Data Warehousing auf AWS

März 2016

Hinweise

Dieses Dokument wird nur zu Informationszwecken zur Verfügung gestellt. Es stellt das aktuelle Produktangebot und die Praktiken von AWS zum Erstellungsdatum dieses Dokuments dar. Änderungen vorbehalten. Kunden sind für ihre eigene unabhängige Einschätzung der Informationen in diesem Dokument und jedweder Nutzung der AWS-Services verantwortlich. Jeder Service wird ohne Gewähr und ohne Garantie jeglicher Art, weder ausdrücklich noch impliziert, bereitgestellt. Dieses Dokument gibt keine Garantien, Gewährleistungen, vertragliche Verpflichtungen, Bedingungen oder Zusicherungen von AWS, seinen Partnern, Zulieferern oder Lizenzgebern. Die Verantwortung und Haftung von AWS gegenüber seinen Kunden werden durch AWS-Vereinbarungen geregelt. Dieses Dokument gehört, weder ganz noch teilweise, nicht zu den Vereinbarungen von AWS mit seinen Kunden und ändert diese Vereinbarungen auch nicht.

Inhalt

Übersicht

Weltweit sind Datenexperten, Datenanalysten und Entwickler in Unternehmen dabei, Data Warehousing in die Cloud zu migrieren, um die Leistung zu steigern und die Kosten zu senken. In diesem Whitepaper wird ein moderner Ansatz für Analysen und Data-Warehousing-Architektur erläutert. Zudem werden Services vorgestellt, die auf Amazon Web Services (AWS) verfügbar sind, um diese Architektur zu implementieren, und gängige Entwurfsmuster für Data-Warehousing-Lösungen, die diese Services verwenden.

Einführung

Heutzutage sind Daten und Datenanalysen für Unternehmen unverzichtbar. Fast alle großen Unternehmen haben für Berichts- und Analysezwecke ein Data Warehouse eingerichtet, in das Daten aus einer Vielzahl von Quellen einfließen, zu denen auch die eigenen Transaktionsverarbeitungssysteme und andere Datenbanken gehören.

Aber es war bisher immer kompliziert und teuer, ein Data Warehouse – ein zentrales Repository für Informationen aus einer oder mehreren Datenquellen – zu erstellen und zu betreiben. Die meisten Data Warehousing-Systeme sind kompliziert einzurichten und erfordern vorab Millionenbeträge für Software und Hardware. Zudem dauern Planung, Beschaffung, Implementierung und Bereitstellung etliche Monate. Nachdem Sie die ersten Investitionen getätigt und das Data Warehouse eingerichtet haben, müssen Sie ein Team von Datenbankadministratoren einstellen, um schnelle Abfragen sicherzustellen und Datenverluste auszuschließen.

Hinzu kommt, dass es schwierig ist, ein traditionelles Data Warehouse zu skalieren. Wenn das Datenvolumen wächst oder Analysen und Berichte weiteren Benutzern zur Verfügung stehen sollen, müssen Sie entweder eine geringere Abfrageleistung in Kauf nehmen oder Zeit und Mühe in eine teure Erweiterung investieren. Es ist schon vorgekommen, dass IT-Teams Erweiterungen des Datenbestands oder zusätzliche Abfragen ablehnen mussten, um Service Level Agreements einzuhalten. Viele Unternehmen ringen um den Erhalt ihres guten Verhältnisses zu traditionellen Datenbankanbietern. Sie sind oft gezwungen, entweder ein Hardware-Upgrade für ein verwaltetes System vorzunehmen oder sich auf langwierige Verhandlungen über eine abgelaufene Lizenz einzulassen. Sobald sie die Skalierungsgrenze einer Data Warehouse-Engine erreicht haben, sind sie zur Migration auf eine andere Engine desselben Herstellers gezwungen, die eine abweichende SQL-Semantik verwendet.

Mit Amazon Redshift hat sich die Ansicht von Unternehmen über das Data Warehousing geändert, da es nun möglich ist, ohne Kompromisse in Bezug auf Funktionsumfang und Leistung die Kosten und den Aufwand zur Bereitstellung von Data Warehouse-Systemen drastisch zu reduzieren. Amazon Redshift ist eine schnelle, vollständig verwaltete und bis in den Petabyte-Bereich skalierbare Data Warehouse-Lösung, mit der Sie im Zusammenspiel mit Ihren vorhandenen Business Intelligence (BI)-Tools große Datenmengen einfach und kosteneffizient analysieren können. Mit Amazon Redshift steht Ihnen die Leistung von spaltenorientierten Data Warehousing-Engines mit massiv-paralleler Verarbeitung (MPP) zu einem Zehntel der Kosten zur Verfügung. Sie können ohne Verpflichtungen mit 0,25 USD pro Stunde klein anfangen und bis hin zu 1.000 USD pro Terabyte und Jahr skalieren.

Seit dem Start im Februar 2013 ist Amazon Redshift einer der am schnellsten wachsenden AWS-Services, mit vielen Tausenden von Kunden aus allen Branchen und Unternehmensgrößen. Unternehmen wie NTT DOCOMO, FINRA, Johnson & Johnson, Hearst, Amgen und NASDAQ sind zu Amazon Redshift migriert. Deshalb steht Amazon Redshift im *Forrester Wave: Enterprise Data Warehouse, Q4 2015*-Bericht an erster Stelle.[1]

In diesem Whitepaper vermitteln wir Ihnen die Informationen, die Sie benötigen, um von der strategischen Verschiebung des lokalen Data-Warehousing-Umfelds in die Cloud zu profitieren:

- Moderne Analysearchitektur
- Verfügbare Data Warehousing-Technologien innerhalb dieser Architektur
- Tiefgehender Einblick in Amazon Redshift und die Unterscheidungsmerkmale
- Eine Vorlage zur Erstellung eines vollständigen Data Warehousing-Systems auf AWS mit Amazon Redshift und anderen Services
- Praktische Tipps für die Migration aus anderen Data Warehousing-Lösungen und Einstieg in unser Partner-Ökosystem

Moderne Analyse- und Data Warehousing-Architektur

Ein *Data Warehouse* ist, wie bereits erwähnt, ein zentrales Repository für Informationen aus einer oder mehreren Datenquellen. Die Daten fließen in der Regel aus Transaktionssystemen oder anderen relationalen Datenbanken in das Data Warehouse. Üblicherweise handelt es sich um strukturierte, halbstrukturierte und unstrukturierte Daten. Diese Daten werden verarbeitet, transformiert und in regelmäßigen Intervallen übernommen. Benutzer wie Datenexperten, Wirtschaftsanalytiker und Entscheider greifen mithilfe von BI-Tools, SQL-Clients und Kalkulationstabellen auf die Daten zu.

