tdwi
WISSEN

DATA FABRIC

E-Book

Digitalisierung zum Self Service 2.0

Autor: Prof. Dr. Carsten Felden

tdwi.eu

Herausgeber

SIGS DATACOM GmbH
Lindlaustraße 2c
53842 Troisdorf

info@sigs-datacom.de
www.sigs-datacom.de

TDWI E-Book in Kooperation mit

denodo
DATA VIRTUALIZATION

Inhalt

Vorwort

Die Verfügbarkeit und Wirtschaftlichkeit der Nutzung von Commodity-Hardware und -Software zur Entwicklung flexibler und skalierbarer Hochleistungsumgebungen hat in den letzten Jahren zugenommen. Das hat die Methoden und Prozesse für Reporting, Business Intelligence und Analytics verändert. Da Unternehmen sich für Cloud-Computing-Plattformen entscheiden und ihre Daten und Anwendungen in eine hybride Cloud-Umgebung migrieren, können sie mehrere Plattformen nutzen, um neue und wachsende Datentypen für Analysen zu unterstützen. Die Datenvirtualisierung, eine Kernkomponente einer Data Fabric, spielt eine wichtige Rolle bei der Unterstützung des Zugriffs auf Daten sowie deren Verwaltung und Analyse – und zwar über unterschiedliche Plattformen für ein traditionelles Berichtswesen und Business Intelligence. Dies gilt auch für moderne Anwendungsfälle wie maschinelles Lernen und künstliche Intelligenz, integrierte Analysen zugunsten einer automatisierten Entscheidungsfindung oder eine Analyse der Kombination ruhender Daten mit Echtzeit-Streaming-Daten. Das vorliegende eBook stellt das Konzept der Data Fabric im Spiegel seiner Anforderungen vor und geht darauf ein, wie Data-Science-Initiativen durch Datenermittlung und Analyse unterstützt werden können, um den Wert des Datenbewusstseins und der Verwendung einer Data Fabric als Unternehmensdatenkatalog herauszustellen.

1 Einleitung

Die Verfügbarkeit und Wirtschaftlichkeit der Nutzung von Commodity-Hardware und -Software zur Entwicklung flexibler und skalierbarer Hochleistungsumgebungen hat in den letzten Jahren zugenommen.

Das hat die Methoden und Prozesse für Reporting, Business Intelligence und Analytics verändert. Da Unternehmen sich zunehmend für Cloud-Computing-Plattformen entscheiden und ihre Daten und Anwendungen in eine hybride Cloud-Umgebung migrieren, können sie mehrere Plattformen parallel nutzen, um neue und wachsende Datentypen für Analysen zu unterstützen.

Ohne bewährte Verfahren und Organisation für ihre hybride Datenumgebung besteht allerdings das Risiko erhöhter Komplexität, wenn Datenverbraucher nach den Datenressourcen suchen, die sie für ihre Arbeit benötigen, und darauf zugreifen. Eine Möglichkeit, Daten in dieser verteilten Umgebung zu verwalten und für Analysen zu verwenden, besteht in der Anwendung eines modernen Ansatzes im Sinne einer logischen Datenstruktur, um unterschiedliche Daten zu verfeinern und auf intelligente Weise zusammenzuführen.

Die Datenvirtualisierung, die eine Kernkomponente der logischen Datenstruktur ist, kann eine wichtige Rolle bei der Unterstützung des Zugriffs auf Daten über unterschiedliche Plattformen für herkömmliche Berichte und Business Intelligence (BI) hinweg sowie deren Verwaltung und die Analyse spielen. Das gilt auch für aktuelle Aufgaben wie Maschinelles Lernen und Artificial Intelligence, integrierte Analysen für automatisierte Entscheidungsfindung und Analysen, die ruhende Daten mit Echtzeit-Streaming-Datenquellen kombinieren.

Im Rahmen dieses eBooks werden sechs relevante Funktionen der logischen Datenstruktur beschrieben, die sich mit modernen Datenmanagement- und Analysebemühungen beschäftigen und mehrere Plattformen, neue Datentypen und -quellen sowie erweiterte Analysen umfassen. Zu diesen Funktionen gehört die Aktivierung von Datenanalysten in der neuen hybriden Multicloud-Datenlandschaft. Dazu kommen umfassende Techniken zur nahtlosen Integration von Daten aus Multicloud-Plattformen. Diese zeigen, wie erweiterte Analysefähigkeit das Datenbewusstsein und die nahtlose Zugänglichkeit bei gleichzeitiger Aufrechterhaltung einer hohen betrieblichen Leistung erleichtert und wie die Leistungssteigerung durch Optimierungen die Verzögerungen im Zusammenhang mit der Datenlatenz reduziert. Darüber hinaus wird erläutert, wie Data-Science-Initiativen durch Datenermittlung und Analytics unterstützt werden können, die ruhende und bewegte Daten kombinieren. Schließlich zeigen die diskutierten Kriterien den Wert des Datenbewusstseins und der Verwendung einer Data Fabric als Unternehmensdatenkatalog auf:

- Datenbereitstellung wird zum Schlüssel, um eine Entkopplung zwischen Datenbereitstellung und Datennutzung zu schaffen.
- Identifikation von Daten zu identifizierten Themen der Digitalisierung.
- Verständnis der Data Fabric als Voraussetzung für die Funktionsfähigkeit des digitalisierten Unternehmens, bei dem eine Dezentralisierung der Datenverwendung mit einhergeht.

Das vorliegende eBook wird im zweiten Kapitel zunächst das Konzept der Data Fabric einführen. Kapitel 3 geht auf das Anwendungsszenario des Self Service ein, das als initiales Szenario für eine Data Fabric gilt. Im vierten Kapitel werden Konsequenzen aus den fachlichen und technischen Anforderungen an eine Data Fabric diskutiert und in Kapitel 5 erfolgt die Präsentation der Ergebnisse von TDWI Research USA zum Thema.

2 Reflektion der Digitalisierung – die Data Fabric im Kontext der Digitalisierung und des Self Service

Die Digitalisierung der Geschäftswelt führt dazu, dass in jedem Augenblick riesige Datenmengen entstehen, die es auszuwerten und zu nutzen gilt. Die traditionellen Maßnahmen greifen hier jedoch zu kurz: Heute wird eine Kombination aus verknüpften Echtzeitdaten, Self Service und einem hohen Maß an Automatisierung, Geschwindigkeit und Intelligenz verlangt. Dazu gibt es weitere Probleme und Herausforderungen: neue Datenquellen, verschiedene Datenmodelle (SQL/noSQL), Batch-Datenbewegung, starre Transformations-Workflows und die Verteilung von Daten über Multi- und Hybrid-Cloud-Umgebungen (hierzu und zum Folgenden: TDWI 2020).

Der Begriff der Data Fabric steht für eine einheitliche Architektur miteinander verknüpfter Technologien, Anwendungen und Dienste, deren Daten übergreifend nutzbar sind. Dazu bedarf es sowohl einer entsprechenden Datenplattform, die sich durch Performance, Interoperabilität, freie Skalierbarkeit und Sicherheit auszeichnet, als auch einer passenden Datenstruktur, sodass sich Daten fließend in Analyseprozesse einbinden lassen.

Eine Datenstruktur ist ein Designkonzept zur Erzielung wiederverwendbarer und erweiterter Datenintegrationsdienste und Datenpipelines unter Nutzung einer entsprechenden Semantik zur flexiblen und integrierten Datenbereitstellung. Diese Struktur ist damit eine Kombination aus Datenmanagement- und Integrationstechnologien, Architekturdesign und Services, die über mehrere Bereitstellungs- und Orchestrierungsplattformen verfügbar gemacht werden. Dadurch entsteht ein schnellerer und in einigen Fällen vollständig automatisierter Datenzugriff mit entsprechender Datenfreigabe.

Grundsätzlich ist eine Datenbereitstellung mit unterschiedlichen technischen Ansätzen nichts Neues. Der Begriff der Data Fabric betont allerdings den Fokus auf umfassende neue und kommende Datenmanagementtechnologien und Datenintegrationsansätze, die durch kollaborative Datenmanagementpraktiken wie zum Beispiel DataOps bereitgestellt werden. Dazu gehören unter anderem Streaming-Datenintegration, Datenvirtualisierung, semantische Anreicherung, von Artificial Intelligence/Machine Learning (AI/ML) unterstützte aktive Metadaten, Wissensdiagramme und graphbasierte Datenbanksysteme (neben anderen nicht relationalen Datenbanken).

Der Hintergrund dieser Betrachtungsweise liegt darin begründet, dass sich das Tempo technischer Entwicklungen und deren Verbreitung in den Unternehmen in den letzten Jahren beschleunigt hat, was die Verfügbarkeit immer neuer Lösungen und das Entstehen neuer Open-Source-Communities begründet. Dies verlangt architektonische Flexibilität. Das Paradigma, dass Daten der entscheidende Wettbewerbsvorteil für jedes Unternehmen sind, um erfolgreich zu sein und zu wachsen, ist perspektivisch gültig. Es ist daher essenziell, dass Unternehmen eine hohe Datenverfügbarkeit aufweisen, um Geschäfts- und Kundenanforderungen vollständig zu erfüllen.

Daten sind mittlerweile ein großer Einflussfaktor auf die Geschäftsmodelle geworden und motivieren Unternehmen zur Digitalisierung ihrer Abläufe. Daher versuchen immer mehr Unternehmen, zusätzlichen Nutzen aus den verfügbaren Daten zu schöpfen. Beispielsweise, um neue Einnahmequellen zu erschließen oder ihre Kosten durch betriebliche Effizienzsteigerungen zu senken (Dittmar et al. 2016). Angesichts der zunehmenden Bedeutung der Cloud und des Internet of Things sowie der immer günstigeren Datenspeicherung und Datenverarbeitung besteht auch eine geringere Bindung an lokale Rechenzentren vor Ort. Dass mehr Daten, mehr Datentypen und mehr Speicherortoptionen als je zuvor verfügbar sind, führt zu enormen Herausforderungen hinsichtlich des Datenmanagements.

Eine Data Fabric wird nun zum Werkzeug für diese Anforderungen an das Datenmanagement. Sie verbindet mehrere Standorte, Datentypen und -quellen miteinander und bietet zahlreiche Methoden, um auf die Daten zuzugreifen. Unternehmen können die Daten verarbeiten, verwalten und speichern, während sie sich innerhalb der Data-Fabric-Architektur bewegen. Sie sind in der Lage, die Daten für interne und externe Anwendungen sowie eine Vielzahl analytischer und betrieblicher Anwendungsfälle abzurufen und zu nutzen. Zu diesen Anwendungsfällen gehören Analysen zur Prognoseerstellung, die Produktentwicklung sowie die Optimierung von Vertrieb und Marketing.

