daten werden komprimiert sowie spaltenbasiert gespeichert und im Arbeitsspeicher verarbeitet.
Bei der Erstellung der Visualisierungen werden die Nutzer durch die Autograph-Funktionalität unterstützt. Diese schlägt auf Basis der Dateneigenschaften, die am besten geeignete Visualisierungsform vor. Neben der bloßen Erstellung von Visualisierungen und Dashboards bietet Quicksight, genau wie einige namhafte Wettbewerber, eine Story-Funktion zur Vermittlung des Kontextes der jeweiligen Analyse.

Bisher bietet Quicksight allerdings noch keinen ähnlich großen Funktionsumfang wie seine historisch gewachsenen Wettbewerber. Dennoch kann sich die Betrachtung des Services vor allem dann lohnen, wenn der Großteil der Analytics-Architektur schon in der AWS Cloud realisiert wird. Zudem kann das „Pay-per-Use"-Preismodell von Quicksight in vielen Nutzungsszenarien für den Anwender relativ kostengünstig sein.

Die Wahl des geeigneten Cloud Services

Die Möglichkeiten, die der Analytics-Markt bietet, sind mit der Verlagerung der Infrastrukturen in die Cloud noch vielfältiger geworden. Gleichzeitig macht die Vielfalt der Modelle, Anbieter und Services die individuelle Entscheidung für ein bestimmtes Tool sehr komplex.

Ein Beispiel dafür sind Preismodelle, die häufig weniger greifbar sind, als eine einmalige Investition in Hardware und Lizenzen zuzüglich Personalkosten. Hierbei können die Preiskalkulatoren der Anbieter einen ersten Anhaltspunkt verschaffen. Doch ist es ratsam, zunächst mögliche Anwendungsfälle zu evaluieren. Andernfalls liefern die Kalkulatoren aufgrund vieler unbekannter Variablen häufig nur bedingt aussagekräftige Ergebnisse.

Eine detaillierte Anforderungsanalyse und ein anschließender Proof of Concepts helfen hier weiter. Mit ihrer Hilfe lassen sich die Eignung eines oder mehrerer evaluierter Services für die definierten Anwendungsfälle valide bewerten. Außerdem kann aus den Ergebnissen geschlossen werden, in welchem Umfang Speicher- und Rechenressourcen im realen Geschäftsleben benötigt werden. Dies hilft zum einen bei der korrekten Skalierung der Ressourcen und kann zum anderen dazu beitragen, die Kosten geringer zu halten.

