LERNEN LEICHT GEMACHT

2. Snowflake Sonderausgabe

Cloud Data Warehousing

für dummies®

Was ist ein Cloud Data Warehouse?

Vergleich von Data-Warehouse-Lösungen

Auswahl eines Cloud Data Warehouse

Präsentiert von:

snowflake®

Joe Kraynak

David Baum

Über Snowflake

Snowflake hatte von Anfang an eine klare Vision: Modernes Data Warehousing für alle Datennutzer effektiv, kostengünstig und zugänglich zu machen. Snowflake macht Unternehmen durch unmittelbare Elastizität, einen sicheren Datenaustausch, eine sekundengenaue Abrechnung und über mehrere Clouds zu datengesteuerten Unternehmen. Da herkömmliche On-Premise- und Cloud-Lösungen in dieser Hinsicht Nachteile aufwiesen, entwickelte Snowflake ein neues Produkt mit einer neuen, für die Cloud geschaffenen Architektur, das die Vorteile von Data Warehousing mit der Flexibilität von Big-Data-Plattformen und der Elastizität der Cloud kombiniert – zu einem Bruchteil der Kosten herkömmlicher Lösungen. Snowflake: Your Data, No Limits.

Für weitere Informationen besuchen Sie **Snowflake unter snowflake.com.**

Cloud Data Warehousing

2. Snowflake Sonderausgabe

Joe Kraynak und David Baum

für dummies®

Cloud Data Warehousing Für Dummies®, 2. Snowflake Sonderausgabe

Veröffentlicht von
John Wiley & Sons, Inc.
111 River St., Hoboken, NJ 07030-5774
www.wiley.com

Allgemeine Informationen zu unseren anderen Produkten und Dienstleistungen oder zur Erstellung eines individuellen *Für Dummies*-Buches für Ihr Unternehmen oder Ihre Organisation erhalten Sie von unserer Abteilung Business Development in den USA unter Tel. 877-409-4177, E-Mail: info@dummies.biz, oder besuchen Sie www.wiley.com/go/custompub. Für Informationen zur Lizenzierung der *Für Dummies*-Marke für Produkte oder Dienstleistungen kontaktieren Sie bitte: BrandedRights&Licenses@Wiley.com.

ISBN 978-1-119-71403-3 (pbk); 978-1-119-71404-0 (ebk)

Hergestellt in den Vereinigten Staaten

10 9 8 7 6 5 4 3 2 1

Danksagung des Verlags

Wir sind stolz auf dieses Buch und auf die Personen, die daran mitgewirkt haben. Für Informationen zur Erstellung eines individuellen *Für Dummies*-Buches für Ihr Unternehmen oder Ihre Organisation kontaktieren Sie bitte info@dummies.biz oder besuchen Sie www.wiley.com/go/custompub. Für Informationen zur Lizenzierung der *Für Dummies*-Marke für Produkte oder Dienstleistungen kontaktieren Sie bitte: BrandedRights&Licenses@Wiley.com.

Die folgenden Personen haben dabei geholfen, dieses Buch auf den Markt zu bringen:

Development Editor: Nicole Sholly

Project Editor: Martin V. Minner

Executive Editor: Steve Hayes

Editorial Manager: Rev Mengle

Business Development Representative: Karen Hattan

Production Editor: Mohammed Zafar Ali

Snowflake Contributors Team: Vincent Morello, Clarke Patterson, Leslie Steere, Kent Graziano

Inhaltsverzeichnis

Einführung

Als Führungskraft, Manager oder Analyst ist Ihnen klar, dass Wissen Macht ist, und dass die genaue und zeitnahe Analyse von Daten die wertvollen Einblicke liefert, die Sie benötigen, um gut informierte Entscheidungen zu treffen und einen Wettbewerbsvorteil zu erzielen. Die meisten Unternehmen verfügen heutzutage über größere Bestände relevanterer Daten als jemals zuvor. Dies umfasst eine Vielzahl interner und externer Quellen wie Data Marts, cloudbasierte Anwendungen und maschinell generierte Daten.

Leider wird die Data-Warehouse-Architektur der letzten 30 Jahre zunehmend von der Last dieser extrem großen und vielfältigen Datenbestände erdrückt. Analysten müssen oft 24 Stunden oder länger warten, bis die von ihnen benötigten Daten in das Data Warehouse gelangen und ihnen zur Analyse zur Verfügung stehen. Mitunter müssen sie sogar noch länger warten, bis sie auf der Grundlage dieser Daten komplexe Abfragen ausführen können. In vielen Fällen sind die für die Verarbeitung und Analyse dieser Daten erforderlichen Speicher- und Rechenressourcen nicht ausreichend. Dies führt dazu, dass Systeme hängenbleiben oder abstürzen. Um dies zu vermeiden, werden Benutzer und Workloads in eine Warteschlange gestellt und müssen weitere Verzögerungen in Kauf nehmen. In jüngerer Zeit sind alternative Ansätze wie Data Lakes entstanden, d. h. große Datenspeicher unterschiedlicher Form. Doch auch diese Lösungen haben ihre Grenzen.

Um effizient und wettbewerbsfähig zu bleiben, müssen Unternehmen in der Lage sein, das Potenzial der riesigen Datenmengen auszuschöpfen, die ständig generiert werden, und mithilfe dieser Daten komplexe Analysen durchzuführen. Glücklicherweise brachte die Kommerzialisierung des Cloud Computing, die vor mehr als zehn Jahren ihren Anfang nahm, zahlreiche Fortschritte in Bezug auf Computerhardware, -architektur und -software mit sich, die Ihrem Unternehmen dabei helfen können, diese Herausforderung zu meistern – und die Ihre Erwartungen wahrscheinlich übertreffen werden.

Über dieses Buch

Willkommen bei der zweiten Ausgabe von *Cloud Data Warehousing Für Dummies*. In diesem Buch erfahren Sie, wie sich Ihr Unternehmen die Macht großer Datenmengen bequem und kostengünstig zunutze machen kann, um seine Effizienz zu steigern und Rohdaten in wertvolle Business Intelligence zu verwandeln.

Eine größere Fülle von Daten öffnet die Tür zu mehr und größeren Möglichkeiten, die jedoch fast immer mit ebenso großen Herausforderungen verbunden sind. Um diese großen Chancen zu nutzen, müssen Sie eine Data-Warehouse-Lösung implementieren, die Daten in unterschiedlichen Formaten speichern und organisieren, einen bequemen Zugriff auf diese Daten ermöglichen und die Geschwindigkeit der Analyse verbessern kann – und das so kostengünstig wie möglich. Dieses Buch zeigt Ihnen, wie es geht.

In diesem Buch verwendete Symbole

In diesem Buch verwenden wir die folgenden Symbole, um Tipps und wichtige Punkte hervorzuheben, die man sich merken sollte:

TIPP

Die hierin enthaltenen Tipps sollen Ihnen zeigen, wie Sie bestimmte Aufgaben einfacher erledigen können, und weisen Sie auf bessere Möglichkeiten zur Nutzung von Cloud Data Warehousing in Ihrem Unternehmen hin.

ERINNERUNG

Dieses Symbol hebt Konzepte hervor, an die Sie sich erinnern sollten, wenn Sie sich in das Thema Cloud Data Warehousing vertiefen.

FALLSTUDIEN

Die Fallstudien in diesem Buch zeigen, wie andere Unternehmen Cloud Data Warehousing eingesetzt haben, um Geld zu sparen und die Geschwindigkeit und Leistung ihrer Datenanalysen erheblich zu verbessern.

Zusätzliche Informationen

Wenn Ihnen gefällt, was Sie in diesem Buch gelesen haben, und wenn Sie mehr zu den darin behandelten Themen erfahren möchten, laden wir Sie zu einem Besuch der Website `www.snowflake.com` ein. Hier können Sie weitere Informationen über das Unternehmen und seine Angebote finden, Snowflake kostenlos testen, Einzelheiten zu verschiedenen Plänen und Preisen erhalten, Webinare ansehen, auf Pressemitteilungen zugreifen, sich über bevorstehende Veranstaltungen informieren, auf Dokumente und andere Unterstützung zugreifen und mit Snowflake Kontakt aufnehmen. Snowflake würde sich freuen, von Ihnen zu hören!

IN DIESEM KAPITEL

- Data Warehousing: damals und heute
- Die Vorteile eines Cloud Data Warehouse
- Wo sich Cloud Data Warehousing in die moderne Wirtschaft einfügt

Kapitel 1
Cloud Data Warehousing: eine Einführung

In der einen oder anderen Form gibt es Cloud Computing und Software-as-a-Service (SaaS) schon seit mehreren Jahrzehnten. Cloud Data Warehouse-as-a-Service (DWaaS) ist allerdings erst seit kurzem eine Alternative zum herkömmlichem On-Premise Data Warehousing und ähnlichen Lösungen. Warum jetzt? Was hat sich verändert? In diesem Kapitel werden wir diese und weitere Fragen beantworten.

Zuerst wollen wir definieren, was ein Data Warehouse ist, und die Entwicklung des Data Warehousing näher beleuchten, um zu zeigen, wie diese Technologie ihren Weg in die Cloud gefunden hat. Dann werden wir uns ansehen, wie Unternehmen von DWaaS in der Cloud profitieren können, und erklären, warum immer mehr Unternehmen auf Cloud Data Warehousing setzen, um in der heutigen datengesteuerten Wirtschaft wettbewerbsfähig zu bleiben.

Was ist ein Data Warehouse?

Ein *Data Warehouse* ist ein Computersystem zur Speicherung und Analyse von Daten, um Trends, Muster und Korrelationen aufzudecken, die Informationen und wertvolle Einblicke liefern. Unternehmen verwenden Data Warehouses zur Speicherung und Integration von Daten aus ihren internen Quellen (in der Regel Transaktionsdatenbanken), u. a. in

den Bereichen Marketing, Vertrieb, Produktion und Finanzen. Das Data Warehouse entstand, als immer mehr Unternehmen erkannten, dass die Analyse von Daten direkt aus diesen Transaktionsdatenbanken ihre Transaktionstätigkeit und die für die Analyse dieser Daten erforderlichen Workloads zu sehr verlangsamte (oder sogar zum Absturz brachte). Man begann, alle diese Daten zur Analyse in einem Data Warehouse zu duplizieren, damit sich die Datenbank auf die eigentlichen Transaktionen konzentrieren konnte.

Im Laufe der Jahre gingen die Datenquellen immer mehr über interne Geschäftsvorgänge und externe Transaktionen hinaus. Heute umfassen sie exponentiell größere Mengen unterschiedlicher Daten mit unterschiedlichen Geschwindigkeiten von Websites, Mobiltelefonen und Mobile Apps, Online-Spielen, Online-Banking-Anwendungen und sogar Maschinen. Seit einiger Zeit erfassen Organisationen auch enorme Datenmengen von Geräten des Internet of Things (IoT).

Die Entwicklung des Data Warehousing

In der Vergangenheit erfassten Unternehmen Daten in klar definierten, hochstrukturierten Formen und in relativ gut vorhersehbarem Tempo und Volumen. Selbst als die Geschwindigkeit älterer Technologien immer weiter zunahm, wurden der Datenzugriff und die Datennutzung sorgfältig kontrolliert und begrenzt, um eine akzeptable Leistung für jeden Benutzer zu gewährleisten, da die Rechenleistung und der Speicherplatz vor Ort knapp waren und diese Ressourcen nicht problemlos erweitert werden konnten. Unternehmen mussten daher sehr lange Analysezyklen in Kauf nehmen.

Die Zeiten haben sich jedoch geändert (siehe Abbildung 1-1). Der technische Fortschritt ermöglicht Unternehmen nun, wichtige Geschäftsentscheidungen auf der Grundlage großer Datenmengen zu treffen.

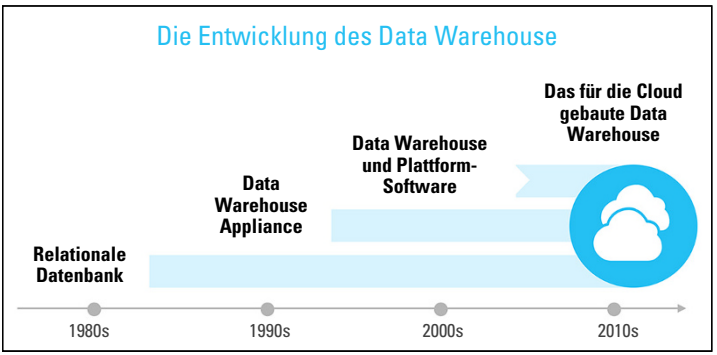

ABBILDUNG 1-1: Herkömmliche Systeme haben zur Entstehung des Cloud Data Warehouse geführt.

Dabei handelt es sich keineswegs nur um Marktführer oder reife Unternehmen. Auch kleinere, agile Markteinsteiger können etablierte Branchen innerhalb weniger Monate oder Jahre verwandeln. Sie verwenden Daten, um neue Möglichkeiten aufzudecken und Produkte und Dienstleistungen zu entwickeln, die die Interaktion von Einzelhändlern und Anbietern mit ihren Kunden verändern.

Die Grenzen des herkömmlichen Data Warehousing

Herkömmliche Data Warehouses waren nicht darauf ausgelegt, die Menge, Vielfalt und Geschwindigkeit der heute verfügbaren Daten zu bewältigen. Neueren, zur Behebung dieser Mängel entwickelten Systemen fällt es schwer, Unternehmen den Datenzugriff und die Analysefunktionen zu bieten, die sie heute benötigen. Die aktuellen Herausforderungen liegen auf der Hand:

- Die Anzahl und Vielfalt der verfügbaren Datenquellen nimmt ständig zu. Daher müssen die unterschiedlichsten Datenstrukturen an einem einzigen Ort koexistieren, um eine umfassende und kostengünstige Analyse zu ermöglichen.
- Bei traditionellen Architekturen ist der Wettbewerb zwischen Benutzern und Datenintegrationsaktivitäten vorprogrammiert und es ist schwierig, gleichzeitig neue Daten in das Data Warehouse zu laden und den Benutzern eine angemessene Leistung zu bieten.
- Es ist immer noch üblich, Daten in Batches und in bestimmten Intervallen zu laden. Viele Unternehmen benötigen jedoch kontinuierliches Laden (*Micro-Batching*) und Streaming von Daten (*Instant Loading*).
- Es ist kostspielig, mühsam und langsam, ein herkömmliches Data Warehouse zu skalieren und an die steigenden Storage- und Workload-Anforderungen anzupassen – manchmal sogar unmöglich.
- Die neueren, alternativen Datenplattformen sind oft komplex und erfordern spezielle Fähigkeiten und viel Anpassung und Konfiguration. Und mit der zunehmenden Anzahl und Vielfalt von Datenquellen, Benutzern und Abfragen wird es nur noch schlimmer.

Die Rettung: Technologie und Design

Die gute Nachricht ist, dass sich Technologie und Data-Warehousing-*Architektur* (das Design und die Bausteine des modernen Data Warehouse) weiterentwickelt haben, um den Anforderungen einer datengesteuerten Wirtschaft gerecht zu werden. Zu diesen Innovationen gehören:

- **Die Cloud:** Die Cloud ist ein Schlüsselfaktor für die Entwicklung des modernen Data Warehouse. Sie bietet nahezu unendlichen,

kostengünstigen Speicherplatz, verbessert die Skalierbarkeit, ermöglicht die Auslagerung von Data-Warehouse-Management und Sicherheit an den Cloud-Anbieter und bietet die Möglichkeit, nur für die tatsächlich genutzten Speicher- und Rechenressourcen zu zahlen.

- **Massively Parallel Processing (MPP):** MPP, bei dem eine einzelne Rechenoperation über eine große Anzahl von separaten Computerprozessoren hinweg aufgeteilt und gleichzeitig ausgeführt wird, entstand Anfang der 2000er Jahre. Diese Arbeitsteilung erleichtert die schnellere Speicherung und Analyse von Daten, wenn Software auf die Nutzung dieses Ansatzes ausgelegt ist.
- **Spaltenorientierte Speicherung:** Traditionell speicherten Datenbanken Datensätze in Zeilen, ähnlich wie bei einem Spreadsheet. Darin waren zum Beispiel alle Informationen über einen Kunden oder eine Einzelhandelstransaktion enthalten. Beim herkömmlichen Abruf von Daten musste das System die gesamte Zeile lesen, um ein Element zu erhalten – ein mühsamer und zeitaufwändiger Vorgang. Bei der spaltenorientierten Speicherung wird jedes Datenelement eines Datensatzes in einer Spalte gespeichert. Bei diesem Ansatz kann ein Benutzer nur ein einziges Datenelement abfragen, etwa die Mitglieder eines Fitnessstudios, die ihre Beiträge bezahlt haben, ohne alles andere in diesem gesamten Datensatz lesen zu müssen, z. B. die ID-Nummer, den Namen, das Alter, die Adresse, die Stadt, das Bundesland oder die Zahlungsinformationen jedes Mitglieds. Mit diesem Ansatz erhält man oft eine viel schnellere Antwort auf derartige analytische Abfragen.
- **Vektorverarbeitung:** Diese Art der Datenverarbeitung für die *Datenanalyse* (die Wissenschaft der Untersuchung von Daten zur Erkenntnisgewinnung) macht sich die neuesten, revolutionären Computerchip-Designs zunutze. Dieser Ansatz liefert eine viel schnellere Performance als ältere Data-Warehouse-Lösungen, die vor Jahrzehnten für ältere, langsamere Hardwaretechnologie entwickelt wurden.
- **Solid State Drives (SSDs):** Im Gegensatz zu Hard Disk Drives (HDDs) speichern SSDs Daten auf Flash-Speicherchips, was die Datenspeicherung, den Datenabruf und die Datenanalyse beschleunigt. Eine Lösung, die sich die Vorteile von SSDs zunutze macht, kann eine deutlich bessere Leistung erzielen.

Weitere Informationen zu technologischen Fortschritten und anderen Trends, die die Entwicklung des Data Warehousing vorantreiben, finden Sie in Kapitel 2.

Das Cloud Data Warehouse: eine Einführung

Cloud Data Warehousing ist eine kostengünstige Möglichkeit für Unternehmen, sich die Vorteile der neuesten Technologie und Architektur zunutze zu machen, ohne die enormen Vorabkosten für den Kauf, die

Installation und die Konfiguration der erforderlichen Hardware, Software und Infrastruktur tragen zu müssen. Die verschiedenen Cloud-Data-Warehouse-Optionen werden im Allgemeinen in drei Kategorien eingeteilt:

- **Traditionelle Data-Warehouse-Software, die auf Cloud-Infrastruktur eingesetzt wird:** Diese Option ähnelt einem herkömmlichen On-Premise- Data Warehouse, da sie die ursprüngliche Codebasis wiederverwendet. Für den Aufbau und die Verwaltung des Data Warehouse wird nach wie vor IT-Fachwissen benötigt. Zwar müssen Sie die Hardware und Software nicht kaufen und installieren, doch es können noch erhebliche Konfigurations- und Tuning-Arbeiten und Prozesse wie regelmäßige Backups erforderlich sein.
- **Herkömmliches Data Warehouse, das in der Cloud von einem Drittanbieter als Managed Service gehostet und verwaltet wird:** Bei dieser Option stellt der Drittanbieter das IT-Fachwissen zur Verfügung, aber Sie werden wahrscheinlich weiterhin mit vielen der Einschränkungen eines herkömmlichen Data Warehouse zu kämpfen haben. Das Data Warehouse wird auf Hardware gehostet, die in einem vom Anbieter verwalteten Rechenzentrum installiert ist. Dies ähnelt dem, was in der Branche als *Application Service Provider (ASP) bezeichnet wird*. Die Kunden müssen weiterhin im Voraus angeben, wie viel Speicherplatz und Rechenressourcen (CPUs und Arbeitsspeicher) sie voraussichtlich nutzen werden.
- **Ein wahres SaaS Data Warehouse:** Mit dieser Option, die oft als *DWaaS* bezeichnet wird, liefert der Anbieter eine komplette Cloud-Data-Warehouse-Lösung, die die gesamte Hardware und Software umfasst und bei der nahezu alle Aufgaben im Zusammenhang mit der Einrichtung und Verwaltung der Leistung, der Governance und der Sicherheit, die bei einem Data Warehouse erforderlich sind, entfallen. Kunden zahlen in der Regel nur für die Speicher- und Rechenressourcen, die sie nutzen – und nur dann, wenn sie sie nutzen. Diese Option sollte bei Bedarf nach oben und unten skalierbar sein, indem jedem Workload eine unbegrenzte Rechenleistung zugewiesen wird, während eine unbegrenzte Anzahl von Workloads gleichzeitig ohne Beeinträchtigung der Leistung betrieben werden können.