3 Herausforderungen im Kontext der Datenbereitstellung

Eine Umfrage von Wrobel zeigt, dass Umsatzsteigerung und Kosteneinsparung die häufigsten durch Big Data verfolgten Ziele sind (Wrobel 2012, S. 35). Davenport kategorisiert die Wertschöpfung in drei Bereiche: Entscheidungsverbesserungen, Kosteneinsparungen sowie Optimierung von Dienstleistungen und Produkten (Davenport 2014, S. 21). Die Analyse kleiner und strukturierter Datenmengen stützt unternehmensinterne Entscheidungen wie die Angebotsunterbreitung an Kunden, die Preisgestaltung der Produkte und das Bestandsmanagement. Die Art der Entscheidungen bleibt gleich, allerdings ermöglicht eine Bereitstellung großer unstrukturierter Datenmengen die aufgabenbezogene Analyse, um neue Erkenntnisse zu erlangen (Davenport 2014, S. 64). Damit verbessern entsprechend verfügbare Daten die Grundlage einer Entscheidung (BITKOM 2015, S. 18).

Um datenorientierte Anwendungen gewinnbringend einzusetzen, empfiehlt es sich, eine entsprechende Strategie aus der Geschäftsstrategie abzuleiten. Daten werden zielgerichtet auf betriebliche Problemstellungen angewendet, damit ein Mehrwert für das Unternehmen entsteht. Dies geschieht beispielsweise durch bessere und schnellere Entscheidungen, die Optimierung der Produktpalette sowie Risikoreduzierung infolge präziserer Prognosen (Dorschel 2015, S. 18). Die in einem solchen Zusammenhang genannten Big Data bieten außerdem das Potenzial technikbasierter Kosteneinsparungen. Ein Hadoop-Cluster speichert und verarbeitet beispielsweise solche Daten in effizienter Art und Weise. Die Kosten für die gleiche Menge an Daten liegen deutlich unter denen einer traditionellen relationalen Datenbank. Allerdings bedingt die Einführung eines neuen Hadoop-Cluster auch Implementierungskosten, zum Beispiel für neues Personal. Daher ist jeder Fall unternehmensspezifisch abzuwägen. Zusätzlich ermöglichen Big-Data-Werkzeuge eine schnellere Datenverarbeitung und somit Zeiteinsparungen. Das führt unter anderem zu besseren Modellen zur Identifikation von Leistungstreibern, der Integration von mehr Variablen für kundenspezifische Angebote oder der schnelleren Reaktion auf Umweltveränderungen (Davenport 2014, S. 58–62). Die Entwicklung neuer beziehungsweise die Verbesserung bestehender Produkte und Dienstleistungen ist ein weiteres Ziel von Big Data. Produktentwicklungen sind Investitionen und wirken dem Ziel der Kosten- und Zeiteinsparung entgegen. Allerdings bieten sie das Potenzial für bedeutende zukünftige Umsatzsteigerungen.

Insgesamt zeigt sich schon durch diese Ausführungen, dass eine effiziente Datenbereitstellung der Schlüssel für eine erfolgreiche digitale Transformation ist. Durch eine solche Datenverfügbarkeit werden Konzepte wie Big-Data-Analytics oder Self-Service-Analytics möglich.

3 Herausforderungen im Kontext der Datenbereitstellung

3.1 Big-Data-Analytics

Eine Vielzahl technologischer Konzepte und Systeme werden heute mit dem Begriff Big Data verbunden. Nicht immer verbergen sich dahinter ausschließlich neuartige Technologien, denn der vieldiskutierte Begriff wird vom Marketing gerne aufgegriffen und mit etablierten Technologien verbunden, sodass die Trennschärfe bei den Anwendern abnimmt (TDWI 2011). Für eine erste Kategorisierung können Technologien zur Erzeugung von Big Data (zum Beispiel Sensorik, Geo-Tracking), zur Verarbeitung und Integration von Big Data (zum Beispiel Streaming, Virtualisierung), zur Speicherung von Big Data (zum Beispiel Analytische Plattformen, Cloud) und zur Auswertung von Big Data (zum Beispiel Visualisierung, Advanced Analytics) unterschieden werden. Für die nachfolgende Betrachtung erfolgt eine Auswahl der Technologiecluster, die maßgeblichen Einfluss auf die Gestaltung klassischer Business-Intelligence-(BI-)Architekturen haben.

Technologien zum Streaming beziehungsweise zur Datenpopulation in Echtzeit erweitern die gängige batchorientierte Datenpopulation darum, Datenquellen mit kontinuierlichem Datenfluss und niedriger Latenzzeit anbinden zu können. Hier werden Einsatzszenarios unterstützt, in denen die Quelldaten (zum Beispiel Daten aus Sensoren) nicht persistent vorgehalten werden, sondern nur kontinuierlich hinsichtlich bestimmter auffälliger Datenkonstellationen untersucht werden müssen, um definierte Aktionen auszulösen.

Bei der Menge und Heterogenität von Big Data sind zunehmend Technologien im Einsatz, die auf die bisher verfolgte Maxime der physischen Integration sämtlicher relevanter Quellen in einer singulären dispositiven Umgebung zugunsten einer Integration über eine logische, auf Metadaten basierende Zwischenschicht verzichten (logische Integration). Somit werden Abfragen über bisher separate oder gar nicht für Analysen verfügbare Systeme ermöglicht. Entsprechende Virtualisierungslösungen ermöglichen den direkten Zugriff auf unterschiedliche Datenquellen für eine integrierte Datenanalyse.

Anwender fordern heutzutage, ohne Unterstützung des IT-Bereichs selbstständig Analysen auf Datenbestände durchführen zu können, welche die Big-Data-Charakteristika aufweisen. Entsprechende Self-Service-BI-Werkzeuge werden heute häufig ebenfalls zu den Big-Data-Technologien gezählt, da dazu unter anderem In-Memory-Technologien eingesetzt werden, die auf dem Desktop-Rechner des Anwenders laufen. Der Funktionsumfang von Self-Service-BI-Werkzeugen, gerade im Bereich Visual Analytics, zeigt erkennbare Überschneidungen zum Funktionsumfang klassischer BI-Suiten. Insofern hat sich ein interessanter Wettbewerb zwischen den erfolgreichen Herstellern von Self-Service-BI-Werkzeugen und klassischen BI-Suiten entwickelt, der eine zunehmende Ausrichtung aller am Markt befindlicher BI-Werkzeuge auf Analytics als Ergänzung der klassischen Fokussierung auf Reporting bedingt.

Self-Service-BI erweitert die klassischen Funktionsklassen wie Standard Reporting und Ad-hoc-Analyse des Presentation Layer. Die Möglichkeit des Nutzers, abgeschottete Analyseräume (Sandboxing) zu bilden, kann im Benutzerwerkzeug entweder an seinem lokalen Rechner oder im DWH ermöglicht werden. Die wesentliche Herausforderung bei der Öffnung der Reporting-Definitionen und der Integration durch Nutzer eingebrachter Daten bleibt jedoch die Ausgestaltung der Governance, um einen konsistenten Rahmen durch organisatorische Instrumente aufrechtzuerhalten.

Die oben aufgezeigten Erweiterungen der traditionellen BI-Architektur lassen sich in einer zukunftsfähigen Gesamtarchitektur integrieren, die sich als analytisches Ökosystem beschreiben lässt. In diesem existieren mehrere analytische Datenhaltungen nebeneinander, da unterschiedliche Granularitäten, Datenquellen oder Analyseziele nicht dogmatisch physisch in einem Data Warehouse zu integrieren sind. Sofern die logische Transparenz besteht, in welchem analytischen Datenpool welche Daten in welcher Form abgelegt sind, lassen sich diese bei konkretem Bedarf durchaus auch nur virtuell und temporär zusammenführen. Das analytische Ökosystem bietet Services bezüglich Datenspeicherung, -veredelung, -distribution, -analyse und -zugriff an. Allerdings ist es nicht für jeden Anwendungsfall sinnvoll, sämtliche Services über alle zur Verfügung gestellten Schichten zu durchlaufen.

3 Herausforderungen im Kontext der Datenbereitstellung

3.2 Self-Service-Analytics

Der Besitz von Daten stellt in Unternehmen nicht automatisch einen Wert dar. Dieser entsteht vielmehr, wenn ebenfalls die Möglichkeit und Fähigkeit vorhanden sind, Informationen aus unübersichtlichen Datenmengen und deren heterogenen Strukturen zu gewinnen und Entscheidungsträgern zur Verfügung zu stellen. Mit der zunehmenden Menge verfügbarer Daten und einer zumindest wahrgenommenen Bewegung hin zu Big Data, also heterogenen, umfangreichen und schnelllebigen Daten, ist zwangsläufig das Thema der Business Analytics und damit die Auswertung in den Mittelpunkt geraten. Big Data machen das beispielsweise per Online-Analytical-Processing-(OLAP-)Anfrage gerichtete Auffinden von Mustern und Informationen zur Suche nach der Nadel im Heuhaufen. Eine zunehmende Datenverfügbarkeit erschwert Informationsarbeitern die im Sinne einer Informationslogistik definierte Bereitstellung von Daten zur rechten Zeit am richtigen Ort in der für den Entscheider notwendigen Qualität (Convertino/ Echenique 2017).

3.2.1 Von der einfachen Ergebnisnutzung zur individuellen Ergebnisgestaltung

Self-Service-Analytics ist ein Ansatz zur Datenanalyse, der es Geschäftsanwendern ermöglicht, autonom auf Unternehmensdaten zuzugreifen und mit diesen zu arbeiten, obwohl sie keinen technischen Hintergrund in den Bereichen statistische Analyse, Business Intelligence oder Data Mining haben. Dieser Ansatz ist insbesondere im Kontext der Big Data prominenter geworden, da Nutzer flexiblere Analysen wünschten. Dass Anwender nun basierend auf ihren eigenen Abfragen und Analysen Entscheidungen treffen können, schafft eine Unabhängigkeit der Mitarbeiter aus der Fachabteilung von den Teams für Business Intelligence und IT, die in der Regel dafür zuständig sind, die meisten Berichte zu erstellen. Somit können sich beide Seiten auf ihre zentralen Aufgaben konzentrieren.

Da nun aber Self-Service-Analytics-Software von Nutzern verwendet wird, die üblicherweise technisch nicht so versiert sind wie ausgebildete Data Scientists, ist es zwingend erforderlich, dass die Benutzerschnittstelle für diese Software intuitiv gestaltet ist und somit ein Dashboard mit benutzerfreundlicher Navigation aufweist. Im Idealfall werden Schulungen geplant, die Benutzern vermitteln, welche Daten verfügbar sind und wie diese abgefragt werden können. So können sie die Software entsprechend nutzen und letztlich datengesteuerte Entscheidungen zur Lösung von Geschäftsproblemen treffen. Dies entbindet IT-Abteilungen aber nicht von der Aufgabe, einen Rahmen zu schaffen, in dem die Anwender agieren können. Unter dem Aspekt, dass Self Service die Benutzer dazu ermutigt, Entscheidungen anhand von Daten zu treffen, statt ihrer Intuition zu folgen, kann die Flexibilität, die sie bietet, unnötige Verwirrung stiften, wenn keine Data-Governance-Richtlinie vorhanden ist. Die Richtlinie sollte unter anderem definieren, welche Schlüsselmetriken für die Erfolgsermittlung verwendet werden, welche Prozesse zum Erstellen und Teilen von Berichten befolgt werden sollten, welche Berechtigungen für den Zugriff auf vertrauliche Daten erforderlich sind und wie Datenqualität, Sicherheit und Datenschutz erhalten werden. Hier sei noch einmal der Unterschied zwischen Self-Service-Reporting und Self-Service-Analytics hervorgehoben. Der Erfolg der Anwendung des Self-Service-Reporting besteht in der Herstellung der Einsicht von Daten, die der Beantwortung einer Frage dienen. Es ist also eine Antwort, die aus fachlichen Daten generiert und auch fachlich gegeben wird. Durch Self-Service-Analytics hingegen kann darüber hinaus auch eine Antwort in Form eines abstrakten Ergebnisses eines Algorithmus entstehen. Dazu müssen beispielsweise Gütemaße für die Ergebnisse der Algorithmen (Vorhersagen) bewertet oder auch das eigentliche Versagen von Algorithmen für neue Datenbestände (Overfitting) als Effekte erkannt werden. Hier geht also das notwendige Verständnis über die Sachkenntnis der Daten und der zugrundeliegenden Geschäftsprozesse hinaus (Schymik et al. 2017).