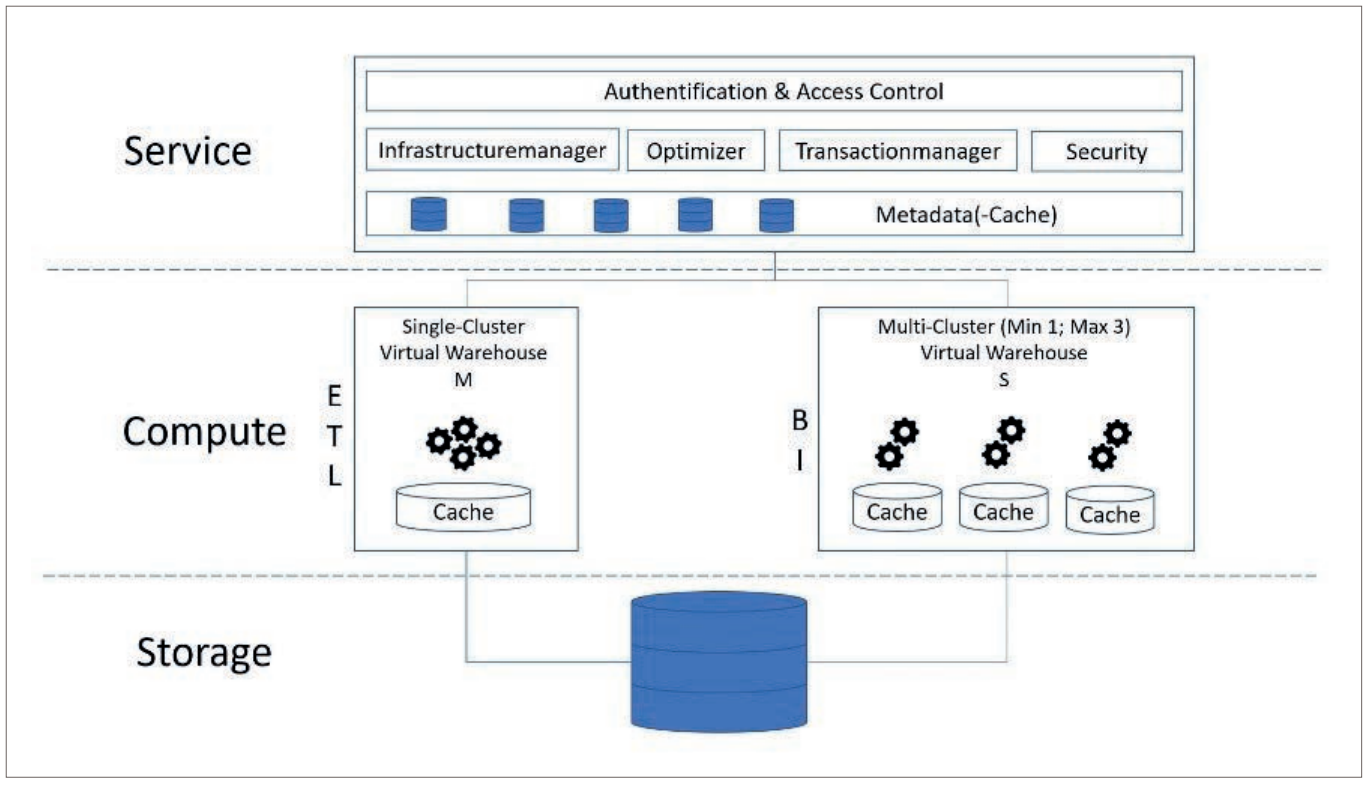

Bild 6: Architektur Snowflake.

Manche Anbieter gewähren Rabatte, wenn eine vorab bestimmte Größe an Ressourcen über einen definierten Zeitraum vom Kunden vertraglich bezogen wird. Nur sollte darauf geachtet werden, die Ressourcen nicht zu großzügig vorzubestellen und damit die Scale-up- und Scale-down-Vorteile der Cloud-Infrastruktur aufs Spiel zu setzen!

Die fachlichen Kriterien der Anforderungsanalyse für den jeweiligen Cloud Service entsprechen denen, die auch für On-Prem-Produkte anzuwenden sind. Außerdem fließen auf technischer Seite die für Cloud Services spezifischen Merkmale in die Bewertung mit ein. Eine Checkliste für die Auswahl liefert Tabelle 1.

Analytics & Cloud – Chance oder Risiko?

„Wo viel Licht ist, ist starker Schatten [...]“

Johann Wolfgang von Goethe in Götz von Berlichingen mit der eisernen Hand

Wie wir gesehen haben ist Cloud Computing mittlerweile auch im Analytics-Bereich mehr als nur ein Trend. Ausschlaggebend dafür sind unter anderem das steigende Interesse und die Wertschätzung moderner Data-Analytics-Anwendungen in Unternehmen. Viele Anwender erkennen die Chancen der Cloud und wollen diese nutzen. Jedoch gehen diese Chancen auch mit Risiken einher.

Merkmal	Mögliche Evaluierungskriterien
Skalierbarkeit	Skaliert der Service automatisch? Skalieren Speicher- und Rechenressourcen unabhängig voneinander? In welche Richtung (horizontal, vertikal) kann skaliert werden? Welche Auswirkungen hat das?
Elastizität	Wie schnell wird das System an vorgenommene Skalierungen angepasst? Geht die Skalierung mit einer Downtime einher?
Kosten	Welches Preismodell liegt vor? Worauf basiert die Berechnung?
Sicherheit	Wie sind die Daten vor Verlust geschützt? Wie sind die Daten verschlüsselt? Welche Kommunikationsprotokolle werden verwendet? Welche Möglichkeiten zur Authentifizierung und zur Zugriffsbeschränkung gibt es?
Infrastruktur	Wird die Infrastruktur in virtualisierter Form (Standard) oder in Form von „Bare Metal“-Ressourcen bereitgestellt? Lassen sich die Ressourcen physisch/logisch von denen anderer Cloud-Kunden abtrennen oder besteht die Möglichkeit dazu?
SLA	Gibt es eine zugesicherte Verfügbarkeit? Welchen Umfang bietet der Kundensupport?
Ein- und Austrittszenarien	Wie kommen bestehende Datenbestände in die Cloud? Welche Folgen und Tücken hat ein Anbieterwechsel oder der Rückzug aus der Cloud?