Für einen detaillierteren Vergleich von Cloud-Data-Warehousing-Lösungen lesen Sie bitte Kapitel 5.

Warum Sie ein Cloud-Data-Warehouse benötigen

Jedes Unternehmen, das auf Daten angewiesen ist, um seinen Kunden besser dienen zu können, seine Abläufe zu optimieren und eine Führungsposition in seiner Branche einzunehmen, wird von einem Cloud Data Warehouse profitieren. Im Gegensatz zu herkömmlichen großen

Data Warehouses können Unternehmen jeder Größe ihr Data Warehouse in der Cloud so dimensionieren, dass es ihren Anforderungen und ihrem Budget genau entspricht. So können sie ihr System dynamisch erweitern oder reduzieren, wenn kurz- oder langfristig Veränderungen auftreten.

Hier sind einige Bereiche, in denen moderne Cloud Data-Warehouse-Technologie die betrieblichen Abläufe eines Unternehmens erheblich verbessern kann:

- **Kundenerfahrung:** Die Überwachung des Endbenutzerverhaltens in Echtzeit kann Unternehmen dabei helfen, Produkte, Dienstleistungen und spezielle Angebote auf die Bedürfnisse einzelner Verbraucher zuzuschneiden. Durch Sentiment-Analyse können Unternehmen die Stimmungslage ihrer Kunden besser verstehen, indem sie große Mengen von Social-Media-Beiträgen, Tweets und anderen Online-Aktivitäten analysieren.
- **Qualitätssicherung:** Unternehmen können Streaming-Daten auch zur Erkennung früher Warnzeichen für Kundendienstprobleme oder Produktmängel verwenden und innerhalb von Minuten oder Stunden anstatt von Tagen oder Wochen die erforderlichen Maßnahmen ergreifen. Das ist nicht möglich, wenn die Beschwerdeprotokolle des Callcenters die einzigen verfügbaren Datenquellen sind.
- **Verbesserung der betrieblichen Effizienz:** Operational Intelligence (OI) umfasst die Überwachung von Geschäftsabläufen und die Analyse von Ereignissen, um zu ermitteln, wie ein Unternehmen Kosteneinsparungen erzielen, Margen erhöhen, Prozesse rationalisieren und schneller auf Marktmechanismen reagieren kann. Wenn Sie Ihrem Unternehmen die Last der Verwaltung eines Data Warehouse abnehmen, können Sie sich besser auf die Analyse von Daten konzentrieren.
- **Innovation:** Anstatt nur in den Rückspiegel zu schauen, um die jüngste Vergangenheit einer Branche zu verstehen, können Unternehmen neue Datenquellen und Datenanalysen (Predictive, Prescriptive, Machine Learning) verwenden, um Trends zu erkennen und zu nutzen und so für Disruption in ihrer Branche zu sorgen, bevor ein unbekannter oder unerwarteter Wettbewerber dies tun kann.

ERINNERUNG

Fast alle Daten eines Unternehmens werden in vielen getrennten Datenbanken gespeichert. Die wichtigsten Fragen, die man sich stellen muss, sind: Wie zugänglich sind diese Daten? Wie viel wird es kosten, alle Daten zu extrahieren, zu speichern und zu analysieren? Was passiert, wenn Sie das nicht tun? An dieser Stelle kommt Cloud Data Warehousing ins Spiel.

IN DIESEM KAPITEL

- **Der steigende Bedarf an Datenzugriff und -analyse**
- **Wie Daten heute erstellt und verwendet werden**
- **Bewältigung der Herausforderungen mit neuen und verbesserten Technologien**

Kapitel 2
Warum das moderne Data Warehouse entstanden ist

Cloud Data Warehousing entstand aus der Konvergenz dreier wichtiger Trends: Veränderungen in Bezug auf Datenquellen, Datenvolumen und Datenvielfalt, einem erhöhten Bedarf an Datenzugriff und -analysen sowie technologischen Verbesserungen, die die Effizienz der Datenspeicherung, des Datenzugriffs und der Datenanalyse erheblich erhöhten. In diesem Kapitel beschreiben wir diese Trends ausführlicher und zeigen, wie ein Data Warehouse die Vorteile der Cloud nutzen kann, um mit diesen Trends Schritt zu halten.

Trends in Daten: Volumen, Vielfalt und Geschwindigkeit

Wenn wir in diesem Buch über Daten sprechen, meinen wir Petabytes. Ein Petabyte entspricht 1 Million Gigabytes. Das sind etwa 500 Milliarden Seiten gedruckter Standardtext oder 58.333 hochauflösende Filme von circa zwei Stunden Länge. Daten strömen aus den unterschiedlichsten Quellen eines Unternehmens herein – aus den täglichen Geschäftsabläufen, durch die Nutzung von Websites und Softwareanwendungen auf Mobilgeräten und aus den täglichen Aktivitäten digitaler und mechanischer Geräte.

In diesem Abschnitt befassen wir uns mit den Veränderungen der Daten und der Datennutzung, die zu einem Bedarf an Cloud Data Warehousing geführt haben.

Der Daten-Tsunami

In nicht allzu ferner Vergangenheit verwalteten Unternehmen hauptsächlich Daten, die von Menschen manuell in das System eingegeben wurden. Möglicherweise verfügten sie auch über Daten aus externen Quellen, z. B. von Kunden, Auftraggebern und Partnern. Die Menge der zu verwaltenden Daten war relativ klein und vorhersehbar und alle Daten wurden im Rechenzentrum des Unternehmens gespeichert, verwaltet und gesichert. Dieses Konzept wird heute als *On-Premise-Methode* bezeichnet.

Gegenwärtig bricht ein Daten-Tsunami über die Geschäftswelt herein: Daten können aus einer Vielzahl von Quellen stammen, die bereits in diesem Buch erwähnt wurden. Daneben gibt es noch weitere Datenquellen, die zu zahlreich und zu vielfältig sind, um hier erwähnt zu werden. Die Menge und Vielfalt dieser Daten kann ein herkömmliches, On-Premise Data Warehouse schnell überfordern. Dies hat oft zur Folge, dass die Datenverarbeitung und -analyse aufgrund der Überlastung durch Benutzer und die von ihnen verarbeiteten Workloads zum Stillstand kommt oder dass das System abstürzt.

Zur Anpassung an die exponentielle Zunahme der Datenmengen ist eine neue Perspektive erforderlich (siehe Abbildung2-1). Die entscheidende Frage ist nicht, wie groß das Data Warehouse eines Unternehmens sein muss, sondern ob es kosteneffizient, reibungslos und in der Größenordnung skaliert werden kann, die für die Verarbeitung gewaltiger Datenmengen erforderlich ist.

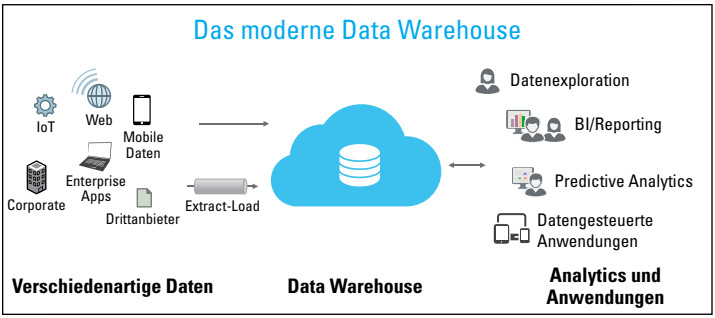

ABBILDUNG 2-1: Das moderne Data Warehouse stellt alle Daten für alle Benutzer zur Verfügung.

ERINNERUNG

Es tauchen immer wieder neue Anwendungsfälle für Cloud Data Warehousing auf. SaaS-Unternehmen und große Konzerne, die die Cloud zur Speicherung ihrer Daten nutzen, monetarisieren (verkaufen) diese Daten häufig. Sie verpacken sie als Service und verkaufen sie an andere Organisationen, die sich möglichst tiefe Einblicke wünschen, um bessere Geschäftsentscheidungen treffen zu können.

In der Cloud entstandene Daten

In vielen Unternehmen hat sich SaaS schnell durchgesetzt, darunter Software für das Kundenbeziehungsmanagement (Customer Relationship Management, CRM), Software-Suites für die Ressourcenplanung (Enterprise Resource Planning, ERP), Werbe- und Einkaufsplattformen und Online-Marketing-Tools, um nur einige zu nennen. Dank der Cloud können neue SaaS-Unternehmen schon um den Preis von ein oder zwei Laptops gegründet werden. Diese SaaS-Produkte erzeugen riesige Mengen wertvoller Daten, die in der Cloud gespeichert werden. Immer mehr Unternehmen erkennen überdies, dass SaaS-Anbieter mehr Sicherheit bieten, als ihre eigenen On-Premise Rechenzentren.

Die Nachfrage nach SaaS-/Cloud-Anwendungen ist ebenfalls gestiegen. Sie lassen sich viel leichter bereitstellen als On-Premise-Lösungen. Früher hatten die meisten Unternehmen vielleicht fünf bis zehn wichtige Unternehmensanwendungen, die Daten generierten. Heute ist es selbst für mittelgroße Unternehmen normal, Hunderte oder sogar Tausende von Anwendungen zu haben, die möglicherweise alle ihr eigenes Datensilo schaffen – Marketingdaten in einem System, Finanzdaten in einem anderen, Produktinformationen in wieder einem anderen – und keines von ihnen ist für eine vollständige und optimale Analyse integriert.

Da sich der Großteil der Daten eines Unternehmens jetzt in der Cloud befindet, ist die Cloud auch der natürliche Ort für die Integration dieser Daten. Mit Cloud Data Warehousing ist es nicht mehr notwendig, dass Sie diese Daten in Ihr Rechenzentrum holen. Angesichts der zunehmenden Menge der in der Cloud gespeicherten Daten wäre dies sehr teuer, zeitaufwändig und kaum sinnvoll.

Verwendung maschinell erzeugter Daten

Maschinell erzeugte Daten sind ein Schlüsselthema im Zusammenhang mit dem *Internet of Things (IoT)* – einer endlosen Sammlung von Geräten, die Daten über das Internet übermitteln, darunter Smartphones, Thermostate, Kühlschränke, Ölplattformen, Haussicherheitssysteme, intelligente Zähler und vieles mehr. Mit den von IoT-Geräten erfassten und analysierten Daten können Produkte und Prozesse verbessert, Geräte überwacht und Vorhersagen über den Wartungsbedarf getroffen werden, um Ausfälle zu vermeiden.

Viele maschinell erzeugte Daten haben ein schlechtes Signal-Rausch-Verhältnis. Sie enthalten wertvolle Daten, aber auch viel „Rauschen“. Um die wertvollen Bestandteile zu finden, ist es deshalb oft nötig, alles zu speichern. Darüber hinaus stammt ein zunehmender Teil dieser Daten aus Quellen außerhalb des Rechenzentrums. Dies macht die Cloud und ihre nahezu unbegrenzte Skalierbarkeit zum natürlichen Ort für die Speicherung und Integration dieser Daten.

Datenexploration

Die Analyse von Daten beginnt mit der Datenexploration: interessante und wertvolle Verbindungen werden identifiziert und den Datennutzern in Form von Berichten und Analysen zur Verfügung gestellt. Die Datenexploration ist zwar kein neues Konzept, doch das zunehmende Datenvolumen macht sie zu einer immer ressourcenintensiveren Aufgabe.

Bei der Datenexploration werden oft große Datenbestände verarbeitet. Außerdem ist der Prozess oft experimenteller Art, was die ROI-Bewertung erschwert, die zur Unterstützung der erheblichen Vorabkosten für die Bereitstellung eines herkömmlichen On-Premise Data Warehouse erforderlich ist. In der Cloud ist es möglich, ein Data Warehouse je nach Bedarf auf- und abwärts zu skalieren. Außerdem kann ein nutzungsbasiertes Modell verwendet werden, sodass sich Unternehmen nicht mit der Frage beschäftigen müssen, ob sie eine teure Vorabverpflichtung eingehen sollten oder nicht.

Data Lakes

Die steigende Notwendigkeit, gewaltige Mengen an Rohdaten in verschiedenen Formaten an einem einzigen Ort vorzuhalten, hat das hervorgebracht, was wir heute als „Legacy Data Lake“ bezeichnen. Den meisten Unternehmen wurde schnell klar, dass diese Lösungen unerschwinglich sind, da es fast unmöglich ist, diese Daten umzuwandeln und wertvolle Erkenntnisse aus ihnen zu gewinnen.

Das ursprüngliche Interesse an Data Lakes machte jedoch deutlich, dass die meisten Unternehmen alle ihre Daten zu vernünftigen Kosten an einem einzigen Ort speichern wollen. Wenn man einem bestehenden Data Lake ein modernes Cloud Data Warehouse hinzufügt oder den Data Lake innerhalb des Data Warehouse aufbaut, lässt sich diese ursprüngliche Vision für den Data Lake leicht realisieren: das kosteneffiziente Laden, Umwandeln und Analysieren unbegrenzter Mengen strukturierter und semistrukturierter Daten – mit nahezu unbegrenzten Speicher- und Rechenressourcen.

Trends in der Berichterstattung und Analytik

Die datengesteuerte Entscheidungsfindung ist nicht länger ein Privileg von Führungsteams oder Datenwissenschaftlern, sondern wird jetzt auch zur Verbesserung nahezu aller betrieblichen Aspekte eines Unternehmens verwendet. Dieser steigende Bedarf an Datenzugriff und -analysen im gesamten Unternehmen kann Systeme jedoch verlangsamen oder zum Absturz bringen, wenn Workloads um Speicher- und Rechenressourcen aus herkömmlichen Data Warehouses wetteifern. Die

Effizienz leidet und Unternehmen müssen mehr Zeit und Geld in zusätzliche Infrastruktur zur Wartung des Systems investieren.

In diesem Abschnitt betrachten wir einige der Trends, die den Datenzugriff und die Datenverwendung verändern, und zeigen, wie diese Trends den Bedarf an modernen, für die Cloud entwickelten Data-Warehouse-Lösungen vorantreiben.

Nutzung von Elastizität für Analysen

Hier sind einige Szenarien, in denen in der Cloud entwickeltes elastisches Data Warehousing eine umfassendere Nutzung von Daten ermöglicht:

- Datenexploration hat viele Vorteile. Allerdings weiß niemand im Voraus genau, welche Rechenressourcen für die Analyse riesiger Datenbestände benötigt werden, sodass sich die elastische Skalierbarkeit bei Bedarf ideal für diese Art von Analyse eignet.
- Die Ad-hoc-Datenanalyse, die immer wieder eingesetzt wird, beantwortet eine einzige, spezifische, geschäftliche Frage. Mit dynamischer Elastizität und dedizierten Ressourcen für jeden Workload können diese Abfragen ausgeführt werden, ohne andere Workloads zu verlangsamen.
- Ereignisgesteuerte Analysen erfordern konstante Daten. Sie verwenden neue Daten, um Berichte und Dashboards kontinuierlich zu aktualisieren, damit Führungskräfte das Geschäft in Echtzeit oder Nahe-Echtzeit überwachen können. Die Aufnahme und Verarbeitung von Streaming-Daten erfordert ein elastisches Data Warehouse, um Schwankungen und Spitzen im Datenfluss zu bewältigen.

Schnelle Iteration statt gründlicher Vorausplanung

Unternehmer haben gewöhnlich zwei Möglichkeiten, um die Marktfähigkeit einer neuen Idee sicherzustellen: gründliche Vorausplanung oder schnelle Iteration. Die erste Option ist ein traditioneller, zeitaufwändiger Prozess. Eine Geschäftschance oder eine neue Produktidee wird durchdacht, Ideen werden erörtert – und man hofft, dass Nachfrage beim Verbraucher entsteht. Bei der schnellen Iteration wird die Idee schnell auf dem Markt getestet und ständig wiederholt, bis eine praktikable und erfolgversprechende Version des Produkts vorliegt. Dann beginnt der Prozess erneut.

Die schnelle Iteration hat sich als effektiverer Prozess zum Verdrängen etablierter Konkurrenten und zur Veränderung der Geschäftsabläufe einer ganzen Branche erwiesen. Um erfolgreich zu sein, ist jedoch eine schnelle Erfassung und Analyse großer Mengen genauer Daten

erforderlich. Fortschritte im Bereich des Cloud Data Warehousing und in der Analytik haben die schnelle Iteration praktikabler gemacht und dafür gesorgt, dass die Datengenauigkeit nicht beeinträchtigt wird.

ERHÖHTER BEDARF AN DATENANALYSEN

FALLSTUDIEN

Jana stellt über 30 Millionen Smartphone-Nutzern in über 15 Schwellenmärkten kostenlosen, uneingeschränkten Internetzugang zur Verfügung. Mit der Android-App mCent verlagert Jana die Kosten des mobilen Internets über gesponserte Inhalte von den Kunden auf mehr als 4.000 Marken.

Wenn neue, markenbezogene Inhalte oder mCent-Funktionen eingeführt werden, analysiert und misst Jana wichtige Kennzahlen, darunter Aufmerksamkeit der Nutzer, User Lifetime Value und Key Performance Indicators (KPIs).

Als Jana größer wurde und seine Daten zunahmen, war die ursprüngliche Analysearchitektur des Unternehmens für seine geschäftlichen Anforderungen nicht mehr effizient genug. Abfragen verlangsamten sich und es konnten keine Tabellen-Scans mehr ausgeführt werden. Um mehr Kapazität und Backup-Systeme hinzuzufügen und Janas Open-Source-Datenspeicher zu verwalten, wurde immer mehr Verwaltungszeit benötigt.

Wie in der Abbildung dargestellt, rüstete Jana die meisten seiner Datenplattform-Komponenten auf und optimierte sein System mit einem in der Cloud erstellten Data Warehouse, um diese Hindernisse zu überwinden und eine Reihe von Vorteilen zu erzielen. Das Unternehmen konnte nun:

- mit den geschäftlichen Anforderungen Schritt halten, die bei der Verarbeitung und Analyse eines schnell wachsenden Stroms unterschiedlicher Daten entstehen
- einen verstärkten Einsatz von Analysen im gesamten Unternehmen unterstützen; 80 Prozent der Mitarbeiter von Jana greifen auf das Data Warehouse zu
- seinen Verwaltungsaufwand erheblich reduzieren.

Janas Umstellung auf ein schnelleres, billigeres und effektiveres Data Warehouse.

Einbetten von Analysen

Für viele Unternehmen ist die Analytik ein separater und eigenständiger Geschäftsprozess. Es wird jedoch immer mehr zum Trend, Analysen in Geschäftsanwendungen zu integrieren, die zunehmend in der Cloud erstellt werden. Diese Anwendungen können eine erhebliche Variabilität in der Anzahl der Benutzer bewältigen, die die Anwendungen abfragen, und in der Anzahl der Abfragen (Workloads), die die Benutzer zur Analyse dieser Daten ausführen. Die Cloud erleichtert den Datentransfer von cloudbasierten Anwendungen zum Cloud Data Warehouse des Unternehmens, wo ihre Skalierbarkeit und Elastizität Benutzer- und Workload-Fluktuationen besser unterstützen kann.

