tdwi
WISSEN

DER DATA LAKE

E-Book

als zentrales Element in Analytics-Architekturen

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

Inhalt

ZUM INHALT

Vorwort

Der Begriff des Data Lake hat inzwischen Prominenz erhalten und wirkt charakterisierend für Ansätze im Spannungsfeld der Advanced Analytics, oder genereller, der Digitalisierung in Unternehmen. In diesem Zusammenhang fällt auf, dass sich Veröffentlichungen zum Thema Data Lake häufig auf die Bewertung der aktuellen Position einer Organisation oder auf technologische Aspekte konzentrieren. Offen bleibt die Frage, wie Unternehmen einerseits bei der Weiterentwicklung ihrer analytischen Architektur und andererseits bei der damit einhergehenden organisatorischen Veränderung zu unterstützen sind. In diesem E-Book wird der Data Lake nicht allein als technisches Architekturkonzept betrachtet, sondern es werden die damit einhergehenden organisatorischen und prozessualen Themen des Unternehmens beleuchtet. Hier kann man Beziehungen zur inzwischen weitverbreiteten Data-Warehouse-Architektur herstellen. Oft enthält diese bereits einen Operational Data Store (ODS), der ebenso wie oft auch ein Data Lake auf die Speicherung von Detaildaten abzielt. Allerdings konzentriert sich ein ODS auf die internen Vorsysteme und ist zumeist spaltenorientiert (SQL-bedienbar) aufgebaut, während der Data Lake unter anderem externe Daten und zeilenorientierte Daten aufnehmen kann. Wie bei den meisten aktuellen IT-Entscheidungen ist zu überlegen, ob dieser unternehmensintern oder -extern angelegt, ob ein klassisches relationales Datenbanksystem oder ein Hadoop Cluster genutzt wird. Daneben sind auch organisatorische Maßnahmen zu ergreifen, um einerseits Verantwortlichkeiten für Datenqualität sowie Datennutzung und -nutzbarkeit zu bestimmen und andererseits die Einhaltung rechtlicher und unternehmensweiter Vorgaben zu unterstützen. Dies verlangt Kompetenzentwicklung aufseiten der Mitarbeiter, um dem generellen Ziel einer sinnhaften Datennutzung im Unternehmen gerecht werden zu können. Diese Themen werden im vorliegenden E-Book zusammengetragen. Ich wünsche Ihnen viel Freude beim Lesen und einen angeregten Austausch zu diesem Thema.

1 Das Entstehen des Data Lake

Die digitale Transformation in Unternehmen ist derzeit ein präsentes Thema. Sie steht zum einen für die Übertragung sowie die sinnhafte Neukonzeption von Prozessen hin zu einer möglichst einfach durch IT unterstützbaren Form – damit liegen viele Daten erstmalig elektronisch verarbeitbar vor. Zum anderen steht Digitalisierung für den Einsatz von Algorithmen des maschinellen Lernens oder der Künstlichen Intelligenz, um eine Lösung im Spannungsfeld von Entscheidungsunterstützung bis hin zu einer Automation zu erzielen. In diesem Umfeld ist es von entscheidender Bedeutung, Daten nach ihrer Erhebung / ihrem Aufkommen so bereitzustellen, dass diese Analysen überhaupt umsetzbar sind. Dies bringt es aber mit sich, dass die zentrale Verfügbarkeit von Daten ein Motor im Kontext der digitalen Transformation ist. Dabei hat sich der Data-Lake-Ansatz als neuer architektonischer Baustein herauskristallisiert, um ein Unternehmen auf digitalisierbare Anwendungsfälle und Geschäftsziele ausrichten zu können. [Durmus 2017]

Ein Data Lake ist ein effektives, datengesteuertes Konstrukt für die Speicherung einer Vielzahl heterogener Datentypen in großem Maßstab. Per Definition ist ein Data Lake zur schnellen Erfassung detaillierter Rohdaten und deren sofortiger Verarbeitung für Erkundungs-, Analyse- und Betriebszwecke optimiert. Darüber hinaus lässt sich dieser aber auch für die Bereitstellung oder Historisierung detaillierter oder aufbereiteter interner Daten nutzen. Die Daten können sowohl aus internen als auch aus externen Quellen stammen. Beide Nutzungsszenarios führen immer zu unternehmensindividuellen Ausgestaltungen der Data Lakes. Architektonisch gibt es keine Festlegung in Bezug auf die Art der Datenhaltung. So reicht die Spanne von SQL- und No-SQL-Datenbanken bis hin zu In-Memory-Technologien, um die Art von Rohdaten bereitzustellen, welche Benutzer für die Datenexploration und entdeckungsorientierte Formen von Daten für eine fortgeschrittene Analytik benötigen.

In Analogie zum Data Warehouse wird auch ein Data Lake zu einem Konsolidierungspunkt, nun aber für typischerweise nur gering vorverarbeitete und detailreiche Daten, um Analysen im Sinne der Anwendung des Maschinellen Lernens oder der Künstlichen Intelligenz zu ermöglichen. Mit den richtigen Endbenutzerwerkzeugen kann ein Data Lake auch einen Self Service ermöglichen, so dass Fachanwender und Data Scientists direkt und eigenständig auf dem Datenbestand agieren können; aber auch insbesondere Analysen durch maschinelle Aufgabenträger werden performant möglich. Der Unternehmenswert, den Big Data, andere neue Datenquellen und eine intensivere Nutzung interner Unternehmensdaten bieten, wird durch die hier schon skizzierten Anwendungs- und Gestaltungsszenarios gewonnen. Dies bringt jedoch Aufwand mit sich, da ein Data-Lake-Aufbau bestehende Landschaften mit Data Warehousing, Online Analytical Processing, Datenintegration und andere datengesteuerte Lösungen erweitert. Dabei werden die Nutzer jedoch nicht auf allen Ebenen des Unternehmens zu finden sein. Die Hauptnutznießer von Data Lakes sind im Analytics-Umfeld und der Data Science zu finden, deren Nutzen aus Big Data und flexiblen Anwendungsszenarios besteht. Durch die geringere Vorverarbeitung haben auch Nutzungsszenarios von Daten mit einer Real-Time-Ausrichtung in einem Data Lake eine gute technologische Basis.

Data-Lake-Umsetzungen sind jedoch auch mit Herausforderungen konfrontiert, die sich aus einer unausgereiften Governance, Integration, Benutzerkenntnissen und Sicherheit zum Beispiel in Hadoop-Umgebungen ergeben. Daher müssen neben technischen Lösungen und der Umsetzung eben diese begleitenden Themen in die Unternehmensdiskussion einfließen. Eine nachträgliche Einführung beispielsweise einer Governance erhöht den Aufwand wesentlich, zumal beispielsweise nicht mehr alle Ansprechpartner vorhanden, Dokumentationen spärlich und nachträgliche Einschränkungen mühselig sind.

Ein Data Lake ist für die Hälfte der Datenverwaltungsbereiche ein zentrales Thema auf der Roadmap, wie eine Studie des TDWI Research ergab [Russom 2019]. Ein Viertel der befragten Unternehmen verfügt bereits über mindestens einen Data Lake

1 Das Entstehen des Data Lake

in der Produktion, der dort aber normalerweise als Data-Warehouse-Erweiterung fungiert. Ein weiteres Viertel wird in einem Jahr in die operative Nutzung gehen. Viele Anwender sind durch die Entwicklung von Datentypen, -strukturen, -quellen und -volumina motiviert, sich intensiver mit Data Lakes auseinanderzusetzen, um der explodierenden Vielfalt und Größe der Daten begegnen zu können. Dabei werden relationale Datenbanken oftmals als limitierend eingeschätzt, weshalb Non-SQL-Datenbanken wie Hadoop als Data-Lake-Plattform in den Vordergrund treten. Anwender, die bereits über einen Data Lake verfügen, stellten zunehmend fest, dass der Data Lake als eine Rohdatenquelle mit eigenen Bereichen eine wichtige Quelle für strukturierte und unstrukturierte Daten wird. Dabei wachsen die Anwendungsbereiche der Nutzer, wenn diese die Datenbereitstellung besser verstehen und einen damit einhergehenden Self Service angemessen umsetzen können.

All die hier bereits skizzierten Aspekte verdeutlichen, dass Data Lakes in einer analytischen Umgebung basierend auf fachlichen Nutzungsszenarios zu integrieren sind, um einen Mehrwert zu schaffen. Da sie andere Aufgaben als ein Data Warehouse übernehmen, ist deren Zusammenspiel zu orchestrieren, um Aufwände im Kontext der Unterstützung datengetriebenen Handelns nicht explodieren zu lassen und gleichzeitig auch eine Beherrschbarkeit im Sinne einer Governance und Einhaltung rechtlicher Rahmenbedingungen zu ermöglichen. Das vorliegende E-Book greift daher die Themen der Abgrenzung sowie des Zusammenspiels zwischen Data Warehouse und Data Lake auf (Kapitel 2) und wendet dann den Blick auf Architekturszenarios (Kapitel 3) sowie organisatorische und rechtliche Aspekte (Kapitel 4), die bei einem Data-Lake-Aufbau zu beachten sind. Weitere Erkenntnisse zu diesem Thema von TDWI Research runden die Betrachtung in Kapitel 5 ab.

2 Das Data Warehouse und der Data Lake – integrierte Parallelwelten

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 auf 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 (Business Intelligence im engen Sinne) versorgen. Diese Mehrschichtenarchitektur hat in bestimmten Anforderungskontexten weiterhin seine Gültigkeit. So zum Beispiel im Unternehmen, in dem übergreifend konsistente Auswertungen über abgestimmte Auswertungssichten zum Beispiel nach Produkt, Kunde oder Region bereitgestellt werden sollen. [Dittmar et al. 2016]

Die Einhaltung dieser idealtypischen klassischen Architektur wird jedoch immer schwieriger, wenn Fachbereiche eine höhere Änderungsdynamik fordern, als sie von einer zentralen Data-Warehouse-Infrastruktur geleistet werden kann. [Dittmar et al. 2016]

An dieser Stelle kommt der Data Lake ins Spiel. Dieser weist einen geringeren Fokus auf Berichtsthemen auf und orientiert mehr auf Detaildaten in deren Rohformat. Gleichzeitig geht dies mit einer wachsenden Flexibilität einher, da Data Lakes einerseits je nach Themenkomplex gebildet werden, andererseits im Unternehmen aufgebaut (on premise) oder außerhalb verortet (cloud) sein können. Im Folgenden werden beide Konzepte vorgestellt, um deren Hintergründe und Ansätze nachvollziehen zu können, so dass ein integrativer Ansatz beider dann zusammenwirkenden Architekturbestandteile nachvollziehbar wird.

2.1 Der Klassiker – das Data-Warehouse-Konzept

Zur Sicherstellung von Ad-hoc-Analysen und einem Drill Down bis auf die Basisdaten wurden in den 1990er-Jahren Architekturkonzepte marktfähig gemacht, die als Data Warehouse (DW) und als Online-Analytical-Processing-Datenbanken (OLAP) in die betriebliche Informationsverarbeitung Einzug hielten. Die Grundidee bestand in der Zusammenführung aller potenziell für einen Entscheidungsfall relevanten Daten in eine unternehmensweite Datenbank, deren struktureller Aufbau sich an den Entscheidungsfeldern, nicht – wie bisher üblich – an den Geschäftsprozessen festmachte. Diese Abkehr vom Paradigma der relationalen Datenbanken zur Transaktionsverarbeitung, dem sogenannten Online Transaction Processing (OLTP), hin zum multidimensionalen Konzept der Informationsspeicherung von Dimensionen und Fakten (OLAP) hatte durchschlagenden Erfolg und führte zu einer Massenbewegung in den Unternehmen. Die Installation von Data-Warehouse-Instanzen versprach einerseits die Lösung des Problems der fehlenden Informationsverfügbarkeit und der unzureichenden Navigation in den Informationsbeständen, verschaffte aber andererseits den Produktanbietern einen Milliardenmarkt, so dass ein starkes Momentum entstand.

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 und zur fünften Phase der Managementunterstützungssysteme (MUS) führte.

2 Das Data Warehouse und der Data Lake – integrierte Parallelwelten

Unter dem Begriff Business Intelligence (BI) wird die Zusammenführung und analytische Auswertung fragmentarischer Unternehmensdaten (intern und extern) verstanden. Die Hinwendung zur aufklärenden Informationsversorgung des Managements mit fachlichen Inhalten und Interpretationshilfen prägte diese Phase. Auf der Basis eines Data Warehouse werden unterschiedliche Nutzergruppen durch BI unterstützt. Dies können Abonnenten von Berichten sein, die lediglich konsumieren, oder Analytiker, die tief in die Datenbestände abtauchen und Hypothesen prüfen beziehungsweise Kausalitäten ableiten. Solche Power-User, die mit Ad-hoc-Analysen die Navigationsmöglichkeiten der OLAP-Würfel nutzen, ziehen den primären Nutzen aus BI. Die Zurechnung der Data-Mining-Verfahren zu BI ist nicht unumstritten, ergänzt aber das Nutzerprofil um die Analysten, die statistische Methoden und maschinelles Lernen einsetzen, um interpretationsfähige Datenmuster zu generieren. Im Folgenden wird aber zunächst das Data-Warehouse-Konzept vorgestellt, da dieses sich als die Keimzelle datengetriebener Entscheidungsunterstützung in Unternehmen verstehen lässt.

2.1.1 Definition und charakterisierende Merkmale

"A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions." [Inmon 1996, 33] Die vier Schlüsselcharakteristiken des Data Warehouse, also die Themenorientierung (subject orientation), die Integration, die Unveränderlichkeit (non-volatility) und die Zeitabhängigkeit (time variancy), werden im Folgenden näher erläutert.

In einem Data Warehouse werden, im Gegensatz zur stark prozess- beziehungsweise applikationsorientierten Datenhaltung operationaler Informationssysteme, Daten themen- beziehungsweise inhaltsorientiert gespeichert. Die Datenhaltung operationaler Systeme orientiert sich zweckmäßigerweise stark an den spezifischen Anforderungen der Online-Transaktionsverarbeitung (OLTP), zum Beispiel optimierter Zugriff auf einzelne Datensätze, redundanzfreie Speicherung von Nicht-Schlüsselelementen, entsprechende Sperrmechanismen etc. Analytisch optimierte Informationssysteme erfordern eine andersartige Datenorganisation. In einem Data Warehouse werden Daten nach unterschiedlichen Themenbereichen organisiert. [Inmon 1996, 33f.] Analytische Anfragen können bei einer derartig organisierten Datenbank oft durch Zugriff auf nur eine Tabelle befriedigt werden. Der Zugriff auf nur eine Tabelle, der gegebenenfalls auch durch gezielte redundante Datenspeicherung erzielt werden kann, ist im Vergleich zu Zugriffen auf mehrere Tabellen bei gleichem Datenvolumen wesentlich performanter.

In einem Data Warehouse werden Daten integriert, das heißt unternehmensweit einheitlich und konsistent gespeichert. Dies ist eine zentrale Eigenschaft des Data-Warehouse-Konzeptes. Die Integration der Daten findet beim periodischen Import von Daten aus den Vorsystemen statt. Durch den Integrationsprozess werden die Daten aus den unterschiedlichen, applikationsspezifischen Darstellungsformen in eine einheitliche Darstellungsform überführt. Die Integration bezieht sich unter anderem auf folgende Eigenschaften der Daten, die in der Tabelle auf der folgenden Seite skizziert sind: [Inmon 1996, 33-35]

2 Das Data Warehouse und der Data Lake – integrierte Parallelwelten

Integrationseigenschaft	Erläuterung	Beispiel
Kodierung und Verschlüsselung	Unterschiedliche Applikationen können dieselben Inhalte unterschiedlich kodieren. So lässt sich beispielsweise das Geschlecht einer Person auf unterschiedliche Weise darstellen.	(m, w), (0, 1) oder (maennlich, weiblich)
Bemaßung von Attributen	Je nach Anwendungszweck können unterschiedliche Applikationen Attribute in unterschiedlichen Maßeinheiten ausdrücken.	Die Ausmaße einer Rohrleitung sind zum Beispiel in cm, Zoll oder m3/h darstellbar.
Format	Die angewendeten Formate können sich in Datenbanken logisch und physikalisch unterscheiden. Auch bei physikalisch gleichem Format kann sich das logische Format der gespeicherten Daten von System zu System unterscheiden.	In diesem Beispiel wird das Datum zwar physikalisch immer gleich (als String), aber in unterschiedlichen logischen Formaten erfasst: Date (yyyymmdd) „20190508“ Date (dd.mm.yyyy) “08.05.2019” date (yyyyjjj) “2019079”
	Entsprechend der ihnen zugrunde liegenden Hard- und Softwarearchitektur können Daten in unterschiedlichen physikalischen Formaten gespeichert sein.	Der Saldo eines Kontos lässt sich unter anderem wie folgt definieren: balance decimal (13, 2) balance decimal (11, 0) balance double precision
Bezeichner	Datenfelder gleichen Inhalts werden in unterschiedlichen Applikationen oft unterschiedlich bezeichnet. Die Beschreibung dieser Datenfelder in der Dokumentation der unterschiedlichen Applikationen wird oft ähnliche Unterschiede aufweisen.	Kundennummer, KundenID, KdNr, Kunde.