Warum ein Data Warehouse erstellen – warum nicht Analyseabfragen direkt in der Datenbank für Online-Transaktionsverarbeitung (OLTP) ausführen, in der die Transaktionen aufgezeichnet werden? Zur Beantwortung dieser Frage wollen wir die Unterschiede zwischen einem Data Warehouse und einer OLTP-Datenbank betrachten. Ein Data Warehouse ist für einzelne, aufeinanderfolgende Schreiboperationen sowie für Operationen optimiert, mit denen große Datenmengen gelesen werden, während eine OLTP-Datenbank für kontinuierliche Schreibvorgänge und eine große Anzahl von Operationen ausgelegt ist, mit denen kleine Datenmengen gelesen werden. Im Allgemeinen verwendet ein Data Warehouse wegen der Anforderungen an einen hohen Datendurchsatz denormalisierte Schemata wie das Sternschema oder das Schneeflockenschema. Dagegen basieren OLTP-Datenbanken auf hochgradig normalisierten Schemata, die besser den Anforderungen an einen hohen Transaktionsdurchsatz entsprechen. Das Sternschema besteht aus einigen großen Faktentabellen, auf mehrere Dimensionstabellen referenzieren. Das Schneeflockenschema ist ein erweitertes Sternschema, in dem die Dimensionstabellen noch weiter normalisiert sind.

Um die Vorteile zu nutzen, die sich aus der Verwendung eines Data Warehouse ergeben, das als separater Datastore mit Ihrem OLTP-Quellsystem oder anderen Quellsystemen verwaltet wird, empfehlen wir den Aufbau einer effizienten Daten-Pipeline. Eine solche Pipeline extrahiert die Daten aus dem Quellsystem, wandelt sie in ein für das Data Warehousing geeignetes Schema um und lädt sie dann in das Data Warehouse. Im nächsten Abschnitt behandeln wir die Bausteine einer Analyse-Pipeline und die verschiedenen AWS-Services, die Sie zum Erstellen der Pipeline verwenden können.

Analysearchitektur

Analyse-Pipelines sind dafür ausgelegt, eine große Anzahl eingehender Datenströme aus unterschiedlichen Quellen – beispielsweise Datenbanken, Anwendungen und Geräte – zu verarbeiten.

Eine typische Analyse-Pipeline führt folgende Schritte aus:

1. Daten erfassen
2. Daten speichern
3. Daten verarbeiten
4. Daten analysieren und visualisieren

Dies verdeutlicht die folgende Abbildung 1.

Abbildung 1: Analyse-Pipeline

Datenerfassung

Beim Erfassen der Daten müssen Sie die unterschiedlichen Arten von Daten berücksichtigen, mit denen Sie arbeiten, beispielsweise Transaktionsdaten, Protokolldaten, Streaming-Daten und Internet of Things (IoT)-Daten. AWS stellt für jede dieser Datenarten entsprechende Datenspeicher bereit.

Transaktionsdaten

Transaktionsdaten wie E-Commerce-Kauftransaktionen und Finanztransaktionen werden in der Regel in relationalen Datenbankmanagementsystemen (RDBMS) oder NoSQL-Datenbanksystemen gespeichert. Die Wahl der Datenbanklösung ist vom jeweiligen Anwendungsfall und der Anwendungscharakteristik abhängig. Eine NoSQL-Datenbank ist geeignet, wenn die Daten nicht so strukturiert sind, dass sie in ein definiertes Schema passen, oder wenn sich das Schema sehr häufig ändert. Dagegen ist eine RDBMS-Lösung dann geeignet, wenn sich Transaktionen über mehrere Tabellenzeilen erstrecken und die Abfragen komplexe Joins erfordern. Amazon DynamoDB ist ein vollständig verwalteter NoSQL-Datenbank-Service, den Sie als OLTP-Speicher für Ihre Anwendungen einsetzen können. Amazon RDS ermöglicht Ihnen die Implementierung einer SQL-basierten relationalen Datenbank für Ihre Anwendung.

Protokolldaten

Die verlässliche Erfassung systemgenerierter Protokolle hilft Ihnen, anhand der gespeicherten Protokollinformationen Probleme zu beheben sowie Audits und Analysen durchführen. Amazon Simple Storage Service (Amazon S3) ist eine beliebte Speicherlösung für nicht transaktionsorientierte Daten wie zum Beispiel Protokolldaten, die für Analysen verwendet werden. Weil der Service eine Datenbeständigkeit von 99,999999999 % (11 Neunen) garantiert, ist Amazon S3 auch eine beliebte Archivierungslösung.

Streaming-Daten

Webanwendungen, mobile Geräte sowie viele Softwareanwendungen und Dienste können beachtliche Mengen von Streaming-Daten generieren (manchmal mehrere Terabyte pro Stunde), die kontinuierlich gesammelt, gespeichert und verarbeitet werden müssen.[2] Mit Amazon Kinesis erledigen Sie das auf einfache Weise mit geringen Kosten.

IoT-Daten

Weltweit senden Geräte und Sensoren kontinuierlich Nachrichten. Für Unternehmen besteht heutzutage immer häufiger die Notwendigkeit, diese Daten zu erfassen und daraus nützliche Informationen zu gewinnen. Mit AWS IoT können verbundene Geräte einfach und sicher mit der AWS Cloud interagieren. AWS IoT macht es einfach, AWS-Services wie AWS Lambda, Amazon Kinesis, Amazon S3, Amazon Machine Learning und Amazon DynamoDB zu verwenden, um Anwendungen zu erstellen, die IoT-Daten sammeln, verarbeiten, analysieren und nutzen, ohne dass dazu eine Infrastruktur verwaltet werden muss.

Datenverarbeitung

Die gesammelten Daten enthalten möglicherweise nützliche Informationen. Sie können die extrahierten Informationen analysieren, um festzustellen, ob sie zum Wachstum Ihres Unternehmens beitragen können. Diese Informationen können Ihnen beispielsweise Hinweise über das Benutzerverhalten und die relative Popularität Ihrer Produkte liefern. Die am besten bewährte Methode zum Sammeln solcher Informationen ist, die Rohdaten in ein Data Warehouse zu laden, um anschließend Analysen durchführen zu können.

Dazu gibt es zwei Arten von Arbeitsabläufen – stapelorientierte Verarbeitung und Echtzeitverarbeitung. Die häufigsten Verarbeitungsformen, Online Analytical Processing (OLAP) und OLTP, verwenden jeweils einen dieser Abläufe. Online Analytical Processing (OLAP)-Verarbeitung ist in der Regel stapelorientiert. Im Gegensatz dazu sind OLTP-Systeme für Echtzeitverarbeitung ausgelegt und im Allgemeinen nicht gut für eine stapelorientierte Verarbeitung geeignet. Wenn Sie die Datenverarbeitung von Ihrem OLTP-System abkoppeln, beeinflusst sie nicht Ihren OLTP-Workload.

Betrachten wir zunächst, welche Prozesse die Stapelverarbeitung umfasst.

ETL-Prozess (Extrahieren, Transformieren, Laden)

ETL ist der Prozess, mit dem Daten aus mehreren Quellen in Data Warehousing-Systeme geladen werden. ETL ist in der Regel ein kontinuierlicher Prozess mit einem wohldefinierten Workflow. Während dieses Prozesses werden zunächst Daten aus einer oder mehreren Quellen extrahiert. Die extrahierten Daten werden dann bereinigt, veredelt, transformiert und in ein Data Warehouse geladen. In einer ETL-Pipeline werden üblicherweise Hadoop-Framework-Tools wie Apache Pig und Apache Hive verwendet, um Transformationen großer Datenmengen durchzuführen.