Technologische Voraussetzungen für ein modernes Data Warehouse

Technologische Innovationen können Data Warehousing und Analysen im Hinblick auf Verfügbarkeit, Einfachheit, Kosten und Leistung erheblich verbessern. In diesem Abschnitt sehen wir uns die wesentlichen Technologien an, die Teil jedes modernen Data Warehouse sein sollten.

Cloud

Die Cloud ist aufgrund ihrer besonderen Eigenschaften besonders gut für Data Warehousing geeignet. Wir haben diese Eigenschaften bereits in anderen Zusammenhängen erwähnt. Es lohnt sich jedoch, noch einmal hervorzuheben, welche Merkmale der Cloud zu verdanken sind:

- **Unbegrenzte Ressourcen:** Cloud-Infrastruktur ermöglicht die Bereitstellung nahezu unbegrenzter Ressourcen – bei Bedarf und innerhalb von Minuten oder Sekunden. Die Abrechnung erfolgt nach Sekunde und Unternehmen zahlen nur für das, was sie tatsächlich nutzen. So kann jede beliebige Größenordnung in Bezug auf Benutzer und Workloads dynamisch unterstützt werden, ohne dass die Performance beeinträchtigt wird.
- **Kosteneinsparungen, Fokus auf Daten:** Unternehmen, die sich für eine cloudbasierte Lösung entscheiden, vermeiden die mit On-Premise-Systemen verbundenen kostspieligen Vorabinvestitionen in Hardware, Software und andere Infrastruktur sowie die Kosten für die Wartung, Aufrüstung und Sicherung. Stattdessen konzentrieren sie sich auf die Analyse von Daten.
- **Natürlicher Integrationspunkt:** Einigen Schätzungen zufolge stammen bis zu 80 Prozent der zu analysierenden Daten aus Anwendungen außerhalb des Rechenzentrums Ihres Unternehmens. Die Zusammenführung dieser Daten in der Cloud ist einfacher und

kostengünstiger als der Aufbau eines internen Rechenzentrums, da Sie nicht im Voraus Hardware und Software im Millionenwert kaufen und dann technisches Personal für die Wartung dieser Ressourcen einstellen müssen.

Spaltenorientierte Speicherung, Verarbeitung

Wie bereits erwähnt, verbessert die spaltenorientierte Speicherung die Effizienz und Leistung der Datenspeicherung, des Datenabrufs und der Analyse erheblich und ermöglicht den Systembenutzern einen schnelleren Zugriff auf die Ergebnisse.

Solid State Drives (SSDs)

Im Gegensatz zu Hard Disk Drives (HDDs) speichern SSDs Daten auf Flash-Speicherchips, die das Speichern, Abrufen und Analysieren der Daten beschleunigen. Diese Verbesserungen erhöhen die Rechenleistung von Data Warehouses, die für die effektive Nutzung von SSDs konzipiert worden sind.

NoSQL

NoSQL steht für *Not Only Structured Query Language (SQL)* und bezeichnet eine Technologie, mit der es möglich ist, neuere Datenformen zu speichern und zu analysieren, z. B. Daten, die von Maschinen und in sozialen Medien generiert werden. Dadurch kann die Datenanalyse eines Unternehmens erheblich bereichert und erweitert werden. Herkömmliche Data Warehouses können mit diesen Datentypen nicht besonders gut umgehen. Deshalb sind in den letzten Jahren neuere Ansätze wie JSON, Avro und XML entstanden, die in der Lage, sind, diese „semistrukturierten" Datenformen zu verarbeiten.

Einige dieser NoSQL-Systeme wurden mit der Absicht entworfen, herkömmliche Data Warehouses zu ersetzen, doch letztendlich ergänzen sie diese nur. Um aus semistrukturierten Daten Nutzen zu ziehen, müssen Unternehmen oft Daten aus einem NoSQL-System extrahieren und umwandeln und sie in ein traditionelles Data Warehouse laden, damit Geschäftsanwender problemlos darauf zugreifen können. Dies führt zu einer weiteren Komplexitäts- und Kostenschicht für Unternehmen (wie Jana; siehe frühere Fallstudie), die versuchen, die Vorteile beider Arten von Systemen zu nutzen.

Das moderne cloudbasierte Data Warehouse muss daher die Aufnahme und Abfrage strukturierter (traditioneller) und semistrukturierter Datenformate integrieren und dafür optimiert werden, damit Unternehmen nicht für zwei Systeme bezahlen und diese verwalten müssen.

IN DIESEM KAPITEL

- Die richtige Data-Warehouse-Lösung
- Ein gutes Preis-Leistungs-Verhältnis
- Priorisierung von Datensicherheit, Datenschutz und Data Governance

Kapitel 3
Auswahlkriterien für ein modernes Data Warehouse

Aufgrund der in Kapitel 2 besprochenen Trends entstand ein Bedarf und eine Möglichkeit zur Schaffung eines neuartigen Data Warehouse, das der Menge, Vielfalt und Geschwindigkeit der heute verfügbaren Daten Rechnung trägt und die neuen Arten berücksichtigt, in denen Daten von Unternehmen genutzt werden. Eine solche Lösung muss sich die Vorteile wichtiger technologischer Innovationen wie der Cloud zunutze machen.

Wenn Sie nach einer Data-Warehouse-Lösung suchen, lohnt es sich, eine Checkliste mit Auswahlkriterien zu verwenden. Dadurch ist es leichter, zu entscheiden, welche Lösung Ihren jeweiligen Anforderungen am besten entspricht. Betrachten Sie dieses Kapitel als Ihre Checkliste, um die beste Data-Warehouse-Lösung für Ihr Unternehmen zu finden.

Es erfüllt aktuelle und zukünftige Anforderungen

Wahre Elastizität hat geschäftliche Vorteile, doch das ist nicht alles. Sie sollten in der Lage sein, sowohl Rechen- als auch Speicherressourcen unabhängig voneinander zu skalieren, damit Sie nicht mehr Speicherkapazität hinzufügen müssen, wenn Sie eigentlich nur zusätzliche Rechenleistung benötigen (und umgekehrt). Dies sind die wesentlichen Funktionen eines elastischen Data Warehouse.

Es speichert und integriert alle Daten an einem Ort

Die in den vorangegangenen Kapiteln besprochenen nicht-traditionellen oder semistrukturierten Daten können die durch Data Analytics gewonnenen Einblicke über die Grenzen traditioneller Daten hinweg erweitern. Dazu ist jedoch ein neuer Ansatz zum Laden und Umwandeln dieser neuen Datentypen erforderlich, der Unternehmen die Analyse dieser Daten ermöglicht. Die meisten herkömmlichen Data Warehouses nehmen Abstriche bei der Leistung oder Flexibilität in Kauf, um diese Datentypen verarbeiten zu können. Mit einem modernen Data Warehouse sollte es nicht mehr erforderlich sein, im Vorfeld starre, traditionelle Strukturen zu entwerfen und zu modellieren, die die Umwandlung semistrukturierter Daten vor dem Laden erfordern. Das Data Warehouse sollte überdies die Abfrageleistung für diese Datentypen optimieren können, während sie sich noch in ihrer ursprünglichen Form befinden. Im Großen und Ganzen sollte das Data Warehouse in der Lage sein, unterschiedliche Daten flexibel zu unterstützen und Performance-Probleme zu vermeiden.

Das effiziente Laden aller Ihrer Daten an einen Ort ist dabei von entscheidender Bedeutung. Die Integration all dieser unterschiedlichen Datentypen für präzisere Analysen ist allerdings eine ganz andere Aufgabe. Ein modernes Data Warehouse sollte Ihre einst auf NoSQL-Systeme beschränkten semistrukturierten Daten automatisch mit den strukturierten Daten einer herkömmlichen relationalen Unternehmensdatenbank integrieren können. Dabei sollte es nicht erforderlich sein, irgendetwas zu installieren oder zu konfigurieren. Tuning und Performance sollten integriert sein. Vor allem aber sollten Sie nicht für zwei separate Systeme zur Verwaltung aller Ihrer Daten zahlen und diese unterhalten müssen.

Es unterstützt vorhandene Fähigkeiten, Tools und Fachwissen

Herkömmliche Data Warehouses sind nur deshalb veraltet, weil die zugrundeliegende Technologie seit vier Jahrzehnten existiert und nicht einfach für die Cloud umgestaltet werden kann. Das bedeutet auch, dass SQL, die Sprache, auf die sich diese Technologie stützt, in der Branche eine tragende Säule bleiben wird. Aus diesem Grund gibt es ein breites Spektrum an ausgereiften und neuen Tools zur Verwaltung, Umwandlung, Integration, Visualisierung und Analyse von Daten sowie Business Intelligence Tools, die mit einem SQL-Data-Warehouse kommunizieren. Aufgrund der seit langem anerkannten wichtigen Rolle von Standard-SQL sind SQL-Kenntnisse zudem weit verbreitet.

Herkömmliche Data Warehouses unterstützen zwar SQL, jedoch nicht die Funktionen, die für die effektive Speicherung und Verarbeitung semistrukturierter Daten erforderlich sind. Viele Unternehmen haben sich daher Alternativen zugewandt, z. B. NoSQL-Lösungen. Die Grenzen dieser Systeme

FALLSTUDIEN

ANALYSE UNTERSCHIEDLICHER DATEN

Chime bietet „Smarter Banking" für Nutzer von Mobilgeräten. Chime erfasst und analysiert Daten über Mobil-, Web- und Backend-Server-Plattformen hinweg, um die Benutzererfahrung seiner Mitglieder zu verbessern und gleichzeitig einen Mehrwert für das Unternehmen zu schaffen.

Zunächst erwies sich die Analyse wichtiger Kennzahlen bei Chime als mühsam, da Daten aus einer großen Anzahl von Services einschließlich Facebook und Google Ads erfasst und analysiert werden mussten. Chime zog auch Ereignisse von anderen Analyse-Tools von Drittanbietern heran, die meist semistrukturierte Daten wie JSON lieferten.

Mit seinem neuen Cloud Data Warehouse erfüllte Chime die folgenden Anforderungen:

- effiziente Bereitstellung von strukturierten und semistrukturierten Daten und deren Bereitstellung für Abfragen nahezu in Echtzeit unter Verwendung von Standard-SQL-Datenbanktabellen.
- Vereinfachung der Data-Pipeline, ohne dass für jeden neuen Datentyp, der in das Data Warehouse geladen wird, ein neues Modell entworfen werden muss.
- Skalierung nach oben und unten, um Workload-Anforderungen zu erfüllen und die Kosten zu kontrollieren.
- schnelle und problemlose Integration mit Datenanalyse-Tools von Drittanbietern.
- Nutzung von SQL anstelle anderer Optionen, die komplizierte Programmiersprachen zum Extrahieren und Analysieren von Daten erfordern.

Die Analysten von Chime modellieren jetzt weitere Szenarien, um die Mitgliederservices zu verbessern, nicht mehr so lange auf Abfrageergebnisse warten zu müssen und der Analyse von Daten mehr Zeit widmen zu können.

stellen allerdings ein weiteres Problem dar. Sie erfordern spezielle Kenntnisse und Fähigkeiten, die nicht allgemein verfügbar sind und sie unterstützen SQL möglicherweise nicht. Ein modernes Data Warehouse sollte mit führender Technologie ausgestattet sein und gleichzeitig auf umfassenden und etablierten Standards (wie SQL) basieren und mit anderen in der Branche allgemein verfügbaren Fähigkeiten und Tools wie Spark, Python und R-Computersprachen kompatibel sein.

Es hilft Ihrem Unternehmen, Kosten einzusparen

Die Kosten eines herkömmlichen Data Warehouse können sich auf Millionen belaufen: Lizenzgebühren, Kosten für Hardware und Services, die für die Einrichtung, Verwaltung, Bereitstellung und das Tuning des Warehouse

erforderliche Zeit und das Fachwissen sowie die Kosten für die Sicherung und das Backup der Daten. Für viele Unternehmen ist der Aufbau eines Data Warehouse, das den Geschäftsanforderungen entspricht und die Menge und Vielfalt der heutigen Daten voll ausnutzen kann, unerschwinglich.

Ein modernes Data Warehouse sollte diese Herausforderungen zu einem viel niedrigeren Preis erfüllen können. Werden beispielsweise Storage und Compute getrennt skaliert, sodass Sie nur für die benötigten Ressourcen zahlen müssen? Werden Workloads und Parallelität ebenfalls skaliert? Unterstützt das Data Warehouse unterschiedliche Datenstrukturen und integriert es unterschiedliche Daten an einem Ort? Wird es nur minimale oder überhaupt keine Ausfallzeiten geben und können Upgrades automatisch oder in gestaffelter Form zur Verfügung gestellt werden? Und schließlich: Kann all dies automatisch durchgeführt werden, ohne die Komplexität, die Kosten und das Kopfzerbrechen, die meist mit der manuellen Anpassung und dem Tuning des Systems verbunden sind, um die beste Leistung zu erzielen? (Siehe Kapitel 5 für einen Vergleich von Cloud Data Warehouses).

ERINNERUNG

Bei Cloud Data Warehousing sollte Ihre Servicegebühr all das für einen Bruchteil der Kosten einer herkömmlichen On-Premise-Lösung abdecken. Allerdings sind nicht alle cloudbasierten Lösungen gleich. Ihre Unterschiede bestimmen auch, wie viel ein Kunde zahlen muss, um wertvolle Dateneinblicke zu erhalten.

Es bietet Datenresilienz und Wiederherstellungsfunktionen

Viele Arten von Data-Warehouse-Fehlern können zu Datenverlusten oder -inkonsistenzen führen. Deshalb muss Ihr Data Warehouse dafür sorgen, dass Ihre Daten sicher, aktuell und jederzeit verfügbar sind. Herkömmliche Data Warehouses schützen Daten in der Regel durch regelmäßige Backups, die wertvolle Rechenressourcen verbrauchen und laufende Workloads beeinträchtigen. Für regelmäßige Backups wird außerdem zusätzlicher Speicherplatz benötigt. Oft werden auch die neuesten Daten ausgelassen, was zu Dateninkonsistenz führt.

Ein modernes Data Warehouse sollte sich selbst verwalten und die Langlebigkeit, Stabilität und Verfügbarkeit des Systems sicherstellen. Es darf keine laufenden Workloads beeinträchtigen, sich nicht nachteilig auf die Performance auswirken oder zu einer Nichtverfügbarkeit von Services führen, weil im Hintergrund Backup-Prozesse laufen. Außerdem sollte es kostengünstig sein und intelligente Möglichkeiten zur sicheren Aufbewahrung Ihrer Daten bieten, ohne dass diese kopiert und an einen anderen Ort verschoben werden müssen. Eine Multi-Cloud-Architektur gibt Ihnen die Möglichkeit, Daten und Workloads zu verlagern, wenn Ihr Unternehmen wächst – sowohl zwischen geografischen Regionen als auch zwischen großen Cloud-Anbietern wie Amazon, Microsoft und Google.

Es sichert Daten im Ruhezustand und bei der Übertragung

Datensicherheit lässt sich in die folgenden zwei Hauptbereiche unterteilen:

- **Vertraulichkeit:** Verhinderung des unberechtigten Zugriffs auf Daten
- **Integrität:** Sicherstellen, dass die Daten nicht verändert oder korrumpiert werden, dass sie ordnungsgemäß verwaltet werden und dass ihre Qualität erhalten bleibt.

Ein modernes Data Warehouse muss auch eine mehrstufige rollenbasierte Zugriffskontrolle (engl. Role-Based Access Control, RBAC) unterstützen. Dadurch wird sichergestellt, dass Benutzer nur auf die Daten zugreifen können, die sie sehen dürfen. Für eine bessere Sicherheit ist Multi-Faktor Authentifizierung (MFA) erforderlich. Bei MFA sendet das System bei der Anmeldung eines Benutzers eine zweite Berechtigungsanfrage, oft an ein Mobiltelefon. Durch die Eingabe des an das Telefon gesendeten Passcodes wird sichergestellt, dass keine unbefugte Person mit einem gestohlenen Benutzernamen und Passwort auf das System zugreifen kann.

Data Governance sorgt für einen ordnungsgemäßen Zugriff auf und die korrekte Nutzung der Daten eines Unternehmens und stellt sicher, dass alle Daten so verwaltet und geschützt werden, dass Verstöße verhindert und detaillierte Vorschriften eingehalten werden. Um die Qualität der Daten zu wahren, die Ihr Unternehmen mit anderen Personen teilt, sind auch strenge Kontrollen erforderlich. Schlechte Daten können zu verpassten Gelegenheiten, schlechten Geschäftsentscheidungen, Einnahmeverlusten und erhöhten Kosten führen. Data Stewards – die mit der Überwachung der Datenqualität beauftragt werden – können erkennen, ob Daten korrumpiert oder unrichtig sind, nicht oft genug aktualisiert werden und daher nicht mehr relevant sind oder ob sie losgelöst vom Kontext analysiert werden.

Die Verschlüsselung der Daten, d. h. die Anwendung eines Verschlüsselungsalgorithmus, um Klartext in Chiffretext zu übersetzen, ist ein weiteres notwendiges Sicherheitsmerkmal. Ein größerer Teil der Lösung ist die „Schlüsselverwaltung". Nach der Verschlüsselung der Daten wird ein Verschlüsselungscode verwendet, um sie wieder zu entschlüsseln. Neben den Daten muss auch der Schlüssel geschützt werden, der zur Entschlüsselung der Daten verwendet wird. Wie lange verwenden Sie den gleichen Schlüssel? Was passiert, wenn der Schlüssel kompromittiert worden ist? All diese Aspekte müssen verwaltet werden. Das Data Warehouse sollte einen hierarchischen Key-Wrapping-Ansatz verwenden, bei dem die Verschlüsselungscodes verschlüsselt werden, sowie ein robustes Schlüssel-Rotationsverfahren, das die Anzahl der Verwendungen eines einzelnen Schlüssels begrenzt.

Darüber hinaus muss der Lösungsanbieter eines modernen Cloud Data Warehouse regelmäßige Sicherheitstests, so genannte Penetrationstests durchführen, um auf proaktive Weise zu prüfen, ob Schwachstellen vorhanden sind. Der Anbieter muss diese Maßnahmen konsistent und automatisch umsetzen, ohne dass die Leistung beeinträchtigt wird.

In Kapitel 8 werden die Sicherheit und Verwaltung von Cloud Data Warehouses ausführlich behandelt.

ERINNERUNG

Wählen Sie ein Data Warehouse mit branchenüblicher, End-to-End-Sicherheit. Halten Sie nach einer Lösung Ausschau, die Sicherheitsaudits wie SOC 1/SOC 2 Typ II und ISO/IEC 27001 bestanden hat.

Es optimiert die Data-Pipeline

Data-Pipeline bezieht sich hauptsächlich auf die Prozesse Extract (Extrahieren), Transform (Umwandeln) und Load (Laden) (ETL), bei denen Daten in einem Format in das Warehouse importiert werden, das Abfragen unterstützt. Bei einer langsamen Data-Pipeline müssen Benutzer, z. B. Analysten, zu lange auf den Datenzugriff warten. Das Problem wird noch durch die Tatsache verschärft, dass die Vielfalt, Anzahl und Größe der nicht-relationalen Daten, die aus verschiedenen Quellen einströmen, so schnell zugenommen hat und immer weiter zunimmt.

Ein modernes Data Warehouse sollte die Komplexität des gesamten Prozesses reduzieren, damit sich Daten schneller durch die Data-Pipeline bewegen können. Moderne Lösungen sollten in der Lage sein, semistrukturierte Daten effizient in ihrem nativen Format zu laden und sie sofort für Abfragen zur Verfügung zu stellen, ohne dass zusätzliche und komplizierte Systeme wie NoSQL zur Umwandlung der Daten benötigt werden. So können Benutzer sofort auf die Daten zugreifen, wie sie es bei der Abfrage einer SQL-Datenbank tun würden. Solche Lösungen können den Zugriff auf neue Daten exponentiell beschleunigen und den Aufnahme- und Umwandlungsprozess von einem Tag auf knapp eine Stunde reduzieren.