Tabelle 1: *Integrationsaspekte im Kontext des Data Warehouse*

Die Datenbestände operationaler Informationssysteme sind typischerweise häufigen datensatzweiten Änderungen (sogenannten Updates) unterworfen. Jeder Datensatz eines operationalen Informationssystems kann nur für den Zeitpunkt seiner Abfrage als korrekt und aktuell betrachtet werden, sein Inhalt kann sofort nach der Abfrage durch eine andere Transaktion geändert worden sein. Dahingegen werden Daten im Data Warehouse nach dem ursprünglichen Laden nicht mehr geändert, es wird nur noch lesend auf sie zugegriffen – sie sind also non-volatile (nicht flüchtig). [Inmon 1996, 35ff.]

Ein Data Warehouse beinhaltet im Gegensatz zur Verwaltung aktueller Daten durch operationale Informationssysteme Daten, die für unterschiedliche, genau definierte Zeitpunkte beziehungsweise Zeiträume gültig sind oder waren. Dies bedeutet, dass die in einem Data Warehouse gespeicherten Daten time variant (zeitvariant) sind; die vollqualifizierten Schlüssel, mit denen auf die Daten eines Data Warehouse zugegriffen wird, beinhalten also ein Zeitelement. Die im Data Warehouse gespeicherten Daten lassen sich somit als Aneinanderreihung von Momentaufnahmen (Snapshots) von Daten der operationalen Informationssysteme betrachten.

Zusätzlich unterscheidet sich der Zeithorizont, über den Daten in den operationalen Informationssystemen und im Data Warehouse vorgehalten sind. Während operationale Systeme möglichst wenige historische Daten speichern, erfasst ein Data Warehouse möglichst viele. Inmon gibt 60 bis 90 Tage für Daten in operativen und fünf bis zehn Jahre für Daten in einem Data Warehouse als typische Verweildauer an. [Inmon 1996, 36f.]

2 Das Data Warehouse und der Data Lake – integrierte Parallelwelten

2.1.2 Struktur und Datenfluss

Die Struktur eines Data Warehouse unterscheidet sich deutlich von der Struktur operationaler Datenbanken. Abbildung 1 veranschaulicht den allgemeinen Aufbau eines Data Warehouse.

Das Zusammenspiel der einzelnen Komponenten basiert auf einem Schnittstellenprinzip, so dass sich operative Datenbanken an die Data-Warehouse-Datenbank anbinden lassen und die Daten schlussendlich für Oberflächenwerkzeuge verfügbar sind. An dieser Stelle soll jedoch nicht vertiefend auf ein Core beziehungsweise Enterprise Data Warehouse eingegangen werden, da es hier lediglich um die grundlegenden Aspekte geht. Das aufgeführte Data Mart ist je nach Entwicklung ein funktions-, abteilungs- oder personenbezogenes Extrakt aus dem Data Warehouse. Allerdings können auch eine Vielzahl von Data Marts zu einem Data Warehouse verbunden werden, wobei auch diese Spielarten hier nicht vertiefend betrachtet werden sollen.

Kern eines Data Warehouse ist die Datenbank, in der die Basisdaten aus den Vorsystemen (OLTP-Quellen) historisch in unterschiedlichen Aggregationsstufen erfasst sind. Die aus den operativen Informationssystemen mittels Extraktions-Transformations-Lade-Prozessen (ETL) extrahierten Daten werden nach der für die integrierte Speicherung der Daten im Data Warehouse notwendigen Transformation als aktuelle Detaildaten abgelegt. Diese Daten spiegeln die aktuellsten Vorfälle wider und befinden sich zunächst auf dem niedrigsten Granularitätsniveau. Damit sind

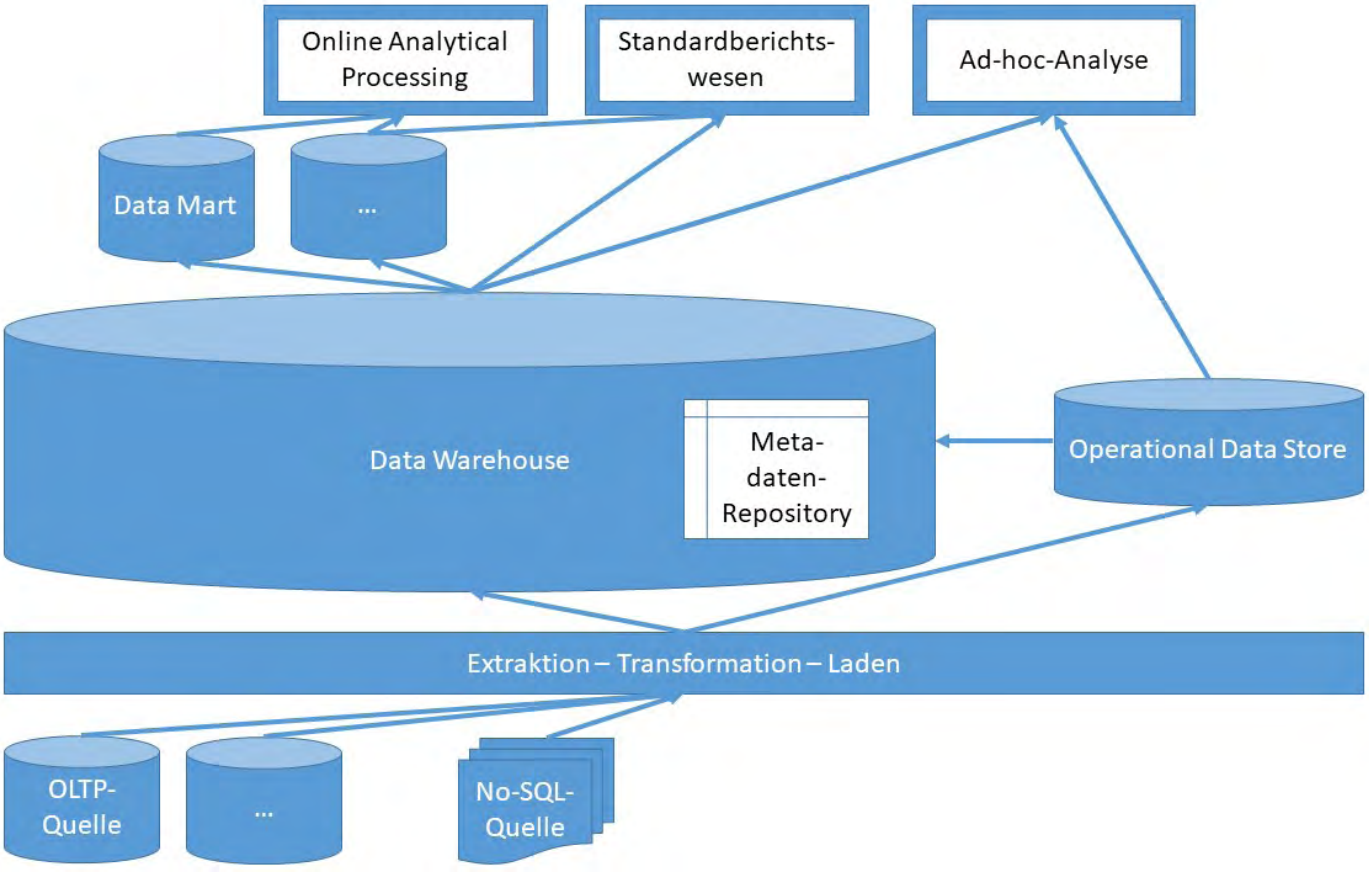


Abb. 1: *Aufbau eines Data Warehouse*

2 Das Data Warehouse und der Data Lake – integrierte Parallelwelten

sie aber auch die unterste Ebene der Dimensionshierarchien. Im Zeitverlauf altern diese aktuellen Daten. Aus diesen Daten werden entsprechend vordefinierter Aggregationsprozesse aggregierte Daten berechnet. Sie sollen einen Teil der häufig gestellten Anfragen an das Data Warehouse befriedigen. So wird der wiederholte ressourcenintensive Zugriff auf die Detaildaten zur Beantwortung von Standardanfragen vermieden.

Eine wichtige Komponente des Data Warehouse sind die Metadaten. Metadaten werden nicht aus den operationalen Informationssystemen extrahiert, sondern knüpfen die Verbindung zwischen den operationalen Informationssystemen und dem Data Warehouse sowie den verschiedenen Aggregationsstufen im Data Warehouse. Die Metadaten beinhalten ein Verzeichnis, das dem Analysten das Auffinden von Daten in der Datenbank ermöglicht. Metadaten beschreiben die Transformationen, denen die Daten beim Transfer von den Vorsystemen zum Data Warehouse unterworfen werden, und die Algorithmen der im Data Warehouse stattfindenden Aggregationsprozesse. [Inmon 1996, 14-15]

Im gegebenen Kontext ist allerdings der ODS von besonderem Interesse. Er ist in seiner Architektur sehr stark auf bestimmte geschäftliche Themen ausgerichtet und stellt damit die Subject-Orientation mehr in den Vordergrund, unterscheidet sich jedoch durch das Granularitätsniveau der Daten vom Data Warehouse. Hier geht es ausschließlich um Detaildaten aus den operativen Prozessen, die für eine kurze Zeit für Analysezwecke zur Verfügung gestellt werden.

Ein Operational Data Store wird für die Integration heterogener operativer Systeme verwendet und enthält im Gegensatz zum Data Warehouse nur die aktuellsten, zeitnahen Daten in sehr hoher Granularität. Um stets das aktuelle Geschehen abbilden zu können, wird er sehr häufig sogar mehrmals am Tag aktualisiert. Die Daten im Operational Data Store werden daher sowohl für das operative Berichtswesen als auch für Echtzeitanalysen verwendet, während in einem Data Warehouse eher historische Daten für strategische Entscheidungen auf breiter Ebene bereitgehalten werden. Bei der Integration der Daten in den ODS werden diese von Redundanzen bereinigt, de-normalisiert und zur Sicherstellung der Datenintegrität geschäftsrelevante Regeln angewendet. Ein ODS ist daher in der Regel eher für einfache Abfragen auf eine kleine Datenmenge optimiert, um beispielsweise den Status von Kundenbestellungen abzufragen. Währenddessen werden im Data Warehouse eher komplexe Abfragen auf eine große Menge von Daten getätigt. Der ODS lässt sich also mit dem menschlichen Kurzzeitgedächtnis vergleichen, das nur aktuelle Informationen und mehr Details enthält, während das Data Warehouse eher dem Langzeitgedächtnis entspricht, das historische Informationen enthält, allerdings oft mit weniger Details. Die Daten aus dem Operational Data Store werden häufig aggregiert für die Langzeitspeicherung und strategische Analysen in das Data Warehouse übertragen. Er kann also als eine Vorstufe für das Data Warehouse dienen. Es kann mehrere Gründe geben, einen ODS einzurichten, wenn beispielsweise nur begrenzte oder nicht ausreichende Möglichkeiten für das Berichtswesen in den Quellsystemen zur Verfügung stehen. Alternativ zum direkten Zugriff auf Quellsysteme kann ein ODS Mitarbeitern das Erstellen von Reports und Analysen auf Detaildaten ermöglichen.

Da nun bereits in den ersten Worten über den Data Lake darauf verwiesen wurde, dass ein solcher Detaildaten für Analysezwecke bereitstellt, erscheint es eigentlich naheliegend, den ODS für diese Zwecke zu nutzen. Dieser kann aber im Rahmen seiner Definition den Anforderungen, die sich aus den zunehmenden Analytics-Szenarios ergeben, nicht gerecht werden, so dass der Data Lake als architektonischer Baustein hinzugekommen ist.

2 Das Data Warehouse und der Data Lake – integrierte Parallelwelten

2.2 Das Umdenken – der Data Lake zur Rohdatenbereitstellung

Der Begriffswandel von Business Intelligence hin zu Business Analytics verspricht einen intensiveren Einsatz von weiterführenden 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. Business Analytics verspricht hingegen 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 Modellen liegen. Die Konvergenz von datenorientierten und modellorientierten Verfahren scheint daher naheliegend und bringt tatsächlich neue Aspekte in die Betrachtung von MUS auf dem Zeitstrahl. Hier treten Algorithmen in den Vordergrund, die (teil)automatisierte Entscheidungsprozesse ermöglichen, welche auf großen polystrukturierten Datenbeständen (Big Data) in Echtzeit Empfehlungen für bestmögliche Entscheidungen geben oder diese Entscheidungen selbst treffen. 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 immer die flexible Grundlage für Data-Analytics-Umgebungen. Dies gilt insbesondere für deren Anforderungen an zusätzliche Informationen: die Metadaten. Metadaten, also die schon zuvor eingeführten 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. Beim Data Lake handelt es sich um einen sehr großen Datenspeicher, der die Daten aus den unterschiedlichsten Quellen in ihrem Rohformat aufnimmt. Dies ist damit bereits ein großer Unterschied zum ODS. Der Data Lake 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 die 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 resultieren aus unterschiedlichen Anforderungen wie Art der Daten, Zeit- und Speicherperformance. Darüber hinaus lassen sich entsprechende Sicherheitsaspekte abbilden, so dass letztlich 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.

Die nachfolgende Tabelle fasst die zentralen Unterschiede zwischen Data Warehouse und Data Lake zusammen.

Kriterium	Data Lake	Data Warehouse
Datenvolumen	Gute Skalierbarkeit auch bei sehr großen Datenmengen	Gute Skalierbarkeit auch bei sehr großen Datenmengen
Datentypen	Polystrukturiert	Strukturiert
Aufbereitung	Daten überwiegend detailliert und original	Daten überwiegend veredelt und teils vorkalkuliert
Flexibilität	Speicherung im Originalformat. Transformation nach Datenablage (Schema-on-Read)	Limitierung durch vorgegebenes DWH-Schema (Schema-on-Write)
Datenquellen	Batch, Realtime, Stream	Batch

Tabelle 2: *Unterschiedsbetrachtung Data Warehouse und Data Lake*

3 Architektonische Aspekte

Es gibt im Wesentlichen zwei große Arten von Data Lakes, auf deren Grundlage die Datenplattform verwendet wird: Hadoop-basierte Data Lakes und relationale Data Lakes. Heute ist Hadoop als Plattform weitaus verbreiteter als relationale Datenbanksysteme. Bei vielen Realisierungen in den Unternehmen hat sich jedoch gezeigt, dass der Data Lake unter Nutzung beider Ansätze realisiert worden ist. Diese Plattformen können sich im Unternehmen selbst, in Clouds oder eben parallel in beiden befinden. Daher sind einige Data Lakes, wie auch die meisten heutigen Data Warehouses, plattformübergreifend und hybrid.

Die Gestaltung einer Data-Lake-berücksichtigenden Architektur lässt sich nicht nur singulär für diese Art des Datenspeichers beantworten, sondern ist im Spannungsfeld der weiteren analytischen Systeme zu verstehen. Hintergrund bilden die polystrukturierten Big Data und deren unterschiedliche Nutzungsszenarios, die ein Zusammenspiel der verschiedenen Architekturelemente bedingen. Nutzungsszenarios können dabei beispielsweise sein:

- Multichannel-Marketing: Datenmix von Websites, Social Media, externen Daten von Fremdanbietern und internen Daten von Customer Touch Points für ein ganzheitliches Bild des Kunden.
- Betrugserkennung (Fraud Detection): Aufdeckung von zum Beispiel Versicherungsbetrug oder Insiderhandel durch Kombination von Daten aus unterschiedlichen Quellen.
- Stimmungsanalyse (Sentiment Analysis): Auswertung subjektiver Aussagen und Meinungen natürlicher Personen vor allem aus Sozialen Medien durch Verfahren des Text Mining.
- Wartung und Instandhaltung (Predictive Maintenance): Kombination von maschinengenerierten Sensordaten mit anderen Daten (zum Beispiel Maschinenstammdaten), um daraus Erkenntnisse über den Zustand der Anlagen und erforderliche Maßnahmen zu gewinnen.