ELT-Prozess (Extrahieren, Laden, Transformieren)

ELT ist eine ETL-Variante, bei der die extrahierten Daten sofort in das Ziel geladen werden. Transformationen werden erst ausgeführt, nachdem die Daten in das Data Warehouse geladen wurden. ELT funktioniert in der Regel gut, wenn das Zielsystem leistungsfähig genug ist, um Transformationen zu handhaben. Häufig wird Amazon Redshift in ELT-Pipelines verwendet, weil dieser Service bei der Durchführung von Transformationen sehr leistungsfähig ist.

Online Analytical Processing (OLAP)

OLAP-Systeme speichern aggregierte historische Daten in multidimensionalen Schemata. OLAP-Systeme werden häufig für das Data Mining verwendet, da sie es ermöglichen, aus mehreren Dimensionen Daten zu extrahieren und Trends zu ermitteln. Für die Erstellung von OLAP-Systemen wird häufig Amazon Redshift verwendet, weil dieser Service für schnelle Joins optimiert ist.

Betrachten wir nun, welche Prozesse bei der Echtzeitverarbeitung von Daten beteiligt sind.

Echtzeitverarbeitung

Wir haben bereits zuvor erwähnt, dass Amazon Kinesis eine Lösung zum Erfassen und Speichern von Streaming-Daten ist. Sie können solche Daten sequenziell Datensatz für Datensatz oder mithilfe gleitender Zeitfenster verarbeiten und die verarbeiteten Daten dann für eine Vielzahl von Analysen verwenden, einschließlich Korrelationen, Aggregationen, Filterungen und Sampling. Diese Art der Verarbeitung wird Echtzeitverarbeitung genannt. Durch die in Echtzeitverarbeitung gewonnenen Informationen erhalten Unternehmen Einblick in viele Aspekte ihrer Geschäfts- und Kundenaktivitäten wie Service-Nutzung (zur Messung oder Abrechnung), Serveraktivität, Website-Klicks sowie Geolokalisierung von Geräten, Personen und physischen Gütern. Zudem werden Unternehmen in die Lage versetzt, schnell auf neue Situationen zu reagieren zu können. Echtzeitverarbeitung erfordert eine Verarbeitungsebene, die hochgradig skalierbar ist und simultane Prozesse ermöglicht.

Zur Verarbeitung von Streaming-Daten in Echtzeit können Sie AWS Lambda verwenden. Lambda kann Daten von AWS IoT oder Amazon Kinesis Streams direkt verarbeiten. Mit Lambda können Sie Code ausführen, ohne dass Sie Server bereitstellen und verwalten müssen.

Eine weitere Möglichkeit zur Verarbeitung von Daten aus Amazon Kinesis Streams bietet Amazon Kinesis Client Library (KCL). KCL gibt Ihnen mehr Flexibilität als AWS Lambda, um eingehende Daten für eine weitere Verarbeitung bereitzustellen. Sie können KCL auch verwenden, um umfangreiche Veränderungen und Anpassungen in der Verarbeitungslogik vorzunehmen.

Die einfachste Methode, um Streaming-Daten in AWS zu laden, bietet Amazon Kinesis Firehose. Damit werden Streaming-Daten erfasst und automatisch in Amazon Redshift geladen, sodass Sie nahezu in Echtzeit Analysen mit BI-Tools und Dashboards vornehmen können, die Sie bereits verwenden. Nachdem Sie Ihre Bereitstellungsregeln mit Firehose definiert haben, übernimmt dieser Service zuverlässig das Bereitstellen der Daten und deren Lieferung an Amazon Redshift.

Datenspeicherung

Im Folgenden wird beschrieben, wie Sie Ihre Daten entweder in einem Data Warehouse oder einem Data Mart speichern.

Data Warehouse

Ein *Data Warehouse* ist, wie bereits erwähnt, ein zentrales Repository für Informationen aus einer oder mehreren Datenquellen. Mit einem Data Warehouse können Sie schnell große Datenmengen analysieren und mithilfe von BI-Tools verborgene Muster in Ihren Daten finden. Datenexperten erstellen Abfragen für ein Data Warehouse, um Offline-Analysen auszuführen und Trends zu ermitteln. Benutzer in der gesamten Organisation verwenden die Daten, um mit Ad-hoc-SQL-Abfragen, regelmäßigen Berichten und Dashboards kritische Geschäftsentscheidungen zu treffen.

Data Mart

Ein *Data Mart* ist ein vereinfachtes Data Warehouse, das auf einen bestimmten Funktions- oder Objektbereich ausgerichtet ist. Zum Beispiel können spezielle Data Marts für jede Abteilung in Ihrer Organisation oder segmentierte Data Marts für jede Region vorhanden sein. Sie können Data Marts aus einem großen Data Warehouse, einem Operational Data Store oder in hybrider Form aus beiden erstellen. Data Marts sind einfach zu entwerfen, zu erstellen und zu administrieren. Da Data Marts jedoch auf bestimmte Funktionsbereiche ausgerichtet sind, können Abfragen über mehrere Funktionsbereiche wegen der Verteilung komplex werden.

Sie können Amazon Redshift verwenden, um Data Marts als Ergänzung zu Data Warehouses einzurichten.

Datenanalyse und -visualisierung

Nachdem Sie die Daten verarbeitet und zur weiteren Analyse bereitgestellt haben, benötigen Sie die richtigen Tools, um die verarbeiteten Daten zu analysieren und zu visualisieren.

In vielen Fällen können Sie die Datenanalyse mit den gleichen Tools ausführen, die Sie für die Verarbeitung von Daten verwenden. Setzen Sie beispielsweise SQL Workbench ein, um Ihre Daten in Amazon Redshift mit ANSI SQL zu analysieren. Auch Amazon Redshift arbeitet gut mit beliebten BI-Lösungen von Drittanbietern zusammen.

Amazon QuickSight ist ein schneller, in der Cloud verfügbarer BI-Service, mit dem Sie mühelos Visualisierungen erstellen, Ad-hoc-Analysen durchführen und umgehend geschäftsbezogene Einblicke in Ihre Daten erhalten. Amazon QuickSight ist in Amazon Redshift integriert, befindet sich derzeit in der Testphase, soll aber noch im Jahr 2016 allgemein verfügbar sein.

Wenn Sie Amazon S3 als primären Speicher verwenden, können Sie einen beliebten Weg gehen und für Analyse und Visualisierung Apache Spark-Notebooks auf Amazon Elastic MapReduce (Amazon EMR) ausführen. Auf diese Weise haben Sie die Wahl, entweder SQL oder benutzerdefinierten Code auszuführen, der in Sprachen wie Python oder Scala geschrieben ist.

Einen anderen Visualisierungsansatz bietet Apache Zeppelin, eine quelloffene BI-Lösung, die Sie auf Amazon EMR ausführen können, um Daten aus Amazon S3 mithilfe von Spark SQL zu visualisieren. Apache Zeppelin ermöglicht auch die Visualisierung von Daten aus Amazon Redshift.

Analyse-Pipeline mit AWS-Services

AWS bietet eine breite Palette von Services zum Implementieren einer End-to-End-Analyseplattform. Abbildung 2 zeigt die bisher genannten Services und wo diese innerhalb der Analyse-Pipeline zum Einsatz kommen.