Es verkürzt Ihre Time-to-Value

Die Bereitstellung einer Lösung sollte kein großes Unterfangen sein. Wichtige Aspekte, die früher manuell ausgeführt wurden, sollten automatisiert werden. Vor allem sollte die gewählte Lösung jederzeit für alle Benutzer verfügbar sein und allen Datentypen zu einem Bruchteil der mit herkömmlichen Systemen verbundenen Kosten gerecht werden. Ein solches System sollte sofortige Dateneinblicke liefern, zur Rationalisierung des Unternehmens beitragen und es dabei unterstützen, seinen Kunden besser zu dienen und eine Führungsposition in seiner Branche einzunehmen.

IN DIESEM KAPITEL

- **Verkürzung des Time-to-Value-Zyklus**
- **Reduzierte Kosten für Speicher- und Rechenkapazitäten**
- **Nutzung dynamischer Elastizität**
- **Auslagerung von Verwaltung und Sicherheit**

Kapitel 4
On-Premise- und Cloud Data Warehousing im Vergleich

Wenn Sie auf der Suche nach einem neuen Data Warehouse sind, sollten Sie zuerst überlegen, wo Ihr Data Warehouse untergebracht werden soll: im Rechenzentrum Ihres Unternehmens oder in der Cloud als Software-as-a-Service (SaaS). Traditionelles On-Premise-Data-Warehousing ist eine ausgereifte und etablierte Technologie, die lange vor der Cloud als praktikable Plattform entwickelt wurde. Mit der zunehmenden Verbreitung von Cloud-Computing steigt der Bedarf an Data-Warehouse-Lösungen, die die Vorteile der Cloud voll ausschöpfen können. In diesem Kapitel befassen wir uns mit den wichtigsten Überlegungen zu Cloud Data Warehousing und vergleichen dieses mit herkömmlichen On-Premise-Systemen.

Evaluierung Ihres Time-to-Value

Die Bereitstellung eines herkömmlichen Data Warehouse (siehe Kapitel 3) dauert oft mindestens ein Jahr und kann sich zu einem mehrjährigen Projekt ausweiten, bevor Sie Erkenntnisse aus Ihren Daten gewinnen können. Aufgrund der Agilität der modernen Geschäftswelt kann es passieren, dass wichtige Stakeholder, die das Projekt unterstützen, und die für den Erfolg des Projekts verantwortlichen Business Enabler und technischen Experten das Team oder das Unternehmen vor der Go-Live-Phase des Projekts verlassen. Bei einem derart langen Zyklus kann das Projekt auch Wirtschaftsabschwüngen und möglichen Umsatzeinbußen des Unternehmens ausgesetzt sein und es besteht das Risiko, dass das Projekt aufgrund von Scope Creep (der schleichenden Erweiterung des Projektumfangs) niemals realisiert wird.

On-Premise-Lösungen sind überdies nicht auf den Umgang mit den heute üblichen semistrukturierten Daten ausgerichtet. Dazu wäre eine quelloffene NoSQL-Plattform erforderlich, die die Komplexität noch weiter erhöhen und die Implementierungsphase eines neuen Data Warehouse verlängern würde.

Wenn man es richtig anpackt, kann ein Cloud Data Warehouse innerhalb weniger Wochen oder Monate einsatzbereit sein. Die meiste Zeit sollte dem Extrahieren von Daten aus Ihren anderen Datenquellen und der Konfiguration eines Front-End-Analysetools gewidmet werden, um schnell Erkenntnisse aus dem Data Warehouse zu gewinnen.

Berechnung von Speicher- und Rechenkosten

On-Premise Data Warehouses sind teuer, was die Kosten für Hardware, Software und ihre Verwaltung anbelangt. Hardwarekosten umfassen zum Beispiel Kosten für Server, zusätzliche Speichergeräte, Platz im Rechenzentrum zur Unterbringung der Hardware, ein Hochgeschwindigkeitsnetzwerk für den Zugriff auf die Daten sowie die für den Betrieb des Systems erforderliche Stromversorgung und redundante Netzteile. Wenn Ihr Data Warehouse geschäftskritisch ist, müssen Sie noch die Kosten für die Konfiguration eines Disaster-Recovery-Standorts hinzufügen. Viele Unternehmen zahlen außerdem hohe Software-Lizenzgebühren für die Data-Warehouse-Software und Zusatzprodukte. Durch zusätzliche Endanwender, einschließlich Kunden und Lieferanten, die Zugang zum Data Warehouse erhalten, können diese Kosten noch weiter ansteigen. Hinzu kommen die laufenden Kosten für jährliche Supportverträge, die oft 20 Prozent der ursprünglichen Lizenzkosten ausmachen. Ein On-Premise Data Warehouse benötigt auch spezielles *IT-Personal* zur Bereitstellung und Wartung des Systems. Wenn Probleme auftreten, kommt es daher oft zu Engpässen und die Verantwortung für das System liegt dann beim Kunden, nicht beim Anbieter.

Mit einem Cloud Data Warehouse werden die anfänglichen Investitionsausgaben (CapEx) und die laufenden Kosten eines On-Premise-Systems in einfache Betriebskosten (OpEx) in der Form nutzungsbasierter Gebühren umgewandelt. Die monatliche Gebühr hängt davon ab, wie viele Speicher- und Rechenressourcen Sie tatsächlich nutzen. Konservativ betrachtet können die jährlichen Kosten für eine Cloud-Data-Warehouse-Lösung ein Zehntel der Kosten eines vergleichbaren On-Premise-Systems betragen.

Dimensionierung, Abgleich und Tuning

Für eine optimale Leistung muss ein On-Premise Data Warehouse modelliert, dimensioniert, abgeglichen und getunt werden, was eine

erhebliche Vorabinvestition sowie laufende Überwachungs- und Verwaltungskosten erfordert. Eine solche Konfiguration beinhaltet oft:

- Anzahl und Geschwindigkeit der Zentraleinheiten (CPUs)
- Menge an Speicherplatz
- Anzahl und Größe der Speicherplatten für die erforderliche Speicherkapazität
- E/A (Eingabe/Ausgabe)-*Bandbreite* (ein Maß dafür, wie viele Daten zu einem bestimmten Zeitpunkt übertragen werden können)
- Ein benutzerdefiniertes Datenmodell, das die Data-Warehouse-Struktur, die enthaltenen Datentypen und die Aktualisierungshäufigkeit definiert

Bei einem On-Premise Data Warehouse dimensionieren Unternehmen ihr System oft für Spitzenauslastungen, die nur einen kleinen Teil des Jahres ausmachen. In vielen Fällen benötigt ein Unternehmen die volle Leistung des Data Warehouse nur am Ende jedes Geschäftsquartals oder -jahres. Trotzdem muss für diese Spitzenkapazität 24 Stunden am Tag bezahlt werden, und zwar jeden Tag, da das System nicht einfach nach oben oder unten skaliert werden kann.

Elastisches Cloud Data Warehousing bietet zwei entscheidende Vorteile:

- Die Komplexität und die Kosten der Kapazitätsplanung und -verwaltung – Dimensionierung, Abgleich und Tuning des Systems – sollten in das System integriert, automatisiert und durch die Kosten Ihres Abonnements abgedeckt werden.
- Das Gleiche gilt für die dynamische Bereitstellung von Speicher- und Rechenressourcen während des Betriebs, um den Anforderungen Ihrer wechselnden Workloads in Spitzenzeiten und bei konstanter Nutzung gerecht zu werden. Kapazität bedeutet, immer das zu haben, was Sie brauchen. Allerdings sind nicht alle Workloads gleich. Mit einem elastischen Cloud Data Warehouse können Sie sehr genau festlegen, welche Ressourcen welchen Benutzern und Workloads zugewiesen werden sollen.

Datenaufbereitungs- und ETL-Kosten

Ein On-Premise Data Warehouse muss Daten aus allen Ihren Datenquellen extrahieren. Dann muss es diese Daten so umwandeln, dass sie sich an die oft starre Datenstruktur innerhalb des Systems anpassen, *bevor* sie in das Data Warehouse geladen werden. Eine wesentliche Herausforderung ist dabei die begrenzte und kostspielige Menge an

Verarbeitungskapazität und Speicherplatz. Die Datenumwandlung muss außerhalb der normalen Geschäftszeiten erfolgen, um nicht mit anderen Datenverarbeitungsaufträgen um Ressourcen wetteifern zu müssen. Das kann teuer werden. Darüber hinaus treffen semistrukturierte Daten nicht in einheitlichen Zeilen und Spalten ein, die den traditionellen Datenstrukturen entsprechen. Die Daten sind zudem hochvolumig und haben eine hohe Geschwindigkeit.

Die besten cloudbasierten Lösungen sind in der Lage, semistrukturierte Daten direkt zu laden, ohne sie umwandeln zu müssen. Diese Lösungen können den Zugriff auf neue Daten bis zu 50 Mal schneller als ein herkömmliches Data Warehouse zur Verfügung stellen. Durch die geringeren Kosten für unbegrenzten Cloud-Storage erhalten Datenanalysten außerdem Zugriff auf alle Daten, anstatt sich mit regelmäßigen Aggregaten dieser Daten begnügen zu müssen.

OPTIMIERUNG EINER DATA-PIPELINE

FALLSTUDIEN

DoubleDown, ein Online-Gaming-Studio, fügte seiner Data-Pipeline ein NoSQL-System hinzu, um Daten zum Laden in sein Data Warehouse aufzubereiten. Dieser Ansatz war jedoch mit langen Verarbeitungszeiten für das tägliche Ereignisprotokoll von DoubleDown (Benutzer-Klicks und andere durch die Aktivitäten der Spieler erzeugte Daten) verbunden. Das Unternehmen konnte erst am nächsten Tag um 15 Uhr auf die Daten des Vortages zugreifen. Schlimmer noch: Wenn einer der Daten-Cluster ausfiel, gingen dem Unternehmen Daten verloren.

DoubleDown entschied sich für ein System, mit dem semistrukturierte Daten direkt ohne vorherige Umwandlung geladen werden konnten und dadurch sofort für Abfragen zur Verfügung standen. Die Qualität und Performance der Data-Pipeline verbesserte sich erheblich, da die Daten fast 100 Mal schneller zu den Analysten gelangten – in 15 Minuten anstatt 24 Stunden. Dadurch wurden fast alle häufig auftretenden Fehler in der bisherigen Pipeline des Unternehmens eliminiert. Die Analysten profitierten von höchster Datengranularität anstelle von periodischen Aggregaten und die Kosten der Data-Pipeline von DoubleDown verringerten sich um 80 Prozent.

Die Analysten von DoubleDown haben jetzt sofortigen Zugriff auf Daten von neuen Produktveröffentlichungen und können schneller datengesteuerte Entscheidungen treffen.

Kosten spezieller Business-Analytics-Tools

Wie in Kapitel 3 erwähnt, sind traditionelle On-Premise Data Warehouses nicht auf den Umgang mit der Menge, Vielfalt und Geschwindigkeit heutiger Daten ausgerichtet. Aus diesem Grund betreiben Unternehmen meist zwei Datenplattformen: ein unternehmensweites SQL-Data Warehouse für die Speicherung herkömmlicher relationaler Daten und eine große NoSQL-Datenplattform, die vor Ort oder in der Cloud ausgeführt werden kann, um nicht-relationale Daten zu speichern.

Leider ist die Verwaltung dieser neueren Systeme sehr komplex und erfordert spezialisierte Tools und Fachkenntnisse, die nicht annähernd so verbreitet sind wie SQL-Tools und -Kenntnisse. Schließlich gibt es SQL schon seit Jahrzehnten, während NoSQL-Systeme noch relativ neu sind.

Die ideale Cloud-Data-Warehousing-Lösung stellt das Beste aus beiden Welten zur Verfügung – die nötige Flexibilität zur Integration relationaler und nicht-relationaler Daten sowie Support für die leicht verfügbaren SQL-Tools und -Fähigkeiten zum Abfragen dieser Daten.


TIPP

Wenn Sie ein neues Data-Warehouse implementieren wollen, sollten Sie die Kosten und die Verfügbarkeit der für die Verwaltung dieses Data Warehouse erforderlichen Fähigkeiten und Fachkenntnisse sowie die zahlreichen Analyse- und anderen Tools berücksichtigen, die in Verbindung mit einem Data-Warehouse benötigt werden.

Skalierbarkeit und Elastizität

Herkömmliche Data Warehouses sind für Systemverlangsamungen und -abstürze anfällig, da Benutzer und Prozesse um begrenzte Ressourcen wetteifern. Bei diesen Systemen sind Speicher- und Rechenressourcen auf einem einzigen Computer-*Cluster* (einer Gruppe von Computern) eng miteinander verbunden. Dadurch wird es kostspielig, nur den einen Aspekt, jedoch nicht den anderen zu vergrößern.

Neuere, in der Cloud erstellte Data-Warehouse-Lösungen bieten eine nahezu unbegrenzte Speicher- und Rechenkapazität; ziehen Sie jedoch ein Data Warehouse in Betracht, bei dem die Speicherkapazität getrennt von der Rechenleistung skaliert werden kann (siehe Abbildung 4-1). Im Idealfall sollte das Cloud-Data-Warehouse auf drei Arten skalierbar sein:

- **Storage:** Cloud-Storage ist von Natur aus skalierbar und die Speichermenge kann leicht an sich ändernde Anforderungen angepasst werden.
- **Compute:** Die für die Verarbeitung von Data Loads und Abfragen verwendeten Ressourcen sollten jederzeit leicht nach oben oder

unten skalierbar sein, wenn sich die Anzahl und Intensität der Workloads ändert.

- **Benutzer und Workloads (Parallelität):** Lösungen mit feststehenden Rechenressourcen werden langsamer, wenn die Benutzerzahl und die Workloads zunehmen. Unternehmen sind oft gezwungen, Daten in separaten Data Marts zu replizieren, einige Workloads außerhalb der normalen Geschäftszeiten zu verschieben und Benutzer in eine Warteschlange zu stellen, um die Leistung aufrechtzuerhalten. Nur in der Cloud kann ein Data Warehouse durch Hinzufügen dedizierter Compute-Cluster auf jede beliebige Größe und für eine nahezu unbegrenzte Anzahl von Benutzern oder Workloads skaliert werden, die alle auf einen Datenbestand zugreifen, ohne dass die Leistung der anderen beeinträchtigt wird.

Halten Sie nach einer Cloud-Lösung Ausschau, die Storage und Compute entkoppelt, sodass beide einfach und unabhängig voneinander skaliert werden können, um die Kosten gering zu halten. Die Lösung sollte auch horizontal skalierbar sein, um mehr Benutzer und Workloads zu unterstützen, ohne die Leistung zu beeinträchtigen.

Skalierung und Elastizität

- **Elastische Skalierung von Speicherressourcen**
  Kostengünstiger Cloud-Storage, vollständig repliziert und robust
- **Elastische Skalierung von Rechenressourcen**
  Auf- und Abwärtsskalierung für Workload-Anforderungen
- **Elastische Skalierung für Parallelität**
  Hinzufügen von Benutzern, Workloads ohne Performance-Einbußen

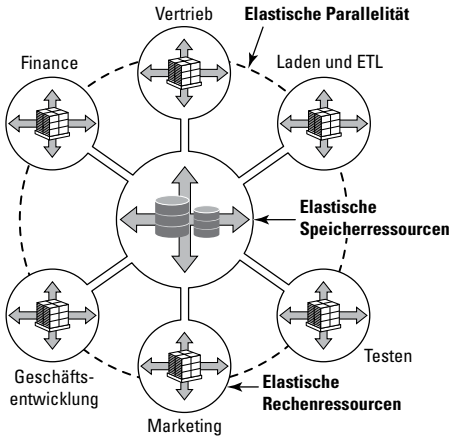

ABBILDUNG 4-1: Das ideale Data Warehouse sollte auf drei Arten skalierbar sein.

Verzögerungen und Ausfallzeiten

Viele Unternehmen mit On-Premise-Lösungen kämpfen vor allem mit zwei Problemen. Sie müssen mehrere Stunden oder sogar länger als einen Tag warten, bevor die am Vortag erfassten Daten im Data

Warehouse verfügbar sind. Genauso lange müssen sie warten, bis eine komplexe Abfrage in einem großen Datenbestand ausgeführt worden ist. In einigen Fällen können mehrere gleichzeitig ablaufende Prozesse dazu führen, dass das System hängen bleibt oder abstürzt, wodurch Verzögerungen und Ausfallzeiten noch verlängert werden.

Mit praktisch unbegrenzten Speicher- und Rechenressourcen sind Cloud-Data-Warehouse-Lösungen, die als dynamisch elastisch konzipiert sind, bestens für die Skalierung nach oben, unten und außen gerüstet, um steigenden Anforderungen gerecht zu werden. Um Verzögerungen zu verringern und ungeplante Ausfallzeiten zu vermeiden, ist jedoch mehr als nur die Aufstockung der Systemressourcen erforderlich. Bessere Lösungen optimieren die Data-Pipeline und speichern Daten zur effizienteren Ausführung von Abfragen ohne manuelles Tuning.

Halten Sie nach Lösungen Ausschau, die diese Arten von Performance-Problemen lösen und Ausfallzeiten minimieren können. Die Geschwindigkeit, mit der Sie auf Ihre Daten und Analysen zugreifen können, kann Ihre Betriebsabläufe und Ihre Fähigkeit, einen Wettbewerbsvorteil zu erhalten, erheblich beeinflussen.

Aus Sicherheitsproblemen resultierende Kosten

Ein einziger Verstoß kann sich schnell zu einem Public-Relations-Alptraum entwickeln und zu Geschäftsverlusten und hohen Bußgeldern seitens der Aufsichtsbehörden führen. Die Angst vor Sicherheitsrisiken in der Cloud ist immer noch weit verbreitet. Dabei kann die Cloud viel sicherer sein als Ihr Rechenzentrum.

Wenn Sie sich für ein On-Premise Data Warehouse entscheiden, sind Sie allein für den Schutz sensibler Daten verantwortlich. Dies erfordert eine sorgfältige und konstante Überwachung des Firewall-Schutzes, Sicherheitsprotokolle, Datenverschlüsselung bei der Speicherung und während der Übertragung, Benutzerrollen und -berechtigungen sowie Überwachung und Anpassung an neue Sicherheitsbedrohungen.

Die Implementierung effektiver Datensicherheit ist komplex und kostspielig, besonders in Bezug auf die Personalressourcen. Mit schlecht implementierten Sicherheitsmaßnahmen können Ihnen bei einer Sicherheitsverletzung noch mehr Kosten entstehen.

Da Anbieter von Cloud-Data-Warehousing-Lösungen zahlreiche Kunden haben, können sie sich das Fachwissen und die Ressourcen leisten, die erforderlich sind, um industrietaugliche, End-to-End-Sicherheit für die Data-Warehouses zur Verfügung zu stellen. Halten Sie nach einem Anbieter Ausschau, der eine dem Industriestandard

entsprechende End-to-End-Verschlüsselung gewährleistet, um Daten sowohl im Speicherzustand als auch während der Übertragung zu sichern.