3 Herausforderungen im Kontext der Datenbereitstellung

3.2.2 Wie weit soll ein Self Service gehen?

Reflektierend auf die vorherigen Ausführungen stellt sich die Frage, wie weit Self Service eigentlich gehen soll. Soll er sich auf die Analyse vorab bereitgestellter Daten konzentrieren oder auch deren Bereitstellung umfassen? Eine Self-Service-Datenvorbereitung benennt den Prozess des Sammelns, Bereinigens und Konsolidierens von Daten in einer Datei oder Datentabelle. Notwendig ist eine solche Datenaufbereitung, da ein Umgang mit qualitativ mangelhaften, inkonsistenten oder nicht standardisierten Daten erforderlich ist, die sowohl aus unterschiedlichen Quellsystemen stammen als auch manuell erfasste unstrukturierte Dokumente beinhalten. Eine solche Datenvorbereitung wird immer dann interessant, wenn Daten ad hoc beziehungsweise flexibel auf zum Beispiel eine Entscheidungssituation bezogen zusammengestellt werden müssen. Dies findet, bedingt durch die Natur der jeweiligen Aufgaben, regelmäßiger im strategischen und taktischen Management statt und weniger im operativen. Dies liegt an den unterschiedlichen Freiheitsgraden.

Auf höheren Managementebenen existieren mehr Freiheitsgrade, da die Arbeit nicht vollständig prozessual strukturiert ist. Es wird lediglich ein Rahmen vorgegeben, beispielsweise die Einführung neuer Produkte oder das Agieren auf neuen Märkten, der dann mit Inhalten zu füllen ist. Dies begründet eine Anforderung der Self-Service-Analytics, da verbunden mit einem höheren Freiheitsgrad der Arbeit sich ein strukturgebender Rahmen nicht mehr definieren lässt. Ein solcher schränkt das dynamische Arbeiten mit Datenbeständen ein und limitiert die Kreativität, die jedoch erforderlich ist, um immer wieder neue Wege durch die unterschiedlichen Kombinationen von Datenbeständen zu finden.

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

Die richtige Architektur für die dispositive Datenverarbeitung war lange klar definiert: Ein idealtypisch singuläres (Enterprise) Data Warehouse sammelt in einem Hub-&-Spoke-Ansatz aus den unterschiedlichen operativen Quellsystemen die relevanten Daten und harmonisiert, integriert und persistiert diese in einem mehrschichtigen Datenintegrations- und Datenveredelungsprozess. Aus diesem vielzitierten Single Point of Truth werden anschließend Datenextrakte in der Regel multidimensional in voneinander fachlich abgrenzbaren Data Marts gehalten, die dann wiederum den Presentation Layer mit seinen spezifischen Berichts- und Analysewerkzeugen versorgen (Business Intelligence im engen Sinne). Diese Mehrschichtenarchitektur hat in bestimmten Anforderungskontexten weiterhin ihre Gültigkeit (Dittmar et al. 2016). Die Einhaltung dieser idealtypischen architekturellen Vorgaben wird jedoch immer schwieriger, wenn Fachbereiche eine höhere Änderungsdynamik fordern, als sie von einer zentralen Data-Warehouse-Infrastruktur geleistet werden kann. Auch in Konzernstrukturen, in denen sehr unterschiedliche Geschäftsmodelle unter einem Dach bestehen, ist ein zentrales Data Warehouse oft nicht die richtige Lösung (Dittmar et al. 2016).

4.1 Zeitgenössische Architektur

Es zeigt sich also, dass die integrierte Analyselandschaft der Zukunft vielfältiger und komplexer sein muss und dabei die Flexibilität im Vordergrund steht. Die Integration erfolgt hier logisch, durch einheitliche Metadaten, Data Governance und Stammdaten. Der physische Integrationsanspruch tritt wieder zurück. Anwenderunternehmen müssen hier ihre Grundsatzentscheidungen bezüglich Make-or-Buy überprüfen. Durch die zunehmende Technologievielfalt ist es eine Strategie, die vorkonfektionierten Lösungen etablierter Anbieter zu adaptieren. Die klassischerweise aus dem Bereich Open Source entstammenden Technologien wirken hier auf den ersten Blick günstiger, es muss aber mehr Basis-Know-how aufgebaut werden, um die Interoperabilität der Architekturkomponenten sicherzustellen. Daher wird zukünftig möglicherweise ein Nebeneinander von selbst erstellten und integrierten vorkonfektionierten Lösungen und extern, zum Beispiel in einer Cloud, bereitgestellten Architekturkomponenten festzustellen sein. Getrennte Welten zwischen operativen und dispositiven Systemen, zwischen disjunkten Werkzeugen für unterschiedliche Analysebedarfe, zwischen backendorientierter Datenintegration und frontendorientierter Datenanalyse, aber auch zwischen Produktiv- und temporären Entwicklungs- und Evaluationsumgebungen konvergieren zunehmend in diesen analytischen Ökosystemen. Einen Überblick über ein entsprechendes analytisches Ökosystem liefert die folgende Abbildung.

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

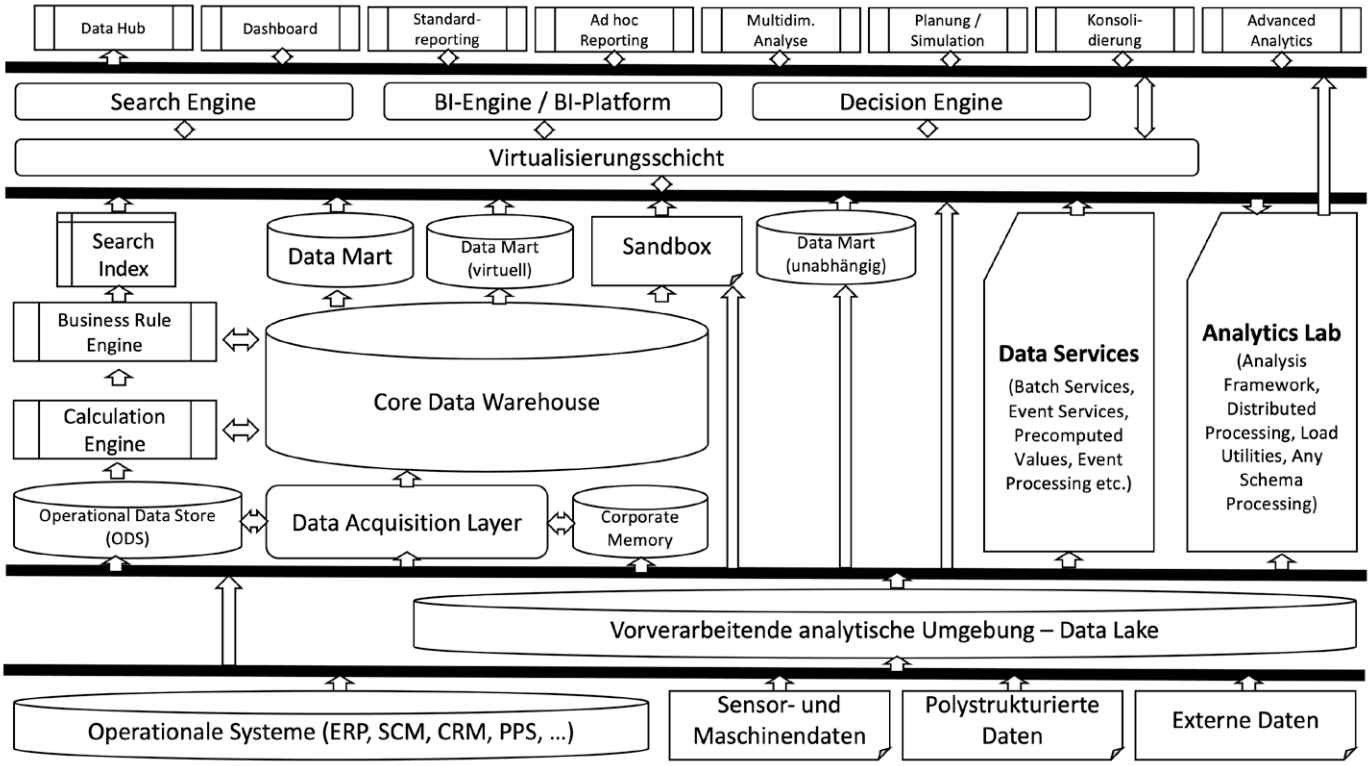

Abbildung 1: *Analytisches Ökosystem im Überblick*

Neben den Komponenten, die von einer klassischen BI-Architektur bekannt sind, sind dort auch Elemente einer Lambda-Architektur für die Realtime-Verarbeitung enthalten, zum Beispiel von Sensordaten mit einem Speed Layer, Batch Layer und Service Layer. Dadurch können Realtime-Daten historisiert abgelegt, aber auch an ein Realtime-Monitoring durchgereicht werden. In einigen Anwendungsfällen kann es dabei sinnvoll sein, den Batch Layer mit den backendorientierten Komponenten der klassischen BI-Architektur zusammenzuführen. Zudem enthält die Architektur auch eine Laborumgebung (Analytics Lab) für Datenanalysen auf Basis flexibler Schemata (Any Schema oder noSQL).

Grundsätzlich ist im Sinne des aktuell propagierten Data-Lake-Konzepts eine Systemumgebung zu schaffen, die zu jedem Zeitpunkt den Sprung zwischen lokalen Ressourcen und dem Objektspeicher ermöglicht. Dadurch wird eine zielorientierte Nutzung möglich und auch noch nicht absehbare Auswertungen werden entsprechend den sich ergebenden Anforderungen abbildbar.

4.2 Data Warehouse

Zur Sicherstellung von Ad-hoc-Analysen und einem Drill Down bis auf die Basisdaten wurden in den 1990er-Jahren Architekturkonzepte unter dem Schlagwort Data Warehouse marktfähig gemacht. Also Systeme, die Daten aus in der Regel operativen Quellensystemen über eine Transformationsschicht in die Datenhaltung übertragen. Die Probleme der Datenbewirtschaftung (Informationslogistik) für das Management mussten gemeinsam technisch wie auch betriebswirtschaftlich gelöst werden, was zu einer neuen Qualität der Unterstützungssysteme führte. Der Aufbau der Technologien entlang der Architekturebenen war nicht trivial, aber sobald die Infrastruktur bereitgestellt war, kamen zunehmend mehr Fragen nach der analytischen Nutzung in den betriebswirtschaftlichen Fachbereichen auf. Dies wiederum führte zu einer neuen Tendenz, die begrifflich stärker auf die fachlichen Domänen abstellte.