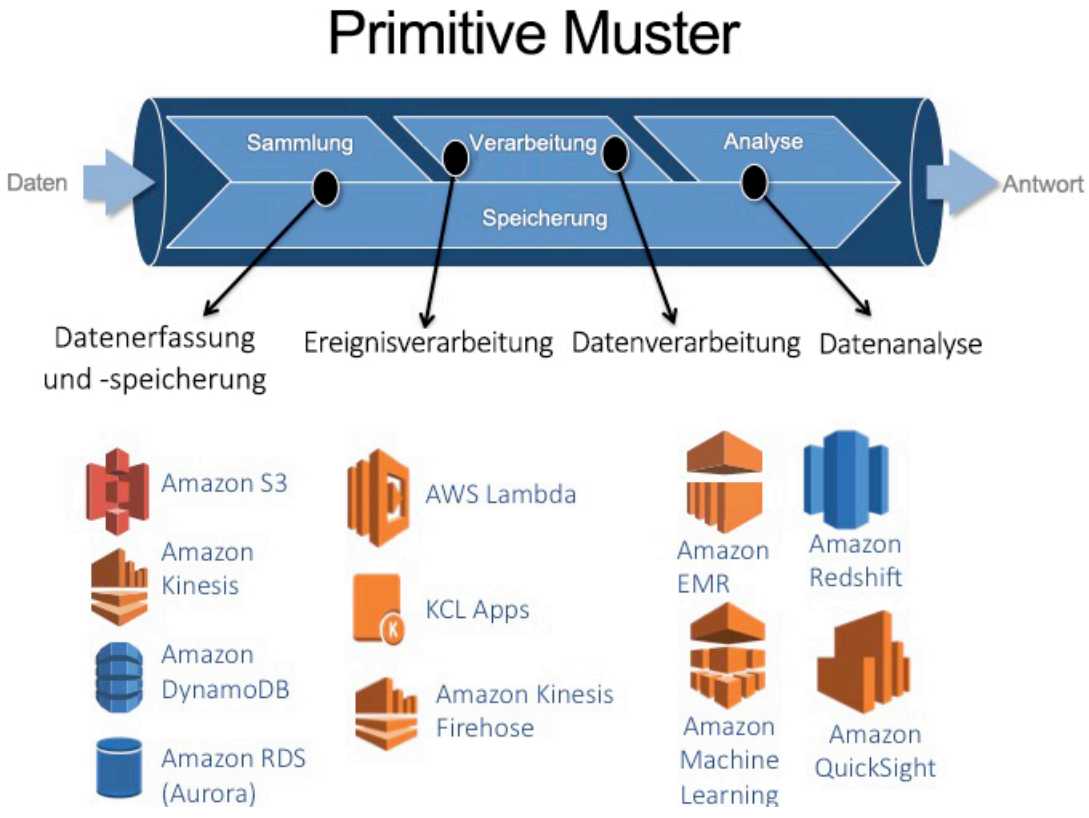

Abbildung 2: Analyse-Pipeline mit AWS-Services

Data Warehouse-Technologieoptionen

In diesem Abschnitt stellen wir Optionen für eine Data Warehouse-Erstellung vor: Zeilenorientierte Datenbanken, spaltenorientierte Datenbanken und Architekturen für massiv-parallele Verarbeitung.

Zeilenorientierte Datenbanken

Zeilenorientierte Datenbanken speichern in der Regel ganze Zeilen in einem physischen Block. Mit Sekundärindizes wird bei Leseoperationen eine hohe Leistung erreicht. Datenbanken wie Oracle Database Server, Microsoft SQL Server, MySQL und PostgreSQL sind zeilenorientierte Datenbanksysteme. Diese Systeme wurden traditionell für Data Warehousing verwendet, sind aber besser für Transaktionsverarbeitung (OLTP) geeignet als für Analysen.

Die Leistung von zeilenbasierten Systemen für Data Warehouses lässt sich mit einer Vielzahl von Techniken optimieren: Materialisierte Ansichten, voraggregierte Rollup-Tabellen, Erstellung von Indizes für jede mögliche Kombination von Prädikaten, Implementierung von Datenpartitionen für den wirksamen Einsatz der Partitionsbereinigung mithilfe des Abfrageoptimierers sowie indexbasierte Joins.

Die Grenzen herkömmlicher zeilenbasierter Datenspeicher werden durch die Ressourcen eines einzelnen Computers bestimmt. Diesem Problem lässt sich zu einem gewissen Grad mit Data Marts nachkommen, da diese funktionelles Partitionieren einsetzen. Ein Data Warehouse lässt sich in mehrere Data Marts unterteilen, die jeweils einen bestimmten Funktionsbereich abdecken. Allerdings verlangsamt sich mit wachsenden Data Marts die Datenverarbeitung.

In zeilenbasierten Data Warehouses werden in den Blöcken, die das Abfrageprädikat erfüllen, sämtliche Spalten aller Zeilen gelesen, darunter auch die Spalten, die vom Benutzer gar nicht ausgewählt wurden. Dieser Ansatz führt in Data Warehouses, deren Tabellen mehr Spalten haben als tatsächlich für die Abfrage benötigt werden, zu erheblichen Leistungsengpässen.

Spaltenorientierte Datenbanken

In spaltenorientierte Datenbanken werden nicht die Zeilen, sondern die Spalten in separaten Gruppen von physischen Blöcken organisiert. Das erhöht die Effizienz des Festplatten- oder Arbeitsspeicherzugriffs schreibgeschützter Abfragen, da nur die Spalten gelesen werden, auf die die Abfrage zugreift. Hierin liegt der Vorteil spaltenorientierter gegenüber zeilenorientierten Datenbanken in Data Warehouses.

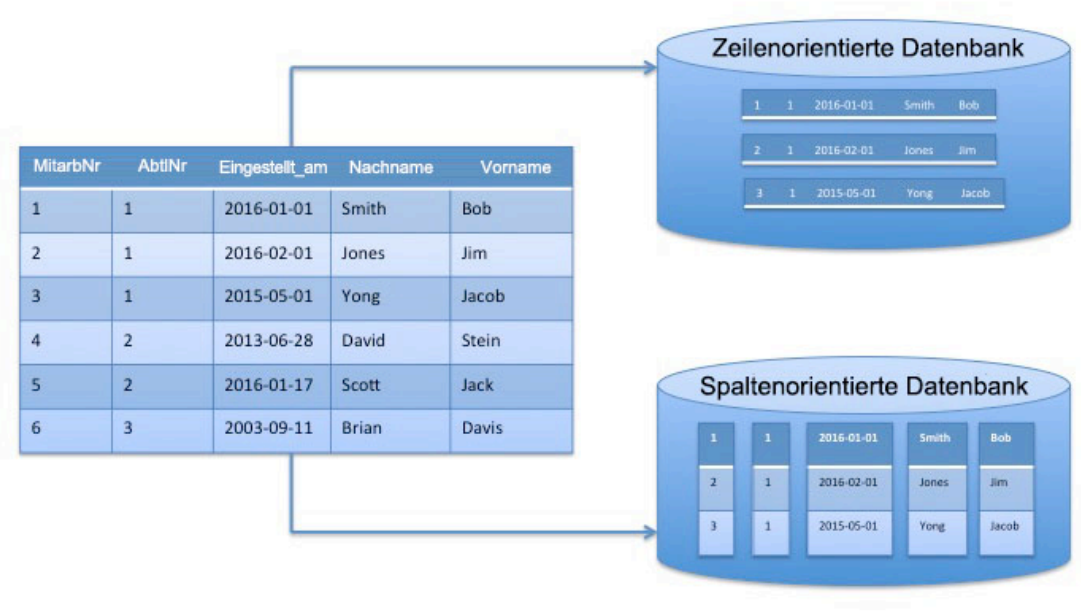

Abbildung 3: Spaltenorientierte Datenbanken im Vergleich zu zeilenorientierten Datenbanken

Abbildung 3 zeigt den Hauptunterschied zwischen zeilen- und spaltenorientierten Datenbanken. In zeilenorientierten Datenbanken werden Zeilen in eigenen Blöcken organisiert; in spaltenorientierten Datenbanken werden Spalten in eigenen Blöcken organisiert.