Kosten für Datenschutz und -wiederherstellung

Bei On-Premise Data Warehouses besteht die Gefahr von Datenverlusten durch Geräteausfall, Stromausfälle oder Überspannungen, Diebstahl oder Vandalismus und Katastrophen (Feuer, Überschwemmung, Erdbeben usw.). Um Ihre Daten zu schützen, müssen Sie diese regelmäßig sichern und Backups an einem entfernten Standort aufbewahren. Eine Backup-Stromversorgung ist ebenfalls erforderlich, um Datenverluste zu verhindern und sicherzustellen, dass Ihr Data Warehouse jederzeit für die Verarbeitung eingehender Daten und Abfragen verfügbar ist. Im Notfall benötigen Sie qualifiziertes Personal, um Daten aus den neuesten Backups wiederherzustellen. Wenn Ihr Data Warehouse geschäftskritisch ist, benötigen Sie möglicherweise auch einen geografisch getrennten Standort für Disaster Recovery (ein zusätzliches Rechenzentrum) sowie Software, Lizenzen und Prozesse, um eine automatische Ausfallsicherung zu gewährleisten, damit es keine Serviceunterbrechungen gibt.

Die Cloud bietet eine ideale Lösung für die Datensicherung und -wiederherstellung. Aufgrund ihrer Beschaffenheit werden Daten außer Haus („Off-Premises“) gespeichert. Bei einigen cloudbasierten Lösungen werden Daten automatisch an zwei oder mehr separaten physischen Standorten gesichert. Wenn die Rechenzentren geografisch voneinander getrennt sind, bieten sie auch eine integrierte Disaster Recovery. Cloud-Rechenzentren verfügen über redundante Stromversorgungen, damit sie auch bei längeren Stromausfällen in Betrieb bleiben. Cloud-Anbieter können diese Schutzmaßnahmen zu wesentlich geringeren Kosten als Sie selbst bereitstellen, da sie die Kosten auf Tausende von Kunden verteilen.


TIPP

Wenn Sie Ihre Daten-Backups nicht selbst verwalten möchten, sollten Sie den potenziellen Anbieter Ihrer Cloud-Data-Warehouse-Lösung fragen, wie er seinen Dienst konfiguriert. Wenn Sie Disaster-Recovery-Schutz benötigen, sollten Sie auch bestätigen lassen, dass die Architektur des Anbieters geografisch getrennte Zentren verwendet. Fragen Sie außerdem, ob Ihr Anbieter seine Lösung über mehrere Cloud-Anbieter hinweg anbietet, falls Sie im Notfall zu einer Instanz Ihres Data Warehouse in einer anderen Cloud wechseln müssen.

IN DIESEM KAPITEL

- **Die Performance beeinflussende Faktoren**
- **Auswahl einer Lösung, die Datenschutz und -sicherheit bietet**
- **Einsparungen bei Verwaltungskosten**

Kapitel 5
Vergleich von Cloud-Data-Warehouse-Lösungen

Die zunehmende Verbreitung der Cloud hat dazu geführt, dass sowohl Anbieter älterer On-Premise-Lösungen als auch neue Markteinsteiger Cloud-Versionen ihrer Data-Warehouse-Produkte anbieten. Natürlich sind keine zwei Lösungen gleich. In diesem Kapitel sehen wir uns einige der Unterschiede an und erläutern, was bei Cloud Data Warehouses besonders zu beachten ist.

Ansätze für Data Warehousing in der Cloud

Die folgenden Cloud-Ansätze sind mit sehr unterschiedlichen Data-Warehouse-Funktionen verbunden:

- **Infrastructure-as-a-Service (IaaS):** Der Kunde muss herkömmliche Data-Warehouse-Software auf den vom Anbieter der Cloud-Plattform bereitgestellten Computern installieren. Der Kunde verwaltet alle Aspekte der Cloud-Hardware und der Data-Warehouse-Software. Die Funktionen des Data Warehouse sind mit denen der Software identisch, die auf On-Premise-Hardware bereitgestellt wird.
- **Platform-as-a-Service (PaaS):** Bei diesem hybriden Ansatz stellt der Data-Warehouse-Anbieter die Hardware und Software als Cloud-Service zur Verfügung. Der Anbieter verwaltet die Hardware-Bereitstellung, die Software-Installation und die Software-Konfiguration. Der Kunde verwaltet die Software und ist für Tuning und Optimierung verantwortlich.

- **Software-as-a-Service (SaaS):** Der Data-Warehouse-Anbieter stellt die gesamte Hard- und Software zur Verfügung. Dies umfasst alle Aspekte der Verwaltung der Hard- und Software. Im Service inbegriffen sind gewöhnlich: Software- und Hardware-Upgrades, Sicherheit, Verfügbarkeit, Datenschutz und Optimierung.

Bei all diesen Szenarien werden der Kauf, die Bereitstellung und die Konfiguration des Rechenzentrums selbst sowie der Hardware zur Unterstützung des Data Warehouse vom Kunden auf den Anbieter übertragen. Abgesehen von diesem Vorteil schneiden diverse Angebote hinsichtlich ihrer Benutzerfreundlichkeit, Sicherheit und Verfügbarkeit sehr unterschiedlich ab.

ERINNERUNG

Wenn ein Data-Warehouse-Anbieter lediglich Zugriff auf sein herkömmliches Data Warehouse über die Cloud bereitstellt, wird die Lösung der ursprünglichen On-Premise-Architektur und Funktionalität wahrscheinlich sehr ähnlich sein.

Vergleich der Architekturen

Viele Anbieter bieten ein Cloud-Data-Warehouse an, das ursprünglich für On-Premise-Umgebungen entwickelt und bereitgestellt wurde. Diese traditionellen Architekturen entstanden, lange bevor die Cloud mit ihren zahlreichen Vorteilen als praktikable Option angesehen wurde. Eine für die Cloud entwickelte Data-Warehouse-Lösung sollte die Vorteile der Cloud optimal ausnutzen (siehe Abbildung 5-1). Halten Sie auf der Suche nach einer auf einer cloudoptimierten Architektur basierenden Lösung Ausschau nach den folgenden Merkmalen:

- zentralisierte Speicherung aller Daten
- unabhängige Skalierung von Rechen- und Speicherressourcen
- nahezu unbegrenzte Parallelität ohne Wetteifern um Ressourcen
- gleichzeitiges Laden und Abfragen von Daten ohne Beeinträchtigung der Performance
- Replizierung von Daten über mehrere Regionen und Clouds hinweg, um die Geschäftskontinuität zu verbessern und die Expansion zu vereinfachen.
- gemeinsame Nutzung von Daten, ohne APIs oder umständliche ETL-Verfahren einzurichten
- ein robuster Metadatendienst für das gesamte System. (*Metadaten* sind Daten über andere Daten wie Dateigröße, Autor und Zeitpunkt der Erstellung). Eine cloudoptimierte Architektur macht sich auch die Vorteile von „Storage-as-a-Service" zunutze, ein Modell, bei dem der Datenspeicherplatz automatisch und für den Benutzer transparent

erweitert und verringert wird. Für ältere Architekturen konzipierte Datenspeicher sind teuer und nur begrenzt skalierbar.

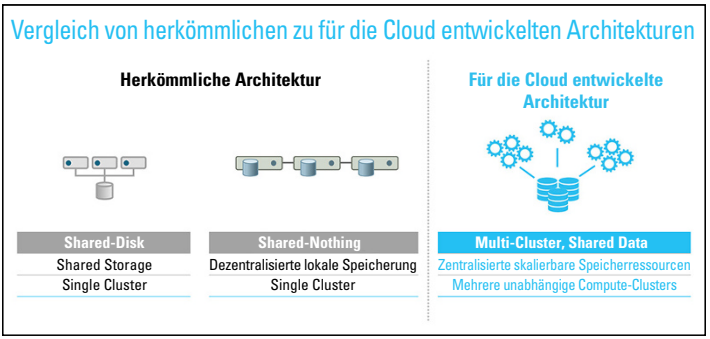

ABBILDUNG 5-1: Wie eine cloudoptimierte Architektur die Performance optimiert.

Verwaltung verschiedenartiger Daten

Ein Schlüsselfaktor, der die Einführung von Cloud Data Warehousing vorantreibt, ist das wachsende Datenvolumen, das aus der Cloud hervorgeht – außerhalb des Rechenzentrums eines Unternehmens. In den meisten Fällen müssen diese nicht-relationalen Daten umgewandelt werden, bevor sie in ein herkömmliches On-Premise Data Warehouse vor Ort oder in der Cloud geladen werden. Dieser Ansatz erhöht die Komplexität erheblich und verzögert den Zugriff auf neue Daten.

Aufgrund dieser größeren Menge und Vielfalt von Daten ist die Cloud zu einem natürlichen Integrationspunkt geworden. Eine ideale Lösung für dieses Problem ist ein Cloud Data Warehouse, das sowohl relationale als auch nicht-relationale Daten verarbeiten kann – und zwar ohne Umwandlung nicht-relationaler Daten und ohne Beeinträchtigung der Performance beim Laden von Daten oder bei der Verarbeitung von Abfragen.

ERINNERUNG

Daten müssen umgewandelt werden, bevor sie in ein herkömmliches, cloudbasiertes Warehouse geladen werden können. Anderenfalls muss das Unternehmen ein zusätzliches System zur Verarbeitung nicht-relationaler Daten anschaffen und betreiben.

Skalierung und Elastizität

Nicht alle Cloud Data Warehouses weisen die gleiche Art von Elastizität auf. Fortschrittliche Lösungen können während des Betriebs nach oben und unten skaliert werden, ohne dass das System offline genommen oder in einen Nur-Lese-Modus versetzt werden muss.

TIPP

Lösungen, die sich nicht gut skalieren lassen, haben die folgenden Nachteile:

- Bei einem Cloud Data Warehouse, das manuell nachkonfiguriert werden muss, ist zur Skalierung von Ressourcen eine sorgfältige Planung und Koordination mit dem Anbieter erforderlich.
- Die Skalierung kann eine Ausfallzeit oder einen Wechsel in den Nur-Lese-Modus erfordern, damit die Daten umverteilt und das System neu konfiguriert werden kann.
- Die meisten Cloud-Data-Warehouse-Angebote bündeln Compute und Storage auf demselben Knoten, sodass Kunden beides skalieren müssen, auch wenn nur eine Ressource erweitert werden muss.
- Bei den meisten dieser Lösungen handelt es sich um „Cloud-Washed"-Versionen von On-Premise-Lösungen; für den potenziellen Spitzenbedarf müssen Sie eine überdimensionierte Konfiguration erwerben, die meist nicht ausgelastet ist. Irgendwann ist das Limit der verfügbaren Ressourcen erreicht und es müssen kostspielige Upgrades durchgeführt werden.

Vergleich von Parallelitätsfunktionen

Parallelität ist die Fähigkeit, zwei oder mehrere Aufgaben gleichzeitig auszuführen bzw. zwei oder mehreren Benutzern Zugriff auf eine Rechenlösung zu ermöglichen. Bei einem herkömmlichen Data Warehouse wird die Parallelität durch feststehende Rechen- und Speicherressourcen beschränkt. In der Cloud sind Compute und Storage jedoch nicht festgelegt. Cloudoptimierte Architekturen unterstützen die Parallelität auf folgende Weise:

- Mehrere Benutzer können dieselben Daten gleichzeitig abfragen, ohne dass die Performance beeinträchtigt wird.
- Lade- und Abfragevorgänge können gleichzeitig erfolgen, sodass mehrere Workloads ohne Ressourcenkonflikte gleichzeitig ausgeführt werden können.

Support für SQL und andere Tools

Fast alle BI (Business Intelligence)-, ETL (Extract, Transform, Load)- und Data-Analytics-Tools können mit einem Data Warehouse kommunizieren, das Standard-SQL unterstützt. Allerdings wird Standard-SQL nicht uneingeschränkt von allen Cloud-Data-Warehouse-Lösungen unterstützt. Bei Big-Data-Lösungen, die sich als „Cloud Data Warehouses" positionieren, handelt es sich zum Beispiel oft um NoSQL-Lösungen mit unvollständigem oder Non-Standard-SQL-Support.

Obwohl die Unterstützung dieser neueren Analysetools wichtig ist, bleibt SQL weiterhin der Industriestandard für die Abfrage von Daten. Ihr Data Warehouse sollte also SQL-Tools für Datenmanagement, Datenumwandlung, Datenintegration, Visualisierung, BI und andere Arten von Analysen unterstützen.

Unterstützung für Backup/Wiederherstellung

Bei On-Premise- und zahlreichen Cloud-Data-Warehousing-Lösungen müssen Kunden ihre eigenen Daten mit Tools zur Datensicherung und Datenreplikation schützen. Bei einigen Cloud-Data-Warehouse-Lösungen ist die Datensicherung jedoch im Service inbegriffen.

ERINNERUNG

Für einen optimalen Schutz sollten Sie nach einer Lösung Ausschau halten, die frühere Datenversionen automatisch speichert oder Daten automatisch zur Verwendung als Online-Backup dupliziert. Die Lösung sollte im Interesse einer vollständigen Geschäftskontinuität auch die Self-Service-Wiederherstellung von verlorenen oder korrumpierten Daten umfassen, und zwar mittels Replikation über verschiedene Regionen bei einem Cloud-Provider oder Replikation über mehrere Cloud-Provider.

Ausfallsicherheit und Verfügbarkeit

Ausfallsicherheit ist die Fähigkeit des Data Warehouse, auch beim Ausfall von Komponenten, Netzwerken oder des gesamten Rechenzentrums den Betrieb automatisch fortzusetzen. *Verfügbarkeit* ist die Fähigkeit der Benutzer, jederzeit auf das System zuzugreifen (auch als „Uptime" bekannt). Die Verantwortung des Kunden für die Verfügbarkeit und Ausfallsicherheit ist nicht bei allen Cloud-Data-Warehouse-Services gleich groß. Grundsätzlich kann ein Cloud-Data-Warehouse-Service vom Kunden verlangen, dass er die Systemüberwachung übernimmt, um Ausfälle zu erkennen und möglicherweise zu verhindern. Möglicherweise muss der Kunde auch für die Datenreplikation sorgen, damit bei einem Ausfall ein Duplikat des Data Warehouse zur Verfügung steht. Am anderen Ende des Spektrums kann der Service des Anbieters die Überwachung, Replikation und automatische Ausfallsicherung umfassen.

Verfügbarkeit ist auch ein wichtiger Faktor für Software-Upgrades. Anbieter haben unterschiedliche Herangehensweisen an Upgrades:

- **Basic:** Kunden verwalten Upgrades und damit verbundene Ausfallzeiten.
- **Besser:** Der Anbieter verwaltet die Upgrades und informiert die Benutzer über bevorstehende Upgrades, damit sie für die Ausfallzeit planen können.

- **Am besten:** Der Anbieter sorgt für transparente Upgrades, ohne dass die Benutzer involviert sind oder Ausfallzeiten in Kauf nehmen müssen. Der Anbieter gibt den Kunden auch die Möglichkeit, sich für oder gegen automatische Upgrades zu entscheiden, sodass sie Upgrades nur dann erhalten, wenn sie sie wünschen.

TIPP

Achten Sie darauf, wie viele „Neunen der Verfügbarkeit" die Cloud-Data-Warehouse-Lösung unterstützt (99,9XX Prozent Betriebszeit).

Optimierung der Performance

Eines der großen Versprechen der Cloud ist ihre Fähigkeit, große Mengen an Ressourcen zur Verfügung stellen zu können, für die man nur dann bezahlen muss, wenn man sie braucht. Halten Sie nach einer Cloud-Data-Warehouse-Lösung Ausschau, die die Performance bei Bedarf optimieren kann und den Verwaltungsaufwand für die Integration neuer Ressourcen beseitigt.

ERINNERUNG

Halten Sie sich von Data Warehouses fern, bei denen Aktivitäten unterbrochen oder verzögert werden, wenn man Ressourcen hinzufügen oder entfernen will. Einige Lösungen sind mit einem erheblichen Verwaltungsaufwand verbunden, z. B. zur Umverteilung von Daten und zur Neuberechnung von Metadaten.

Datensicherheit in der Cloud

Die Cloud wird oft als weniger sicher als On-Premise-Speicherlösungen angesehen. Nach mehreren Einbrüchen in „sichere" On-Premise-Rechenzentren haben sich Cloud-Lösungen jedoch immer mehr durchgesetzt. Diese Vorfälle haben gezeigt, dass Unternehmen nur begrenzt in der Lage sind, ihre eigenen Daten zu sichern. Mit Cloud-Data-Warehousing-Lösungen wird die Verantwortung für die Sicherheit des physischen Rechenzentrums auf den Lösungsanbieter verlagert. Dabei ist jedoch Vorsicht geboten: Die Sicherheitsmerkmale unterscheiden sich von Anbieter zu Anbieter:

- Grundlegende Cloud-Data-Warehouse-Lösungen verfügen nur über einige Sicherheitsfunktionen und überlassen dem Kunden die Verschlüsselung, Zugriffskontrolle und Sicherheitsüberwachung.
- Andere Lösungen umfassen Funktionen wie Verschlüsselung und Zugriffskontrolle, die der Kunde auf Wunsch aktivieren kann. Wenn diese Funktionen nicht aktiviert werden, ist das System jedoch für Sicherheitsverletzungen anfällig.
- Cloud-Data-Warehouse-Angebote, die serviceorientierter sind, enthalten Sicherheitsfunktionen und bieten Verschlüsselung,

Schlüsselverwaltung, Schlüsselrotation, Einbruchserkennung und andere Funktionen im Rahmen des Service an.

Verwaltung

Herkömmliche Data Warehouses erfordern einen erheblichen Zeit- und Arbeitsaufwand und Fachkenntnisse von Seiten des Kunden. Ein oder mehrere Datenbankadministratoren (DBAs) müssen Software-Patches und -Upgrades, die Datenpartitionierung und Neupartitionierung, Index-Management, Workload-Management, Statistikaktualisierungen, Sicherheitsmanagement und -überwachung, Backups und Replikation, das Feinabstimmen und Umschreiben von Abfragen und viele andere Aufgaben durchführen.

Grundsätzlich muss der Kunde bei einer Cloud-Data-Warehouse-Lösung, die auf älterer On-Premise-Technologie basiert, alle diese Aspekte weiterhin verwalten. Bei neueren Data-Warehousing-Angeboten wird ein Großteil dieses Verwaltungsaufwands durch neue Designs und Automatisierung reduziert oder beseitigt.

Sicherer Datenaustausch (Data-Sharing)

Viele Unternehmen können ihre betrieblichen Abläufe durch die Nutzung von Datenspeichern, -services und -strömen von Drittanbietern verbessern. Bei herkömmlichen Datenaustauschmethoden wie FTP, APIs und E-Mail müssen Sie Daten kopieren und an Verbraucher senden. Diesen umständlichen, kostspieligen und risikobehafteten Methoden liegt die gemeinsame Nutzung statischer Daten zugrunde, die schnell veraltet sind und ständig durch neuere Versionen aktualisiert werden müssen. In Kapitel 6 erfahren Sie, wie ein cloudbasiertes Data Warehouse einen sicheren und verwalteten Datenaustausch in Echtzeit ermöglicht.

TIPP

Mithilfe der heute verfügbaren robusten Datenaustauschmethoden ist es möglich, Live-Daten auszutauschen, ohne sie von einem Ort zum anderen bewegen zu müssen.

Globale Datenreplikation

Bei der *Datenreplikation* werden mehrere Kopien Ihrer Daten in der Cloud erstellt. Diese Art globaler Präsenz ist nicht nur für die Disaster Recovery und die Geschäftskontinuität unerlässlich: Sie ist auch nützlich, wenn Sie Daten mit einem globalen Kundenstamm teilen möchten, ohne ETL-Pipelines zwischen den Regionen einrichten zu müssen. Mit den Lösungen führender Data-Warehouse-Anbieter können Daten

problemlos über mehrere geografische Regionen und über mehrere Clouds hinweg genutzt werden, einschließlich Amazon Web Services (AWS), Microsoft Azure und Google Cloud Platform (GCP). Diese globalen Replikationsfunktionen können Ihnen dabei helfen, Ihre Märkte zu erweitern, besser mit Partnern zusammenzuarbeiten und ein umfassenderes Ökosystem für Analysen und Datenaustausch zu entwickeln.