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

4.3 Data Lake

Der Begriffswandel der BI hin zu Big-Data-Analytics verspricht einen intensiveren Einsatz weiterführender Datenanalysen, verbunden mit direkten Handlungsempfehlungen, die aus den Analyseergebnissen abgeleitet werden. Dabei wird BI nicht diskreditiert, sondern eher in den Kontext der performanten Informationslieferung und aktiven Analyse gesetzt. Hier verspricht Analytics eine Aufklärung mittels Algorithmen über bestmögliche zukünftige Handlungen, womit bekannte Prognoseverfahren und Optimierungsrechnung erneut in den Fokus rücken. Die neue Qualität von Business Analytics wird in der sinnvollen Kombination von Methoden der Datenanalyse (Data Science) und Analysemodellen liegen. Die Konvergenz datenorientierter und modellorientierter Verfahren scheint daher naheliegend und bringt neue Aspekte in die Betrachtung von MUS auf dem Zeitstrahl. Hier treten Algorithmen in den Vordergrund, die (teil-)automatisierte Entscheidungsprozesse ermöglichen, die auf großen polystrukturierten Datenbeständen (Big Data) in Echtzeit Empfehlungen für bestmögliche Entscheidungen geben oder selbst entscheiden. Noch ist nicht eindeutig, welche Überschrift dieser Phase später zugeordnet wird, jedoch scheint sich der allgemeine Begriff der Digitalisierung herauszukristallisieren.

Traditionelle Dateisysteme wie das Data Warehouse sind zwar stabil und bekannt, liefern aber nicht die immer die flexible Grundlage für Data-Analytics-Umgebungen. Dies gilt insbesondere für deren Anforderungen an zusätzliche Informationen, die Metadaten. Metadaten, also Daten über die Daten, sind erforderlich, um die nun im Sinne der Big Data heterogenen und unstrukturiert vorliegenden Daten in einem sogenannten Objektspeicher derart zu beschreiben, dass sie maschinell nutzbar werden. Hier greift dann auch das Konzept des Data Lake. Dabei handelt es sich um einen sehr großen Datenspeicher, der die Daten aus unterschiedlichsten Quellen in ihrem Rohformat aufnimmt. Er kann sowohl unstrukturierte als auch strukturierte Daten enthalten und lässt sich für Big-Data-Analysen einsetzen (Litzel 2018). Durch das zugrunde gelegte Verständnis der Objektorientierung werden in einem Data Lake die Dateien und dazugehörigen Metadaten gegebenenfalls erweitert, um weitere benutzerdefinierte Informationen in einem Objekt gekapselt und entsprechend in verschiedenen variablen Storage-Klassen im Sinne eines Objektspeichers abgelegt. Solche Storage-Klassen basieren auf unterschiedlichen Anforderungen wie der Art der Daten oder der Zeit- und Speicherperformance. Darüber hinaus lassen sich entsprechende Sicherheitsaspekte abbilden, sodass eine Unterteilung in kleinere Virtual Object Stores stattfindet, die je nach Service Level unterschiedlich konfigurierbare Attribute erlauben und damit eine Anpassung an die Nutzungsszenarios der Anwender ermöglichen. Subsummiert ist die Verwendung des Objektspeichers mit seinen Metadaten ein wichtiger Baustein im Kontext der Digitalisierung, um das Change Management zu initiieren und datengetriebenes Agieren zu ermöglichen.

4.4 Data Lakehouse

Als Unternehmen begannen, große Datenmengen aus verschiedenen Quellen zu sammeln, begannen gleichzeitig Systemarchitekten, sich ein einziges System vorzustellen, um Daten für viele verschiedene Analyseprodukte und Workloads zu beherbergen (hierzu und im Folgenden Lorica et al. 2020). Vor etwa einem Jahrzehnt begannen Unternehmen mit dem Bau von Data-Lake-Repositories für Rohdaten in einer Vielzahl von Formaten. Obwohl sie für die Speicherung von Daten geeignet sind, fehlen Data Lakes einige kritische Funktionen: Sie unterstützen keine Transaktionen, sie erzwingen keine Datenqualität, und ihre mangelnde Konsistenz/Isolation macht es fast unmöglich, Anhänge und Lesevorgänge sowie Batch- und Streaming-Aufträge zu mischen. Aus diesen Gründen haben sich viele der Versprechungen der Data Lakes nicht erfüllt und führen in vielen Fällen zum Verlust der Vorteile von Data Warehouses.

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

Die Notwendigkeit eines flexiblen, leistungsstarken Systems hat aber nicht nachgelassen. Unternehmen benötigen Systeme für verschiedene Datenanwendungen, einschließlich SQL-Analysen, Echtzeitüberwachung, Data Science und Machine Learning. Die meisten der jüngsten Fortschritte in der KI lagen in besseren Modellen, in denen auch unstrukturierte Daten wie Texte, Bilder, Video oder Audio Verarbeitung fanden. Dies sind genau die Datentypen, für die ein Data Warehouse nicht optimiert ist. Ein gängiger Ansatz ist die Verwendung mehrerer Systeme: eines Data Lake, mehrerer Data Warehouses und anderer spezialisierter Systeme wie Streaming, Zeitreihen, Diagramm- und Bilddatenbanken. Eine Vielzahl von Systemen führt zu Komplexität und vor allem zu Verzögerungen, da Datenexperten immer wieder Daten zwischen verschiedenen Systemen verschieben oder kopieren müssen.

Es entstehen neue Systeme, welche die Grenzen von Data Lakes austesten, was zu den Überlegungen eines Data Lakehouse führt. Ein solches Data Lakehouse ist eine offene Architektur, welche die besten Elemente von Data Lakes und Data Warehouses kombiniert. Data Lakehouses werden durch ein neues offenes und standardisiertes Systemdesign ermöglicht: Implementierung ähnlicher Datenstrukturen und Datenmanagementfunktionen wie in einem Data Warehouse, direkt auf der Art eines so genannten low cost Storage, der für Data Lakes verwendet wird. **Ein Data Lakehouse hat die folgenden Hauptmerkmale**:

1. **Transaktionsunterstützung:** In einem Enterprise Lakehouse lesen und schreiben viele Datenpipelines gleichzeitig Daten. Die Unterstützung für ACID-Transaktionen stellt die Konsistenz sicher, da mehrere Anwender Daten gleichzeitig lesen oder schreiben, dies in der Regel mit SQL.
2. **Schemadurchsetzung und -steuerung:** Das Lakehouse besitzt eine Möglichkeit, die Schemadurchsetzung und -entwicklung zu unterstützen und DW-Schemaarchitekturen wie Stern-/Schneeflockenschemas zu unterstützen. Das System sollte in der Lage sein, Datenintegrität sicherzustellen, und es sollte über robuste Governance- und Auditing-Mechanismen verfügen.
3. **BI-Unterstützung:** Data Lakehouses ermöglichen die Verwendung von BI-Tools direkt auf den Quelldaten. Dies reduziert die Verbreitung veralteter Daten und verbessert die wechselseitige Abhängigkeit zwischen Datenbereitstellung und -nutzung, verringert die Latenz und senkt die Kosten für die Operationalisierung zweier Datenkopien sowohl in einem Data Lake als auch in einem Data Warehouse.
4. **Speicher ist vom Rechencomputer entkoppelt:** Speicher und Computer verwenden separate Cluster, sodass sich diese Systeme auf viel mehr gleichzeitige Benutzer und größere Datengrößen skalieren lassen.
5. **Offenheit:** Die verwendeten Speicherformate sind offen und standardisiert, zum Beispiel Parkett, und bieten eine API, sodass eine Vielzahl von Werkzeugen und Engines, einschließlich Machine Learning und Python/R-Bibliotheken, effizient direkt auf die Daten zugreifen zu können.
6. **Unterstützung für verschiedene Datentypen, von unstrukturierten bis hin zu strukturierten Daten:** Es kann zum Speichern, Verfeinern, Analysieren und Zugreifen auf Datentypen verwendet werden, die für viele neue Datenanwendungen benötigt werden, einschließlich Bilder, Video, Audio, halbstrukturierte Daten und Text.
7. **Unterstützung für unterschiedliche Workloads:** Dies gilt auch für Data Science, Machine Learning sowie SQL und Analytics. Möglicherweise sind mehrere Werkzeuge erforderlich, um all diese Aufgaben zu unterstützen, die alle auf demselben Datenrepository basieren.
8. **End-to-End-Streaming:** Echtzeitberichte sind in vielen Unternehmen die Norm. Durch die Unterstützung des Streamings entfallen separate Systeme zur Bereitstellung von Echtzeitdatenanwendungen.

Systeme der Enterprise-Klasse erfordern zusätzliche Funktionen. Werkzeuge für die Sicherheit und Zugriffskontrolle sind grundlegende Anforderungen. Data-Governance-Funktionen, einschließlich Auditing, Aufbewahrung und Abstammung, sind insbesondere im Hinblick auf die jüngsten Datenschutzbestimmungen von entscheidender Bedeutung. Werkzeuge, welche die Datenermittlung ermöglichen, wie Datenkataloge und Datennutzungsmetriken, sind ebenfalls erforderlich. Bei einem Data Lakehouse sind solche Unternehmensfunktionen aber nun in einem System implementiert, getestet und verwaltet, was den gesamten Administrationsaufwand reduziert.

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

4.5 Konsequenzen für eine Data Fabric

Die Datenstruktur im Sinne eines Datenmanagement-Designkonzepts ist eine direkte Antwort auf die langjährigen Herausforderungen bei der Datenintegration, mit denen Daten- und Analyseleiter in einer stark verteilten und vielfältigen Datenlandschaft konfrontiert sind (hierzu und zum Folgenden: Gartner 2019). Zu diesen Herausforderungen gehören:

1. Vielzahl der Datenquellen und -typen
2. Steigende Anzahl von Datensilos
3. Zunehmende Komplexität der Datenintegration
4. Steigende Nachfrage nach Echtzeit- oder ereignisgesteuertem Datenaustausch
5. Steigende Nachfrage nach geschäftsgeführter Datenmodellierung und Schema- und Semantikzuweisung
6. Bedarf an Information und Automatisierung von Teilen der Datenintegration, was schließlich zu einem erweiterten Datenmanagement führt

Herkömmliche Datenintegrationsarchitekturen und -werkzeuge, die sich ausschließlich auf das Replizieren und Verschieben von Daten konzentrierten, liefern oft nur langsam semantisch angereicherte und integrierte Daten, die direkt in Analysen nutzbar oder zur Ausführung bereit sind. Die Herausforderung besteht darin, für diese einen Entwurf zu gestalten, sodass die Bereitstellung immer aktueller Daten im Sinne einer zeitgerechten Integration automatisiert möglich ist. Unternehmen benötigen eine umfassende architektonische und auch datenorientierte Struktur, um ein dynamisches Integrationsdesign aufzubauen, das sich an die sich schnell ändernden Anforderungen eines verteilten Datenökosystems anpassen kann. Das verlangt Wissen über den invertierten Datenprozess – von der Senke, also der aktiven Nutzung, bis zur Quelle, der Datenherkunft. Um dies zu erreichen, müssen Unternehmen aktiv-metadatenangereicherte Wissensdiagramme basierend auf einem Data Lineage verwenden, um ein für die jeweilige Organisation stimmiges Data-Fabric-Design umzusetzen.