Tabelle 1: Evaluierungsmerkmale und -kriterien für Cloud Analytics-Services

Diese steigen häufig mit dem Grad der Nutzung von Cloud Services an. Daher gilt es, die Risiken fortlaufend zu evaluieren und mit den möglichen Chancen der Nutzung abzuwägen. Nachfolgend werden etwaige Chancen und Risiken beschrieben und gegenübergestellt.

Bestehende On-Prem-Architekturen lassen sich nur unter sehr großem Aufwand an neue und wechselhafte Anforderungen anpassen. Hingegen kann zum Beispiel eine Testumgebung mithilfe von Cloud-Ressourcen schnell und unabhängig vom häufig ausgelasteten internen IT-Betrieb bereitgestellt werden.

Diese simple und schnelle Bereitstellung von Cloud Services fördert unbestritten die Agilität in der Entwicklung neuer analytischer Anwendungen. Doch Vorsicht: Bleibt hierbei die Frage der Überführung von Entwicklungs- und Testumgebungen in den produktiven Betrieb ungeklärt, kann dies zu nachgelagerten Barrieren führen. Außerdem kann die gewonnene Agilität zu einem Wildwuchs der Systemarchitektur führen, wenn keine regulatorischen Maßnahmen einen Rahmen vorgeben. Um Analytics in der Cloud erfolgreich einzuführen und umzusetzen braucht es also eine dedizierte IT- und Cloud-Strategie.

Das Mehr an Flexibilität ist mit Sicherheit ein großer Pluspunkt der Cloud Services. Durch die Anpassung der Ressourcen an den vorliegenden Bedarf können sowohl Hochlast-Szenarien performant bedient als auch unterausgelastete On-Prem-Rechenzentren vermieden werden.

Daraus folgt eine weitere Chance auf Seiten der Kosten. Die an die vorliegenden Bedarfe gekoppelten Kosten können besonders bei schwankenden Auslastungen zu Nutzungspotenzialen und Einsparungen führen. Umgekehrt ist die Cloud häufig dann teurer, wenn die Ressourcen rund um die Uhr vollumfänglich bereitstehen müssen.

Bei dem Blick auf die Kosten sollte allerdings miteinbezogen werden, dass der Cloud-Anbieter den Anwender von verschiedenen administrativen Aufgaben entbindet. Dabei hilft ein Blick über die bloßen Hard- und Softwarekosten hinaus auf die Gesamtkosten des Betriebs: Hier greift das Konzept der Total Cost of Ownership (TCO).

Datenschutz

Auf die häufig angeführten Bedenken bezüglich Datenschutz, Sicherheit und (Hoch-)Verfügbarkeit haben die großen Anbieter passende Antworten gefunden. Anwender können wählen, in welcher geografischen Region sich das Rechenzentrum befinden soll, in dem die Daten verarbeitet und gespeichert werden. Zusätzlich sind die physischen und logischen Sicherheitskonzepte heute umfänglicher als in den meisten Unternehmen. Und auch wenn die Hersteller keine hundertprozentige Verfügbarkeitsgarantie geben, ist die Verfügbarkeit der Services durch verschiedene Ausfallsicherheitsmechanismen doch sehr hoch. Für die meisten eigenen Rechenzentrum klingt dies ohnehin noch utopisch. Die Cloud Services sind hier also durchaus im Vorteil.

Auch hinsichtlich der Performance gewähren die Public-Cloud-Anbieter leider in der Regel keine vertraglichen Zusicherungen. Zudem sind die Gründe für eine schwankende Performance kundenseitig häufig nicht eindeutig feststellbar. Daher sind mehrere Testläufe konkreter Anwendungsfälle im Rahmen eines PoCs empfehlenswert, um einen Mittelwert der realen Performance bestimmen zu können.

Nicht von der Hand zu weisen sind die entstehenden Abhängigkeiten gegenüber dem Anbieter. Bei auftretenden Problemen, die nicht auf Anwenderseite gelöst werden können, besteht eine Abhängigkeit vom Anbietersupport und bei den Pay-per-Use-Bezahlmodellen ist der Anwender abhängig von der Preispolitik des Anbieters.