Isolierung von Workloads

Ein wesentlicher Faktor, der die Geschwindigkeit und die Performance eines Data Warehouse beeinflusst, ist die Fähigkeit, Workloads zu isolieren. Das Cloud-Data-Warehouse sollte in der Lage sein, problemlos mehrere Pools von Rechenressourcen (unterschiedlicher Größe) zu konfigurieren, um die Workloads von Benutzern und Prozessen zu trennen, die gleichzeitig ausgeführt werden müssen. Dadurch werden nicht nur Konflikte vermieden, sondern Ressourcen werden auch in angemessener Größe für die jeweiligen Workloads bereitgestellt. Im Idealfall sollten diese getrennten Workloads gleichzeitig auf dieselben Daten zugreifen können und sich je nach Bedarf leicht ein- und ausschalten lassen.

Abdecken aller Anwendungsfälle

In herkömmlichen Umgebungen werden unterschiedliche Anwendungsfälle in verschiedenen Datensystemen bearbeitet – ein Data Warehouse für die operative Berichterstattung, Data Marts für die Berichterstattung und Analyse auf Abteilungsebene, Data Lakes für die Datenexploration und spezialisierte Tools für Aktivitäten wie Predictive Analytics. Jedes dieser Systeme benötigt Hardware, eine Kopie der Daten, seine eigene Verwaltung usw.

Um diese unterschiedlichen Anwendungsfälle in der Cloud zusammenzubringen, sollte ein Data Warehouse in der Lage sein, mehrere Kopien von Tabellen, Schemata und Datenbanken schnell und effizient zu klonen, jedoch ohne all das Kopfzerbrechen und die Kosten, welche die mit herkömmlichen Formen der Datenduplizierung verbundene Speicherung gewöhnlich mit sich bringt. Das Cloud-Data-Warehouse sollte dem Unternehmen dabei helfen, sich schnell von Fehlern oder Problemen zu erholen, die durch Datenumwandlungsprozesse verursacht wurden. Dies kann durch Funktionen wie Time-Travel erreicht werden, die einen einfachen Zugriff und das „Zurücksetzen“ (Rollback) auf frühere Datenversionen ermöglichen.

IN DIESEM KAPITEL

- Die Bedeutung des Datenaustauschs
- Aufbau einer effizienten Architektur für den Datenaustausch
- Nutzung von Möglichkeiten zum Datenaustausch

Kapitel 6
Nutzung von Data Sharing (Datenaustausch)

Datenaustausch (Data Sharing) bedeutet, Zugang zu Daten zur Verfügung zu stellen – sowohl innerhalb eines Unternehmens als auch zwischen Unternehmen, die wertvolle Ressourcen miteinander teilen möchten. Das Unternehmen, das seine Daten zur Verfügung stellt oder teilt, ist ein *Datenlieferant*. Das Unternehmen, das die Daten nutzen möchte, ist ein *Datenkonsument*. Jedes Unternehmen kann ein Datenlieferant, ein Datenkonsument oder beides sein.

Neben den Daten, die Unternehmen intern generieren und gemeinsam nutzen, machen sich viele auch die Datenbestände, Services und Datenströme von Drittanbietern zunutze, um ihre betrieblichen Abläufe zu verbessern. Ein Finanzdienstleistungsunternehmen kann beispielsweise verschiedene Markt-, Finanz- und Wirtschaftsindikatoren zur Erstellung besserer Datenmodelle nutzen, die wiederum die Schaffung neuer Produktangebote für seine Kunden ermöglichen.

Die weltweit zunehmenden Datenquellen bieten eine Fülle potenzieller Werte, die auszuschöpfen sich lohnt – sowohl intern als auch durch externe Marktplätze und Börsen. Bis vor kurzem gab es noch keine Technologie für den Datenaustausch, die nicht mit erheblichen Risiken, Kosten, Kopfzerbrechen und Verzögerungen verbunden war. Obwohl Data Sharing seit fast einem Jahrhundert kommerziell genutzt wird, waren die Möglichkeiten bisher begrenzt. Stellen Sie sich vor, was möglich wäre, wenn alle Unternehmen bei Bedarf Zugriff auf gebrauchsfertige Live-Daten hätten und diese sofort nutzen könnten. Die Daten müssten nicht mehr vom Datenlieferanten dekonstruiert, zum Datenkonsumenten verschoben und dann vom Datenkonsumenten

rekonstruiert werden. Sie wären sofort zugänglich und innerhalb einer sicheren, verwalteten Umgebung einsatzbereit.

Technische Herausforderungen

Bei herkömmlichen Methoden zum Datenaustausch wie File Transfer Protocol (FTP), Cloud Storage (Amazon S3, Box, Dropbox u. a.), Application Programming Interfaces (APIs) und E-Mail muss eine Kopie der gemeinsam zu nutzenden Daten erstellt und an den Datenkonsumenten gesendet werden. Diese umständlichen, kostspieligen und risikobehafteten Methoden erzeugen statische Daten, die schnell veralten und andauernd durch neuere Versionen ersetzt und aktualisiert werden müssen, wozu die ständige Bewegung und Verwaltung von Daten erforderlich ist.

Mithilfe neuer Datenaustauschtechnologie können Unternehmen Segmente ihrer Daten problemlos und auf sichere, geregelte Weise austauschen und Daten zur gemeinsamen Nutzung empfangen. Dazu sind keine Datenbewegungen, keine ETL-Technologien (Extract, Transform, Load) und keine ständigen Updates erforderlich, um die Aktualität der Daten zu gewährleisten. Es ist nicht nötig, Daten über FTP zu übertragen oder APIs zur Verknüpfung von Anwendungen zu konfigurieren. Da die Daten nicht kopiert, sondern gemeinsam genutzt werden, ist kein zusätzlicher Cloud-Storage erforderlich. Mit dieser neuen Architektur können Datenanbieter Daten einfach und sicher veröffentlichen, damit sie von Datenkonsumenten sofort ermittelt, abgefragt und angereichert werden können, wie in Abbildung 6-1 dargestellt.

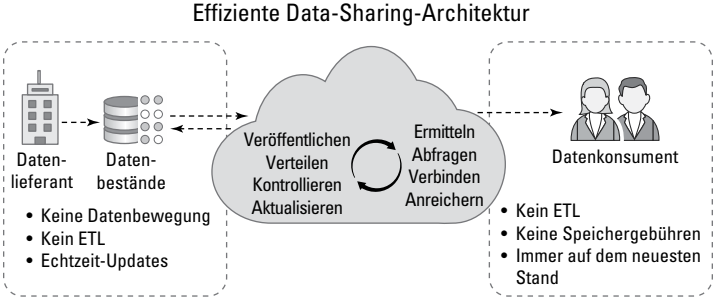

ABBILDUNG 6-1: Eine effiziente Architektur für den Datenaustausch in Echtzeit.

Ein mandantenfähiges Cloud-Data Warehouse ist eine ideale Plattform für einen Data-Sharing-Service, da es befugten Mitgliedern eines Cloud-Ökosystems das Abrufen schreibgeschützter Live-Versionen der Daten ermöglicht. Datenlieferanten können Daten mit Anbietern, Lieferketten- und Logistikpartnern, Kunden und vielen anderen Beteiligten austauschen. Diese cloudbasierten Lösungen machen sich die neuesten Entwicklungen im Bereich Cloud Computing und Data

Warehousing zunutze. Anstatt die Daten physisch an interne oder externe Konsumenten zu übertragen, ermöglicht das Data Warehouse den schreibgeschützten Zugriff auf einen bestimmten Teil des Live-Datensatzes über SQL.

Erfolgreicher Datenaustausch

Die meisten Unternehmen, die Datenaustausch betreiben wollen, folgen einem bekannten Prozess:

1. **Interne Zusammenarbeit:** Daten werden innerhalb des Unternehmens unter Geschäftsbereichen und Zweigniederlassungen ausgetauscht, wodurch die Zusammenarbeit verbessert und Datensilos aufgelöst werden.
2. **Geschäftserkenntnisse:** Vollständigere Daten verbessern die Zusammenarbeit und bieten bessere Geschäftseinblicke, da die gemeinsame Nutzung von Daten zur Norm wird.
3. **Kundenanalysen:** Das Unternehmen erstellt kundenorientierte Analysen, um den Wert eines Produkts oder einer Dienstleistung zu verbessern – der erste Schritt zur Monetarisierung der Daten.
4. **Erweiterte Analysen:** Wenn Kunden mehr Daten anfordern, entwickelt das Unternehmen kundenspezifische Analysedienste, um den Kunden reichhaltige Informationen aus seinen Daten zur Verfügung stellen zu können.
5. **Datenservices:** Das Unternehmen nutzt interne Datenbestände, um seinen Kunden auch Services zur Datenaugmentation wie Datenmodellierung, Datenanreicherung und Datenanalyse anzubieten.
6. **Datenbörse:** Das Unternehmen sucht nach Möglichkeiten, seine Datenprodukte zu verbessern, indem es externe Daten beschafft und seine Datenprodukte einem breiteren Publikum anbietet, in der Regel über einen Datenmarktplatz oder eine Datenbörse.

Monetarisierung von Daten

Die meisten Unternehmen betreiben bereits Datenaustausch oder haben dies vor. Dabei ist ihnen möglicherweise nicht bewusst, wie sie ihre Daten monetarisieren können. Es gibt einen großen, schnell wachsenden Marktplatz zur Monetarisierung von Daten. In dem IDC-Bericht „2019 Predictions for Digital Transformation" prognostiziert das Forschungsunternehmen, dass 80 Prozent der Unternehmen bis 2020 Möglichkeiten für das Datenmanagement und die Monetarisierung von Daten entwickeln werden und dass bis 2023 95 Prozent der Unternehmen neue digitale Key Performance Indicators (KPIs) integrieren werden.

TIPP

Mit der richtigen Architektur für den Datenaustausch können Sie problemlos einen größeren Teil Ihrer Daten analysieren, um neue Produkte, Dienstleistungen und Marktchancen zu entdecken.

FALLSTUDIEN

AXIMIERUNG DER UMSATZCHANCEN

Environics Analytics ist eines der führenden Datenanalyse-Unternehmen in Nordamerika. Um mehr als 3.000 Kunden datengesteuerte Einblicke bieten zu können, nimmt Environics große Mengen an demographischen, Standort- und Verbraucherdaten auf und analysiert sie.

Environics hat diese Analysetätigkeiten vor kurzem in ein cloudbasiertes Data Warehouse verlagert, das jede Datenmenge und jede Anzahl von Workloads bewältigen kann. Mithilfe eines integrierten Datenaustauschservice können Kunden neue Daten finden und sofort erhalten. Laut Sean Howard, Senior Vice President of Product Development bei Environics, stellt ein sicherer Datenaustauschservice einen praktischen Mechanismus zur Datenlieferung dar und bietet enorme Möglichkeiten zur Steigerung der Geschäftseinnahmen. Die Cloud-Plattform lässt sich schnell nach oben oder unten skalieren, um die Analyseanforderungen jedes einzelnen Benutzers zu erfüllen – ohne die Hilfe des IT-Teams.

Früher speicherten die Datenwissenschaftler von Environics Datenbestände auf ihren Computern und tauschten fertige Produkte über FTP mit ihren Kunden aus. Dies führte intern oft zu Verwirrung und behinderte das Geschäftswachstum. Bei der Exploration riesiger Datenbestände mit Milliarden von Ereigniszeilen wurde ständig die Unterstützung des IT-Teams benötigt – um Hardware zu installieren, SQL-Server-Umgebungen aufzubauen, die Abfrageleistung zu optimieren und die Nutzung von Speicher- und Rechenressourcen zu überwachen.

Mit einer Analyseumgebung, die bei Bedarf skaliert werden kann, können die Datenwissenschaftler nun bedenkenlos große Datenbestände aus jeder Branche, jeder Quelle und jedem Dateityp prototypisieren. Sie können Milliarden von Rohdatenpunkten in brauchbare Datenprodukte umwandeln. Der sichere Data-Sharing-Service erhöht die Kundenbindung, reduziert die Kosten für die Auftragsabwicklung und beseitigt unnötige Dateiübertragungen. Gleichzeitig wird die Versionsverwaltung erheblich vereinfacht.

Einzelhändler, Banken, Kreditgenossenschaften, Immobilienunternehmen, gemeinnützige Organisationen und Regierungsbehörden nutzen den Datenaustausch, um fundierte Entscheidungen über Verbraucher und Märkte zu treffen. Dank eines neuen Service, der für eine kontinuierliche Datenaufnahme sorgt, das Laden von Daten beschleunigt und Analysen in Nahe-Echtzeit ermöglicht, experimentiert Environics jetzt mit Daten aus dem Internet of Things (IoT) und aus anderen großen Datenquellen. „Die Teilnahme am Datenaustausch wird unser Geschäftswachstum fördern und uns dabei helfen, mehr potenzielle Kunden mit unseren Daten zu erreichen“, sagte Howard.

IN DIESEM KAPITEL

- **Disaster Recovery und Geschäftskontinuität**
- **Portabilität zwischen Clouds – ohne Anbieterbindung**
- **Globale Expansionsinitiativen**
- **Vereinfachung von Sicherheit und Verwaltung in Multi-Cloud-Umgebungen**

Kapitel 7
Eine Multi-Cloud-Strategie

Ein Data Warehouse, das sich über mehrere Regionen und Clouds erstrecken kann, bietet enorme Vorteile in Bezug auf den Datenaustausch, die Geschäftskontinuität und die geografische Durchdringung. Laut dem Bericht „2019 State of the Cloud“ von Flexera haben 84 Prozent der Unternehmen eine Multi-Cloud-Strategie, die die Realitäten des Marktes widerspiegelt. Ob Amazon Web Services, Microsoft Azure oder Google Cloud Platform – jeder Cloud-Service erfüllt andere Bedürfnisse, auch wenn die Abweichungen oft nur gering sind.

Für Unternehmen, die mit ihrem Data Warehouse globale Reichweite anstreben, ist eine Cross-Cloud-Strategie sinnvoll: Sie ermöglicht die freie und sichere Bewegung von Daten – überall in der Welt. Gleichzeitig können Sie jene Cloud-Storage-Anbieter auswählen, die Ihren Anforderungen am besten entsprechen. Vielleicht hat jede Abteilung in Ihrem Unternehmen ihre eigenen speziellen Cloud-Anforderungen. Anstatt zu verlangen, dass alle Geschäftsbereiche denselben Anbieter nutzen, kann mit einer Multi-Cloud-Strategie jeder Bereich diejenige Cloud nutzen, die für ihn am besten geeignet ist. Wenn diese Flexibilität für Sie wichtig ist, sollten Sie nach einem Anbieter Ausschau halten, der mehrere Cloud-Umgebungen unterstützt und Cross-Cloud-Support anbietet.

Cross-Cloud

Multi-Cloud bedeutet, dass Sie Ihre Daten in mehreren unterschiedlichen Clouds speichern können. *Cross-Cloud* bedeutet, dass Sie gleichzeitig auf Daten aus all diesen Clouds zugreifen, analytische Abläufe nahtlos von einer Cloud in eine andere migrieren und Daten zwischen den Clouds austauschen können. Dies ist der heilige Gral des Cloud Data Warehousing, da Sie nicht an einen einzigen Cloud-Anbieter gebunden sind. Warum ist das so wichtig?

- Es ist ein strategischer Vorteil für globale Unternehmen, da nicht alle Cloud-Anbieter in allen Regionen tätig sind.
- Es kann bei der Übernahme eines Unternehmens nützlich sein, das eine andere Cloud als Standard übernommen hat als die von Ihnen genutzte Cloud.
- Wenn Sie vorhaben, Ihre Daten zu teilen oder zu monetarisieren, können Sie mit einer einheitlichen Datenmanagement-Plattform, die sich über Regionen und Clouds hinweg erstreckt, Ihren adressierbaren Markt erweitern.

In den folgenden Abschnitten gehen wir näher auf die Technologien ein, die einem cloudübergreifenden Data Warehouse zugrunde liegen.

TIPP

Arbeiten Sie mit einem Data-Warehouse-Anbieter zusammen, der bereits viel Arbeit investiert hat, um das Problem unterschiedlicher Cloud-Konfigurationen zu lösen, und dessen Lösung auf einer gemeinsamen Codebasis aufbaut, die sich über alle Clouds erstreckt.

Globale Replikation

Datenreplikation ist ein Prozess, bei dem Daten an mehr als einem Ort gespeichert werden, um die Datenverfügbarkeit während eines räumlich begrenzten Ausfalls sicherzustellen. Es ist auch die grundlegende Technologie, die den Datenaustausch über Regionen und Clouds hinweg ermöglicht. Data Warehouses benötigen eine fortschrittliche Datenreplikationstechnologie, um regionale Bereitstellungsoptionen zu maximieren, Geschäftskontinuität zu gewährleisten und den Betrieb weltweit auszuweiten.

Ihre Data-Warehouse-Plattform sollte eine regionen- und cloudübergreifende Replikation ermöglichen, ohne die betriebliche Leistung in Bezug auf Ihre Primärdaten zu beeinträchtigen.

Minimierung von Serviceunterbrechungen

Die cloudübergreifende Data-Warehouse-Replikation ist für geschäftskritische Disaster-Recovery-Szenarien wichtig. Sie sorgt bei einem Ausfall dafür, dass Sie die Datenverarbeitung sofort wieder aufnehmen können, ohne dass es zu Ausfallzeiten kommt (siehe Abbildung 7-1). Ohne die richtige Datenreplikationstechnologie kann die Wiederherstellung von Geo-Backups für große Data Warehouses Stunden oder sogar Tage dauern. Entspricht dies Ihren Vorstellungen in Bezug auf die Wiederherstellungszeit?

Fragen Sie Ihren Data-Warehouse-Anbieter, ob er den sofortigen Zugriff und die Wiederherstellung von Datenbanken jeder Größe, in jeder Cloud und in jeder Region unterstützt. Wenn es irgendwo auf der Welt zu einem Notfall kommen sollte, können Sie sofort auf die in einer anderen Region oder einem anderen Cloud-Service replizierten Daten zugreifen. Finden Sie heraus, ob Ihr Data-Warehouse-Anbieter Datenbanken repliziert und diese über Cloud-Plattformen und Regionen hinweg synchron hält.

Regionen- und cloudübergreifende Replikation

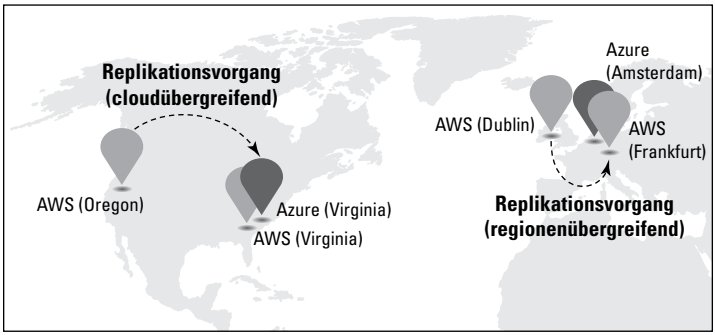

ABBILDUNG 7-1: Die globale Datenreplikation sorgt bei Ausfällen für Geschäftskontinuität.

Unterstützung mehrerer Clouds

Datenportabilität ist eine Herausforderung für alle Unternehmen, die über große Datenmengen verfügen. Jeder Public-Cloud-Anbieter hat einen unterschiedlichen Grad der regionalen Durchdringung. Mit einer cloudübergreifenden Architektur ist es einfacher, Daten und Workloads zwischen geografischen Regionen und Clouds zu verschieben.

Die Datenportabilität vereinfacht die Einhaltung von Vorschriften, wenn Ihre Branche vorschreibt, dass Ihre Daten in einem bestimmten Land

oder in einer bestimmten Region verbleiben müssen. Eine Fusion oder Übernahme eines anderen Unternehmens, das einen anderen Cloud-Anbieter verwendet, ist in diesem Fall ebenfalls einfacher.