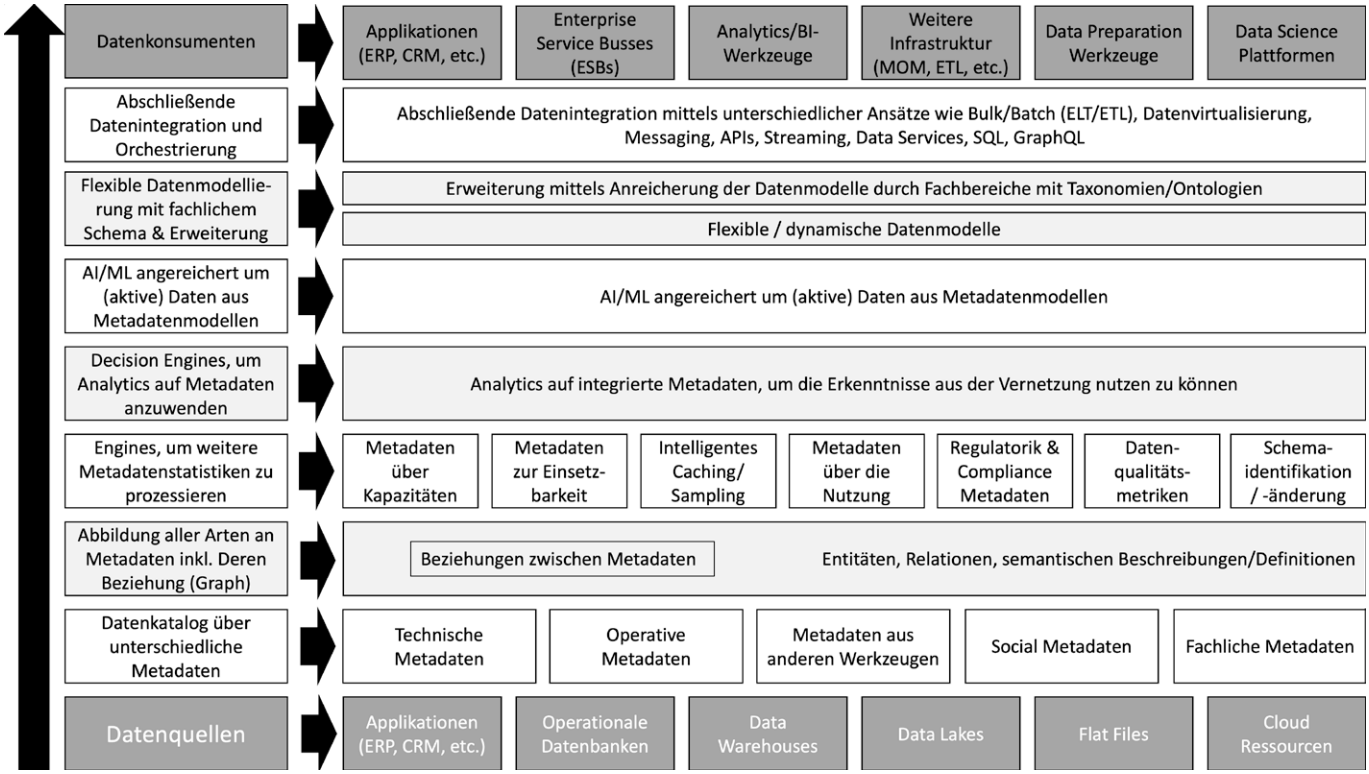

Abbildung 2: *Ein mit aktiven Metadaten und semantischem Wissensdiagramm angereichertes Data-Fabric-Design (entnommen und modifiziert aus: Gartner 2019)*

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

4.5.1 Eine Data Fabric muss die Möglichkeit haben, alle Formen von Metadaten zu sammeln und zu analysieren

Der Übergang zu einem dynamischeren Data-Fabric-Design erfordert grundsätzlich ein umfassendes Verständnis des Kontexts von Daten (hierzu und zum Folgenden: Gartner 2019). Eine Datenintegrationsebene kann den Anspruch eines reibungslosen und automatisierten Integrationsdesigns nicht einlösen, ohne die Möglichkeit zu haben, alle Formen von Metadaten zu identifizieren, zu verbinden und zu analysieren. Solange es keine gut vernetzte Bereitstellung von Metadaten gibt, aus denen sich Rückschlüsse ziehen lassen, wird es der Data Fabric nicht möglich sein, zukünftige Anforderungen oder Änderungen im Datenintegrationsdesign entsprechend zu erkennen und umzusetzen. Data Fabrics sind mit allen Arten von Metadaten anzureichern und nicht nur mit technischen Metadaten (was für sich genommen zwar wichtig, aber nicht ausreichend ist). Organisationen müssen die Analyse aller Metadatentypen unterstützen, einschließlich:

1. Technische Metadaten – Schemata, Datentypen, Datenmodelle, Konfigurationen, Funktionen
2. Betriebs- oder Laufzeitmetadaten – Ausgabe von Prozessen, ETL/andere Aktionen zu Daten, Datenherkunft, Metadaten zur Leistung
3. Geschäftsmetadaten – Ontologie (Klassifizierung und Kennzeichnung von Daten), Geschäftsbeziehungen zugeordnete Metadaten, Taxonomie, Richtlinien
4. Soziale Metadaten – beschreibende Fakten zu Datenbeziehungen, nutzergenerierten Inhalten, Bereichswissen, Bewertungen/Kommentaren, Zertifizierung.

4.5.2 Eine Data Fabric muss die Möglichkeit haben, passive Metadaten zu analysieren und in aktive Metadaten zu konvertieren

Um die Automatisierung bei der Entwicklung und Bereitstellung von Integrationen zu unterstützen, müssen Unternehmen, die eine reibungslose gemeinsame Nutzung von Daten anstreben, über Metadatenfunktionen verfügen, die über die passiven Metadatenpraktiken hinausgehen (hierzu und zum Folgenden: Gartner 2019). Passive Metadaten sind statische Metadaten. Diese entstehen in der Regel zur Entwurfszeit und erfordern oft menschliche oder manuelle Aktualisierungen für neue Modellversionen. Passive Metadaten bestehen meist aus einer einfachen Dokumentation oder technischen Metadaten zur Entwurfszeit. Das Aktivieren passiver Metadaten erfordert, dass die Data Fabric folgende Möglichkeiten hat:

1. Entdecken und Verbinden aller Formen von Metadaten sowie deren Verknüpfung. Diese basiert auf den einzigartigen und sich ändernden Beziehungen, um eine leicht verständliche Graphdarstellung von Metadaten zu erzeugen.
2. Schaffung eines kontinuierlichen Zugriffs auf das erstellte Metadatendiagramm und Ausführung von Analysen auf diesem Metadatendiagramm, um wichtige Metriken und Statistiken zu erzeugen. Beispielsweise lassen sich Metriken durch die Analyse von Metadaten zu Parametern wie Zugriffshäufigkeit, Datenherkunft, Leistungsoptimierung und Datenqualität (erreicht aus anderen Datenqualitäts-/Daten-Governance-/Informationsverwaltungslösungen) erhalten.
3. Verwendung wichtiger Metriken und Statistiken, um AI/ML-Algorithmen anzureichern, die letztendlich Empfehlungen zur Information und Automatisierung von Teilen des Entwurfs und der Bereitstellung von Datenintegration enthalten. Dieser gesamte Prozess hilft bei der genannten Aktivierung passiver Metadaten, um die dynamische und zunehmend automatisierte Datenintegration voranzutreiben.

Die Data Fabric nutzt die aktiven Metadaten, um diese in angereicherten AI/ML-Algorithmen zu verwenden. Dadurch lässt sich im Laufe der Zeit Wissen ansammeln, um genauere Vorhersagen und Entscheidungen in Bezug auf Schlüsselaspekte der Datenverwaltung und -integration zu treffen. Zu nennen sind beispielsweise:

1. Dynamische Schemaerkennung und Korrektur von Schemaveränderungen
2. Automatische Integration nächstbester Datenquellen, empfehlungsoptimaler Transformationen und automatische Ausführung repetitiver Transformationen
3. Förderung von Self-Service-Integrationsflüssen (erstellt von Mitarbeitern im Fachbereich) in Produktionsumgebungen
4. Automatische Konvertierung virtueller Ansichten von Daten bei Bedarf in physische Ansichten (vice versa)

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

5. Erstellung eines überlegenen Abfrageausführungsplans
6. Automation der Verteilung und Ausführung von Integrationsworkloads über die optimale Infrastruktur hinweg

Für Architekten und Lösungsdesigner sind diese Rückkopplungen zum automatisierten Integrationsdesign hilfreich. Dies unterstützt sie dabei, sich auf Design- und strategische Datenmanagementinitiativen zu konzentrieren, anstatt sich in wiederholenden Datenintegrationstransformationen, Pipelinebereitstellungen und -ausführungen zu verlieren.

4.5.3 Data Fabric muss die Möglichkeit haben, ein Wissensdiagramm zu erstellen, welches das Data-Fabric-Design operationalisieren kann

Die Nutzung der aktivierten Metadaten ist relevant, um die Erstellung flexibler Datenmodelle voranzutreiben und diese dabei mit Semantik anzureichern, sodass sie dann für verschiedene Datenkonsumenten und Anwendungen bereitgestellt werden können (hierzu und zum Folgenden: Gartner 2019). Dies wird durch die Erstellung von Wissensdiagrammen unterstützt, die einheitliche Metadaten nutzen, die mit kontextbezogener und semantischer Relevanz über Datensilos angereichert werden.

Wissensdiagramme kombinieren die Funktionen von Datenspeichern mit einem wissensbasierten System für die Datenvereinheitlichung und -automatisierung, die ideal für die Darstellung und Beibehaltung verbundener Metadatenmodelle sind (normalerweise werden Diagrammdatenspeicher für diese Anforderungen genutzt). Zusammen bieten diese einen ganzheitlichen Überblick über die Daten in einem Unternehmen. Wissensdiagramme können Verbindungen durchlaufen, Beziehungen zwischen Knoten ermitteln und ein Netz von n:n-Beziehungen aufdecken. Sie betten Wissensressourcen ein, die es Mitarbeitern aus den Fachbereichen ermöglichen, Datenmodelle mit umfassender Semantik und Kontext hinzuzufügen. Dies unterstützt auch ein automatisiertes Maschinenverständnis und eine Erklärbarkeit durch eingebettete AI/ML-Engines. Wissensdiagramme müssen Abfragen durch entsprechende Abfragesprachen auf den zugrunde liegenden Daten- und Datenmodellen unterstützen und Ebenen der Maschinenverständlichkeit hinzufügen, indem eine umfangreichere Semantik in den Datenmodellen eingebettet wird.

Data Fabrics, die durch ein Wissensdiagramm unterstützt werden, bewegen sich in Richtung eines vollständig metadatengesteuerten Data-Fabric-Designs. Dabei sollten Metadaten in Bezug auf Folgendes beibehalten werden:

1. Rate des Datenzugriffs
2. Plattformen, die auf die Daten zugreifen
3. Benutzer, die auf die Daten zugreifen
4. Die physische Kapazität und Nutzung von Infrastrukturkomponenten
5. Anwendungsfall, Anreicherung und Integrität in den Daten

Diese Metadaten-Metriken sollten in den AI/ML-Algorithmen als Teil des Wissens verfügbar sein, die dann viel dynamischere und genauere Designs bereitstellen, um die Datenintegration und -bereitstellung zu erweitern und zu automatisieren.