Big Data als Analyse und Echtzeitverarbeitung großer, unstrukturierter und kontinuierlich fließender Datenmengen aus einer Vielzahl unterschiedlicher Datenquellen zur unmittelbaren Entscheidungsunterstützung fordert im Gegensatz zu den klassischen Ansätzen der dispositiven Datenverarbeitung keinen Single Point of Truth. Es sollen vielmehr nur die jeweiligen Datenquellen verbunden werden, die für den jeweiligen Big-Data-Anwendungsfall benötigt werden, um Datenvolumen-, Datentypen- und Datenquellenvielfalt sowie Anforderungen an die Verarbeitungsgeschwindigkeit zu erfüllen. Damit setzt Big Data den Trend zur Dezentralisierung von dispositiven Architekturen fort und führt zur Bildung von Multi-Platform Environments oder analytischen Ökosystemen. Traditionelle Data Warehouses mit konsistentem Datenbestand bilden zwar in solchen analytischen Ökosystemen nach wie vor die ideale Plattform für das Berichtswesen, Dashboarding, Performance Management und OLAP, doch neue Stand-Alone-Datenplattformen entstehen in Ergänzung zur bestehenden Architektur und dienen der Verarbeitung und der Analyse von Big-Data-Anwendungsfällen.

Die Abbildung auf der folgenden Seite bringt zunächst die unterschiedlichen Ebenen zusammen, die im analytischen Umfeld relevant sind, und positioniert den Data Lake.

3 Architektonische Aspekte

Abb. 2: *Data Lakes in einer hybriden Architektur*

Ein Data Lake nimmt Daten in ihrer unbearbeiteten, originalen Form mit keiner oder wenig Bereinigung, Standardisierung, Neumodellierung oder Veränderung direkt von den Datenquellen auf. Die Transformation der Inhalte des Data Lakes erfolgt zur Laufzeit im Rahmen der Auswertung (wie zum Beispiel im Rahmen von Ad-hoc-Analysen) oder als Zwischenschritt zur Vorbereitung für wiederkehrende Aufgaben (wie beispielsweise das Berichtswesen) [Russom 2017]. Der Data Lake stellt eine wichtige Komponente einer hybriden Datenarchitektur dar, die sich dadurch auszeichnet, dass in den unterschiedlichen Datenhaltungsbereichen oder -schichten eines analytischen Ökosystems verschiedene Speichertechnologien zum Einsatz gelangen und dadurch die jeweiligen individuellen Stärken einbringen können [Hardt/Lenzhölzer 2017]. Werden in einem gesondert kontrollierten und verwalteten Bereich des Data Lake zusätzlich aufbereitete und qualitätsgesicherte Inhalte vorgehalten, so lässt sich hierfür die Bezeichnung des Data Reservoir verwenden, zu dem auch ein Katalog mit Metadaten zu den gespeicherten Objekten gehört, der grundlegende Angaben über die Daten wie zum Beispiel deren Herkunft, Ursprungsformat, fachliche Bedeutung etc. beinhaltet [Kromer 2015].

Es zeigt sich also, dass die integrierte Analyselandschaft der Zukunft vielfältiger und komplexer wird und dabei die Flexibilität in den Vordergrund stellt. Die Integration erfolgt hier eher logisch, durch einheitliche Metadaten, Data Governance und Stammdaten. Der physische Integrationsanspruch tritt wieder zurück. Anwenderunternehmen müssen hier auch 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 von Architekturkomponenten sicherzustellen. Daher wird möglicherweise zukünftig ein Nebeneinander von selbst erstellten und integrierten vorkonfektionierten

3 Architektonische Aspekte

Lösungen und extern zum Beispiel in einer Cloud bereitgestellten Architekturkomponenten festzustellen sein. Klassische Grenzen zwischen operativen und dispositiven Systemen, zwischen disjunkten Werkzeugen für unterschiedliche Analysebedarfe, zwischen Backend-orientierter Datenintegration und Frontend-orientierter 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.

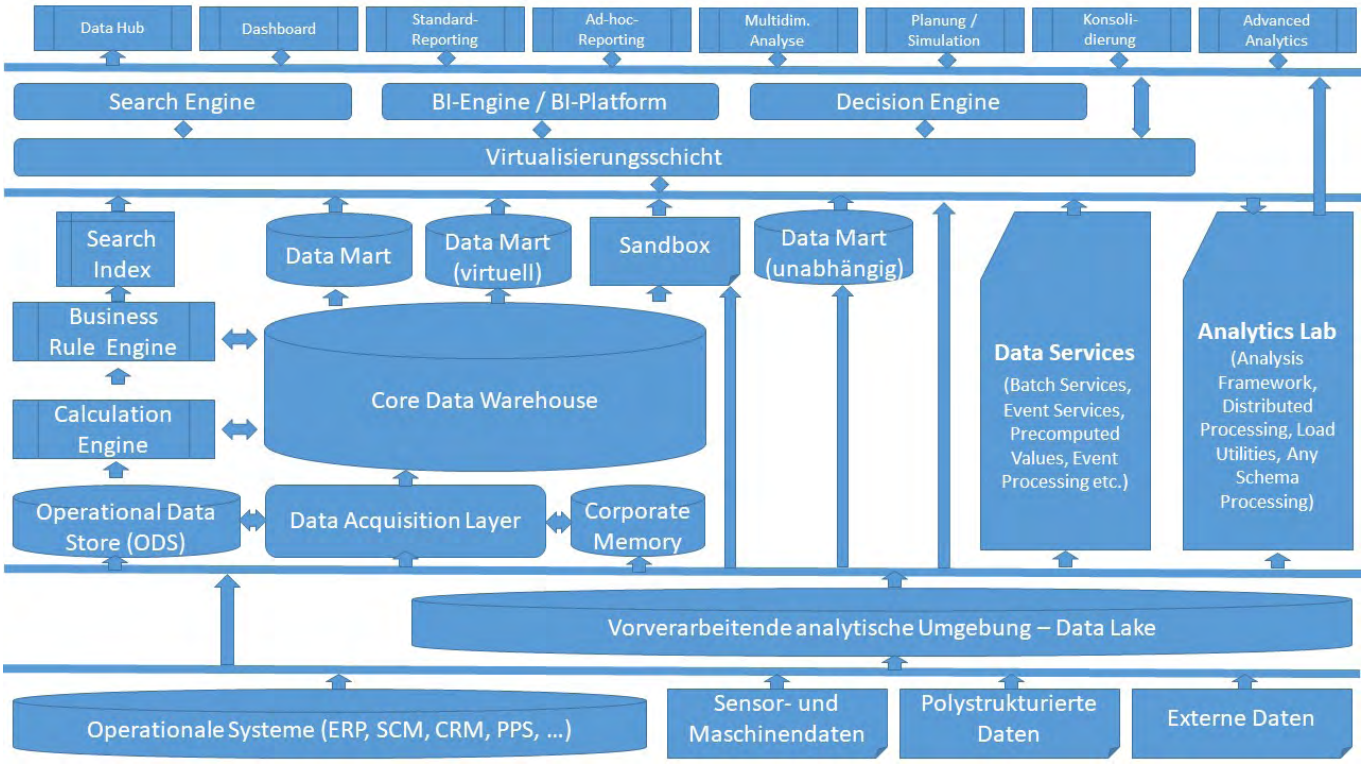


Abb. 3: *Analytisches Ökosystem im Überblick*

Neben den Komponenten, die aus einer klassischen BI-Architektur bekannt sind, sind dort auch Elemente einer Lambda-Architektur für die Realtime-Verarbeitung zum Beispiel von Sensordaten mit einem Speed Layer, Batch Layer und Service Layer enthalten. 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 Backend-orientierten 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 zuvor beschriebenen Data-Lake-Konzepts eine Systemumgebung zu schaffen, die zu jedem Zeitpunkt den Sprung zwischen lokalen Ressourcen und dem Objektspeicher ermöglicht, so dass eine zielorientierte Nutzung möglich ist und auch noch nicht absehbare Auswertungen entsprechend den sich ergebenden Anforderungen abbildbar sind.

4 Datenflussorientierte Betrachtungen zum Data Lake

Wie nun zu sehen war, ist ein Data Lake immer eine weitere Komponente in der analytischen Architektur, die eine Vielzahl relevanter Schnittstellen zu unterstützen hat, um die Kompatibilität zu den bereits bestehenden Umgebungen aufrechtzuerhalten. Zu nennen sind da HTTP(S), WebDAV, CIFS, NFS, SMTP, NDMP oder die S3-Schnittstelle (Simple Storage Services) von Amazon. Darüber hinaus ist zu beachten, dass kein Unternehmen seine Daten vollständig in der Cloud ablegen wird, aber viele werden versuchen, die Vorteile des Cloud-Speichers für einen Teil ihres Datenbestandes zu nutzen. Ein möglicher Ansatz, dies zu realisieren, ist ein hybrides Modell, bei dem lokale Speicherressourcen (das sogenannte On-Premises) und die Cloud nahtlos miteinander integriert werden. Eine solche Cloud-Tiering-Integration lässt sich mit dafür dedizierter Software, Cloud-Applikationen oder entsprechend ausgestatteten Speichersystemen und Cloud-Gateways umsetzen, so dass eine flexible und skalierbare Umgebung entsteht. Im Folgenden werden zwei Szenarios exemplarisch herausgegriffen. Diese sind das Metadata Driven Onboarding und die Streamlined Data Refinery. [Felden 2019]

4.1 Metadata Driven Onboarding

Am Anfang eines jeden Data Lake gilt es zunächst, Daten in diesen zu übertragen, so dass analytische Anwendungen möglich werden. Ziel ist es einerseits, möglichst viele Datenquellen zu integrieren, um daraus wiederum möglichst viele Erkenntnisse zu gewinnen, andererseits aber auch eine gewisse Struktur zu wahren, damit der Lake nicht zum Sumpf verkommt. Dabei sei noch einmal auf die Bedeutung der Metadaten hingewiesen, die letztlich die Transparenzschaffung über die Daten ermöglichen. Zudem wächst die Anzahl der Datenquellen mit zunehmendem Projektfortschritt exponentiell, was eine manuelle Integration aufwändig werden lässt und eine entsprechende Governance schwierig macht.

Beim Data Lake sollen die Rohdaten für gewöhnlich ohne größere Bearbeitung im ersten Schritt zunächst in die Speicherstrukturen überführt werden, um dort dann weiterverarbeitet werden zu können (Extraktion-Laden-Transformation [ELT] anstatt Extraktion-Transformation-Laden [ETL]). Die nachfolgende Abbildung zeigt zunächst grundsätzlich den ELT-Ansatz.

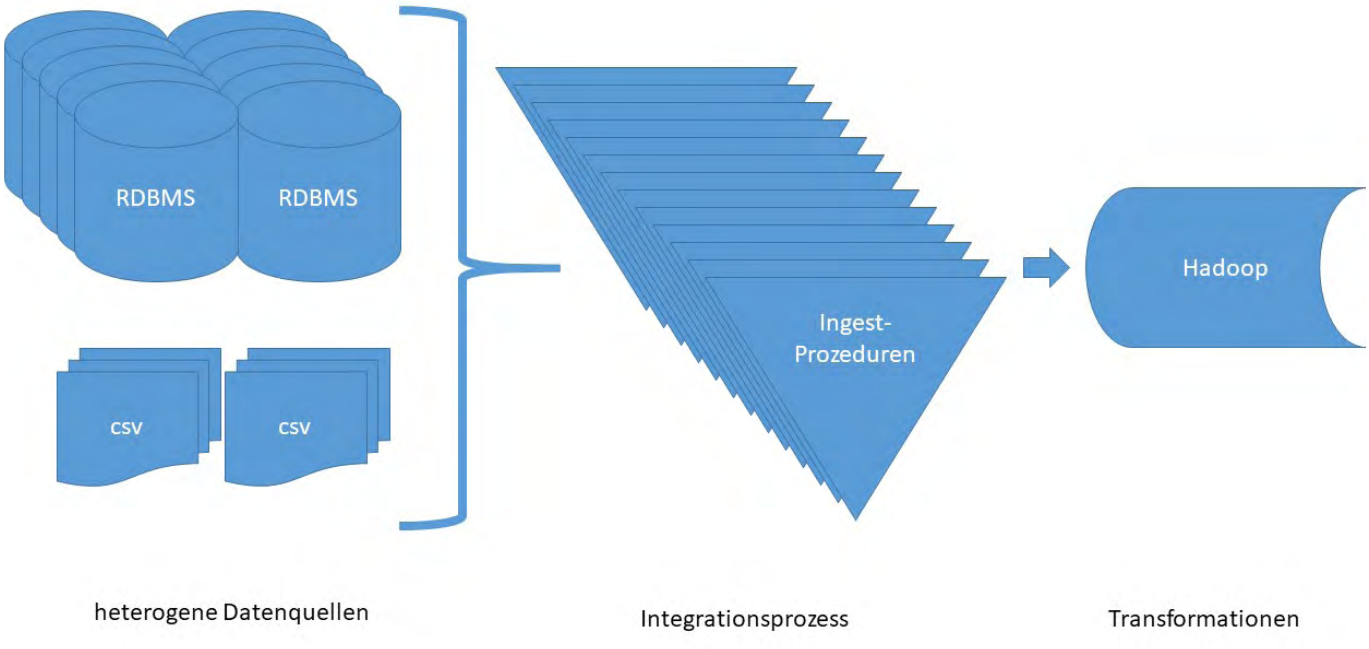


Abb. 4: *Ansatz nach dem Prinzip: Extraktion-Laden-Transformation*

4 Datenflussorientierte Betrachtungen zum Data Lake

Aus heterogenen Datenquellen werden anhand entsprechender Prozeduren Daten geladen und mittels Hadoop transformiert. Ein Großteil der Onboarding-Prozeduren ist hierbei zunächst einfach gehalten, wobei diese aber auch beliebig komplex werden können. Ingest bezeichnet dabei Standardprozeduren aus dem S3-Umfeld, die generell aus zwei Teilen bestehen. Der erste Part ist für das Abrufen der Eingabeposition der Daten in S3 sowie für das Festlegen von Eigenschaften verantwortlich, die vom wiederverwendbaren Teil der Vorlage genutzt werden. Der zweite Teil ist die wiederverwendbare Vorlage. Diese überträgt die Daten zunächst in AVRO oder in ein ähnliches natives HDFS-File-Format. Bei AVRO handelt es sich um einen zeilen- und objektorientierten Container und Remote Procedure Call und Serialisierungs-Framework von Apache, HDFS ist das Hadoop Distributed File System, die beide hier Einsatz finden. Die dann folgende Verarbeitung führt zu einer weiteren Übertragung in Data Marts wie Hive oder Impala. Außerdem werden die Daten zusammengeführt, validiert, profiliert und indiziert.

Typischerweise lassen sich mit einem solchen Ansatz nicht alle, in der Regel aber bis zu 80 Prozent der Integrationsprozesse abdecken. Um der großen Anzahl von Quellen Herr zu werden, fängt hier die Generierung beziehungsweise Nutzung der Metadaten an. Abbildung 5 zeigt diesen Ansatz.