Datenhoheit

Wenn Ihr Unternehmen wächst, möchten Sie Ihre Datenverarbeitungsprozesse möglicherweise in den Regionen ansiedeln, die Sie bedienen. Eine Multi-Cloud-Strategie gibt Ihnen die nötige Flexibilität, um in jeder Region die stärkste Cloud auszuwählen. So können Sie eine Architektur aufbauen, die die Latenzzeiten auf ein Minimum reduziert, Geo-Residency-Anforderungen entspricht und die Vorgaben der Datenhoheit erfüllt. Sie können den Betrieb in entlegene Regionen ausdehnen, ohne den Datenzugriff aufs Spiel zu setzen, und den Wert einer Single Source of Truth, d. h. einer einzigen relevanten Datenquelle für Ihr gesamtes Unternehmen, schätzen lernen.

Die Datenreplikation erleichtert den Austausch und die Monetarisierung von Daten sowie die Einbindung von Partnern in den Austausch. Das Grundprinzip des Data Sharing bleibt dabei gewahrt: Daten existieren vor Ort in einer einzigen Datenquelle, von der aus auf sie zugegriffen werden kann, anstatt sie zu verschieben.

Vereinfachung von Sicherheitsfunktionen

Wie können Sie bei der Arbeit mit mehreren Clouds sicherstellen, dass die gleichen Sicherheitskonfigurationen und -techniken für alle Ihre Cloud-Anbieter gelten? Müssen Sie Unterschiede bei Audit-Trails und Ereignisprotokollen beheben? Müssen sich Ihre Cybersicherheitsexperten mit unterschiedlichen Regelsätzen befassen oder an mehreren Schlüsselverwaltungssystemen zur Verschlüsselung von Daten herumbasteln? Eine einheitliche Codebasis, die alle Cloud-Plattformen umfasst, vereinfacht all diese Vorgänge. Sie müssen keine Mitarbeiter mit speziellen Fähigkeiten einstellen oder mit den Nuancen mehrerer Clouds vertraut sein.

ERINNERUNG

Dank fortschrittlicher Replikationstechnologie können Sie Daten problemlos über mehrere Regionen und die Clouds verschiedener Anbieter hinweg teilen – ohne Daten-Pipelines einrichten, Daten kopieren oder Unterschiede bei Sicherheitskonfigurationen beheben zu müssen. So können Sie Ihre Märkte erweitern, Partner leichter einbeziehen und von einem robusten Ökosystem für die Analyse und den Austausch von Daten profitieren.

IN DIESEM KAPITEL

- Umfassende Datensicherheit
- Einhaltung von Datenschutzbestimmungen
- Bescheinigungen und Zertifizierungen
- Verbesserte Datenhaltung, Datenschutz und Verfügbarkeit

Kapitel 8
Sicherung Ihrer Daten

Fakten über die Sicherheit in der Cloud: In den meisten Fällen sind Ihre Daten in der Cloud sicherer als in Ihrem eigenen Rechenzentrum. Laut einer 2019 von Deloitte unter IT-Führungskräften durchgeführten und von einem Team verfassten Studie, dem u. a. Tom Davenport, Ashish Verma und David Linthicum angehörten, bewahren über 90 Prozent aller Unternehmen ihre Daten hauptsächlich auf Cloud-Plattformen auf. Datensicherheit und Governance seien die wichtigsten Faktoren, die Unternehmen dazu veranlassen, ihre Daten in die Cloud zu migrieren, so das Ergebnis der Umfrage.

SaaS-Cloud-Anbieter bedienen Tausende oder sogar Millionen von Kunden. Sie können sich die Ressourcen leisten, die für eine branchentaugliche End-to-End-Datensicherheit erforderlich sind. Allerdings machen sich nicht alle Cloud-Anbieter die Mühe, Ihre Daten ausreichend zu sichern. Wenn Sie genau hinschauen, werden Sie feststellen, dass sich die Sicherheitsfunktionen oft sehr stark voneinander unterscheiden.

Die Grundvoraussetzungen

Der Schutz Ihrer Daten und die Einhaltung einschlägiger Vorschriften muss für die Architektur, bei der Implementierung und beim Betrieb eines Cloud-Data-Warehouse-Services von grundlegender Bedeutung sein. Alle Aspekte des Service müssen auf den Schutz Ihrer Daten im Rahmen einer mehrschichtigen Sicherheitsstrategie ausgerichtet sein, die sowohl aktuelle als auch sich entwickelnde Sicherheitsbedrohungen

berücksichtigt. Diese Strategie muss externe Schnittstellen, Zugriffskontrolle, Datenspeicherung und physische Infrastruktur in Verbindung mit umfassender Überwachung, Warnmeldungen und überprüfbaren Cybersicherheitsverfahren berücksichtigen.

Standardmäßige Datenverschlüsselung

Zur Verschlüsselung von Daten wird ein Verschlüsselungsalgorithmus verwendet, der Klartext in Chiffretext übersetzt. Dies ist eine grundlegende Voraussetzung für die Sicherheit Ihrer Daten. Verschlüsseln Sie Daten ab dem Zeitpunkt, an dem sie Ihre Räumlichkeiten verlassen, durch das Internet und in das Warehouse: wenn sie auf der Festplatte gespeichert werden, wenn sie in einen Staging-Bereich gebracht werden, wenn sie in ein Datenbankobjekt eingefügt werden und wenn sie in einem virtuellen Data Warehouse zwischengespeichert werden. Abfrageergebnisse sollten ebenfalls verschlüsselt werden. All dies sollte in die Lösung integriert sein und nicht nur als Option angeboten werden.

Der Anbieter sollte auch die Entschlüsselungscodes schützen, mit denen Ihre Daten entschlüsselt werden. Die besten Service Provider verwenden AES-256-Bit-Verschlüsselung mit einem hierarchischen Schlüsselmodell. Diese Methode verschlüsselt die Verschlüsselungscodes und veranlasst eine Schlüsselrotation. Dadurch wird die Zeit begrenzt, in der ein einzelner Schlüssel verwendet werden kann.

ERINNERUNG

Ihre Daten befinden sich wahrscheinlich an vielen verschiedenen Orten. Sie müssen den Datenfluss an jedem Punkt schützen und kontrollieren. Alle Daten müssen durchgängig und automatisch verschlüsselt werden, sowohl während der Übertragung als auch im Ruhezustand.

Zugriffskontrolle

Die Sicherung von Daten ist nur ein Aspekt umfassender Sicherheit. Verletzungen der Datensicherheit sind oft auf die Auswahl schwacher Passwörter durch Benutzer in Verbindung mit rudimentären Authentifizierungsverfahren zurückzuführen. Ein Cloud Data Warehouse Service sollte immer den Zugriff von Benutzern autorisieren, ihre Anmeldeinformationen durch Benutzerauthentifizierung bestätigen und ihnen nur den Zugriff auf jene Daten gewähren, für die sie eine Berechtigung haben.

Dies beginnt mit *rollenbasierter Zugriffskontrolle*, durch die sichergestellt wird, dass Benutzer nur auf die Daten zugreifen können, die sie sehen dürfen. Die Zugriffskontrolle sollte auf alle Datenbankobjekte einschließlich Tabellen, Schemata und alle virtuellen Erweiterungen des Data Warehouse angewendet werden. Für maximale Sicherheit sollte Ihr Cloud Data Warehouse auch *Multi-Faktor-Authentifizierung* bieten, die eine zweite Verifizierung erfordert, z. B. einen einmaligen Sicherheitscode, der an das Mobiltelefon des Benutzers gesendet wird.

Single-Sign-On-Verfahren und föderierte Authentifizierung erleichtern es Benutzern, sich direkt von anderen sanktionierten Anwendungen aus beim Data Warehouse Service anzumelden. Bei der *föderierten Authentifizierung* werden Identitätsmanagement- und die Zugriffskontrollverfahren zentralisiert, was Ihrem Team die Verwaltung von Benutzerzugriffsrechten erleichtert.

TIPP

Ihr Cloud-Data-Warehouse-Anbieter sollte nicht auf unverschlüsselte Kundendaten zugreifen können, es sei denn, Sie geben Ihm die ausdrückliche Erlaubnis dafür.

Patching, Updates und Netzwerküberwachung

Software-Patches und Sicherheitsupdates müssen auf allen relevanten Softwarekomponenten installiert werden, sobald sie verfügbar sind. Der Anbieter sollte auch regelmäßige Sicherheitstests (so genannte Penetrationstests) durch eine unabhängige Sicherheitsfirma durchführen lassen, um proaktiv nach Schwachstellen zu suchen.

Zu den physischen Sicherheitsmaßnahmen im Rechenzentrum sollten biometrische Zugangskontrollen, bewaffnetes Sicherheitspersonal und Videoüberwachung gehören, damit sich niemand unbefugt Zugang verschaffen kann. Alle physischen und virtuellen Maschinen müssen außerdem durch strenge Software-Verfahren zur Prüfung, Überwachung und Alarmierung kontrolliert werden. Zusätzlichen Schutz bieten Tools zur Datei-Integritätsüberwachung (FIM-Tools), die dafür sorgen, dass kritische Systemdateien nicht manipuliert werden können. Mit Whitelists für IP-Adressen können Sie den Zugriff auf das Data Warehouse auf vertrauenswürdige Netzwerke beschränken. (Eine Whitelist ist eine Liste von E-Mail-Adressen oder Domainnamen, deren Nachrichten von einem E-Mail-Blockierungsprogramm zugelassen werden).

Sicherheits-„Events“, die von Cybersicherheits-Überwachungssystemen generiert werden, welche das Netzwerk überwachen, sollten automatisch in einem manipulationssicheren SIEM-System (Security Information and Event Management) protokolliert werden. Das Sicherheitspersonal sollte automatische Warnmeldungen erhalten, wenn verdächtige Aktivitäten festgestellt werden.

Datenschutz, Datenhaltung und Redundanz

Wenn etwas Unvorhergesehenes passieren sollte, müssen Sie in der Lage sein, frühere Versionen Ihrer Daten in einer Tabelle oder Datenbank innerhalb einer bestimmten Aufbewahrungsfrist, die in Ihrem Service-Level-Agreement (SLA) mit dem Cloud-Data-Warehouse-Anbieter festgelegt ist, sofort wiederherzustellen oder abzufragen. Eine umfassende Datenhaltungsstrategie sollte über das Duplizieren von Daten in derselben Cloud-Region oder -Zone hinausgehen: Sie sollte diese Daten zwischen mehreren Verfügbarkeitszonen replizieren, um eine

geografische Redundanz zu erreichen. Automatisches Failover auf diese anderen Zonen kann den kontinuierlichen Geschäftsbetrieb sicherstellen.

Mandantenisolierung

Wenn Ihr Data-Warehouse-Anbieter eine mandantenfähige Cloud-Umgebung verwendet, in der viele Kunden dieselbe physische Infrastruktur nutzen, müssen Sie sicherstellen, dass jeder Kunde über ein virtuelles Data Warehouse verfügt, das von allen anderen Data Warehouses getrennt ist. Für die Speicherung sollte sich diese Isolierung bis auf VM-Ebene (virtuelle Maschine) erstrecken: Die Datenspeicherumgebung jedes Kunden sollte von der Umgebung jedes anderen Kunden getrennt und durch unabhängige Verzeichnisse und einzigartige Verschlüsselungscodes verwaltet werden. Einige Anbieter bieten auch dedizierte Virtual Private Networks (VPNs) und Brücken von den Systemen eines Kunden in das Cloud Data Warehouse an. Diese dedizierten Dienste stellen sicher, dass die sensibelsten Komponenten Ihres Data Warehouse vollständig von denen anderer Kunden getrennt sind.

Aufrechterhaltung von Governance und Compliance

Data Governance stellt sicher, dass der Zugriff auf und die Nutzung von Unternehmensdaten ordnungsgemäß erfolgen und dass beim täglichen Datenmanagement alle einschlägigen gesetzlichen Bestimmungen befolgt werden. Governance-Richtlinien legen Regeln und Verfahren zur Kontrolle der Eigentumsrechte und der Zugänglichkeit Ihrer Daten fest. Zu den Informationen, die üblicherweise unter diese Richtlinien fallen, gehören Kreditkarteninformationen, Sozialversicherungsnummern, Geburtsdaten, IP-Netzwerkinformationen und Geolokalisierungskoordinaten.

Bescheinigungen und Zertifizierungen

Bei der Einhaltung von Vorschriften geht es nicht nur um robuste Cybersicherheitsverfahren. Ihr Data-Warehouse-Anbieter muss auch nachweisen können, dass er die erforderlichen Sicherheitsverfahren anwendet. Datenverstöße und die zugehörigen Abhilfemaßnahmen können Kosten in Millionenhöhe nach sich ziehen und die Beziehungen zu Ihren Kunden dauerhaft schädigen.

Branchenübliche Bescheinigungsberichte belegen, dass Cloud-Anbieter geeignete Sicherheitskontrollen anwenden. So muss ein Cloud-Data-Warehouse-Anbieter nachweisen, dass er Bedrohungen und Sicherheitsvorfälle angemessen überwacht und darauf reagiert und dass er angemessene Verfahren implementiert hat, um mit diesen Vorfällen umzugehen (siehe Abbildung 8-1).

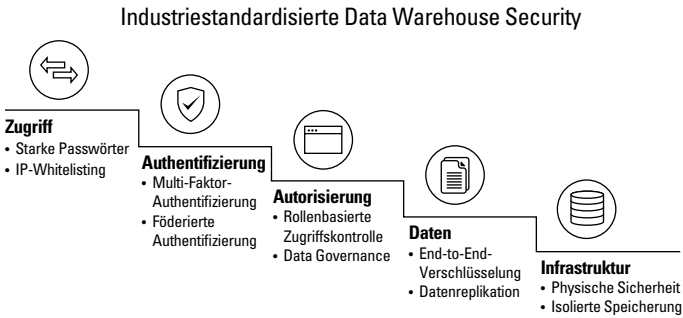

ABBILDUNG 8-1: Vergewissern Sie sich, dass der gesamte Datenverkehr verschlüsselt und sicher ist und dass Ihre Cloud-Anbieter über alle relevanten Zertifizierungen verfügen.

Stellen Sie auch sicher, dass Ihr Cloud-Anbieter neben den branchenüblichen Technologie-Zertifizierungen wie ISO/IEC 27001 und SOC 1/SOC 2 Typ II auch alle geltenden staatlichen und branchenspezifischen Vorschriften erfüllt. Dies kann PCI-, HIPAA/Health Information Trust Alliance (HITRUST) und FedRAMP-Zertifizierungen umfassen, je nachdem, was für Ihr Unternehmen relevant ist.

Verlangen Sie Nachweise und stellen Sie sicher, dass Ihre Anbieter eine Kopie des gesamten Berichts für jeden relevanten Standard und nicht nur Begleitschreiben vorlegen. Der SOC-2-Typ-II-Bericht bestätigt zum Beispiel, dass in den letzten 12 Monaten angemessene technische und administrative Kontrollen durchgeführt wurden. Die Bescheinigung über die PCI-DSS-Konformität gibt Aufschluss darüber, ob Ihr Anbieter Kreditkartendaten ordnungsgemäß speichert und verarbeitet. Wenn Sie geschützte Gesundheitsinformationen verarbeiten, muss Ihr Anbieter HIPAA-Richtlinien einhalten.

TIPP

Durch Konformität und Bescheinigungen kann Ihr Data-Warehouse-Anbieter nachweisen, dass er die Sicherheit ernst nimmt und transparent ist.

Cloud-Anbieter sollten auch Nachweise darüber erbringen, dass die Softwareanbieter, mit denen sie zusammenarbeiten, alle Vorschriften einhalten und regelmäßige Sicherheitsaudits durchführen. Ihre Daten sind nur so sicher wie das schwächste Glied in der Technologiekette. Stellen Sie daher sicher, dass alle Beteiligten über robuste Sicherheitskontrollen verfügen und die branchenüblichen Sicherheitsverfahren einhalten. Wenn ein Konformitätsnachweis fehlen sollte, beschaffen Sie sich die entsprechende Dokumentation.

ERINNERUNG

Arbeiten Sie nur mit Cloud-Anbietern zusammen, die nachweisen können, dass sie die in der Branche geltenden Sicherheitsverfahren einhalten und dass dies von unabhängigen Prüfern bestätigt wurde. Ihre Mindestanforderungen für dieses wichtige Daten-Repository sollten in diese Compliance-Erwägungen einbezogen werden.

Eine umfassende Sicherheitspraxis

Gute Sicherheit ist teuer und erfordert Spezialwissen. Geräteausfälle, Netzwerkeinbrüche und Wartungsfehler können zu Datenverlusten und zur Inkonsistenz Ihrer Daten führen. Eine umfassende Sicherheitspraxis umfasst zahlreiche Aspekte. Ihr Cloud-Data-Warehouse-Anbieter sollte über Verfahren zum Schutz vor versehentlicher oder absichtlicher Zerstörung verfügen. Einige Anbieter bieten rudimentäre Sicherheitsfunktionen an und überlassen die Verschlüsselung, Zugriffskontrolle und Sicherheitsüberwachung Ihnen, dem Kunden. Die Sicherheit sollte eine Grundlage des Data-Warehouse-Service sein, damit Sie keine zusätzlichen Maßnahmen zur Sicherung Ihrer Daten ergreifen müssen.

TIPP

Ein Anbieter, der für Transparenz in Bezug auf seine Sicherheitszertifizierungen sorgt, hat mit großer Wahrscheinlichkeit ein solides Sicherheitsprogramm.

IN DIESEM KAPITEL

- Schaffung einer kosteneffektiven Speicherumgebung
- Mehrwert und beste Performance durch Architektur und Preisgestaltung

Kapitel 9
Minimierung der Data-Warehouse-Kosten

In diesem Kapitel sehen wir uns an, wie Sie ein cloudbasiertes Data Warehouse betreiben können und wie Ihnen Ihr Data-Warehouse-Anbieter bei der langfristigen Minimierung Ihrer Kosten helfen kann.

Minimierung der Speicherkosten

Je mehr Daten Sie speichern können, desto tiefere Einblicke können Sie gewinnen. Glücklicherweise ist Cloud-Storage von Amazon, Microsoft und Google mittlerweile relativ kostengünstig, sodass Sie wahrscheinlich keinen Einschränkungen in Bezug auf die Menge und Art der von Ihnen gespeicherten Daten unterliegen werden. Prüfen Sie die Vertragsbedingungen, um sicherzustellen, dass Ihr Cloud-Data-Warehouse-Anbieter diese Rohdatenspeicherkosten nicht in die Höhe treibt. Der Anbieter sollte die Listenpreise direkt an Sie weitergeben. Ihr Data-Warehouse-Anbieter kann Ihnen einen Mehrwert bieten, indem er Ihre Daten drei- bis *fünffach* komprimiert. Mit dreifacher Komprimierung müssen Sie nur noch ein Drittel der Datenmenge speichern – und das zu einem Drittel der Kosten.

Prüfen Sie die Bedingungen der Nutzungsvereinbarung: Sie sollten nur für den von Ihnen genutzten Speicherplatz zahlen, nicht für überschüssige oder „reservierte" Speicherkapazität. Ebenso wenig sollten Sie für das Klonen von Datenbanken in Ihrem Data Warehouse zu Entwicklungs- und Testzwecken zahlen. Sie sollten in der Lage sein, Ihre Daten mehrfach zu referenzieren – nicht zu kopieren – sodass keine zusätzlichen Kosten für die Speicherung anfallen.

Ihr Cloud-Data-Warehouse sollte es Ihnen auch ermöglichen, strukturierte und semistrukturierte Daten wie JSON zu speichern und abzufragen. Halten Sie nach einem Anbieter Ausschau, der *Multi-Cloud-Fähigkeiten* bietet, da Sie dadurch Kosten sparen können, wenn Sie Ihr Data Warehouse in der Zukunft in eine andere Cloud-Storage-Umgebung migrieren.