4 Data Warehouse und Data Lakehouse – Basis der Data Fabric

4.5.4 Data Fabric muss es Mitarbeitern im Fachbereich ermöglichen, Datenmodelle mit Semantik anzureichern

Die semantische Ebene des Wissensdiagramms unterstützt bei Aufgaben wie Abfragen für Analysen oder dem Erstellen einer semantischen virtuellen Ebene, um Daten mit anderen Werkzeugen für Data Science, zur Assoziationsanalyse oder einfach dem Data Retrieval zu teilen (hierzu und zum Folgenden: Gartner 2019). Wenn sich das Unternehmen weiterentwickelt und immer mehr Datenquellen und Anwendungsfälle anzubinden sind, absorbieren Wissensdiagramme kontinuierlich die neuen Metadaten ohne negative Folgen für die Administration und Verfügbarkeit.

Es ist wichtig, dass die Datenstruktur es den Fachbereichsmitarbeitern ermöglicht, beim Erstellen von Datenmodellen zu helfen und diese im Weiteren durch Hinzufügen von Semantik zu ergänzen. Die vom Wissensdiagramm bereitgestellten Datenmodelle (innerhalb einer Datenstruktur) veranschaulichen die Struktur der Datenbeziehungen anhand semantischer Standards, zum Beispiel Ontologien oder Resource Description Framework (RDF) Triples, insofern Diagrammdatenbanken von der Datenstruktur verwendet werden. Die semantische Ebene ist eine Ebene von Metadaten, die dem Diagramm insgesamt Tiefe und Bedeutung hinzufügt, sodass Abfragen und Algorithmen diese Informationen für Analysen und andere betriebliche Anwendungsfälle verwenden können. Das Ergebnis ist ein nützliches und selbsterklärendes Wissensdiagramm, das es den Mitarbeitern in Fachbereichen ermöglicht, flexible Datenmodelle zu erstellen. Diese können dann Änderungen der Anforderungen gerecht werden und werden mit Semantik und Bedeutung für den fachlichen Einsatz angereichert.

4.5.5 Empfehlungen für Datenanalyseverantwortliche

Da das Thema der Metadaten eine bereits hohe, aber eine weiter zunehmende Bedeutung hat, sind Data Catalogs ein wichtiges Thema. ML-Algorithmen nutzen verschiedene Arten von Metadaten zusammen mit den zugehörigen Beziehungen in einem geeigneten Modell (hierzu und zum Folgenden: Gartner 2019). Dies ist im Wesentlichen der erste Schritt beim Erstellen einer Data Fabric. Einige Data Catalogs modellieren automatisch Enterprise-Metadaten-Assets in einer Ontologie. Neue Datenquellen lassen sich so schnell und entsprechend unkompliziert in das Gesamtgefüge integrieren. Die Veranschaulichung dieser verbundenen Metadaten in einem Diagramm erleichtert es den Mitarbeitern in Fachbereichen, einen Beitrag zur Datenmodellierung inklusive angereicherter Semantik zu leisten. Da Datenquellen immer vielfältiger und komplexer werden, erhalten Data Scientists eine Unterstützung, wertvolle Datenressourcen mit entsprechender Performance umfänglich zu nutzen.

Wichtig erscheint ein einfacher Zugriff auf die bereitgestellten Daten eines Wissensdiagramms. Die darin enthaltene Datenstruktur ist so abzubilden, dass von Data Engineers und anderen Datenintegrationsexperten bekannte Integrationsstandards und werkzeuge verwendbar sind. Parallel müssen angemessene Fähigkeiten und Schulungen geschaffen werden, um das Auftreten unbekannter Standards und Abfragetechniken zu verhindern oder zumindest frühzeitig aufzudecken und so das Risiko des Auftretens von Shadow IT zu reduzieren. Dies beinhaltet auch das Auflösen komplexer Definitionen von Security- und Governance-Regeln.

5 TDWI Research: Sechs Fähigkeiten einer logischen Data Fabric

In den folgenden Abschnitten werden sechs wichtige Funktionen der logischen Datenstruktur beschrieben, die sich mit modernen Datenmanagement- und Analysebemühungen befassen, die mehrere Plattformen, neue Datentypen und -quellen sowie erweiterte Analysen beinhalten (hierzu und zum Folgenden: TDWI 2020). Zu diesen Funktionen gehören die Aktivierung von Datenanalysten in einer aktuell diskutierten hybriden Multicloud-Datenlandschaft und Techniken für die nahtlose Datenintegration über Multicloud-Plattformen. Zu nennen sind außerdem eine erweiterte Analysefähigkeit, das Datenbewusstsein sowie die nahtlose Zugänglichkeit bei gleichzeitiger Aufrechterhaltung einer hohen betrieblichen Leistung und die Leistungssteigerung durch Optimierungen der Latenzen im Zusammenhang mit der Datennutzung. Darüber hinaus wird im Folgenden erläutert, wie Data-Science-Initiativen durch unterstützt werden können Datenermittlung und Analysen, die statische und dynamische Daten kombinieren. Abschließend wird auf das Datenbewusstsein fokussiert und die Verwendung einer Data Fabric als Unternehmensdatenkatalog betrachtet.

5.1 Integration von Daten in Multicloud-Umgebungen

In Anerkennung der praktischen Vorteile einer Cloud-Migration und der potenziellen wirtschaftlichen Vorteile wechseln viele Unternehmen in die Cloud, um das Datenmanagement sowie die Umgestaltung und Modernisierung ihrer Reporting- und Analyseplattformen und Analyseanwendungen zu adressieren (hierzu und zum Folgenden: TDWI 2020). Eine TDWI-Umfrage in den USA zu Daten und Analysen im Jahr 2020 zeigt eine wachsende Akzeptanz: 46 % der Befragten gaben an, dass sie die Cloud bereits nutzen, und weitere 34 % gaben Pläne zu deren Nutzung an.

In zunehmendem Maße erkennen Unternehmen jedoch Vorteile einer Strategie auf Basis einer zweistufigen Multicloud-Architektur. In dieser Struktur werden mehrere virtuelle Cloud-Umgebungen mit demselben Cloud-Anbieter eingerichtet (zum Beispiel virtuelle Plattformen, die in verschiedenen geografischen Regionen eingerichtet sind) oder sogar potenziell über verschiedene Cloud-Anbieter hinweg. Abgesehen von den niedrigeren Kosten ermöglicht diese Multicloud-Strategie den Benutzergemeinschaften, eine breite Palette an Funktionen und Diensten für Datenmanagement und -analyse zu nutzen und gleichzeitig Einschränkungen bei der Sperrung von Anbietern zu vermeiden. Dieser Ansatz ermöglicht hybride Lösungen, die mehrere Cloud-Plattformen umfassen und den optimalen Entwurf unterstützen, um die Anwendungs-, Analyse- und Datenmanagementanforderungen eines Unternehmens zu erfüllen. TDWI USA sieht derzeit, dass Unternehmen einen oder zwei Cloud-Anbieter für ihre Datenverwaltungsanforderungen verwenden. Dabei ist jedoch zu erwarten, dass die Anzahl der Organisationen, die zwei oder mehr Cloud-Anbieter verwenden, zunehmen wird.

Allerdings widerspräche die Wahl einer Multicloud-Umgebung dem Gedanken der vereinfachten Datenbereitstellung, da so zusätzliche Komplexität für Datennutzer entsteht. Die Multicloud-Datenmanagement-Strategie eines Unternehmens bedarf einer logischen Datenstruktur, die transparent auf Daten aus verschiedenen cloudbasierten Quellen zugreift, diese integriert und Daten aus verschiedenen Cloud-Umgebungen in einer logischen Ansicht aggregieren kann. Dabei helfen Werkzeuge, die grundlegende Datenvirtualisierungstechniken einsetzen, in denen separate Instanzen in jeder Clouddomäne (beispielsweise AWS, Azure oder GCP) platziert werden. Diese Werkzeuge können auf die Daten innerhalb dieses Cloudanbieters zugreifen und diese aggregieren. Diese Instanzen können wiederum eine Verbindung zu einer Schicht herstellen, die den Zugriff koordiniert und die Daten aus den verschiedenen Clouds aggregiert, um eine ganzheitliche Datenansicht im gesamten Hybridunternehmen zur einheitlichen Analyse bereitzustellen.

5 TDWI Research: Sechs Fähigkeiten einer logischen Data Fabric

5.2 Automation manueller Aufgaben mit ergänzender Analysefähigkeit

Die Annahme einer Multicloud-Strategie beeinflusst die Plattformauswahl und den Umgebungsentwurf, um cloudbasierte Services zu nutzen. Diese tragen dazu bei, die Datenreplikation zu reduzieren. Zum Teil können sie zudem die Berechnung auf die Daten übertragen, um Knoten zu berechnen, anstatt dafür Daten zu ziehen. Davon abgesehen gibt es drei Herausforderungen, denen sich jede logische Datenstruktur stellen muss, um sicherzustellen, dass die Erwartungen der nachgelagerten Datenverbraucher erfüllt werden:

1. **Datenbewusstsein und -verfügbarkeit:** Die schnelle Einführung von Data-Lake-Architekturen, die sich über mehrere Plattformen erstrecken, wird zu einem höheren Beitrag von Datenbeständen in den Data Lake sowie zu einer zunehmenden Integration von Datenquellen von außerhalb des Unternehmens inspirieren. Data Scientists und Analysten sind sich möglicherweise nicht aller verfügbaren Datenquellen bewusst, ganz zu schweigen von den Quellen, die ihren Analyseanforderungen am besten entsprechen.
2. **Nahtlose Durchgängigkeit:** Die zunehmende Anzahl externer Datenquellen stellt ein Risiko dar, das sich aus der Dynamik der Datenbereitsteller ergibt. Datenobjektmodelle und deren Layouts unterliegen ungeplanten oder unangekündigten Änderungen. Eine Enterprise Data Fabric muss diese Änderungen berücksichtigen, ohne dass eine übermäßige Entwicklerbelastung entsteht.
3. **Sicherstellung der operativen Leistungsfähigkeit:** In allen Fällen kann der Zugriff auf Daten von verschiedenen Plattformen mit unterschiedlichen Leistungsmerkmalen und Datenlatenzen zu Engpässen führen, welche die Leistung beeinträchtigen.

Eine logische Datenstruktur auf Unternehmensstufe muss diese Probleme, die traditionell manuell behandelt wurden, automatisch beheben. Fortschrittliche Analysetechniken, die Automatisierung unterstützen, beispielsweise maschinelles Lernen, werden schnell in das Unternehmen integriert. In einer kürzlich durchgeführten TDWI-Umfrage zu Daten und Analysen gaben 71 % der Befragten an, dass die Nachfrage nach maschinellem Lernen steigt. Das deutet auf den wachsenden Wunsch hin, die Vorteile der erweiterten Intelligenz zu nutzen. Machine-Learning-Techniken lassen sich in eine logische Datenstruktur integrieren, um das Bewusstsein für Datenquellenstrukturen zu erhalten, Änderungen zu überwachen und die logischen, das heißt virtualisierten, Modelle automatisch anzupassen und die Methoden für den Zugriff auf und die Bereitstellung von Daten automatisch zu optimieren.