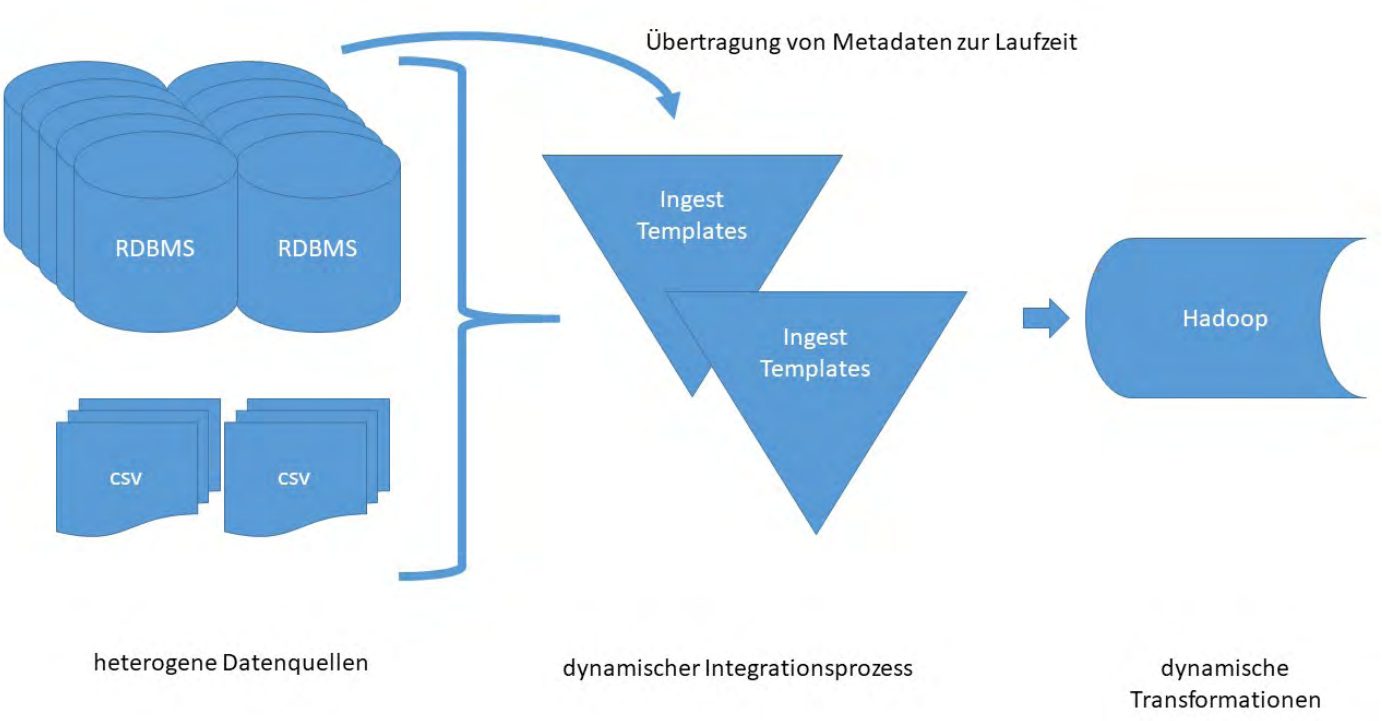


Abb. 5: *Nutzung von Templates und einer Metadata Injection*

Aus den zu integrierenden Datenquellen sind die Metadaten auszulesen, sei es beispielsweise aus Metadatentabellen im Falle relationaler Datenbanksysteme oder über ein Profiling für Files. Die so gewonnenen Metadaten werden direkt in den zentralen Datenkatalog überführt, in dem damit transparent wird, aus welchen Quellen mit welchem Inhalt die Daten stammen und an welchem Ort im Data Lake sie nun vorliegen. Insgesamt reduziert dies auch die Entwicklungszeit und -kosten dieser Datenintegrationsprozesse.

4 Datenflussorientierte Betrachtungen zum Data Lake

Im nächsten Schritt werden die so gewonnenen Metadaten in Onboarding-Templates übertragen, also pro Datenquelle aus dem Template ein Prozess mit den Metadaten der entsprechenden Quelle generiert, der dann diese Quelle in den Data Lake lädt. Abbildung 6 zeigt die Nutzung von Ingestion Templates im dynamischen Integrationsprozess.

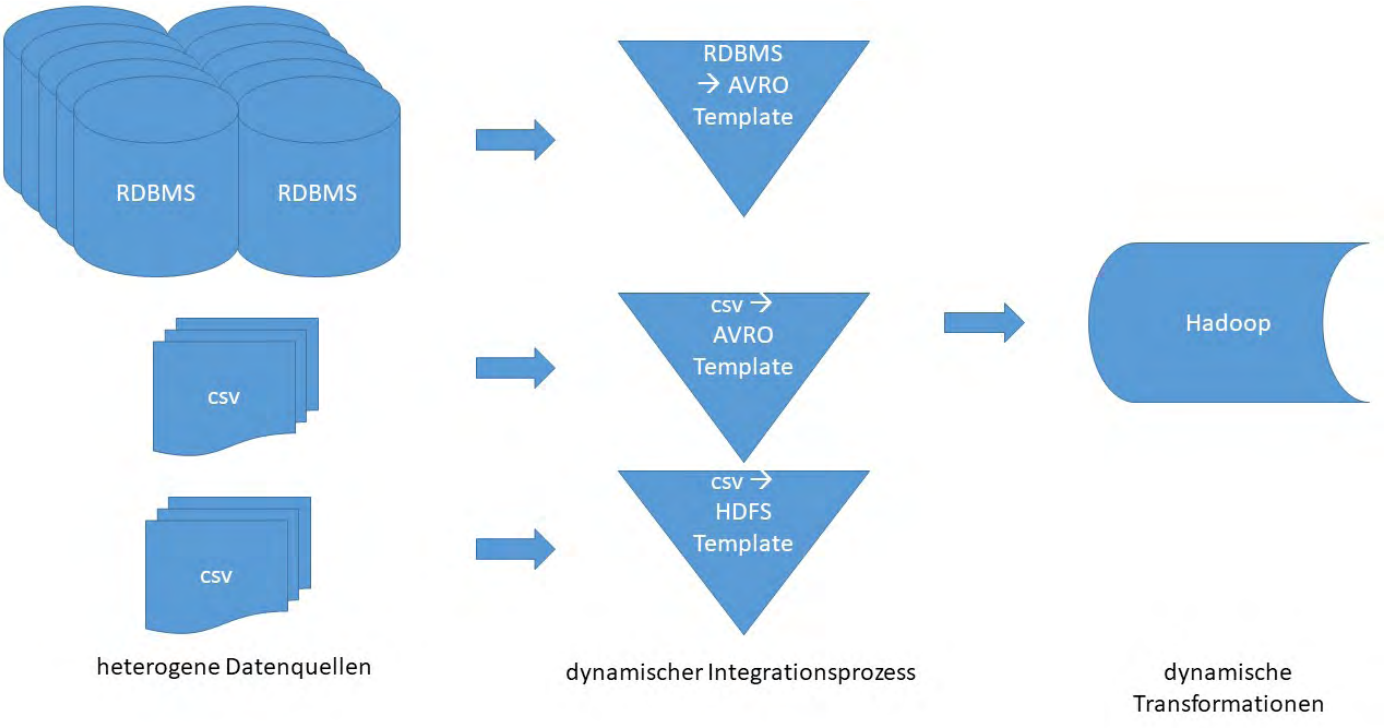


Abb. 6: *Beispielhafte Nutzung von Ingestion Templates*

Die Art der zu ladenden Daten ist oft recht ähnlich, die Anzahl der Quellen dafür sehr hoch. Deshalb kann man hier mit nur wenigen Templates eine größere Anzahl an Datenquellen integrieren und dadurch Effizienzgewinne realisieren, da so ein zentraler Datenkatalog geschaffen wird. Dabei ist es egal, ob sich der Data Lake aus nur einer Lösung, zumeist Hadoop, oder mehreren Systemen wie einem DWH, Hadoop, NoSQL, HCP usw. zusammensetzt.

Diese Herangehensweise bietet die Grundlage für eine Data Governance und ein Data Lineage, für ein intelligentes Zusammenführen kombinierbarer Datenquellen im Data Lake sowie für unterschiedliche analytische Nutzungen der Daten.

4 Datenflussorientierte Betrachtungen zum Data Lake

4.2 Streamlined Data Refinery

Im Kern befasst sich dieser Ansatz mit der umgekehrten Herangehensweise an die Datenverwaltung in einem Data Lake. Im Vergleich zum klassischen DWH, das auf das sogenannte Schema-on-Write setzt, nutzt man für die Rohdaten im Data Lake typischerweise das Schema-on-Read. Statt also wie im DWH die Daten in ein vordefiniertes relationales Korsett zu zwängen, werden die Daten in Rohform abgelegt und erst beim Lesen, also beim Konsum, beziehungsweise bei der Vorbereitung der Datennutzung in eine strukturierte und dabei meist relationale Form umgewandelt.

Nun ist es allerdings nicht zu erwarten, dass die zu definierenden Views beziehungsweise zu berechnenden Aggregate bekannt und definiert und dass spezifische Fragestellungen direkt abbildbar sind. Vielmehr liegt die konstante Veränderung auch in der Datenanalyse in der Natur der Sache. Schon im Data Warehouse ergibt es keinen Sinn, alles vordefiniert zur Verfügung zu stellen, da dadurch kaum wiederverwendbare Aggregate in unbegrenzter Anzahl erstellt würden. Im Data Lake ist dieses Problem aufgrund der beschriebenen Bandbreite an Anwendungen und Anforderungen exponentiell größer. Somit wurde mit der Streamlined Data Refinery ein Ansatz geschaffen, der in einem Projekt täglich Petabytes an Rohdaten abarbeiten kann. Die nachfolgende Abbildung 7 zeigt den schematischen Ablauf dieses Ansatzes.

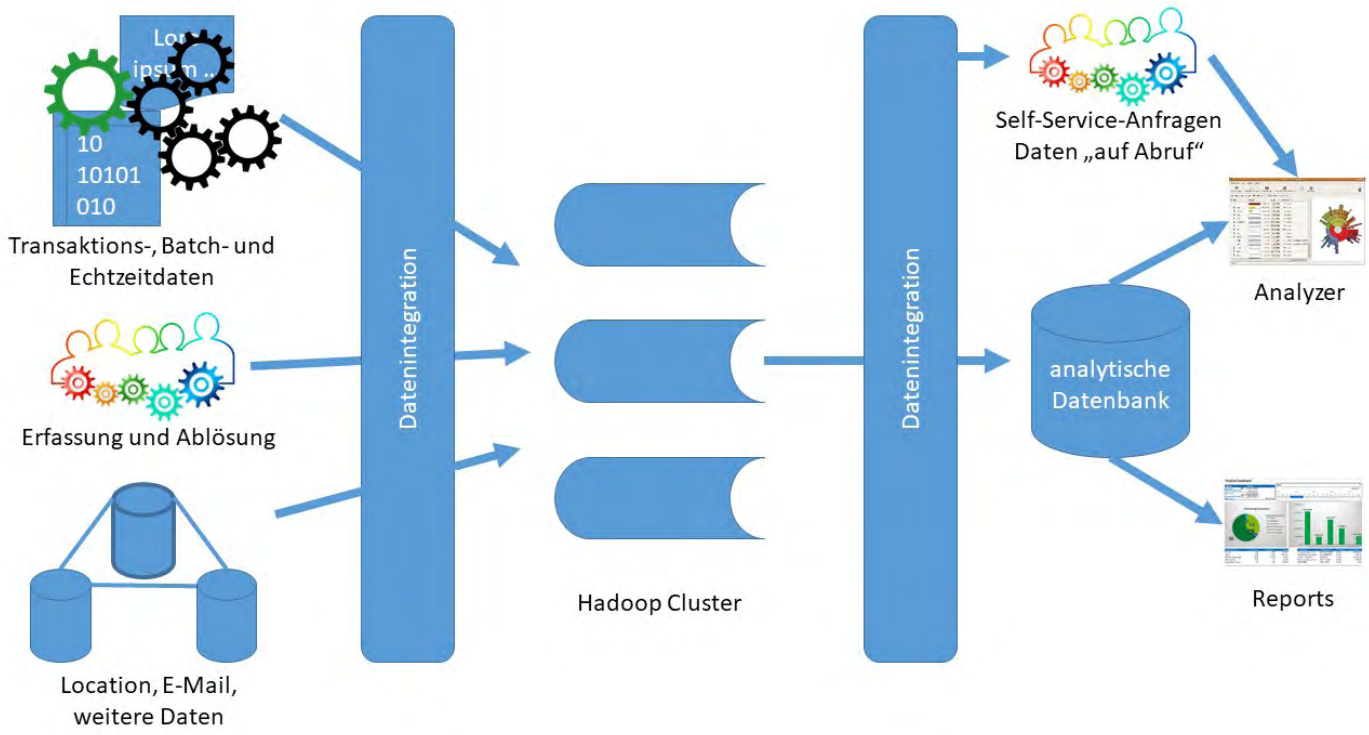


Abb. 7: *Konzept der Streamlined Data Refinery*

Der Lösungsansatz sieht den Einsatz von Datenintegrations-Templates in Kombination mit einem Webinterface vor. Dem Endbenutzer wird eine Oberfläche angeboten, auf der er sich aus einem Katalog von Daten (metadatenbasiert) die gewünschte Untermenge an Daten herauspickt, wobei quellenübergreifend selektiert werden kann. Dieses Interface startet ein entsprechendes Template, das unter Berücksichtigung entsprechender darin verankerter Governance die selektierten Daten einsammelt, aufbereitet und in einem Ad-hoc-Data-Mart bereitstellt. Gleichzeitig lässt sich das Metadatenmodell für eine jeweilige Analyse automatisch generieren. Voraussetzung für diesen Ansatz ist allerdings, dass Datenintegration und Analyse in einer integrierten Plattform vorliegen, die auch bei der Entwicklung den zuvor beschriebenen dynamischen Ansatz unterstützen.

5 Rechtliche und organisatorische Aspekte

Schon durch die Europäische Datenschutz-Grundverordnung sind Unternehmen dazu aufgefordert, Privacy by Design sicherzustellen. Das bedeutet, eine Architektur und einen Datenfluss aufzusetzen, der die rechtskonforme Nutzung von Daten nicht nur ermöglicht, sondern alternativlos macht. Das bringt einerseits die Umsetzung rechtlicher Anforderungen mit sich, die sich in der gestalteten Architektur widerspiegeln. Andererseits ergeben sich organisatorische Aspekte, dabei insbesondere die Zuordnung von Verantwortlichkeiten.

5.1 Rechtliche Perspektive

Insbesondere die Europäische Datenschutz-Grundverordnung wirkt sich auf die Daten aus, die Unternehmen in einem Data Lake bereitstellen und aus diesem nutzen wollen. Die Grundverordnung ist bereits am 25. Mai 2016, zwanzig Tage nach der Veröffentlichung im EU-Amtsblatt, in Kraft getreten. Entsprechend der darin geregelten Übergangsfrist kam sie allerdings erst zwei Jahre nach Inkrafttreten zur Anwendung. Das bedeutet, dass sie seit dem 25. Mai 2018 für alle gilt und ihre Einhaltung durch die EU-Datenschutzaufsichtsbehörden und Gerichte überprüfbar ist. Die nachfolgende Übersicht stellt einige Aspekte zusammen, die im Kontext des Data Lake relevant sind.

Aspekte
Der Anwendungsbereich der Verordnung ist auf alle Verarbeitungen ausgeweitet, die sich an EU-Bürger richten und personenbezogene Daten von EU-Bürgern verarbeiten.
Grundsätzlich ist eine Datenminimierung zu betreiben, die eine redundante Speicherung der Daten ausschließt. Darüber hinaus muss die gewählte Architektur die Ziele des Datenschutzes unterstützen (Privacy by Design).
Die Verarbeitung zu anderen Zwecken als den ursprünglichen Erhebungszwecken ist anders geregelt als im BDSG – Weiterverarbeitung nur bei kompatiblen Zwecken zulässig. Zwar bleibt der Grundsatz der Zweckbindung erhalten – der Wortlaut der allgemeinen Regelung für die Datenweiterverarbeitung ändert sich jedoch. Die Regelung zur Weiterverarbeitung für einen anderen Zweck als den, zu dem die Daten ursprünglich erhoben wurden, findet sich im BDSG in § 28 Abs. 2, während in der Verordnung Art. 6 Abs. 4 maßgeblich ist.
Die Erteilung der Einwilligung erfordert eine freiwillige, spezifisch informierte und eindeutige Handlung. Keine Einwilligung stellen laut Erwägungsgrund 32 zur DS-GVO ein stilles Einverständnis, standardmäßig angekreuzte Kästchen oder Untätigkeit des Betroffenen dar. Zudem fordert die DS-GVO, dass in verschiedene Datenverarbeitungsvorgänge jeweils gesondert eingewilligt werden muss. Andernfalls soll es an der Freiwilligkeit fehlen.
Unternehmen müssen zukünftig dem Betroffenen eine Reihe an weiteren Informationen bereitstellen. Dazu gehören u. a. Informationen zu der Rechtsgrundlage, auf welche sie die Datenverarbeitung stützen, und Angaben zur Dauer der Speicherung oder, falls dies nicht möglich ist, über die Kriterien zur Festlegung der Dauer. Zudem müssen sie neuerdings vor jeder Weiterverarbeitung der Daten zu einem anderen Zweck den Betroffenen erneute Informationen nach Art. 13 und 14 DS-GVO bereitstellen.
Unternehmen sollten seit Mai 2018 in der Lage sein, auf Anfrage personenbezogene Daten, die der Betroffene selbst bereitgestellt hat, in einem gängigen und elektronischen Format dem Betroffenen bereitzustellen.
Unternehmen sollten dokumentieren, welche personenbezogenen Daten sie verarbeiten, woher sie die Daten haben und an wen sie die Daten weitergegeben.
Nach Art. 82 Abs. 1, 4 DS-GVO haftet im Gegensatz zur bisherigen Rechtslage nicht nur der für die Verarbeitung Verantwortliche, sondern auch der Auftragsverarbeiter gegenüber dem Betroffenen im Außenverhältnis gesamtschuldnerisch auf Schadensersatz.
Für die Festlegung der angemessenen technisch-organisatorischen Maßnahmen sind nach Art. 32 DS-GVO verschiedene Faktoren der Datenverarbeitung sowie die Eintrittswahrscheinlichkeit und Schwere des Risikos für die Rechte und Freiheiten natürlicher Personen zu berücksichtigen. Um nachweisen zu können, dass man die technisch-organisatorischen Maßnahmen aufgrund einer solchen umfassenden Betrachtung ausgewählt hat, muss diese Prüfung beziehungsweise ihr Ergebnis dokumentiert werden.
Einer Datenschutzfolgeabschätzung sollte ein adäquates Risikomanagement vorausgehen. Sollte man bei der Risikoabschätzung der einzelnen Datenverarbeitung zu dem Ergebnis kommen, dass diese ein hohes Risiko für die Rechte und Freiheiten des Betroffenen darstellt, muss eine Datenschutz-Folgeabschätzung durchgeführt werden – insbesondere dann, wenn es um eine automatisierte Entscheidung für den Betroffenen geht, massenhaft sensible Daten verarbeitet werden oder systematisch öffentlich zugängliche Bereiche massenhaft beobachtet werden. Auch bei der Einführung neuer Technologien ist eine Datenschutzfolgeabschätzung notwendig.