Neben den kürzeren Zugriffszeiten liegt ein weiterer großer Vorteil spaltenorientierter Datenbanken in der verbesserten Datenkomprimierung. Jede Spalte wird zu einer eigenen Gruppe von Blöcken zusammengefasst, deren Daten jeweils denselben Datentyp aufweisen. Dadurch kann die Datenbank äußerst effiziente Komprimierungsalgorithmen einsetzen, wodurch im Vergleich zu zeilenorientierten Datenbanken zum einen der Speicherbedarf und zum anderen

die Anzahl der Zugriffe sinkt, weil eine bestimmte Datenmenge in weniger Blöcken gespeichert wird.

Zu den spaltenorientierten Datenbanken für den Einsatz in Data Warehouses gehören unter anderem Amazon Redshift, Vertica, Teradata Aster und Druid.

MPP-Architekturen

Eine MPP-Architektur (Massively Parallel Processing Architecture, Architektur mit Massenparallelrechnern) ermöglicht die Nutzung aller im Cluster verfügbaren Ressourcen für die Datenverarbeitung, wodurch sich die Leistung von Data Warehouses in Petabyte-Größe signifikant erhöht. Der Leistungsgewinn in MPP-Data-Warehouses wird durch Hinzufügen zusätzlicher Knoten zum Cluster erreicht. Beispiele für Data Warehouses, die auf der MPP-Architektur basieren, sind Amazon Redshift, Druid, Vertica, GreenPlum und Teradata Aster. Auch Open-Source-Frameworks wie Hadoop und Spark unterstützen MPP.

Weitere Informationen zu Amazon Redshift

Amazon Redshift ist eine MPP-Technologie für Datenbanken mit Spaltenorientierung, die wesentliche Vorteile für ein leistungsstarkes und kostengünstiges Data Warehousing mit effizienter Kompression, weniger Zugriffen und geringerem Speicherbedarf bietet. Sie basiert auf ANSI SQL, sodass vorhandene Abfragen mit geringfügigen Änderungen oder auch unverändert ausgeführt werden können. Daher nutzen heute immer mehr Unternehmen diese Technologie für ihre Data Warehouses und Data Marts. Im vorliegenden Abschnitt werden die Funktionen von Amazon Redshift detaillierter beschrieben.

Amazon Redshift bietet kurze Abfragezeiten und hohe Zugriffsraten für Datensätze nahezu jeder Größe. Dies wird durch die spaltenweise Speicherung und die parallele Verteilung von Abfragen über mehrere Knoten erreicht. Der Service erleichtert Verwaltungs- und Wartungsarbeiten und verringert deren Kosten durch die Automatisierung vieler alltäglicher administrativer Aufgaben, die mit der Bereitstellung, Konfiguration, Überwachung, Sicherung und dem Schutz eines Data Warehouse einhergehen. Mithilfe dieser Automatisierung können Sie in Minutenschnelle ein Data Warehouse in Petabyte-Größe aufsetzen – eine Aufgabe, die früher bei einer lokalen Implementierung Wochen oder Monate in Anspruch genommen hat.

Leistung

Mithilfe von Techniken wie der spaltenweisen Speicherung, Datenkomprimierung und Zone Maps reduziert Amazon Redshift die Anzahl der für Abfragen erforderlichen Ein- und Ausgaben. Die überlappende Sortierung sorgt für eine hohe Leistung ohne die Notwendigkeit der Verwaltung von Indizes oder Projektionen.

Amazon Redshift basiert auf einer MPP-Architektur, die es ermöglicht, SQL-Operationen parallel durchzuführen und zu verteilen, damit alle verfügbaren Ressourcen ausgenutzt werden. Die zugrunde liegende Hardware ist für Hochleistungsdatenverarbeitung ausgelegt und verwendet lokal angeschlossene Speicher, um den Durchsatz zwischen den CPUs und den Festplatten zu maximieren. Der Durchsatz zwischen den Knoten wird mithilfe eines 10GbE-Mesh-Netzwerks maximiert. Die Leistung kann den Bedürfnissen eines Data Warehouse angepasst werden: Dafür bietet AWS die Dense Compute (DC)-Technologie mit SSD-Festplatten sowie Dense Storage (DS)-Optionen an. Softwareupgrades werden kontinuierlich bereitgestellt. Dadurch profitieren die Benutzer ohne eigenes Zutun von ständigen Leistungsverbesserungen.

Beständigkeit und Verfügbarkeit

Amazon Redshift bietet das höchstes Maß an Datenbeständigkeit und -verfügbarkeit, da ausgefallene Knoten in einem Data-Warehouse-Cluster automatisch erkannt und ersetzt werden. Der Service stellt den Ersatzknoten unmittelbar zur Verfügung und lädt die Daten mit dem häufigsten Zugriff, sodass die Abfrage so schnell wie möglich fortgesetzt werden kann. Da Amazon Redshift innerhalb des Clusters eine Datenspiegelung durchführt, können ausgefallene Knoten mit den Daten eines anderen Knotens wiederhergestellt werden. Bis ein Ersatzknoten bereitgestellt und der Datenbank hinzugefügt wurde, ist der Cluster schreibgeschützt. Dieser Vorgang dauert in der Regel nur wenige Minuten.

Alle Amazon Redshift-Cluster befinden sich dabei in einer Availability Zone.[3] Wenn Sie Amazon Redshift-Implementierungen mit mehreren Availability Zones bevorzugen, können Sie eine Datenspiegelung vornehmen und die Replikation und das Failover selbst übernehmen.

In der Amazon Redshift Management Console, der Verwaltungskonsole von Amazon Redshift, lässt sich mit wenigen Klicks eine zuverlässige Umgebung für die Notfallwiederherstellung einrichten. Kopien der Datensicherungen können in mehreren AWS-Regionen gespeichert werden. Im Falle einer Serviceunterbrechung in einer AWS-Region lässt sich ein Cluster anhand der in einer anderen AWS-Region gespeicherten Sicherung wiederherstellen. Sie erhalten innerhalb weniger Minuten nach Einleitung des Wiederherstellungsvorgangs Lese-/Schreibzugriff für den betreffenden Cluster.