Maximierung der Recheneffizienz

Rechenressourcen sind teurer als Speicherressourcen. Daher sollte Ihnen Ihr Data-Warehouse-Service die Möglichkeit bieten, jede Ressource unabhängig zu skalieren. Es sollte einfach sein, im Rahmen eines nutzungsabhängigen Preismodells nur die Rechenressourcen im benötigten Ausmaß aufzustocken. Der Anbieter sollte Ihnen nur die von Ihnen genutzten Ressourcen in Rechnung stellen – auf die Sekunde genau. Ungenutzte Rechenressourcen sollten automatisch terminiert werden, um unkalkulierbare Kosten zu vermeiden. Mit der nutzungsbasierten Preisgestaltung können Sie selbst entscheiden, wie Sie Ressourcen verbrauchen.

Flexible Bedingungen sorgen dafür, dass Sie Ihre Rechencluster für jeden Workload richtig dimensionieren können. Wenn Sie einen ETL-Auftrag (Extract, Transfer, Load) mit geringen Rechenanforderungen ausführen, können Sie diesem Workload einen kleinen Cluster zuordnen, anstatt die Kosten für einen überdimensionierten Cluster zu tragen. Zum Testen neuer Machine-Learning-Module können Sie einen großen Cluster verwenden. Dadurch erhalten Sie eine feingranulare Skalierbarkeit für jeden Workload und minimieren gleichzeitig Ihre Nutzungskosten. Der Betrieb Ihres Warehouses wird kostengünstiger sein als der Betrieb von On-Premise-Warehouses und deren Cloud-Versionen, die langsam sind, enorme Ressourcen verbrauchen und nur begrenzte Ergebnisse liefern. Da jeder Workload seinen dedizierten Rechencluster hat, werden die Workloads nicht langsamer bzw. kommen nicht völlig zum Stillstand.

IN DIESEM KAPITEL

- **Erfolgskriterien und Anforderungen, die Ihr Data Warehouse erfüllen muss**
- **Berücksichtigung aller Faktoren der Total Cost of Ownership**
- **Testen des Data Warehouse vor dem Kauf**

Kapitel 10
Sechs Schritte zum Einstieg in das Cloud Data Warehousing

In diesem Kapitel beschreiben wir sechs wichtige Schritte zur Auswahl eines Cloud Data Warehouse für Ihr Unternehmen. Der Prozess beginnt mit der Ermittlung Ihrer Anforderungen an ein Data Warehouse und endet mit dem Testen Ihrer bevorzugten Wahl. So erhalten Sie einen Plan, der Ihnen bei der Auswahl Ihrer Lösung hilft.

Schritt 1: Evaluierung Ihrer Anforderungen

Das für Sie geeignete Data Warehouse sollte Ihren aktuellen Bedürfnissen entsprechen und gleichzeitig in der Lage sein, Ihre zukünftigen Anforderungen zu erfüllen. Berücksichtigen Sie daher die Art Ihrer Daten, die bereits vorhandenen Fähigkeiten und Tools, Ihre Nutzungsanforderungen und die Zukunftspläne Ihres Unternehmens. Stellen Sie sich vor, wie ein Data Warehouse Ihr Unternehmen voranbringen kann – in einer Art, die Sie sich bisher nicht vorgestellt haben:

- **Daten:** Welche Arten von Daten muss das Data Warehouse enthalten? In welchem Tempo werden neue Daten erstellt? Wie oft werden Daten in das Warehouse geladen? Auf welche wichtigen Daten können Sie derzeit nicht zugreifen?
- **Anpassung an vorhandene Fähigkeiten, Tools und Prozesse:** Welche Tools und Fähigkeiten Ihres Teams werden bei den verschiedenen Cloud-Data-Warehouse-Optionen angewendet? Auf welche Prozesse wird sich ein Cloud Data Warehouse auswirken?

- **Nutzung:** Welche Benutzer und Anwendungen werden Zugriff auf das Data Warehouse haben? Welche Arten von Abfragen werden Sie ausführen? Auf wie viele Daten werden die Benutzer zugreifen müssen und wie schnell? Wie werden sich Workloads im Laufe der Zeit ändern? Welche Performance benötigen Ihre Benutzer und Anwendungen? Wie viele Benutzer sollten derzeit auf das Data Warehouse zugreifen, tun dies aber aufgrund von Ressourceneinschränkungen nicht?
- **Data Sharing:** Beabsichtigen Sie, Daten innerhalb Ihres Unternehmens und mit Kunden und/oder Partnern sicher auszutauschen? Wenn ja, welche Arten von Daten werden Sie austauschen und werden Sie einen Datenmarktplatz oder eine -börse einrichten, um Daten auch zu monetarisieren? Werden Sie diesen Datenkonsumenten den Zugriff auf Rohdaten gestatten oder werden Sie diese Daten anreichern, indem Sie auch Datenservices wie Analysen anbieten?
- **Globaler Zugang:** Planen Sie die Speicherung von Daten in einem öffentlichen Objektspeicher wie Amazon S3, Microsoft Azure oder Google Cloud Platform? Haben Sie spezifische funktionale, regionale oder auf die Datenhoheit bezogene Anforderungen, die die Aufrechterhaltung dieser Beziehungen erforderlich machen? Benötigen Sie eine cloudübergreifende Architektur, um regionale Bereitstellungsoptionen zu maximieren, Disaster Recovery zu unterstützen oder die globale Geschäftskontinuität zu gewährleisten?
- **Ressourcen:** Welche personellen Ressourcen stehen zur Verwaltung des Data Warehouse zur Verfügung? Wie viel möchten Sie investieren, um die Verfügbarkeit, Performance und Sicherheit zu überwachen und zu verwalten? Verfügen Sie über spezielle Fachkenntnisse im Bereich der Data-Warehouse-Entwicklung und -Prüfung oder über ein DevOps-Team zu dessen Optimierung?

Schritt 2: Migrieren oder neu beginnen

Am Anfang jedes Cloud-Data-Warehouse-Projekts sollte bestimmt werden, welcher Anteil Ihrer bestehenden Umgebung auf das neue System migriert werden kann und was für ein Cloud Data Warehouse neu aufgebaut werden sollte. Diese Entscheidungen können sich auf alle Aspekte beziehen – vom Design der ETL-Prozesse (Extract, Transform, Load) bis hin zu Datenmodellen und Methoden des Softwareentwicklungs-Lebenszyklus. Stellen Sie sich die folgenden Fragen:

- **Ist dies ein brandneues Projekt?** Wenn ja, ist es oft sinnvoll, das Projekt so zu gestalten, dass alle Möglichkeiten eines Cloud Data Warehouse voll ausgeschöpft werden, anstatt eine bestehende Implementierung mit Einschränkungen fortzusetzen.

- **Welcher Teil Ihres gegenwärtigen System bereitet Ihnen das meiste Kopfzerbrechen?** Bei einer gut geplanten Migration können Sie sich beispielsweise darauf konzentrieren, die problematischsten Workloads zuerst in das Cloud Data Warehouse zu verlagern. Vielleicht möchten Sie auch einfache Workloads migrieren, um schnelle Erfolge zu erzielen.
- **Welche Aspekte Ihres gegenwärtigen Systems führen zu bestimmten Einschränkungen, die bei einem Cloud Data Warehouse nicht mehr vorhanden sind?** Tools und Prozesse, die darauf ausgelegt sind, Ressourcenbeschränkungen zu umgehen, den Aufwand für die Kapazitätserweiterung zu vermeiden oder die Kosten zu optimieren, können mit der richtigen Cloud-Lösung unnötig sein.
- **Wie greifen aktuelle Benutzer und Anwendungen auf das Data Warehouse zu?** Benutzer und Anwendungen, die sich auf Industriestandard-Schnittstellen wie SQL verlassen und Standard-ETL- und Business-Intelligence-Tools verwenden, werden bei der Anpassung an einen neuen Ansatz weniger Veränderungen erleben.
- **Wie werden sich Ihre Daten- und Analyseanforderungen in Zukunft ändern?** Eine Lösung, die für die kontinuierliche Weiterentwicklung konzipiert wurde, wird voraussichtlich länger als andere Lösungen bestehen und neue Möglichkeiten bieten, die erweiterte Funktionen wie sicheres Data Sharing und globalen Datenzugriff sich zunutze machen.

ERINNERUNG

Wenn Sie ein großes und komplexes herkömmliches Data Warehouse haben, können Sie zunächst einen kleinen Teil des Systems migrieren, um sich mit der Nutzung eines Cloud Data Warehouse vertraut zu machen. Dann können Sie Ihren Cloud-Footprint interaktiv erweitern.

Schritt 3: Festlegen von Erfolgskriterien

Wie messen Sie den Erfolg der Umstellung auf ein neues Cloud Data Warehouse? Wählen Sie wichtige geschäftliche und technische Anforderungen aus. Die Kriterien sollten sich auf Performance, Concurrency, Einfachheit und Total Cost of Ownership (TCO) konzentrieren.

ERINNERUNG

Wenn Ihr neues Cloud Data Warehouse über Funktionen verfügt, die bei Ihrem vorherigen System nicht vorhanden waren, und wenn diese Funktionen für die Bewertung des geschäftlichen und technischen Erfolgs Ihrer neuen Lösung relevant sind, sollten Sie diese unbedingt mit einbeziehen.

Legen Sie bei der Festlegung der Erfolgskriterien Ihrer neuen Lösung fest, wie Sie diesen Erfolg messen wollen. Bestimmen Sie, welche Kriterien quantifizierbar und welche qualitativ sind, wie Sie die quantifizierbaren Kriterien messen und wie Sie die qualitativen Kriterien bewerten wollen.

FALLSTUDIEN

LÖSUNG VON LATENZPROBLEMEN

White Ops ist ein führender Anbieter von Cybersicherheitsservices. Im Gegensatz zu herkömmlichen Ansätzen, bei denen statistische Analysen verwendet werden, bekämpft White Ops kriminelle Aktivitäten, indem es überprüft, ob diese von Bots oder von Menschen ausgehen. Das Unternehmen arbeitet daran, neue Betrugsmuster aufzudecken und zu charakterisieren. Dieser fortlaufende Prozess erfordert die Speicherung und Verarbeitung großer Datenmengen.

White Ops hatte sich zuvor auf NoSQL-Systeme verlassen, um diese Daten zu speichern und zu verarbeiten. Je nach Workload betrug die Latenzzeit für die Ergebnisse jedoch mindestens 24 Stunden. Je mehr Abfragen, desto länger die Verzögerungen.

Um die Produktivität und Performance zu steigern, implementierte White Ops ein Cloud Data Warehouse mit SQL als Kernsprache, welches als Service bereitgestellt wurde. Mit diesem Data Warehouse kann White Ops alle Daten an einem Ort speichern, elastisch skalieren, unterschiedliche Daten mit Standard-SQL abfragen und schneller neue Lösungen zur Betrugsprävention entwickeln.

White Ops kann nun riesige Datenmengen konsolidieren und skalieren, Datenzugriff ermöglichen, ohne sich auf Spezialisten mit fundierten Programmierkenntnissen verlassen zu müssen, und seinen Kunden dabei helfen, die verheerenden Auswirkungen von Online-Betrug zu vermeiden.

Schritt 4: Evaluierung von Lösungen

Sobald Sie Ihre Erfolgskriterien und die Anforderungen bestimmt haben, die Ihr Data Warehouse erfüllen muss, können Sie mit der Evaluierung von Lösungen beginnen. In diesem Buch werden die Unterschiede zwischen den verfügbaren Optionen ausführlich beschrieben (siehe Kapitel 3, 4 und 5). Achten Sie beim Vergleich darauf, dass die jeweilige Lösung die folgenden Kriterien erfüllt:

- berücksichtigt aktuelle und zukünftige Bedürfnisse
- integriert strukturierte und semistrukturierte Daten, speichert sie alle an einem Ort und vermeidet die Erstellung von Datensilos
- unterstützt vorhandene Fähigkeiten, Tools und Fachwissen
- schützt vor Datenverlust und ermöglicht eine einfache Datenwiederherstellung
- schützt Ihre Daten mit branchenüblichem Passwortschutz und Verschlüsselung

- stellt sicher, dass Daten und Analysen jederzeit verfügbar sind
- optimiert die Daten-Pipeline, sodass neue Daten in kürzester Zeit zur Analyse zur Verfügung stehen
- optimiert die Time-to-Value, damit Sie die Vorteile Ihres neuen Data Warehouse so schnell wie möglich nutzen können.
- stellt Ressourcen für isolierte Workloads zur Verfügung
- ermöglicht den Datenaustausch, ohne dass Live-Daten kopiert oder verschoben werden müssen, und verbindet Datenlieferanten und Datenkonsumenten auf einfache Weise
- repliziert Datenbanken und hält sie über Konten, Cloud-Plattformen und Regionen hinweg synchron, um die Geschäftskontinuität zu verbessern und die Expansion zu rationalisieren
- bietet Zero-Copy-Clone-Funktionen für Datenbanken für die Entwicklungs- und Testzwecke und zur Unterstützung von Anwendungsfällen mit Mehrfachnutzung, z. B. Berichterstattung, Datenexploration und Predictive Analytics
- sorgt für eine einfache Wiederherstellung von Daten, die aufgrund von Fehlern oder Angriffen verloren gegangen sind, da auf frühere Datenversionen zurückgegriffen werden kann
- ermöglicht die unabhängige und automatische Skalierung von Rechen- und Speicherressourcen und somit die Parallelität, ohne die Leistung zu verlangsamen.

Schritt 5: Berechnung der TCO

Wenn der Preis für Sie das entscheidende Kriterium bei der Auswahl eines Cloud Data Warehouse ist, sollten Sie die TCO für ein herkömmliches Data Warehouse in Betracht ziehen. Dazu gehören die Kosten für die Lizenzierung, die in der Regel auf der Anzahl der Benutzer basieren, die Kosten für Hardware (Server, Speichergeräte, Netzwerk), das Rechenzentrum (Büroräume, Strom, Verwaltung, Wartung und laufendes Management), Datensicherheit (Passwortschutz und Verschlüsselung), Lösungen zur Gewährleistung von Verfügbarkeit und Ausfallsicherheit, die Unterstützung für Skalierung und Parallelität sowie die Schaffung von Entwicklungs- und Staging-Umgebungen.

Bei einigen Lösungen müssen Sie zusätzliche Kosten berücksichtigen, z. B. für den Aufbau und die Verwaltung mehrerer Data Marts, die Verwendung mehrerer Datenkopien in verschiedenen Data Marts, die Schulung von Mitarbeitern, die Verwendung mehrerer Systeme (z. B. SQL und NoSQL) für unterschiedliche Daten usw.

Bei Cloud-Data-Warehouse-Lösungen ist die Berechnung der Kosten in der Regel einfacher. Sie unterscheiden sich jedoch je nach den Dienstleistungen des jeweiligen Anbieters. Wenn Sie alles an den Anbieter auslagern wollen und sich deshalb für ein Data-Warehouse-as-a-Service (DWaaS) entscheiden, können Sie die TCO auf der Grundlage der monatlichen Abonnementgebühr berechnen. Entscheiden Sie sich für eine Infrastructure-as-a-Service (IaaS)- oder Platform-as-a-Service (PaaS)-Lösung (siehe Kapitel 5), müssen Sie die Kosten für die Software, die Verwaltung und die Dienstleistungen hinzurechnen, die nicht in der Lösung enthalten sind.

TIPP

Unternehmen berechnen die TCO gewöhnlich für die erwartete Lebensdauer des Data Warehouse, die in der Regel ein bis drei Jahre beträgt. Ein wichtiger Vorbehalt: Oft wird angenommen, dass ein Cloud-System rund um die Uhr bei hoher Kapazität läuft. Dabei wird übersehen, welche Einsparungen erzielt werden können, weil die Cloud-Lösung je nach Bedarf dynamisch nach oben oder unten skaliert wird und die Abrechnung nach Sekunden erfolgt.

Schritt 6: Durchführung eines Proof of Concept

Nachdem Sie verschiedene Cloud-Data-Warehouse-Optionen untersucht, sich Demos angesehen, Fragen gestellt und sich mit dem Team jedes Anbieters getroffen haben, sollten Sie einen Proof of Concept (PoC) durchführen, bevor Sie sich schließlich zum Kauf entscheiden. Mithilfe eines PoC wird eine Lösung getestet, um zu bestimmen, wie gut sie Ihren Anforderungen entspricht und Ihre Erfolgskriterien erfüllt. Betrachten Sie es als eine Testfahrt. Der Test dauert in der Regel ein oder zwei Tage, kann aber auch über mehrere Wochen hinweg durchgeführt werden. Sie beantragen einen PoC bei einem potenziellen Anbieter unter der Annahme, dass Sie das Produkt kaufen werden, wenn die Lösung zufriedenstellend funktioniert. Im Falle von Cloud Data Warehousing abonnieren Sie den Service.

TIPP

Führen Sie bei der Durchführung Ihres PoCs alle Anforderungen und Erfolgskriterien auf – nicht nur die Probleme, die Sie lösen wollen, sondern alles, was mit einer Cloud-Lösung möglich ist.

Erstellen Sie eine umfassende Checkliste mit Erfolgskriterien und Anforderungen, die Ihr Data Warehouse erfüllen soll. Stellen Sie sicher, dass das neue Data Warehouse alles tut, was Ihr aktuelles Data Warehouse leistet – nur besser – und es die Nachteile des aktuellen Systems beseitigt. Wenn Sie ein PoC mit mehreren Anbietern durchführen, sollten Sie für jeden Anbieter die gleiche Checkliste verwenden.

Mit Cloud Data Warehousing einen Wettbewerbsvorteil erzielen

Unternehmen haben heute Zugang zu exponentiell größeren Datenmengen, die sie analysieren können, um möglichst tiefe Einblicke zu gewinnen. Darüber hinaus möchten Unternehmen auf sichere Weise Daten austauschen und auf gemeinschaftlich genutzte Daten zugreifen – über Geschäftsbereiche hinweg, in ihren Business-Ökosystemen und darüber hinaus – und Datenbörsen zur Monetarisierung ihrer Daten verwenden. Der Zugriff auf diese Daten ist jedoch mit noch größeren Herausforderungen verbunden, mit denen herkömmliche Data-Analytics-Plattformen nach wie vor zu kämpfen haben. Viele moderne Unternehmen haben erkannt, dass Cloud Data Warehousing die effektivste und kostengünstigste Art ist, um alle ihre Daten für alle ihre Geschäftsanwender zu speichern und zu analysieren. Dieses Buch zeigt, welche Lösungen Ihnen zur Verfügung stehen und wie Ihr Unternehmen aus dieser neuen und spannenden Technologie Nutzen ziehen kann.

Im Buch...

- Warum das Cloud Data Warehouse entstanden ist
- Wie das Cloud Data Warehouse im Vergleich zu anderen Lösungen abschneidet
- Wie unterschiedliche Warehouses bewertet werden sollten
- Warum Sicherheit und Governance wichtig sind
- Die Vorteile einer Cross-Cloud-Lösung
- Wie der moderne Datenaustausch noch tiefere Einblicke ermöglicht
- Fallstudien aus der Praxis

snowflake®

Joe Kraynak ist ein erfahrener Dummies-Autor, der Dutzende von Büchern zu unterschiedlichen Themen verfasst und mitverfasst hat. **David Baum** ist ein freischaffender Business-Autor, der sich auf Wissenschaft und Technologie spezialisiert hat.

Besuchen Sie Dummies.com®
für Videos, step-by-step Beispiele, Anleitungen oder die zum Einkaufen!

ISBN: 978-1-119-71403-3

für dummies®

Auch als E-Book erhältlich

9 781119 714033

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.