Tabelle 3: *Ausgewählte Aspekte der Europäischen Datenschutz-Grundverordnung*

5 Rechtliche und organisatorische Aspekte

Die hier genannten ausgewählten Aspekte machen bereits deutlich, dass insbesondere die detailreiche und echtzeitnahe Datenhaltung, wie sie eben im Data Lake vorliegt, durch die Grundverordnung betroffen ist. Zur Abbildung der rechtlichen Perspektive geraten sogenannte Privacy Enhancing Technologies (PET) in den Fokus [hierzu und im Folgenden Dorschel 2015]. Diese PET-Werkzeuge sollen zunächst die Identifizierbarkeit einzelner Individuen verhindern und andererseits die Kontrolle über die Nutzung personenbeziehbarer Daten erhöhen. Im Weiteren betrifft dies dann auch Werkzeug- und Netzwerkanalyse bezüglich der Datentransfers.

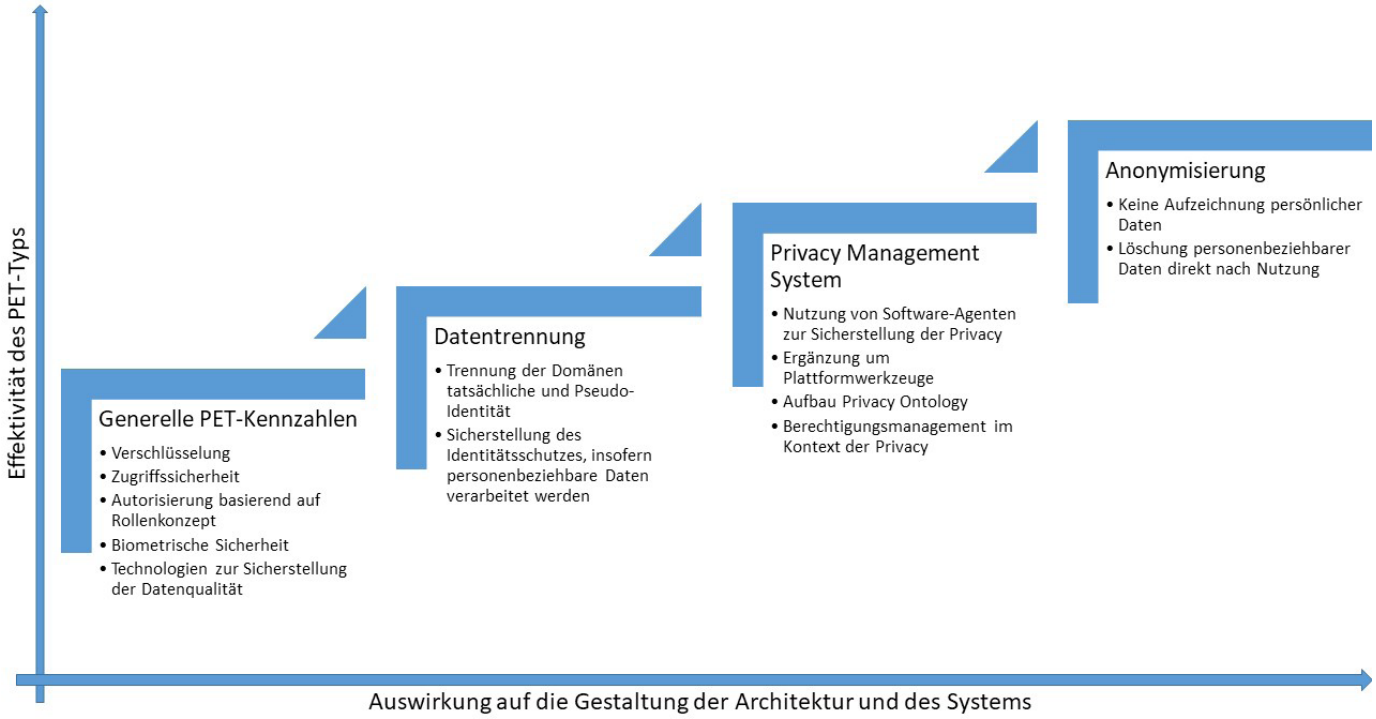


Abb. 8: *Stufen der Privacy Enhancing Technologies*

Die Abbildung zeigt ein PET-Treppendiagramm, in welchem die Wirksamkeit des Schutzes personenbeziehbarer Daten vom verwendeten PET-Typ abhängt. Die PET-Treppen stellen kein Wachstumsmodell im engeren Sinne dar und müssen nicht bis zum Ende und damit der höchsten Stufe betrieben werden, so dass also, wenn eine Organisation allgemeine PET-Maßnahmen implementiert hat, dies nicht bedeutet, dass die Organisation zu höheren PET-Ebenen heranwachsen muss. Die Eignung verschiedener PET-Typen hängt hauptsächlich von der Art und Komplexität des Informationssystems, den angestrebten Ambitionen und der Sensibilität personenbezogener Daten ab und ist somit immer eine Individualentscheidung in einem Unternehmen.

Im Rahmen analytischer Aktivitäten besteht grundsätzlich beispielsweise die Möglichkeit, mittels des sogenannten Anonymize-then-Mine-Ansatzes Daten in einer Art und Weise zu erheben, dass kennzeichnende Merkmale von Anfang an nicht erfasst werden. Dies bedeutet, dass lediglich anonyme oder pseudonyme Daten erhoben werden. Somit ist ein Extraktionsprozess aufzustellen, der von Anfang an nur die Daten hochlädt, die der zuvor genannten Anforderung genügen, also im Umkehrschluss Daten weglässt. Dies steht allerdings konträr zur ursprünglichen Idee des Data Lake, zunächst Daten zu sammeln, da man nie wissen kann, welche dieser Daten in einem dann relevanten Kontext tatsächlich hilfreich sein werden. Ein Hauptvorteil des Anonymize-then-Mine-Ansatzes

5 Rechtliche und organisatorische Aspekte

ist allerdings die Entkopplung von Datenschutz und Data Mining. Eine solche Entkopplung hat zwei Vorteile:

- Sie garantiert eine „sichere" Data-Mining-Praxis, da sie mit dem anonymisierten Datensatz durchgeführt wird und per Definition die Data-Mining-Ergebnisse nicht die Anonymität der Originaltabelle verletzen können.
- Sie bietet ein gewisses Maß an Flexibilität, da das Data Mining von anderen Nutzern (also nicht dem Dateninhaber) durchgeführt werden kann. Dies entspannt den Zustand, wenn der Dateninhaber kein Data-Mining-Experte ist.

Mit diesem Ansatz sind jedoch zwei Herausforderungen verbunden:

- Lassen sich solche anonymisierten Daten effektiv verwenden, um genaue Modelle zu erstellen?
- Könnte das resultierende Modell, also das mit anonymisierten Daten erstellte Modell, selbst die Privatsphäre verletzen?

Bei einigen Data-Mining-Aufgaben ist es möglich, effektive Antworten auf die erste Frage zu finden. Bei der Erstellung eines Entscheidungsbaums ist es beispielsweise möglich, die Domäne der Quasi-Bezeichnerattribute beim Hinzufügen eines verallgemeinerten Werts zu erweitern. Zum Beispiel ist die Top-Down-Spezialisierung eine Anonymisierungstechnik zur Erstellung genauer Entscheidungsbäume auf der Grundlage anonymisierter Datensätze. Bei diesem Ansatz werden die Daten gemäß des klassenbedingten Entropiemaßes anonymisiert. Natürlich lassen sich auch direkt anonymisierte Daten verwenden, um eine Klassifizierung durchzuführen. Trotzdem wird bei einigen anderen Data-Mining-Aufgaben die Mustererkennung in anonymisierten Daten schwieriger. Ein Beispiel für eine solche Schwierigkeit ist der Fall, in dem eine Distanzberechnung zwischen Attributen erforderlich ist.

Zur Beantwortung der zweiten Herausforderung gibt es bisher nur wenige Studien, die sich mit potenziellen Verstößen gegen die Privatsphäre im Fall von Assoziationsregel-Mining beschäftigen. [Aggarwal et al. 2006] argumentieren, dass die Unterdrückung der sensiblen Werte, die von einzelnen Datenlieferanten ausgewählt werden, unzureichend ist, da Assoziationsregeln verwendet werden könnten, die aus den Daten entstanden sind, um die unterdrückten Werte zu schätzen. In ihrer vorgeschlagenen Lösung führten sie einen heuristischen Algorithmus ein, um einen minimalen Satz von Einträgen auszublenden und die Verletzung der Privatsphäre durch solche Angriffe zu verhindern. [Verykios et al. 2004a] schlugen Algorithmen vor, um sensible Zuordnungsregeln in einem bestimmten Transaktionsdatensatz zu verbergen. Ihre allgemeine Idee war es, jeweils die Regel zu verbergen, die das Vertrauen oder die Unterstützung für die Ergebniseinschätzung verringert. Daher werden die Regeln entfernt, die keine bestimmte Mindestunterstützung und kein Mindestvertrauen erfüllen.

Bei der Methode Mine-then-Anonymize werden die Data-Mining-Modelle auf der Grundlage der ursprünglichen Daten erstellt, anschließend wird der Anonymisierungsprozess auf den Data-Mining-Ergebnissen angewendet. Dabei wird es im Allgemeinen als ausreichend angesehen, die Muster und nicht die Daten selbst zu anonymisieren, wenn das Ziel darin besteht, Data-Mining-Ergebnisse zu veröffentlichen. Dieses Vorgehen führt zu einem besseren Informationsnutzen als die Durchführung des Data Mining mit anonymisierten Daten. Bei der Erstellung der Modelle unter Verwendung der Originaldaten muss die Anonymität entweder nach Abschluss des Data Mining (eine zweistufige Methode) oder innerhalb des Data-Mining-Prozesses selbst (eine einstufige Methode) erfüllt sein [Ciriani et al. 2008]. In beiden Fällen sollten die gewonnenen Ergebnisse keinen Rückschluss auf die Existenz über Quasi-Identifikatorwerte ermöglichen. Die einstufige Methode erfordert jedoch eine Änderung des Data-Mining-Prozesses. Dies bedeutet, dass die genutzten Algorithmen geändert werden müssen, um die Anonymität direkt umzusetzen. In der Regel ergeben sich jedoch Leistungsvorteile [Ciriani et al. 2008]. [Friedman et al. 2008] identifizieren zwei Vorteile, wenn die Data-Mining-Techniken als Grundlage für die Anonymisierung verwendet werden:

- Zunächst werden die Anonymisierungsalgorithmen optimiert, um bestimmte Datenmuster basierend auf der Data-Mining-Technik beizubehalten.
- Im Weiteren können, wenn die Anonymisierungsalgorithmen auf der Grundlage von Data-Mining-Techniken verwendet werden, anstelle der

5 Rechtliche und organisatorische Aspekte

gleichen Generalisierung für alle Tupel verschiedene Generalisierungen für mehrere Tupel-Gruppen angewendet werden. Dies führt dazu, dass nützlichere Informationen erhalten bleiben.

Daraus folgt, dass einer der Hauptvorteile dieses Ansatzes eine höhere Qualität der Data-Mining-Ergebnisse und eine höhere Effizienz ist. Dieser Ansatz wurde in Bezug auf Assoziationsregeln, Entscheidungsbaum und Clustering untersucht [Friedman et al. 2008]. Ein Hauptnachteil dieses Ansatzes ist, dass Data Mining nur vom Dateninhaber durchgeführt werden kann [Ciriani et al. 2008]. Eine solche Anforderung wirkt sich negativ auf die Anwendbarkeit des Verfahrens aus, da der Dateninhaber implizit ein Data-Mining-Experte sein muss. Die folgende Tabelle fasst die Unterschiede zusammen.

Ansatz	Vorteil	Nachteil
Anonymize-then-Mine	Flexibel: Ermöglicht anderen Parteien als dem Dateneigentümer das Durchführen von Data Mining	Anonymisierung bewirkt vor allem in spärlichen und hochdimensionalen Datensätzen Informationsverlust durch Anonymisierung, der sich negativ auf die Qualität der Data-Mining-Ergebnisse auswirken kann.
Mine-then-Anonymize	Bessere Modelle werden erhalten, da Data-Mining-Algorithmen auf Originaldaten angewendet werden. Es ist möglich, verschiedene Verallgemeinerungen für mehrere Tupelgruppen zu haben.	Nicht flexibel: Data Mining kann nur vom Dateneigentümer durchgeführt werden.

Tabelle 4: *Prinzipielle Ansätze zur Umsetzung von Datenschutzanforderungen in der Analyse*

Alternativ besteht mit dem Privacy Preserving ein weiterer Ansatz. Bei diesem werden die Daten vollständig aus den Quellen geladen und im Data Lake abgelegt. Erst zur Nutzungszeit wird entschieden, welche für die Analyse gewählten Daten ein datenschutzrechtliches Problem darstellen und dann entsprechend zur Sicherstellung des Datenschutzes zu bearbeiten sind. Dieses Bearbeiten kann ein einfaches Weglassen der Daten sein, genauso aber auch eine Pseudonymisierung zur Laufzeit. Dieser Ansatz bedingt aber, dass der Analyseprozess und das Analyseziel klar sind, so dass eine solche Entscheidung getroffen werden kann. Zur Dokumentation ist im Rahmen eines Data Lineage entsprechend auch zu vermerken, welche Daten einbezogen wurden und welche nicht. Das bedeutet allerdings, dass das Data Lineage eine dynamische Natur erhält, da es letztlich analyseabhängig ist.

Die vorherige Darstellung zeigt auf, dass Big-Data-Analysen nur dann datenschutzkonform sind, wenn sie in Vereinbarkeit mit den gesetzlichen Vorgaben bei Erhebung und Speicherung umgesetzt werden – also wenn unter anderem der Data Lake bereits angemessen konzipiert wurde. Die Datenverarbeitung ist also zum Beispiel unter Einsatz der vorgenannten Privacy Preserving Technologies so zu modellieren, dass entweder anonyme, anonymisierte oder pseudonymisierte Daten erhoben beziehungsweise (weiter-)verarbeitet werden. Alternativ besteht nur die Möglichkeit, dass für jeden Verarbeitungsschritt ein gesetzlicher Erlaubnistatbestand vom jeweiligen Betroffenen eingeholt wird, was in der Praxis jedoch nicht realistisch ist. Zur Veranschaulichung sollen nachfolgend anhand der jeweils genannten Datenverarbeitungsschritte geeignete Maßnahmen skizziert werden.