Skalierbarkeit und Elastizität

Mit nur wenigen Klicks in der Konsole oder über einen API-Aufruf lässt sich die Anzahl der Knoten im Data Warehouse anpassen, wenn sich die Leistungs- oder Kapazitätsanforderungen ändern.[4] Mit Amazon Redshift lässt sich ein einzelner Knoten mit einer Kapazität von 160 GB problemlos auf eine Infrastruktur in Petabyte-Größe mit komprimierten Benutzerdaten und zahlreichen Knoten hochskalieren. Weitere Informationen finden Sie unter About Clusters and Nodes im *Amazon Redshift Cluster Management Guide*.[5]

Während der Skalierung versieht Amazon Redshift den bestehenden Cluster mit einem Schreibschutz, stellt einen neuen Cluster in der von Ihnen gewählten Größe bereit und kopiert dann die Daten aus dem alten Cluster in den neuen. Sie müssen in diesem Zeitraum nur für den aktiven Amazon Redshift-Cluster bezahlen. Während der neue Cluster angelegt wird, kann der alte weiterhin Abfragen beantworten. Nachdem die Daten auf den neuen Cluster kopiert wurden, leitet Amazon Redshift Abfragen automatisch an den neuen Cluster um und löscht den alten.

Mithilfe von Amazon Redshift-API-Aktionen lassen sich programmgesteuert Cluster starten und skalieren sowie Sicherungen anlegen und wiederherstellen. Die API-Aktionen können wahlweise in einen vorhandenen Automatisierungsstack integriert werden, oder Sie können eine benutzerdefinierte Automatisierung ganz nach Ihrem Bedarf erstellen.

Schnittstellen

Amazon Redshift ist mit benutzerdefinierten Java Database Connectivity (JDBC)- und Open Database Connectivity (ODBC)-Treibern ausgestattet, die sich über die Registerkarte **Connect Client** der Konsole herunterladen lassen, d. h. Sie können unter mehreren vertrauten SQL-Clients auswählen. Sie können auch die JDBC- und ODBC-Standardtreiber von PostgreSQL verwenden. Weitere Informationen über Amazon Redshift-Treiber finden Sie unter Amazon Redshift and PostgreSQL im *Amazon Redshift Database Developer Guide.* [6]

Amazon Redshift ist mit vielen BI- und ETL-Lösungen bekannter Anbieter kompatibel.[7] In diesen Integrationen werden Lade- und Entladevorgänge im Verarbeitungsknoten parallel ausgeführt, was zu einer Maximierung der Aufnahme- und Exportrate von Daten in oder aus mehreren Ressourcen führt, darunter Amazon S3, Amazon EMR und Amazon DynamoDB. Mit Amazon Kinesis Firehose können Streaming-Daten erfasst und automatisch in Amazon Redshift geladen werden. Sie erhalten damit nahezu in Echtzeit Analysedaten für vorhandene BI-Tools und -Dashboards. Metriken für die Datenverarbeitungs-, Arbeitsspeicher- und Speichernutzung sowie für den Lese-/Schreibdatenverkehr im Amazon Redshift Data Warehouse-Cluster stehen in der Konsole oder über Amazon CloudWatch-APIs zur Verfügung.

Sicherheit

Für mehr Datensicherheit können Sie Amazon Redshift in einer virtuellen privaten Cloud ausführen, die auf dem Amazon Virtual Private Cloud Service (Amazon VPC) basiert. Mithilfe des softwaredefinierten Netzwerkmodells der VPC lassen sich Firewallregeln zur Beschränkung des Datenverkehrs konfigurieren.[8] Amazon Redshift unterstützt SSL-fähige Verbindungen zwischen Clientanwendung und Data-Warehouse-Cluster, wobei die Daten bei der Übertragung verschlüsselt werden.

Die Daten werden im Amazon Redshift-Datenverarbeitungsknoten gespeichert, können aber nur vom Führungsknoten des Clusters aus abgerufen werden. Diese Art der Isolation trägt zusätzlich zu mehr Sicherheit bei. Amazon Redshift ist kompatibel mit AWS CloudTrail, sodass alle Amazon Redshift-API-Aufrufe geprüft werden können.[9] Zum Schutz von Ruhedaten verschlüsselt Amazon Redshift jeden Block während des Schreibvorgangs auf den Datenträger mit einer hardwarebeschleunigten AES-256-Methode. Diese Verschlüsselung findet auf einer unteren Ebene des E/A-Untersystems statt. Dabei werden alle auf einen Datenträger geschriebenen Daten verschlüsselt, darunter auch Zwischenabfrageergebnisse. Die Blöcke werden im Ist-Zustand gesichert, das bedeutet, dass Sicherungen ebenfalls verschlüsselt werden. Amazon Redshift übernimmt standardmäßig die Schlüsselverwaltung. Wahlweise können Sie Ihre Schlüssel aber auch mit eigenen Hardware-Sicherheitsmodulen (HSM) oder über den AWS Key Management Service verwalten.[10,11]

Kostenmodell

Für ein Amazon Redshift-Data-Warehouse-Cluster müssen Sie keine langfristige vertragliche Bindung eingehen oder Vorauszahlungen leisten. Durch dieses Preismodell entsteht Ihnen keinerlei Investitionsaufwand und komplexe Vorausplanungen zum Erwerb von Data-Warehouse-Kapazitäten entfallen. Die Gebühren fallen entsprechend der Größe und Anzahl der Knoten im Cluster an.

Sicherungsspeicher im Umfang von bis zu 100 % des bereitgestellten Speichers ist kostenlos. Verfügen Sie beispielsweise über einen aktiven Cluster mit 2 XL-Knoten und insgesamt 4 TB Speicher, erhalten Sie von AWS kostenlos bis zu 4 TB Sicherungsspeicher für Amazon S3. Weiterer Sicherungsspeicher sowie Sicherungen, die nach Beendigung Ihres Clusters weiterhin gespeichert werden, werden zu den üblichen Amazon S3-Preisen abgerechnet.[12] Es entstehen keine Datenübertragungskosten für die Kommunikation zwischen Amazon S3 und Amazon Redshift. Weitere Informationen dazu finden Sie unter Amazon Redshift – Preise.[13]

Optimale Nutzungsmuster

Amazon Redshift ist optimal geeignet für Online Analytical Processing (OLAP) mithilfe Ihrer vorhandenen BI-Tools. Unternehmen setzen Amazon Redshift für folgende Aufgaben ein:

- BI und Reporting
- Analysieren globaler Verkaufsdaten für unterschiedliche Produkte
- Speichern von Aktienhandelsverlaufsdaten
- Analysieren von Ad Impressions und Klicks
- Sammeln von Daten aus Spielen
- Analysieren sozialer Trends
- Messen von klinischer Qualität, Betriebseffizienz und finanzieller Leistungsfähigkeit im Gesundheitswesen

Nicht empfohlene Anwendungsfälle

Amazon Redshift ist für folgende Anwendungsfälle in der Regel nicht geeignet:

- **Kleine Datenbestände** – Amazon Redshift ist auf die parallele Verarbeitung innerhalb eines Clusters ausgerichtet. Bei Datenbeständen unter 100 GB werden Sie nicht von allen Vorteilen profitieren, die Amazon Redshift bietet. In diesem Fall ist Amazon RDS die geeignete Lösung.
- **OLTP** – Amazon Redshift wurde für ausgesprochen schnelle und kostengünstige Analysemöglichkeiten für Data-Warehouse-Arbeitslasten entwickelt. Wenn Sie ein schnelles Transaktionssystem benötigen, sollten Sie auf ein konventionelles Amazon RDS-System mit einer relationalen Datenbank oder eine NoSQL-Datenbank wie Amazon DynamoDB zurückgreifen.
- **Unstrukturierte Daten** – Die Daten müssen für Amazon Redshift anhand eines festgelegten Schemas strukturiert sein. Ein beliebiges Schema für jede Zeile wird von Amazon Redshift nicht unterstützt. Sind Ihre Daten unstrukturiert, können Sie sie mit ETL-Prozessen (Extrahieren, Transformieren und Laden) über Amazon EMR für die Übermittlung an Amazon Redshift vorbereiten. Bei JSON-Daten können Sie Schlüssel-/Wertpaare speichern und in Abfragen die systemeigenen JSON-Funktionen verwenden.[14]

- **BLOB-Daten** – Wenn Sie BLOB-Dateien (Binary Large Object) wie Video-, Bild- und Musikdateien speichern möchten, sollten Sie Amazon S3 verwenden und in Amazon Redshift auf den Speicherort der Daten verweisen. In diesem Anwendungsfall werden in Amazon Redshift die Metadaten (z. B. Elementname, Erstellungsdatum, Eigentümer und Speicherort) der binären Dateien verwaltet, während die großen Dateien selbst in Amazon S3 gespeichert werden.

Migration zu Amazon Redshift

Die Strategie für die geplante Migration eines vorhandenen Data Warehouse zu Amazon Redshift hängt von mehreren Faktoren ab:

- Größe der Datenbank und der darin enthaltenen Tabellen
- Netzwerkbandbreite zwischen Quellserver und AWS
- Sollen Migration und Umstellung auf AWS in einem oder in mehreren Schritten erfolgen?
- Datenänderungsrate im Quellsystem
- Umwandlungen während der Migration
- Partnertool für Migration und ETL

Migration in einem Schritt

Eine Migration in einem Schritt eignet sich besonders für kleine Datenbanken, die keinen kontinuierlichen Betrieb erfordern. Sie können vorhandene Datenbanken in CSV-Dateien (Comma-Separated Value, Durch Trennzeichen getrennt) exportieren, die Datenbestände dann mit AWS Import/Export Snowball oder einem ähnlichen Service für Amazon S3 vorbereiten und schließlich in Amazon Redshift laden. Anschließend kann die Amazon Redshift-Zieldatenbank auf Datenkonsistenz mit der Quelle getestet werden. Nach erfolgreichem Abschluss aller Validierungen wird die Datenbank schließlich auf AWS umgestellt.

Migration in zwei Schritten

Eine Migration in zwei Schritten eignet sich in der Regel für alle Datenbankgrößen:

1. **Migration von Ursprungsdaten:** Die Daten werden aus der Quelldatenbank extrahiert. Dies erfolgt vorzugsweise außerhalb der Spitzenzeiten, um die Beeinträchtigungen möglichst gering zu halten. Anschließend werden die Daten unter Befolgung der zuvor beschriebenen Schritte zu Amazon Redshift migriert.
2. **Migration von geänderten Daten:** Hierbei handelt es sich um Daten, die in der Quelldatenbank nach Übergabe der Ursprungsdatenmigration an die Zieldatenbank, aber vor der Umstellung geändert wurden. Bei diesem Schritt werden Quell- und Zieldatenbank synchronisiert. Nach der Migration aller geänderten Daten können Sie die Daten in der Zieldatenbank validieren und testen und bei bestandenen Tests die Umstellung auf das Amazon Redshift Data Warehouse durchführen.

Tools für die Datenbankmigration

Für die Migration von Daten steht eine Reihe von Tools und Technologien zur Verfügung. Einige dieser Tools sind austauschbar. Daneben gibt es auch noch weitere Produkte von Drittanbietern oder Open-Source-Tools.

1. Der AWS Database Migration Service unterstützt die zuvor beschriebenen Migrationsverfahren in einem oder in zwei Schritten.[15] Für eine Migration in zwei Schritten muss eine zusätzliche Protokollierung aktiviert werden, damit die Änderungen am Quellsystem erfasst werden. Die Protokollierung erfolgt auf Tabellen- oder Datenbankebene.
2. Zusätzlich stehen folgende Partnertools für die Datenintegration zur Verfügung:
   - Attunity
   - Informatica
   - SnapLogic
   - Talend
   - Bryte

Weitere Informationen zur Datenintegration und zur Kontaktaufnahme mit Partnern finden Sie unter Amazon Redshift-Partner.[16]

Entwerfen von Data-Warehousing-Workflows

In den vorherigen Abschnitten haben Sie Amazon Redshift und die Funktionen kennen gelernt, die den Service zur idealen Lösung für Data Warehouses machen. Im Folgenden werden nun die gängigsten Muster für den Entwurf von Data-Warehousing-Workflows mit Amazon Redshift sowie entsprechende Anwendungsfälle vorgestellt.

Angenommen, ein internationales Unternehmen in der Bekleidungsbranche mit mehr als tausend Filialen vertreibt bestimmte Bekleidungslinien über Kaufhäuser und Discounter und verfügt über eine Internetpräsenz. Vom technischen Standpunkt her sind diese drei Vertriebskanäle zurzeit voneinander unabhängig. Sie verfügen über unterschiedliche Managementteams, POS-Systeme und Buchhaltungsabteilungen. Es gibt kein zentrales System, in dem alle Datenbestände zusammenlaufen und die der Vorstandsvorsitzenden einen transparenten Einblick in das gesamte Geschäft bieten würden.

Weiterhin angenommen, sie wünscht sich einen unternehmensweiten Überblick über die Vertriebskanäle und möchte bei Bedarf Sofortanalysen durchführen können, z. B.:

- Welche Trends gibt es in den einzelnen Vertriebskanälen?
- Welche Regionen entwickeln sich in den Vertriebskanälen besser?
- Wie effektiv sind die Werbemaßnahmen und Angebote des Unternehmens?
- Welche Trends gibt es in den einzelnen Bekleidungslinien?
- Welche externen Faktoren, wie die Arbeitslosenquote oder das Wetter, könnten sich auf den Umsatz des Unternehmens auswirken?
- Welche Auswirkungen auf den Umsatz haben Filialattribute wie die Dienstzugehörigkeit von Mitarbeitern und Management, die Lage in einer Ladenzeile oder einem Einkaufszentrum, die Platzierung der Waren im Geschäft, Werbeaktionen, Gondelköpfe, Werbeschreiben, Laden-Displays usw.?