Ebenfalls ist ein Rückzug aus der Cloud zurück in die On-Prem-Welt oder ein Umzug in eine andere Cloud-Plattform neben technischen Unwägbarkeiten meist auch mit finanziellen Nachteilen verbunden. Während das Befüllen der Cloud-Plattform mit Daten häufig kostenfrei ist, verlangen Anbieter in der Regel Gebühren für das Extrahieren von Daten, sogenannte „Data-Egress-Gebühren".

Da der Aus- beziehungsweise Umstieg also meist weniger leicht von der Hand geht, als der Einstieg, sollte die Entscheidung, den Aufbau eines unternehmensweiten analytischen Systems mittels Cloud Services umzusetzen, immer wohl überlegt sein. Auch aus diesem Grund sollte die Auswahl der Komponenten des analytischen Systems auf einer unternehmensweiten Cloud-Strategie basieren. Die Abhängigkeit von einem einzelnen Anbieter kann dabei übrigens zum Beispiel mithilfe von Hybrid- oder Multi-Cloud-Strategien reduziert werden.

Schlussendlich bleibt festzuhalten, dass sich der Einsatz von Cloud Services im Analytics-Kontext durchaus positiv auswirken kann. Die erfolgreiche Umsetzung ist allerdings von vielen verschiedenen Gesichtspunkten abhängig und bedarf zusätzlicher flankierender Maßnahmen, wie der Entwicklung einer geeigneten Strategie und Governance. Optimaler Weise wird die IT- und Cloud-Strategie dabei unternehmensweit getroffen und nicht siloartig im Analytics-Bereich.

Chance	Risiko
Erhöhte Agilität, die Innovationen fördert	Neue organisatorische Aufwände
Gesteigerte Flexibilität durch eine bessere Skalierbarkeit	Anbieterabhängigkeit, die sich in Preisen und Support bemerkbar machen kann.
Kosteneinsparungen durch Pay-per-Use-Bezahlmodell	Anbieterwechsel ist nicht immer so einfach möglich und bedeutet einen hohen Migrationsaufwand
Minimierung des Aufwands für IT-Administration	Fehlende Transparenz und Sicherheit bei Performance und Latenz

Tabelle 2: Chancen und Risiken der Cloud-Nutzung.

Fazit

Die Cloud bietet zahlreiche Vorteile für BI- und Analytics-Anwendungen – dennoch ist sie nicht die Patentlösung für jede Organisation und jeden Anwendungsfall. Die Möglichkeiten sind durch die Cloud vielfältiger geworden, die Auswahl der richtigen Lösung damit zugleich komplexer. Und wie so häufig liegt die Kunst darin, die beste Lösung in einer ganzheitlichen Betrachtung zu finden.

Für den Fall, dass Ressourcenbedürfnisse oder verschiedene Nutzungsszenarien noch Unbekannte in der Betrachtung sind, lohnt sich auf jeden Fall der Blick auf Cloud-Lösungen. Dort wird das Ausprobieren erleichtert und der Anwender kann die Wahl anhand einer erprobten Lösung treffen.

Jan-Hendrik Groth

Quellen

[1] Gartner Hype Cycle 2005 – 2018: https://www.computerwoche.de/a/gartner-trends-im-reality-check,3070089 (abgerufen am 08.11.2019)

[2] Entwicklung der Nutzung von cloud-basierten BI & Analytics Lösungen in Unternehmen: https://www.forbes.com/sites/louiscolumbus/2018/04/08/the-state-of-cloud-business-intelligence-2018/#7789ffce2180 (abgerufen am 08.11.2019)

[3] Vgl. Baars, H.: Die Cloud als Agilitätshebel für Business Intelligence & Analytics in BI & Analytics in der Cloud; dpunkt.verlag; 2018; S. 38

[4] https://www.heise.de/newsticker/meldung/Bundespolizei-speichert-Bodycam-Aufnahmen-in-Amazons-AWS-Cloud-4324689.html (abgerufen am 08.11.2019)

[5] https://barc.de/news/die-wichtigsten-trends-fur-bi-anwender-2019-sind-keine-bi-trends (abgerufen am 08.11.2019)

[6] https://datagravitas.com/2010/12/07/data-gravity-in-the-clouds/ (abgerufen am 08.11.2019)

[7] Vgl. Inmon, W.H.: Building the Data-Warehouse, 3. Auflage; New York; 2002; S. 31