Im Allgemeinen hat jedes Unternehmen zu beachten, dass zur Gewährleistung der Datenschutzanforderungen schon die Datenerhebung auf das zwingend erforderliche Maß zu beschränken ist. Dabei gilt der Grundsatz, dass nur anonyme Daten zu erheben oder die Daten unmittelbar nach der Erhebung zu anonymisieren oder zu pseudonymisieren sind. Ausnahme bildet jedoch an dieser Stelle, dass die Betroffenen, wie schon zuvor erwähnt, einer weiteren Analyse zugestimmt haben können und damit dann ein alternatives Analyseziel verfolgt wird. Um späteren Nachweispflichten gerecht zu werden, sollte allerdings

5 Rechtliche und organisatorische Aspekte

bereits bei Erhebung die Zweckbindung und damit auch Nichtverkettbarkeit als Kriterium berücksichtigt werden und damit die Speicherung determinieren. Eine getrennte Erhebung sollte daher in der Konsequenz zu einer getrennten Speicherung führen. Dies würde, wie bereits benannt, auch die Transparenz gegenüber den Betroffenen erhöhen bzw. die Informationsberichterstattung über die erfassten Daten vereinfachen.

Nicht nur für das Sammeln von Daten ist eine Befugnis für die verantwortliche Stelle einzuholen, auch für die weiteren Verarbeitungsschritte und damit insbesondere für die Weiterverarbeitung ist eine solche erforderlich. Im Zusammenhang mit der Speicherung sind die zuvor dargestellten Aspekte und demnach eine dezentrale Datenhaltung, Anonymisierungs- oder Pseudonymisierungsmaßnahmen oder andere Privacy Enhancing Technologies und auch eine frühestmögliche Löschung nicht (mehr) benötigter Daten zu erwägen. Eine Ausnahme davon besteht, wenn ein legitimiertes Interesse vorhanden ist oder aber die Daten aus einer allgemein zugänglichen Quelle stammen, da für diese dann keine Zweckbindung mehr vorliegt. In beiden Fällen ist das Speichern aber nur dann zulässig, wenn es zusätzlich auf einen bestimmten neuen Zweck begrenzt ist und die schutzwürdigen Interessen des Betroffenen nicht überwiegen.

Die Nutzung von Daten aus dem Data Lake dient nun dem Analysezweck, Maschinelles Lernen oder Künstliche Intelligenz anzuwenden und so bestimmte Fragestellungen zu beantworten oder für den Fall relevante Eigenschaften und Informationen zu einem entsprechenden Profil zusammenzufügen. Typische Beispiele sind das Verfolgen des Nutzungsverhaltens im Internet oder eine Kreditwürdigkeitsprüfung. Da diese Auswertungen oft die Zielsetzung verfolgen, das Verhalten der Person vorhersagen und beeinflussen zu können, und damit in einem besonderen Spannungsverhältnis zum Recht auf informationelle Selbstbestimmung stehen, können entsprechend personenbezogene Analysen nur auf Grundlage einiger weniger Legitimationsbestände rechtskonform begründet werden. Letztlich sind es Einzelfallentscheidungen, die sich durch die Zustimmung des Betroffenen ergeben. Deshalb muss der jeweilige Datenschutzbeauftrage hier intensiv miteinbezogen werden, so dass die gesamte Prozesskette, beginnend bei der Datenerhebung bis hin zur Datennutzung verbunden mit den dazugehörigen Auskunftspflichten, angemessen aufgestellt ist. Ergänzend ist zu beachten, dass die Nutzung der Daten einer Entscheidungsunterstützung dient, jedoch nur eingeschränkt einer Entscheidungsautomation. Unzulässig wäre diese im Falle, dass Entscheidungen zum Nachteil des Betroffenen getroffen werden. Dies bezieht sich auf alle diejenigen Entscheidungen, die eine Rechtsfolge begründen und auf der Bewertung von Persönlichkeitsmerkmalen beruhen. Somit liegt im Umkehrschluss eine zulässige Entscheidungsautomation vor, wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung basierend auf den expliziten Daten einer natürlichen Person stattgefunden haben.

Damit zeigt sich, dass die abstrakte Auswertung von Daten unproblematisch ist und der Data Lake, aus rechtlicher Sicht und der Einfachheit halber, bereits von Beginn an so gestaltet werden sollte. Dadurch wird das Hauptaugenmerk bei den Analysen auf allgemeinen Entwicklungen, Strukturen und Mustern liegen, die dann von einem menschlichen Aufgabenträger im jeweiligen Kontext zu interpretieren sind. Beziehen sich die Ergebnisse also auf ausreichend große Gruppen von Betroffenen, so dass einzelne Betroffene nicht erkannt werden können, ist diese Form der Auswertung mangels Personenbezug datenschutzrechtlich zulässig [Roßnagel 2013]. Dabei ist im Sinne eines Informationsmanagements durch rechtliche, technische oder organisatorische Maßnahmen sicherzustellen, dass nicht über eine statistische Auswertung doch noch ein Personenbezug herstellbar ist.

Im Allgemeinen haben sich durch die Europäische Datenschutz-Grundverordnung vielfältige Aufgaben ergeben, die im Kontext analytischen und datengetriebenen Handelns in einem Unternehmen zu beachten sind. Insgesamt ergeben sich die nachstehenden zentralen Tätigkeiten, die in einem Unternehmen umzusetzen sind, um den Anforderungen gerecht zu werden. Dabei können diese im Einzelfall umfassender sein, so dass die nachstehende Auflistung als Information über allgemeingültige Herausforderungen zu verstehen ist.

5 Rechtliche und organisatorische Aspekte

- Dokumentation der Datenverarbeitungsprozesse im Unternehmen (insbesondere Erweiterung der Dokumentationspflichten bei Auftragsverarbeitern, möglicherweise zusätzliche Dokumentationserfordernisse für Risk und Privacy Impact Assessment)Datenschutzerklärungen (Erweiterung der Informationspflichten)Einwilligungserklärungen (Verschärfung der formalen Vorgaben), Prozess für Widerruf der EinwilligungAnpassung der Betriebsvereinbarungen an DS-GVOProzesse zur Umsetzung von WidersprüchenVereinbarungen zur Auftragsverarbeitung (Haftungsregelung, Dokumentation) Prozess bei Datenpannen entsprechend der neuen Vorgaben überarbeitenVerfahren, um Daten in einem gängigen elektronischen Format übertragen zu könnenDurchführung von zielgruppengerechten Schulungen zu den Neuerungen der DS-GVO und den eigenen ProzessenEinführung von Risk Assessment zur Festlegung geeigneter technisch-organisatorischer MaßnahmenEinführung eines Privacy Impact AssessmentMonitoring nationaler Gesetzgebung und Fortbildung

5.2 Organisatorische Perspektive

Data Lakes sind durch ihr oftmals zugrunde liegendes technisches Verständnis bei den Data-Warehouse-Verantwortlichen beziehungsweise in IT-Abteilungen verortet. Letztlich sind sie aber immer in die Rollenkonstrukte der Verantwortlichkeiten miteinzubeziehen, da auch hier, in der bekannten Analogie zum Data Warehouse, unter anderem Datenquellen-, Datenqualitätsverantwortliche, Datenbank- und Berichtsentwickler etc. zu bestimmen sind, um die Nutzungsqualität eines solchen Systems langfristig sicherzustellen. Zu den Data-Lake-Mitarbeitern gehören Data Engineers, Data Architects, Business Analysts, Developer und Data Scientists. Ein Drittel davon haben eher eine Beratungsfunktion, die anderen eine Realisierungsfunktion. TDWI Research zeigte in seiner Studie auf, dass primär Datenverwaltungsfachleute im Data-Lake-Kontext relevant sind, die in den Bereichen Big Data, Hadoop und Advanced Analytics geschult sind. Die meisten Data Lakes konzentrieren sich generell auf die Unterstützung analytischer Aufgaben, andere fallen jedoch in Kategorien, die auf ihren Eigentümern oder Anwendungsfällen basieren – zum Beispiel Vertrieb, Gesundheitswesen und Betrugserkennung – und für diese explizit Daten bereitstellen. Die meisten Anwendungsfälle für Data Lakes erfordern Geschäftsmetadaten, Self-Service-Funktionen, SQL, mehrere ETL-Methoden und mehrschichtige Sicherheit. Hadoop ist in diesen Bereichen schwach aufgestellt, daher füllen Benutzer die Lücken von Hadoop mit mehreren Werkzeugen von Anbietern und/oder Open-Source-Communities. So oder so zeigt sich aber, dass Kompetenzen verbunden mit Verantwortlichkeiten vorhanden sein müssen, um einen Data Lake in einer analytischen Architektur zielorientiert und langfristig erfolgreich betreiben zu können. In diesem Zusammenhang hat das Analytics Cooperation Center (ACC) eine Prominenz, da es als Ansatz basierend auf dem Business Intelligence Competence Center (BICC) Themen zusammenträgt und somit ein Management der Gesamtstruktur zugunsten eines strategischen Handelns ermöglicht.

Nachfolgende Abbildung zeigt mögliche Gestaltungsformen des ACC, deren Auswahl klar in Bezug zu existierenden Strukturen, organisatorischen Rahmenbedingungen und strategischer Ausrichtung des jeweiligen Unternehmens auszugestalten ist. Ein Pol ist die Zentralisierung der analytischen Kompetenz, die dann auch verantwortlich die Themen in einem Unternehmen betreibt. Der andere Pol ist die dezentrale Ansiedlung der Kompetenz in den Fachbereichen. Dazwischen ergeben sich unterschiedliche Varianten. Der Center-of-Excellence-Ansatz betont beispielsweise, dass die Themen aus den Fachbereichen kommen und dort in Projekten beheimatet sind. Das bedeutet auch, dass Datenbeschaffung in den Data Lake dort zu steuern ist, da die Fach- oder Funktionsbereiche ihre eigenen Themen aufsetzen. Data Scientists wirken als Service in die zu schaffenden Projektteams hinein und unterstützen damit den Fachbereich, dessen Themen in operativ nutzbare Ergebnisse zu übertragen. Bei der unterstützenden Variante entsteht das Projekt aus dem Funktionsbereich und ist dann dem Fachbereich zugeordnet, um die Themenbindung zu

5 Rechtliche und organisatorische Aspekte

erzielen. Somit erhält der Funktionsbereich in diesem Zusammenhang die entsprechende Prominenz. Auch hier wird analytische Kompetenz als Service hinzugezogen. Bei der funktionalen Alternative existiert die analytische Kompetenz im Funktionsbereich, der damit auch die Fachbereiche unterstützt und somit die steuernde Hoheit über die Themenentwicklungen in einem Hause erhält.

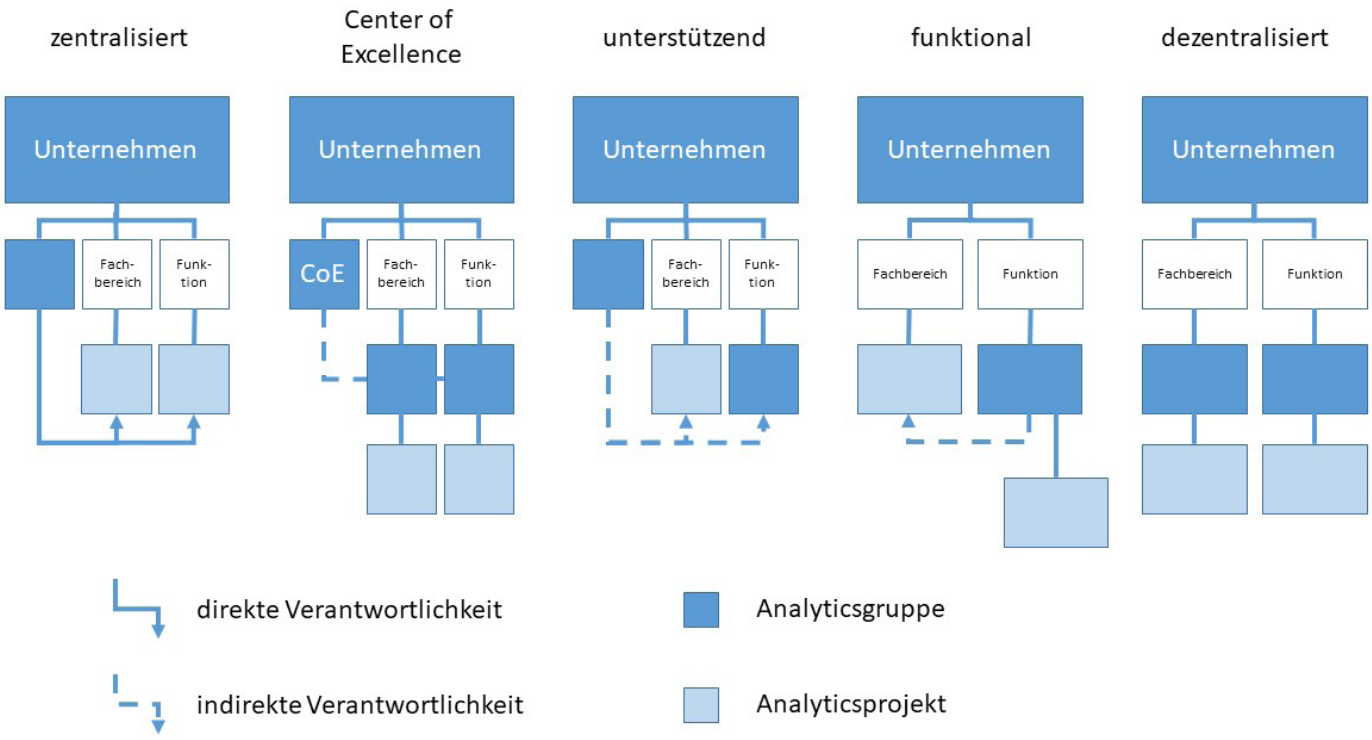


Abb. 9: *Organisatorische Varianten des ACC*

Grundsätzlich kann das Analytics Cooperation Center als eigene organisatorische Einheit aufgebaut werden. Es besteht dabei exklusiv in dem zuvor genannten definitorischen Themenbereich und bewegt sich im Spannungsfeld der Digitalisierung und der Business Intelligence. Dabei ist es normalerweise keine Weiterentwicklung eines eventuell vorhandenen BICC, welches nach wie vor für die Themen der Business Intelligence zuständig sein wird. Die organisatorische Einheit des ACC besteht aus einer definierten Anzahl von Data Scientists, wobei einer die verantwortliche Rolle des ACC-Koordinators (Leiter ACC) übernimmt. Darüber hinaus gibt es einen technischen und einen fachlichen Datenadministrator. Der Erstgenannte ist für die technische Betreuung des Data Lake und damit der zur Verfügung stehenden Rohdaten zuständig. Der fachliche Vertreter betreut Datenquellen und schafft die Verbindung zwischen der technischen Ablage und der fachlichen Nutzung. Insgesamt bindet der Themenkomplex Datensammlung, -aufbereitung und -bereitstellung in analytischen Projekten intensiv Ressourcen, da er circa 80 Prozent des Zeitaufwands darstellt. Durch die Zentralität im ACC lassen sich an dieser Stelle Skaleneffekte bewirken. So kann in den Fachbereichen die Konzentration auf der Auswertung liegen. Im Allgemeinen gilt, dass im Fachbereich ein Mitarbeiter pro Thema vorgesehen wird, unter der Rahmenbedingung, dass Themen seriell bearbeitet werden und die Servicefunktion des ACC wahrgenommen wird. Das bedeutet, dass Fachbereiche die Kompetenz nicht zwingend initial aufbauen müssen, sondern diese auch zentral abzurufen ist.

5 Rechtliche und organisatorische Aspekte

Das ACC als solches fungiert folglich insgesamt als Service Center, indem es einerseits Themen aus den Fachbereichen aufnimmt und für diese auch bearbeitet, andererseits aber auch Kapazitäten für die Fachbereiche zur Verfügung stellt, die diese zur Ergänzung der eigenen analytischen Kompetenz benötigen. Auch eine Unterstützung bei der Auswahl und Zusammenarbeit mit externen Anbietern kann erfolgen. Damit bietet das ACC den Service des Lastspitzenausgleichs für Abteilungen mit eigenen Kapazitäten beziehungsweise eine Teamressource für Abteilungen ohne dauerhafte Themen oder kann als zentraler Bearbeiter für relevante Themen agieren. Damit wird die Möglichkeit geschaffen, Know-how an zentraler Stelle im Unternehmen aufzubauen und von dort aus in die Abteilungen der Gesellschaften auszubreiten und Synergien zu erzeugen. Die analytische Kompetenz der Fachbereiche hat parallel die Aufgabe der Missionarstätigkeit, die bereits aus den allgemein verfügbaren BICC-Empfehlungen bekannt ist. So sollen Fachbereichsvertreter/innen befähigt werden, Ideen zu entwickeln und schneller in eine operative Nutzung zu gelangen. Gleichzeitig unterbindet dieser Ansatz nicht die Möglichkeit, externe Ressourcen einzukaufen, da dies letztlich in Analogie zur Einbindung interner Ressourcen stattfindet und somit auch konform zur IT-Strategie im Sinne der Reduktion der eigenen Wertschöpfungstiefe zu sehen ist. Letztlich verlangt jeder Ansatz, dass für seine Umsetzung ein entsprechendes Budget zur Verfügung steht.