Dieses Problem lässt sich mit einem Enterprise Data Warehouse lösen. Dieses erfasst die Daten aus den unterschiedlichen Systemen der drei Vertriebskanäle sowie die Wetter- und Wirtschaftsdaten aus öffentlichen Quellen. Jede Datenquelle lädt ihre Daten täglich in das Data Warehouse. Da sich die Strukturen der Datenquellen voneinander unterscheiden können, wird ein ETL-Prozess ausgeführt, der die Datenstruktur vereinheitlicht. Anschließend können Daten aus allen Quellen gleichzeitig analysiert werden. Zu diesem Zweck wird die folgende Datenverkehrsarchitektur eingesetzt:

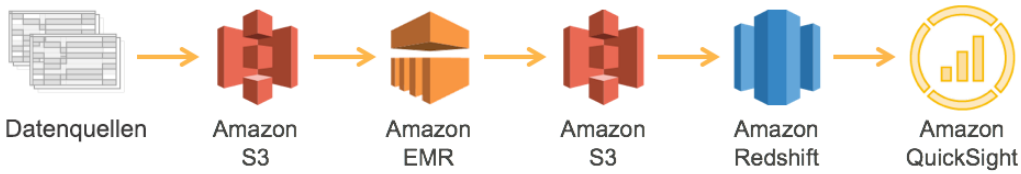

Abbildung 4: Enterprise Data Warehouse – Workflow

1. Im ersten Schritt des Prozesses müssen die Daten aus vielen verschiedenen Quellen in Amazon S3 gelangen. Amazon S3 ist eine höchst robuste, kostengünstige und skalierbare Speicherplattform, auf der zu sehr günstigen Kosten parallele Schreibvorgänge aus vielen verschiedenen Quellen erfolgen können.
2. Amazon EMR kommt zur Bereinigung und Umwandlung der Daten vom Quell- in das Zielformat zum Einsatz. Die werkseitige Integration von Amazon EMR in Amazon S3 ermöglicht parallele Durchsatz-Threads von jedem Knoten im Amazon EMR-Cluster zu und von Amazon S3.

Arbeitslasten in Data Warehouses fallen in Regel nachts an. Da die Analysen nicht mitten in der Nacht erfolgen müssen, besteht die einzige Anforderung an diesen Umwandlungsprozess, dass er bis morgens abgeschlossen sein muss, wenn die Vorstandsvorsitzende und andere Unternehmensbenutzer Zugriff auf Berichte und Dashboards benötigen. Daher können Sie Amazon EC2 Spot Market einsetzen, um hier die Kosten für die Extraktion, die Transformation und das Laden der Daten weiter zu senken.[17] Eine gute Spot-Strategie wäre es, zunächst um Mitternacht einen sehr geringen Preis zu bieten und diesen im Lauf der Zeit zu steigern, bis die notwendige Kapazität gesichert ist. Falls kein Erfolg mit den Spot-Geboten erzielt wurde, können Sie bei nahendem Fristende auf die On-Demand-Preise zurückgreifen, um sicherzustellen, dass Sie die Aufgabe rechtzeitig erfüllen. Der Umwandlungsprozess der einzelnen Quellen auf Amazon EMR kann sich unterscheiden, aber mit dem bedarfsbasierten Zahlungsmodell von AWS können Sie für jede Umwandlung einen eigenen Amazon EMR-Cluster erstellen und diesen zum geringstmöglichen Preis genau auf die jeweiligen Erfordernisse abstimmen, ohne mit den Ressourcen für andere Aufgaben in Konflikt zu geraten.

3. Nach jeder Umwandlungsaufgabe werden die formatierten und bereinigten Daten auf Amazon S3 bereitgestellt. Amazon S3 kommt hier erneut zum Einsatz, da Amazon Redshift diese Daten parallel aus Amazon S3 laden kann; dies geschieht über mehrere Threads von jedem Cluster-Knoten aus. Amazon S3 bietet zudem einen historischen Datensatz und bildet somit die formatierte Informationsquelle zwischen den Systemen. Wenn im Laufe der Zeit zusätzliche Anforderungen hinzugefügt werden, können die Daten auf Amazon S3 von anderen Tools für Analysen genutzt werden.

4. Amazon Redshift lädt, sortiert, verteilt und komprimiert die Daten in Tabellen, damit Analyseabfragen effizient und parallel ausgeführt werden können. Wenn das Unternehmen und mit ihm die Datenmenge im Lauf der Zeit wächst, lässt sich die Kapazität leicht durch Hinzufügen weiterer Knoten ausbauen.

5. Zur Visualisierung der Analysen können Sie Amazon QuickSight oder eine der vielen Partnerplattformen zur Visualisierung nutzen, die sich über ODBC oder JDBC mit Amazon Redshift verbinden lassen. Hier können die Vorstandsvorsitzende und ihre Mitarbeiter Berichte, Dashboards und Diagramme anzeigen. Die Mitglieder der Unternehmensleitung können nun anhand dieser Daten fundierte Entscheidungen zum Einsatz der Unternehmensressourcen treffen, die letztendlich den Umsatz und den Shareholder Value steigern können.

Diese Architektur ist sehr flexibel und lässt sich erweitern, wenn das Unternehmen wächst, neue Vertriebskanäle eröffnet, zusätzliche mobile Anwendungen für Kunden veröffentlicht oder weitere Datenquellen importiert. Dafür sind jeweils nur wenige Klicks in der Amazon Redshift Management Console oder einige API-Aufrufe erforderlich.

Zusammenfassung

Data Warehouses unterliegen derzeit einem strategischen Wandel, da die Unternehmen ihre lokalen Analysedatenbanken und -lösungen vermehrt in die Cloud verlagern, um so von der Einfachheit, der hohen Leistung und der Kosteneffizienz profitieren zu können. Dieses Whitepaper analysiert umfassend den aktuellen Status des Data-Warehousing-Angebots auf AWS. AWS bietet umfangreiche Services und ein starkes Partnernetzwerk, mit dem Sie Ihre Data Warehouses in die Cloud verlagern können. Das Ergebnis ist eine hochleistungsfähige, kosteneffiziente Analysearchitektur, die im Gleichklang mit Ihrem Unternehmen in der globalen AWS-Infrastruktur wächst.

Mitwirkende

Dieses Dokument ist unter der Mitarbeit folgender Personen und Organisationen entstanden:

- Babu Elumalai, Solutions Architect, Amazon Web Services
- Greg Khairallah, Principal BDM, Amazon Web Services
- Pavan Pothukuchi, Principal Product Manager, Amazon Web Services
- Jim Gutenkauf, Senior Technical Writer, Amazon Web Services
- Melanie Henry, Senior Technical Editor, Amazon Web Services
- Chander Matrubhutam, Product Marketing, Amazon Web Services

Weitere Informationen

Zusätzliche Informationen finden Sie in den folgenden Ressourcen:

- Apache Hadoop-Softwarebibliothek[18]
- Bewährte Methoden für Amazon Redshift[19]
- Lambda-Architektur[20]

Anmerkungen

1 https://www.forrester.com/report/The+Forrester+Wave+Enterprise+Data+Warehouse+Q4+2015/-/E-RES124041

2 http://aws.amazon.com/streaming-data/

3 http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html

4 http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html

5 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-about-clusters-and-nodes

6 http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-and-postgres-sql.html

7 http://aws.amazon.com/redshift/partners/

8 https://aws.amazon.com/vpc/

9 https://aws.amazon.com/cloudtrail/

10 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-HSM.html

11 https://aws.amazon.com/kms/

12 http://aws.amazon.com/s3/pricing/

13 http://aws.amazon.com/redshift/pricing/

14 http://docs.aws.amazon.com/redshift/latest/dg/json-functions.html

15 https://aws.amazon.com/dms/

16 https://aws.amazon.com/redshift/partners/

17 http://aws.amazon.com/ec2/spot/

18 https://hadoop.apache.org/

19 http://docs.aws.amazon.com/redshift/latest/dg/best-practices.html

20 https://en.wikipedia.org/wiki/Lambda_architecture