Daraus lässt sich die Anforderung einer logischen Datenstruktur ableiten, die erweiterte Analysen wie maschinelles Lernen enthält, um das wachsende Volumen und die Komplexität von Daten durch die Automatisierung manueller Aufgaben bewältigt. Eine logische Datenstruktur kann maschinelles Lernen verwenden, um Datennutzungsmuster der Anwender zu analysieren und automatisch Datensätze für neue Analysen zu empfehlen. Darüber hinaus kann die Überwachung von Datennutzungs- und Zugriffsmustern in der Multicloud-Umgebung die Data Fabric informieren, Datenzugriffsanforderungen zu antizipieren und einen intelligenten Caching-Mechanismus zu verwenden, um erwartete Service-Level-Vereinbarungen für Performance-Service zu erfüllen.

5.3 Steigerung der Analytics-Performance durch eine schnelle Datenbereitstellung

TDWI-Untersuchungen zeigen, dass 80 % der Befragten es wichtig finden, Lösungen, Cloud-Services und Cloud-Praktiken zu haben, um schnellere Analysen zu ermöglichen (hierzu und zum Folgenden: TDWI 2020). 39 % gaben an, dies sei extrem wichtig. Gleichzeitig sagen 77 % der Unternehmen, dass Near-Realtime oder echte Realtime-Daten, BI-Dashboards und Analysen für den Erfolg ihres Unternehmens wichtig sind. Darunter sagen 30 %, dass sie sehr wichtig sind. Da Unternehmensdatensätze über eine immer vielfältigere verteilte Konfiguration verstreut sind, muss die logische Datenstruktur die Optimierung über die gesamte Datenlandschaft hinweg unterstützen. Dadurch wird die Datenlatenz nach Möglichkeit reduziert und Datenzugriff, -aggregation und -bereitstellung werden vorhersehbar schnell ermöglicht.

5 TDWI Research: Sechs Fähigkeiten einer logischen Data Fabric

Dies wird noch wichtiger, wenn Unternehmen versuchen, Live-Streaming-Daten in Echtzeitberichte und entsprechende Analysen zu integrieren. Die Daten stammen aus mehreren Quellen und liegen in verschiedenen Formaten, Latenzen und Strukturen vor. Also wird eine logische Datenstruktur benötigt, die nicht nur diese Unterschiede berücksichtigt, sondern auch Optimierungstechniken nutzen kann, um den Datenzugriff über die hybride Konfiguration hinweg anzupassen. Zu diesen Optimierungen gehören:

1. **Pushdown Optimierung:** Diese Optimierung basiert auf den zugrunde liegenden Systemen, um Komponenten der Abfrage auszuführen. Beispielsweise kann die Aggregation auf die zugrunde liegenden cloud-gehosteten Datenbanksysteme übertragen werden und die Teilergebnisse lassen sich innerhalb der Datenstruktur kombinieren und dem Benutzer präsentieren.
2. **Caching:** Dieser Ansatz zielt darauf ab, die Unterschiede zwischen Datenlatenzen aus verschiedenen Quellen auszugleichen, indem lokale Kopien von Datensätzen und Ergebnissen gespeichert werden, auf die häufig zugegriffen wird. Dies geschieht unter Verwendung der leistungsstärksten Plattform, um die Abfrageausführung zu beschleunigen.
3. **Datenbewegung:** Wenn Komponenten einer JOIN-Abfrage über verschiedene Cloudquellen hinweg verbunden werden müssen, ist die Anzahl der Datensätze in einer der Quellen deutlich größer als die Anzahl in den anderen Quellen. Ein naiver Föderationsansatz würde effektiv beide Datenquellen abrufen und versuchen, die JOIN innerhalb des temporären Speicherplatzes des Verbundtools auszuführen. Wenn die Datenstruktur jedoch den Unterschied im Datenvolumen erkennen kann, kann der kleinere Datensatz in eine temporäre Tabelle mit der größeren Datenquelle übertragen werden. Der JOIN kann lokal ausgeführt werden und das Ergebnis lässt sich über die Datenvirtualisierungsschicht an den Analysten weiterleiten.

Eine unternehmensweite logische Datenstruktur, die dynamische Abfrageoptimierungstechniken mit Unterstützung für die Nutzung massiv paralleler Verarbeitungsmodule mithilfe von In-Memory-Datenmanagement zusammen mit der Verwendung von Zusammenfassungstabellen kombiniert, steigert die Leistung und beschleunigt die Bereitstellung von Abfrageergebnissen.

5.4 Unterstützung der Datenentdeckung und Datenauswertung sowie von Data-Science-Aktivitäten

Aktuelle wissenschaftliche Untersuchungen zeigen, dass Organisationen die Anwendung der Data Science vertiefen. Eine TDWI-Umfrage ergab, dass 68 % der befragten Organisationen bereits Data Scientists eingestellt haben. 22 % gaben an, eine Einstellung zu planen, um ihre Analyseinitiativen voranzubringen (hierzu und zum Folgenden: TDWI 2020). Entsprechend haben Unternehmen eine starke Nachfrage nach Technologien, die Data Science und Advanced Analytics unterstützen. Vier grundlegende Datenmanagement-Funktionen sind erforderlich, um Data Scientists und die fortschrittliche Analyse und Strategie des Unternehmens zu unterstützen:

1. **Organisationales Datenbewusstsein:** Die iterative Natur des Erstellens und kontinuierlichen Testens und Verfeinerns von Analysemodellen bedeutet, dass Data Scientists wissen möchten, welche Datenressourcen zur Verfügung stehen, die in ihren Prozess integriert werden können.
2. **Demokratisierung der Datenverfügbarkeit:** Die Folge des Datenbewusstseins ist die Datendemokratisierung, bei der Datenkonsumenten mit entsprechenden Berechtigungen Methoden zur Self-Service-Konfiguration des Zugriffs auf verfügbare Datenbestände zur Verfügung gestellt werden.
3. **Transparente Zugriffsbereitschaft:** Um die Komplexität zu überwinden, die verschiedenen Methoden des Zugriffs auf eine Vielzahl von Datensätzen in einer breiten Hybridumgebung zu verfolgen, muss es eine Möglichkeit zur Bereitstellung einer einheitlichen Ebene für den Zugriff geben.

5 TDWI Research: Sechs Fähigkeiten einer logischen Data Fabric

4. **Flexibilität des Datenmodells:** Datenwissenschaftler leiten aus der Kombination und Integration von Informationen aus einer Vielzahl von Quellen einen Wert ab. Diese Analysten profitieren von der Möglichkeit, alternative Datenmodellansichten über verschiedene Datensätze aufzuerlegen, um die Einschränkungen der Quelle zu lindern.

Die logische Datenstruktur unterstützt Data Science, indem sie diese grundlegenden Funktionen bereitstellt. Eine Data Fabric erlaubt den Zugriff auf die gesamte Unternehmensdatenlandschaft und ermöglicht die Lieferung aller Unternehmensdatensätze an die verschiedenen Data-Science-Projekte über eine Vielzahl von Kanälen, zum Beispiel über ein BI-Frontend, über APIs etc.

Die nahtlose Transparenz der Datenstruktur und die Möglichkeit, logische Modelle über Quelldaten aufzuerlegen, gestatten verschiedenen Data Scientists, dieselben Quelldatensätze in ihren eigenen Anwendungskontexten zu nutzen. Es ermöglicht interaktive Entdeckung, um dem Analytiker zu helfen, die zu untersuchende Anfragemenge zu verfeinern. Dadurch wird die iterative Natur der Entwicklung und Erstellung mehrerer Analysemodelle und deren Veröffentlichung für die Produktion für andere Data Scientists unterstützt.

5.5 Analyse ruhender und dynamischer Daten

Ein wesentlicher Teil der Anwendung der Business Intelligence und Analytics umfasst ruhende Daten. Dies sind entweder Daten, die aus betrieblichen oder Transaktionsverarbeitungssystemen extrahiert wurden, oder Datensätze, die aus anderen Quellen stammen und in der Speicherumgebung der Organisation gespeichert sind (hierzu und zum Folgenden: TDWI 2020). Die wachsende Migrationswelle in Cloud-Umgebungen verringert die Hindernisse zur Integration einer viel größeren Vielfalt statischer, dynamischer und Streaming-Datenquellen aus zwei Gründen:

- Erstens gibt es praktisch unbegrenzte Speicherkapazität, sodass Fenster von Streaming-Daten erfasst und in Reporting- und Analyseanwendungen integrierbar sind.
- Zweitens unterstützt die elastische Rechenleistung, die in einer Cloud-Computing-Umgebung bereitgestellt wird, die Verarbeitung zahlreicher Datenströme in Echtzeit.

Etwa 20 % der Befragten der TDWI-Umfragen sagen, dass sie derzeit Maschinendaten verwenden. Dies wird in Zukunft zunehmen, wenn die Nutzer an ihren Plänen festhalten. Da Unternehmen beginnen, verschiedene Arten von Social-Media-Daten, gestreamte Maschinendaten (zum Beispiel Sensorinformationen) und andere Arten kontinuierlich fließender Informationen, zum Beispiel Nachrichten oder Wetterfeeds, in ihre Datenverwaltungs- und Analysebemühungen einzubeziehen, passen Datenanalysten herkömmliche Daten im Ruhezustand schnell an.

Die Kombination ruhender und dynamischer Daten wird für integrierte Analysen genutzt. Beispielsweise kombinieren IoT-Anwendungen (Internet of Things) gesammelte historische Daten aus Streaming- und ruhenden Datenquellen, um Analysemodelle zu entwickeln, welche die Entscheidungsfindung beeinflussen. Das Einbetten dieser Analysemodelle an verschiedenen Punkten der Unternehmensdatenpipelines reduziert manuelle Eingriffe und optimiert die vertrauenswürdige automatisierte Entscheidungsfindung.

Zu gestalten ist eine logischen Datenstruktur, welche die Integration und Verwendung von Daten in Bewegung mit ruhenden Daten unterstützen kann. Zu bewerten ist beispielsweise, wie die Data Fabric die Verwendung von Datenstreaming-Tools, zum Beispiel Apache Kafka, mit den strukturierten Daten vereinfachen kann, die sich im Data Warehouse befinden, um eine integrierte Analyse sowohl der ruhenden Daten als auch der Echtzeit-Streaming-Daten zu unterstützen.

5 TDWI Research: Sechs Fähigkeiten einer logischen Data Fabric

5.6 Katalogisierung aller Daten zur Discovery, Lineage und Verknüpfung

Es ist wichtig, die Vorteile des Datenbewusstseins bei der Unterstützung der Data Science zu stärken und die Prozesse der Entwicklung und Veröffentlichung von Analysen zu vereinfachen (hierzu und zum Folgenden: TDWI 2020). Ein wachsender Bestand an Datensätzen, die aus einer Vielzahl von Datenquellen stammen, wirft zwangsläufig Fragen zur Semantik und Konsistenz der genutzten Definitionen auf. Beispielsweise möchte ein Analyst das Kundenverhalten verstehen, zum Beispiel, welche Produkte Kunden haben, über welche Kanäle sie gekauft haben, welche Kaufneigung eines Kunden besteht, welche Arten von Garantien gewünscht werden usw. Wenn jedoch Informationen aus einer Vielzahl von Datensätzen angesammelt werden, gibt es in der Organisation zwangsläufig unterschiedliche Definitionen des Kunden. Das bedeutet, dass Geschäftsbegriffe dokumentiert und diesen Metadaten den entsprechenden Datenelementen zuzuordnen sind, sodass die Beziehungen zwischen den verschiedenen logischen Begriffen zu verstehen sind.