5.2.1 Der Head of ACC

Der Leiter hat die Personalverantwortung für die Mitglieder des ACC. Weiterhin verantwortet er die Organisation der unternehmensindividuellen Analytics Community und des Analytics Governance Gremiums. Er sollte die Projektauswahl begleiten und auch organisationsweit aktiv sein, um Analytics als Methode zu etablieren und Mehrwerte für Fachbereiche deutlich zu machen. Bei der Projektauswahl bzw. der vorangegangenen Bewertung ist es wichtig, ein Bewertungsrahmenwerk im Sinne von Leitlinien zu haben, so dass bei knappen Ressourcen objektiv und nachvollziehbar eine Auswahl erfolgen kann.

5.2.2 Das Project Lab

Das Project Lab benennt das eigentliche Team für ein Analytics-Projekt. Dieses Team stellt eine personelle Ressourcenkombination aus den Fachbereichen und dem ACC dar. Hier wird das eigentliche analytische Artefakt entwickelt. Um eine Überlappung mit Linientätigkeiten im Arbeitsalltag zu vermeiden, können getrennte physische Räume etabliert werden.

5.2.3 Die Data Science Community

Die Analytics Community betreibt den inhaltlichen Austausch auf operativer Ebene, so dass die Wissensbasis im Unternehmen in Bezug auf Analytics selbst und die Anwendung des verfügbaren Wissens und der Kompetenz in Analytics-Projekten verbreitert werden kann. Eine Zielstellung ist es, dass Erfahrungen mit Modellen ausgetauscht oder erfolgreiche Umsetzungen auf mögliche Synergieeffekte geprüft werden können. Ein weiterer wesentlicher Aspekt des Austausches ist die Transparenzschaffung über Themen, die vielleicht in mehr als einer Fachabteilung existieren.

5 Rechtliche und organisatorische Aspekte

5.2.4 Die Fachbereiche

Die Fachbereiche sind und bleiben der zentrale Ort der Ideenfindung und können Inspiration für Analytics-Projekte aus dem Tagesgeschäft generieren. Auch die Nutzung und damit die eigentliche Weiterentwicklung der analytisch erzielten Ergebnisse findet im Fachbereich statt.

5.2.5 Die IT

Die IT-Abteilung ist grundsätzlich für die Bereitstellung der technischen Rahmenbedingungen zuständig. Sie umfasst den Betrieb und die Weiterentwicklung der datenorientierten Infrastruktur (Data Warehouse und Data Lake) sowie der Entwicklungsumgebungen (beispielsweise R und Python, aber auch grafische Werkzeuge). Dabei kann es sich auch um die Übertragung und Betreuung der erstellten Ergebnisartefakte in den operativen Betrieb handeln. Gleichzeitig kann die IT-Abteilung Nachfrager der angebotenen Service-Leistungen des ACC sein, um eigene Themen anzugehen.

5.2.6 Das Analytics-Governance-Gremium

Begleitet wird das ACC von einem Governance-Gremium, in dem die unterschiedlichen Bereiche des Unternehmens vertreten sind. Das Gremium selbst ersetzt dabei nicht den IT-Bereich bezüglich technischer Aufgabendefinitionen. Es evaluiert regelmäßig die Leitlinien des analytischen Handelns im Unternehmen und betrachtet dazu die Entwicklung des Transformationsprozesses. Es werden Leitplanken definiert, welche die Priorisierung der Themen-/Projektvorschläge ermöglichen, und damit wird auch ein Einfluss auf die Projekte des ACC genommen. Bezüglich der Definition und der Priorisierung sowie der Abarbeitungsreihenfolge des Analytics-Projektportfolios gibt es zwei mögliche Ausgestaltungen. Einerseits kann der ACC-Leiter basierend auf den Leitlinien des Gremiums das Analytics-Portfolio definieren und priorisieren. Andererseits kann das Governance-Gremium diese Aufgabe übernehmen. Für die Übernahme durch die ACC-Leitung spricht die Geschwindigkeit und der zu vermeidende Overhead.

Darüber hinaus erfolgt hier die Abstimmung mit anderen Gremien im jeweiligen Unternehmen wie dem BICC-Governance-Gremium. Durch die Beteiligung aller Fachabteilungen entsteht Transparenz über den Umsetzungsstand analytischer Initiativen beziehungsweise im Einsatz befindlicher Ansätze, um so in der Community Lerneffekte und Recycling bestehender Lösungen zu ermöglichen. Die Treffen des Gremiums sollten alle vier Wochen erfolgen. Die Größe des Gremiums resultiert aus der Anzahl der beteiligten Bereiche, die jeweils einen Vertreter entsenden sollten, so dass Transparenz über die einzelnen Aktivitäten gewahrt bleibt. Der ACC-Leiter sollte ein weiteres ständiges Mitglied des Gremiums sein und hier auch über den Erfolg der Projekte berichten.

5 Rechtliche und organisatorische Aspekte

Die in der folgenden Abbildung dargestellten Funktionsblöcke fassen zusammen, wie die Verantwortlichkeiten im Analytics-Ökosystem aussehen sollten.

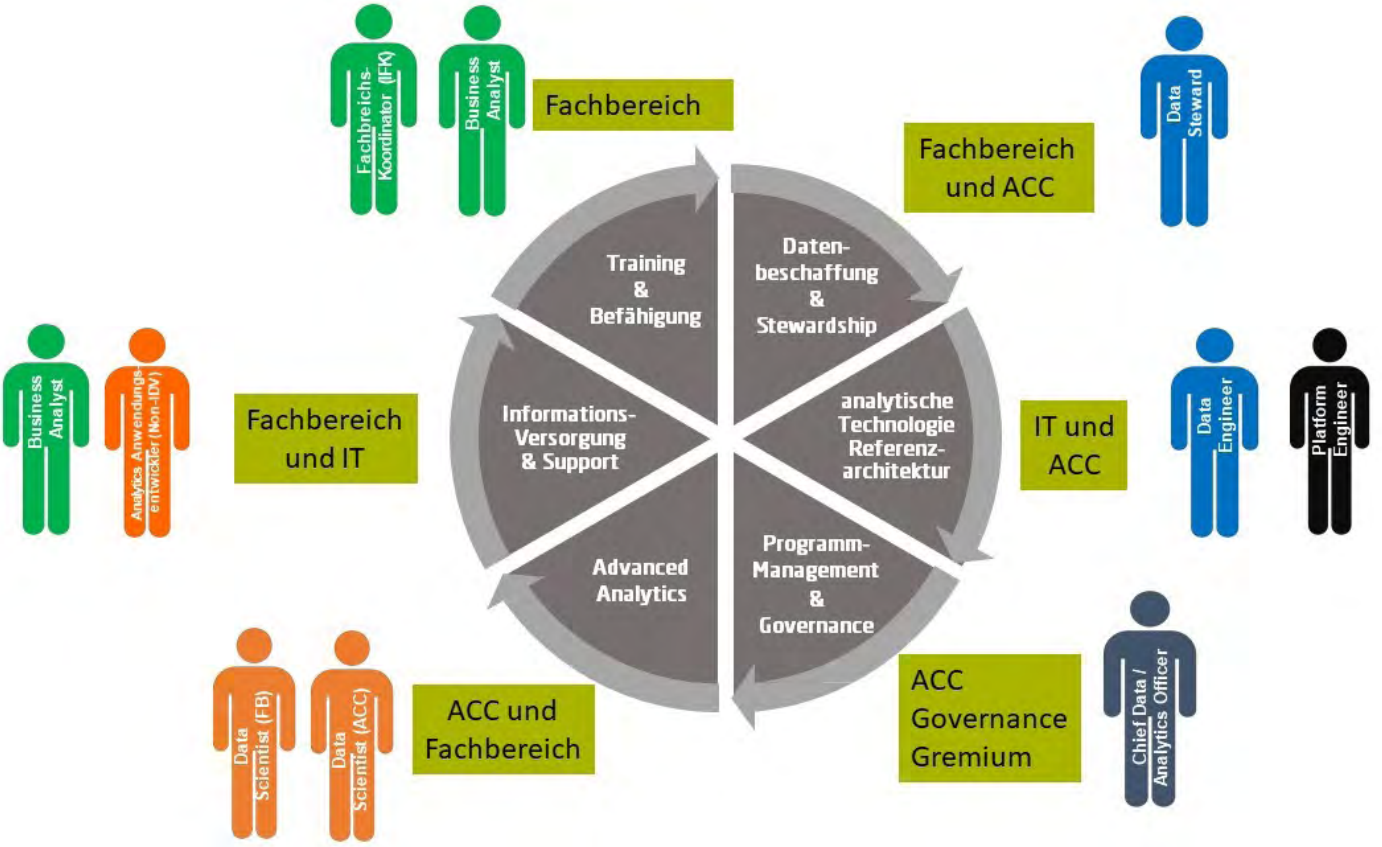


***Abb. 10**: Funktionale Verantwortlichkeiten im ACC-Ökosystem*

Es ist also sinnvoll, bestimmte ACC-Funktionen in den Geschäftseinheiten zu belassen. Jedoch sind dann Rollen und Verantwortlichkeiten klar zu definieren sowie ein Prozess zu etablieren, der die reibungslose Zusammenarbeit gewährleistet.

6 TDWI Research – Prioritäten für Data Lakes

Russom fasst in seinem TDWI Report die zwölf Prioritäten für Data Lakes zusammen, die sich aus seiner empirischen Studie im Rahmen der TDWI-Marktforschungsaktivitäten ergeben haben. Ergänzend werden diese ermittelten Prioritäten kommentiert, um Empfehlungen, Anforderungen oder Regeln zu benennen, die Unternehmen bei einer erfolgreichen Umsetzung eines Data Lakes unterstützen können.

6.1 Verständnis über den Data Lake

Ein Data Lake soll einen geschäftlichen Mehrwert liefern. Für Geschäftskunden dreht sich bei einem Data Lake alles um Analytik. Selbst, wenn ein Unternehmen über einige Formen der Analytik verfügt (zum Beispiel OLAP), sind nach und nach fortgeschrittenere Formen sinnvoll, zum Beispiel Predictive Analytics, Data Mining im Allgemeinen oder Visual Analytics, um mit den sich entwickelnden Märkten, Kundenbasen, Partnern und Wettbewerbern Schritt halten zu können. In ähnlicher Weise verlangen eine wachsende Zahl versierter Anwender den Zugriff im Sinne von Self Service, Exploration und Visualisierung. Data Lakes unterstützen eine schnelle Datenerfassung, die es einem Unternehmen ermöglicht, Informationen früher zu identifizieren und basierend darauf zu reagieren. Ein gut konzipierter Data Lake mit den richtigen Endnutzer-Werkzeugen kann diesen geschäftlichen Anforderungen gerecht werden.

6.2 Technologische Vorteile

Ein Data Lake bietet technologische Vorteile. Für Technologie-Nutzer dreht sich alles um die freie Datenzusammenstellung. Das liegt daran, dass die entdeckungsorientierte Exploration und Analytik, die Unternehmen heutzutage anstreben, umfangreiche Datensätze benötigen, welche leicht umstrukturiert (wenn überhaupt) aus zahlreichen Quellen aggregiert werden. Das ist eine zentrale Aufgabe, die durch Data Lakes in allen Größenordnungen unterstützt werden kann und soll.

6.3 Plattform

Wichtig ist ein Bewusstsein für die bestehenden Datenanforderungen, so dass eine passende Plattform gewählt werden kann. Eine aktuelle Möglichkeit ist die Nutzung von Hadoop. Dabei sind wiederum die Anforderungen zu definieren, um die Sinnhaftigkeit einer Lösung wie Hadoop bestimmen zu können, ergänzt um die Aufgabe, Werkzeuge zu definieren, um SQL-Statements durch Hadoop performanter auszuführen oder direkt Daten in einem Hadoop-Cluster zu erfassen und zu bearbeiten. Beide Aspekte sind relevant und sollten daher anforderungsgemäß in der Architektur abgebildet sein. Strukturierte Anforderungslisten bilden eine Wissensbasis zur Definition und späteren Evaluation der Nutzungs-und Hadoop-Anforderungen.

6 TDWI Research – Prioritäten für Data Lakes

6.4 Hybride Architektur

Zu betrachten ist die zuvor vorgestellte hybride Architektur für den Data Lake. Es darf nicht außer Acht gelassen werden, dass der Data Lake als Erweiterung einer bestehenden komplexen Datenumgebung (zum Beispiel für Lagerhaltung, Marketing, Lieferkette etc.) am wertvollsten ist – nicht so sehr als eine unabhängige Datenerhebung. Aus diesem Grund zeigen fast alle Anwendungsfälle einen Data Lake, der Bestandteil eines größeren Datenökosystems ist. Der Data Lake trägt zu einem bereits hybriden Ökosystem bei, wobei die Umsetzung derzeit eher als logisches Konstrukt stattfindet, welches physisch über mehrere Plattformen verteilt ist (analog zu aktuellen Data-Warehouse-Systemen). In diesem Sinne wird der Data Lake selbst hybrid, was ihm eine breitere Palette von Datentypen und Analysen ermöglicht. Wie bei Data Warehouses kombiniert der Hybrid-Data-Lake Hadoop und ein relationales Datenbanksystem sowie gegebenenfalls weitere Plattformen, um diese Angebotspalette zu erreichen.

6.5 Hadoop ergänzen

Ergänzend ist zu prüfen, inwieweit vorhandene architektonische Lücken von Hadoop mit zusätzlichen Werkzeugen gefüllt werden müssen. Einige Werkzeuge ergänzen Hadoop, wie zum Beispiel die für die Datenintegration. Andere beheben die Versäumnisse von Hadoop, nämlich Werkzeuge für Metadaten, Sicherheit und SQL-Unterstützung zur Verfügung zu stellen. Um das Portfolio der verschiedenen Werkzeuge zu vereinfachen, sollten Anbieter Berücksichtigung finden, die eine breite Funktionsunterstützung bieten, um die Vielfalt der Lösungen in einem Unternehmen nicht zu groß werden zu lassen. Dabei sollte ein sukzessives Werkzeugwachstum geplant sein, um die Architektur schrittweise wachsen zu lassen und die dafür erforderliche Ressourcenbindung möglichst gering zu halten.

6.6 Graphical User Interface (GUI)

Es sind Endbenutzer-Werkzeuge zu wählen, die eine entsprechende fachliche Unterstützung liefern. Die meisten Fachanwender und einige technisch versierte Mitarbeiter werden den Wert des Data Lake über die GUI der Werkzeuge für die Analyse, Datenaufbereitung, Visualisierung und andere Analyseaufgaben wahrnehmen. Es ist zu identifizieren und regelmäßig zu evaluieren, was die Endnutzer in ihrem jeweiligen Bereich benötigen. Diese Nutzer sollten Hilfe erhalten, Werkzeuge zu finden, die ihnen einen Mehrwert bieten, so dass sie ihre jeweiligen Aufgaben besser erledigen können. Es ist sicherzustellen, dass Werkzeuge Datenschutzmaßnahmen unterstützen, so dass die Inhalte des Data Lake direkt erkundet und analysiert werden können.

6 TDWI Research – Prioritäten für Data Lakes

6.7 Vorsicht vor Datendumping

Erste Aussagen über Data Lakes lauteten, dass man beliebig große Datenmengen in diesen speichern könne und Endanwender alles darin einfach nutzen könnten. Eine Reihe von bereits bekannten Fehlern zeigt auf, dass diese Annahme falsch war. Diese Art des Datendumping führt zu überflüssigen Daten, welche die Analyseergebnisse verzerren, zu nicht nachvollziehbaren Daten, denen niemand vertrauen wird, und zu einer schlechten Abfrageleistung, die das primäre Ziel des Data Lake eliminiert. Im schlimmsten Fall stellt der Zugriff auf den Data Lake eine Compliance- oder Datenschutzverletzung dar. Wichtig ist ein definierter Ordnungsrahmen, der genau festlegt, welche Daten in den Data Lake gehen, basierend auf den Arten der Erkundung und Analysen, die für prioritäre Nutzer und Anwendungen erforderlich sind, sowie auf Datenladung und Übertragung in das Data Warehouse und damit verbundenen Prozessen. Daten außerhalb dieser Prozesse sollten nicht im Data Lake bereitliegen, sondern zunächst Gegenstand einer Evaluierung dahingehend sein, ob diese Daten tatsächlich einen Mehrwert für das unternehmerische Handeln liefern.