Da die logische Datenstruktur alle Daten im Unternehmen zusammenstellt, wird sie zur zentralen Quelle für das Verständnis des Wissens über Daten in der gesamten Organisation. Damit fungiert sie als zentraler Katalog, um den Speicherort, die verwendeten Datentypen und das Format der Daten zusammen mit den Zuordnungen eines Datensatzes mit einem anderen zu dokumentieren. Mithilfe dieses Data Catalog können Analysten die Datenermittlung an einem zentralen Ort durchführen, anstatt separate BI-Werkzeuge verwenden zu müssen, die je nach den Quellsystemen, mit denen sie verbunden sind, nur eine partielle Ansicht der Daten unterstützen.

Mithilfe einfacher googleähnlicher Suchvorgänge sollten die Analysten nicht nur die Daten, sondern auch ihre Beziehung zu verwandten Daten anzeigen können. Zu nennen ist die Bereitstellung von Transparenz für die Datensätze, mit denen ermittelt werden kann, welche Kunden welche Produkte gekauft haben. Eine Data Fabric, die eine intelligente Suche anbietet, kann einen Mehrwert schaffen, indem sie das Verständnis und die Dokumentation der Daten für Ermittlungen, Governance und eine bessere Unterstützung für fachliche Anwender verbessert.

6 Fazit

Für alle Unternehmen kommt es darauf an, in Bezug auf die eigenen Daten für Interoperabilität zu sorgen. Nur so ist es möglich, im Sinne der Digitalisierung alle verfügbaren Daten auf einmal zu überblicken, miteinander zu verknüpfen und entsprechend auszuwerten. Eine Data Fabric dient dazu, die digitale Transformation von und in Unternehmen voranzutreiben. Ein Data Catalog enthält Informationen über alle angebundenen Datenquellen. Die vorhandenen Metadaten kategorisieren, beschreiben und erleichtern das Auffinden und Verstehen verfügbarer Daten. Eine solche Datenplattform schafft in Unternehmen eine Struktur aus Verbindungen zwischen Datenquellen. Mit ihr gehören Datensilos und damit Informationslücken der Vergangenheit an. Wichtig ist zudem, dass diese modernen Technologien sowohl lokal als auch in der Cloud zu betreiben sind und sich damit problemlos in bestehende IT-Infrastrukturen einbinden lassen.

Obwohl Unternehmen aktuell die Migration in die Cloud nutzen, verringert sich das Erfolgspotenzial ohne einen geregelten Ansatz zur Verwaltung des Datenbewusstseins und zur Vereinfachung der Datenzugriffsfähigkeit. Eine Organisation kann versuchen, eine Reihe von Werkzeugen zusammenzufügen, die auf einer Vielzahl von Cloud-Hosts und entsprechenden Konfigurationen geschichtet sind. Aber wenn sich die Hybridumgebung erweitert, wird dieser Ansatz nicht einfach skaliert, sondern wird schnell herausfordernd zu verwalten und zu warten.

Sinnhaft erscheint die Data Fabric als logische Datenstruktur, die den Datenzugriff in Multicloud-Umgebungen vereinfacht und die Leistung optimiert. Eine logische Datenstruktur unterstützt dabei, moderne Datenmanagement- und Analysebemühungen zu bewältigen, die mehrere Plattformen, neue Datentypen und Datenquellen sowie erweiterte Analysen umfassen. Die Integration eines logischen Datengefüges, das die sechs aufgezählten kritischen Funktionen umfasst, wird Datenanalysten und Data Scientists stärken und die Entwicklung und Bereitstellung von Reporting, Business Intelligence und integrierten erweiterten Analysen optimieren.

Die Implementierung einer Data Fabric unterstützen Unternehmen die Erfassung, Governance, Integration, Steuerung sowie den Austausch von Daten innerhalb der Organisation. Eine Data-Fabric-Architektur ist keine singuläre Lösung für ein bestimmtes Problem der Datenintegration oder Datenverwaltung, sondern stellt eine dauerhafte und skalierbare Lösung dar, mit der sich alle Daten in einer einheitlichen Umgebung verwalten lassen.

Letztendlich kann die Implementierung einer Data Fabric einem Unternehmen dabei helfen, die Herausforderungen im Zusammenhang mit Datenmanagement zu bewältigen und gegenüber der Konkurrenz einen Wettbewerbsvorteil zu erhalten. Zusammenfassend ergeben sich die nachstehenden Vorteile einer Data Fabric (Talend 2021):

1. **Bereitstellung einer einzigen Umgebung,** über die alle Daten erfasst und zugänglich sind. Dies unabhängig davon, wo und wie diese gespeichert sind. Mit einer Data Fabric lassen sich nachhaltig Datensilos beseitigen und eine Optimierung verteilter Queries durch einen Cost-based-Optimizer realisieren.
2. **Einfacheres und einheitliches Datenmanagement** inklusive Verbesserung der Datenqualität, Datenintegration, Data Governance sowie Datenfreigabe und Datenaustausch. Der Einsatz mehrerer Werkzeuge erübrigt sich und Unternehmen können schneller auf vertrauenswürdigere Daten zugreifen.
3. **Größere Skalierbarkeit,** wodurch die Anpassung an wachsende Datenmengen, Datenquellen und Anwendungen gewährleistet ist.
4. **Einfachere Nutzung der Cloud** durch Unterstützung von On-Premises-, Hybrid- und Multi-Cloud-Umgebungen und schnellere Migration zwischen diesen Umgebungen.
5. **Geringere Abhängigkeit** von früher geschaffenen und immer noch vorhandenen Infrastrukturen und Lösungen.
6. **Zukunftssicherheit der Datenmanagementinfrastruktur,** da sich neue Datenquellen und Endpunkte sowie neue Technologien zur Data-Fabric-Architektur hinzufügen lassen, ohne bestehende Verbindungen oder Bereitstellungen zu unterbrechen beziehungsweise zu stören. Das schafft eine sogenannte source-agnostische Datenarchitektur zu Gunsten einer Zukunfts- und Investitionssicherheit.

6 Fazit

Im Sinne einer Datenvirtualisierung bietet eine Data-Fabric-Infrastruktur das Schlüsselelement der Digitalisierung. Das aktuelle Thema des Self Service ist ein initiales Nutzungsszenario, das die Notwendigkeit, aber auch den erreichbaren Nutzen aufzeigen kann. Zukünftig wird das Thema Data Governance insbesondere im Kontext des Risk Management effizient und nutzenstiftend über die Data Fabric implementiert werden. Dies als Baustein einer nahtlosen Hybrid- und Multi-Cloud Integration beim Aufbau einer AI/ML-ready Dateninfrastruktur. Damit wird die Data Fabric zum Enabler für Advanced Analytics für eine einfache Bereitstellung von Daten für Self Services unter Berücksichtigung aller notwendigen Security- und Governance-Regeln. Dies schafft eine Unterstützung verschiedener User-Gruppen: Power-User können Daten erforschen und neue Modelle erstellen, Business-User können auf vertrauensvolle Daten zugreifen und Data Scientists werden bei der Suche nach relevanten Daten durch einen Data Catalog unterstützt, sodass effizientes Arbeiten möglich ist.

Literatur

BITKOM-Arbeitskreis Big Data (2015): Big Data und Geschäftsmodell-Innovationen in der Praxis: 40+ Beispiele. BITKOM, Berlin.

Convertino, G.; Echenique, A. (2017): Self-Service Data Preparation and Analysis by Business Users: New Needs, Skills, and Tools. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17). ACM, New York, NY, USA, S. 1075–1083.

Davenport, T. H. (2014): Big Data @ work, Harvard Business School Publishing Corporation. Cambridge.

Dittmar, C.; Felden, C.; Finger, R.; Scheuch, R.; Tams, L. [Hrsg.] (2016): Big Data – Ein Überblick. dpunkt-Verlag, Heidelberg.

Dorschel, J. (2015): Praxishandbuch Big Data – Wirtschaft-Recht-Technik. Springer, Wiesbaden.

Gartner (2019): Zaidi, E.; Thoo, E.: Data Fabric Add Augmented Intelligence to Modernize Your Data Integration. https://www.gartner.com/doc/reprints?id=1-251MKCBH&ct=210119&st=sb&utm_campaign=TY. Abruf am: 26.05.2021.

Litzel, N. (2018): Was ist ein Data Lake?, https://www.bigdata-insider.de/was-ist-ein-data-lake-a-686778/, Abruf am 16.03.2021.

Lorica, B.; Armbrust, M.; Ghodsi, A.; Xin, R.; Zaharia, M. (2020): Was ist ein Lakehouse? https://databricks.com/de/blog/2020/01/30/what-is-a-data-lakehouse.html, Abruf am 16.03.2021.

Schymik, G.; Corral, K.; Schuff, D.; St Louis, R. D. (2017): Designing a Prototype for Analytical Model Selection and Execution to Support Self-Service BI. In: Proceedings of the Twenty-third Americas Conference on Information Systems, Boston.

Talend (2021): Data Fabric: Architektur, Funktionen und Vorteile. https://www.talend.com/de/resources/data-fabric. Abruf am: 26.05.2021.

TDWI (2011): Big Data Analytics, Best Practices Report 4.

TDWI (2020): Halper, F.; Loshin, D.: TDWI Checklist Report – Six Capablities of a Logical Data Fabric.

Wrobel, S. (2012): Big Data – Vorsprung durch Wissen. Fraunhofer Institut für intelligente Analyse- und Informationssysteme, Sankt Augustin.

Über unseren Sponsor

denodo
DATA VIRTUALIZATION

Kontaktadresse

Denodo Technologies GmbH
Karlstraße 10
80333 München

Telefon: +49 89 59990450

E-Mail: info@denodo.com
URL: www.denodo.de

Denodo ist der führende Anbieter im Bereich der Datenvirtualisierung und bietet Unternehmen agile und hochleistungsfähige Datenintegration, Datenabstraktion und Datendienste in Echtzeit aus einer Vielzahl verschiedener Quellen wie Unternehmensdaten, Cloud-Daten oder Big Data und unstrukturierten Daten an, und das zur Hälfte der Kosten herkömmlicher Datenintegrationsansätze. Denodo´s Kunden aus allen wichtigen Branchen konnten durch den schnelleren und einfacheren Zugriff auf einheitliche Geschäftsinformationen für ihre agile BI, Big Data Analytics, Web- und Cloud-Integration, Single-View-Anwendungen sowie Unternehmensdatendienste ihre Flexibilität und Rentabilität erheblich steigern. Denodo ist kapitalkräftig, profitabel und in Privatbesitz. Für weitere Informationen besuchen Sie https://www.denodo.com/de

E-Book

tdwi.eu