6.8 Gestaltung des Data Lake

Sobald eingehende Daten strukturiert geplant sind, ist im Weiteren festzulegen, wie Volumen, Partitionen und Zonen innerhalb des Data Lake zu organisieren sind. Gesammelte Best Practices machen deutlich, dass die typischen Zonen die Bereiche Datenladung, Dateninszenierung, Datendomains (zum Beispiel Kundendaten), Abteilungsdomänen (zum Beispiel Daten, die von Marketingunternehmen verwendet werden), Analysearchive und Analysesandboxes sind. Sobald die Zonen bekannt sind, sind die Datenflüsse für die zu ladenden Daten in diese Bereiche zu modellieren. Für dieses ETL ist zu beachten, dass ein Data Lake kein Data Warehouse ist, und eine solche Zone ist nicht so stark strukturiert wie ein Themenbereich oder eine Dimension im Data Warehouse. Innerhalb jeder Zone befinden sich die Daten noch in ihrem Rohzustand oder sind nur leicht standardisiert, was im Einklang mit dem Fokus des Data Lake auf detaillierte Quelldaten zur Erkundung und wiederholten Neunutzung steht.

6.9 Konzentration auf Rohdaten

Zwar liegen im Data Lake Rohdaten vor, dennoch sollte eine Strukturierung mit zunehmendem Datenwachstum geplant werden. In Analogie zu den Erfahrungen mit Datenbanken gibt es bestimmte Daten, die immer wieder aufgerufen und genutzt werden. Für diese Daten sollte entsprechend geplant werden, diese so aufzubereiten und bereitzustellen, dass eine schnelle und einfache Nutzung durch die Anwender möglich ist. So werden die Zugriffsleistung und die Datenkonsistenz verbessert. In Bezug auf das Data-Lake-Design führen Restrukturierungsdaten in der Regel zu eher einfachen Erfassungs- oder tabellarischen Strukturen, die in der Regel über eine Metadatenstandardisierung erreicht werden. Letztlich gilt hier die gleiche Anforderung wie schon im Data Warehouse, nämlich dass ein gemeinsames Sprachverständnis vorliegen sollte, um das allgemeine Datenverständnis zu vereinfachen. Es ist ja auch Aufgabe des Data Lake, für andere Datenbanken als Quelle zu dienen.

6 TDWI Research – Prioritäten für Data Lakes

6.10 Keine generelle, sondern eine individuelle Steuerung des Data Lake

Idealerweise liegt bereits ein Data-Governance-Konzept vor, welches eine Bibliothek von Richtlinien für die konforme Nutzung von Unternehmensdaten darstellt, sowie Datenstandards beinhaltet, um die Datenqualität und -struktur zu steuern. Wie bei jeder neuen Datenerhebung sollte das Data-Governance-Gremium für die Datenverwaltung eines neuen Data Lake prüfen und festlegen, welche bestehenden Richtlinien gelten, und definieren, ob alte Richtlinien zu überarbeiten sind, um den Data Lake zu aktualisieren, oder ob gegebenenfalls neue Richtlinien erforderlich sind. Wenn Hadoop neu in einer Organisation ist, bedarf es möglicherweise einer gesonderten Prüfung. Und schließlich ist Governance am besten, wenn sie kooperativ ist. Neue Governance-Funktionen in den jeweils genutzten Werkzeugen können das Wissen über Daten und deren originäre Quellen erfassen und austauschen und dabei auch als Dokumentation über die Nutzung der Daten dienen

6.11 Spezialisten für Datenmanagement

Wie bereits erwähnt, gibt es nur sehr wenige Datenmanagement-Experten, die bereits Erfahrung mit Data Lakes und Hadoop haben. Die Mitarbeiter, die zur Verfügung stehen, sind in der Regel mit hohen Gehältern ausgestattet. Aus diesem Grund ziehen es Organisationen vor, bestehende Mitarbeiter in diesen Fähigkeiten weiterzubilden, anstatt neue Mitarbeiter einzustellen. Diese Strategie zeigt sich aktuell erfolgreich, weil die Datenmanagement-Mitarbeiter diese Weiterbildungsmöglichkeiten gerne annehmen.

6.12 Vernetzung der Stakeholder

Grundsätzlich ist eine Vernetzung der Mitarbeiter mit Beratern mit Data-Lake-Erfahrung sinnvoll. Da es schwierig ist, neue Mitarbeiter mit Hadoop- und Data-Lake-Fähigkeiten zu finden, erscheint es sinnhaft zu sein, externe Beratungskompetenz hinzuziehen, um die Fähigkeiten in einem Unternehmen positiv zu entwickeln. Dadurch werden Projektrisiken reduziert, die Lieferzeit verkürzt und ein wertvoller Wissenstransfer von Beratern an die Mitarbeiter ermöglicht.

7 Fazit – Data Oceans und Data Swamps

Der Aufbau eines Data Lake ist im Rahmen der digitalen Transformation eine wesentliche Komponente, um das datengetriebene Handeln in Unternehmen ermöglichen zu können. Datengetriebenes Handeln erfordert einen höheren Grad an Flexibilität unter Beachtung einer Data Governance und entsprechender Umsetzung von Sicherheitsvorschriften. Der digitale Wandel soll es aber nun ermöglichen, dass Endbenutzer dynamisch und bei Bedarf unterschiedliche Kombinationen von Daten anfordern können, ohne dass die IT diese manuell und zeitaufwendig bereitstellen muss. Ein Template-orientierter Ansatz ermöglicht die Abbildung entsprechender Sicherheitsmaßnahmen, so dass diese nicht mehr gegebenenfalls redundant in nachgelagerten Systemen geführt werden. Dies reduziert Komplexität und Verwaltungsaufwand, was sich sonst konfliktär zur geforderten Flexibilitätserhöhung verhalten würde. So können allerdings der erforderliche Aufwand und die Antwortzeiten durch einen hohen Automationsgrad reduziert werden, so dass der Nutzer schneller erste Ergebnisse und damit Daten zur Bearbeitung der eigentlichen analytischen Aufgabe erzielen kann. Allgemein lässt sich konstatieren, dass eine dynamische Herangehensweise bei Aufbau und Betrieb des Data Lake relevant ist, da häufig verwendete Kombinationen nahezu ohne Aufwand aus der Ad-hoc-Welt in eine dauerhafte Form überführbar sind. Das Nutzerverhalten wird nachvollziehbarer, da einzelne Anfragen auswertbar sind. Dies führt wiederum zu besseren Planungen für die weitere Datenbereitstellung im Sinne des Change-Managements der digitalen Transformation.

Data Lakes haben das Potenzial, existierende Daten-Ökosysteme zu modernisieren und analytische Programme zu erweitern. Ein Bereich tritt dabei in den Vordergrund: Advanced Analytics. Der eigentliche Treiber für die meisten heutigen Trends im IT- und Datenmanagement ist die wachsende Anzahl von Unternehmen, Regierungsbehörden und anderen Organisationen, die eine breitere Palette von Analysen benötigen, um wettbewerbsfähig zu sein oder in ihrem Geschäft zu wachsen, Kunden zu binden und andere organisatorische Ziele zu erreichen. Selbst, wenn OLAP und ältere Formen der Analyse vorhanden sind, benötigen Unternehmen prädiktive und entdeckungsorientierte Analysen, die auf fortschrittlichen Technologien wie Data Mining, Clustering, Visual Analytics, künstlicher Intelligenz oder maschinellem Lernen basieren. Nutznießer sind fortschrittliche Analysen, neue datengetriebene Geschäftspraktiken, die Nutzung von Big Data und die Modernisierung von Data Warehouses.

Der Data Lake kann eine skalierbare Sandbox zum Durchsuchen von Daten aus mehreren Quellen bereitstellen, um neue Fakten über das Unternehmen und seine Kunden, Partner und Produkte zu ermitteln. Damit sie sowohl alte als auch neue Daten untersuchen können, fordern sowohl fachliche als auch technische Benutzer die Erkundung von Daten, zusammen mit anderen neuen Methoden aus dem Bereich des Self Service und der Datenvisualisierung. Wichtig ist aber, zu verstehen, dass nicht die Datensammlung selbst den Erfolg gibt, sondern die Nutzung der erfassten Daten. Diese können letztlich nur in einem für das Unternehmen definierten Kontext wirken und sollten daher auch entsprechend zweckorientiert zusammengestellt sein. Dann wird ein Data Lake auch zu der großen Datenquelle für Analysen. Dabei ist Hadoop zur bevorzugten, aber nicht ausschließlichen Plattform für Big Data und Data Lakes geworden, da Anwender kostengünstige Hardware und Software und extreme Skalierbarkeit erwarten.

Modernisierung ist weiterhin ein starker Trend im Bereich Data Warehousing. Data Lakes, ob auf Hadoop oder relationalen Systemen, werden im Rahmen des Modernisierungsprozesses regelmäßig zu plattformübergreifenden Data-Warehouse-Umgebungen (DWEs) hinzugefügt. Dabei zeigt sich, dass ein Data Lake als Erweiterung des Data-Warehouse-Speichers, als Datenlandung und -bereitstellung sowie als Strategie für die Entlastung und Kostensenkung eines Data Warehouse fungieren kann. Ebenso kann der Data Lake eine Erweiterung der Datenintegration darstellen, häufig durch eine sogenannte Push-down-Verarbeitung. Ein weiterer Vorteil, der von einem Hadoop-basierten Data Lake erwartet wird, ist die Fähigkeit, unterschiedlichste Datenstrukturen und Dateitypen zu erfassen und zu verarbeiten, einschließlich Maschinendaten von Internet of Things (IoT), Robotern, Sensoren, Zählern etc.

Ein Data Lake hat seine Vorteile, dennoch zeigte unter anderem TDWI Research mit seinen Umfrageergebnissen auch viele potenzielle Hindernisse auf. Die Probleme erstrecken sich dabei insbesondere über die Aspekte einer Data Governance, die ein Schlüsselelement des erfolgreichen Einsatzes von

7 Fazit – Data Oceans und Data Swamps

Daten in Unternehmen darstellt. Wie bereits erwähnt, kann das unkontrollierte Ablegen von Daten in einen Data Lake zu einem sogenannten Datensumpf (Data Swamp) führen. Die Umfrageteilnehmer sind sich dieses potenziellen Problems voll und ganz bewusst, wobei ein Mangel an Data Governance als Hauptproblem eingestuft wird. Dies führt zu den zuvor im Kontext des ACC genannten Komponenten eines strategischen Ordnungsrahmens, mit dem die wichtigsten Hindernisse in Governance, Integration, mangelnder Erfahrung, Datenschutzproblemen sowie unausgereiften Technologien und Praktiken im Bereich der Datenintegration adressiert werden. Die Datenaufnahme und ihre Steuerung sind, wie bereits erläutert, wichtige Erfolgsfaktoren für einen Data Lake. Viele Unternehmen klagen über einen Mangel an Datenintegrationswerkzeugen und -fähigkeiten insbesondere in Bezug auf Hadoop. Aktuell zeigt sich aber, dass Softwareanbieter und die Open-Source-Community ihre Datenintegrationstools ergänzt und Hadoop eingebunden haben, so dass die erforderlichen Schnittstellen-, Speicher- und Verarbeitungsmethoden unterstützt werden. In einem verwandten Bereich müssen Data Lakes demokratisiert werden, indem Mitarbeiter aus Fachbereichen und weniger technische Benutzer in einer einfachen Art und Weise Zugang zu den Daten erhalten. Dabei gehen die Anwender jedoch davon aus, dass ein unterstützter Self Service Probleme eines vormals zu technischen Zugriffs löst und sie nun direkt Big Data in ihren eigenen Prozessen nutzen können. Die Bereitstellung von Big Data ist der Anlass für die meisten Unternehmen, sich für Data Lakes und Hadoop zu interessieren. Allerdings ist aktuell die Erfahrung mit Big Data, Data Lakes und Hadoop noch nicht ausreichend vorhanden, um schlüsselfertige Mehrwerte anbieten und erzielen zu können.

Daher sind unzureichende Fähigkeiten für Big Data, unzureichende Fähigkeiten für Hadoop, unzureichende Fähigkeiten für das Entwerfen von Big-Data-Analysesysteme und unzureichende Kenntnisse für das Design von Data Lakes Anlässe zur Sorge. Organisationen stellen sich derzeit auf, indem sie vorhandene Datenverwaltungsmitarbeiter schulen, Berater mit Big-Data-Erfahrung von extern engagieren und seltener neue Mitarbeiter mit Big-Data-Kenntnissen einstellen. Die Entwicklung der eigenen Fähigkeiten in den vorhandenen Kapazitäten steht im Vordergrund. Data Lakes sind noch recht neu, und sowohl die Mitarbeiter aus den Fachabteilungen als auch das IT-Personal lernen immer noch etwas über sie. Da dieses Lernen noch nicht abgeschlossen ist, ist in Unternehmen, die datengetrieben agieren möchten, ein überzeugendes Geschäftsmodell oder ein Sponsoring für die Mitarbeiter zu etablieren.

Ein genereller und überzeugender Business Case ist offensichtlich unwahrscheinlich, wenn die Organisation keinen Data Lake benötigt. Erfolgreiche Business Cases fokussieren derzeit fachliche Anforderungen mit Nutzung von Advanced Analytics, also umfassender Datenexploration unter Nutzen von Big Data bei Beachtung von Datenschutz und Compliance. Dabei wird die mangelnde Einhaltung des Datenschutzes in einem Data Lake als Risiko gesehen, nämlich als das Risiko, sensible Daten wie personenbezogene Daten preiszugeben. Es ist für Unternehmen wichtig, solche potenziellen Probleme mit unternehmensweiten Programmen für Data Governance und/oder Stewardship auf dem Data Lake sowie mit Richtlinien zur Datenaufnahme und für die Verwendung von Daten im Data Lake anzugehen. Vieles ist jedoch aktuell noch im Fluss, und Data Lakes und deren Governance sind als unreif zu konstatieren. Dazu trägt bei, dass der Markt durch Open-Source-Entwicklungen bestimmt ist und die Sicherheit fehlt, welches Werkzeug dauerhaft eine angemessene Unterstützung für all die bestehenden Herausforderungen liefert. Deshalb wird vielerorts eine abwartende Position eingenommen. Insbesondere die Unreife von Hadoop in Bezug auf Datensicherheit, Metadatenverwaltung und SQL nach ANSI-Standard trägt dazu bei. Auch in diesen Bereichen setzen die Softwareanbieter und die Open-Source-Community regelmäßig Fortschritte um. Da aktuell aber auch eine Reduktion der Vorbehalte gegenüber der Nutzung von Cloud-Ressourcen festzustellen ist, ist davon auszugehen, dass dies bei den technischen Konzepten analog sein wird, da sowohl die Verfügbarkeit als auch das Know-how in den Unternehmen wachsen werden.

Letztendlich gilt, bewusst nicht nur technisch, sondern auch organisatorisch mehrere Data Lakes zu entwickeln und damit einen unternehmensweiten Data Ocean bereitzustellen. Das Organisatorische hat aber eine dominierende Rolle, da Data Swamps zu verhindern sind, in denen es vielleicht noch Moorleichen, aber wenig Nutzbares gibt.

Literatur

Aggarwal, C. C.; Pei, J.; Zhang, B.: (2006): On privacy preservation against adversarial data mining. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM: 510-516.

Ciriani, V.; Di Vimercati, S. D. C.; Foresti, S.; Samarati, P (2007): Microdata protection. Secure data management in decentralized systems, Springer: 291-321.

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

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

Durmus, M. (2017): Wann sollten Unternehmen in einen Data Lake (Data Lake) investieren?, https://www.aisoma.de/wann-sollten-unternehmen-in-einen-data-lake-Data Lake-investieren/, Abruf am 2019-01-14.

Felden, C. (2019): Digitale Transformation – Mehr als nur ein Technologie-Update – Wie Unternehmen ihre Digitalisierungsprojekte zum Erfolg führen, TDWI E-Book. SIGS DATACOM GmbH, Troisdorf, 2019.

Friedman, A.; Wolff, R.; Schuster, A. (2008): Providing k-anonymity in data mining. The VLDB Journal 17(4): 789-804.

Hardt, F.; Lenzhölzer, C. (2017): Wie Lakes, Labs und Governance das DWH beeinflussen, in: BI-Spektrum, 12. Jg., 2017, Nr. 3, 22.

Inmon, W. H. (1996): Building the Data Warehouse, 2. Auflage, New York, 1996.

Kromer, M. (2015): Modern Hybrid Big Data Warehouse Architectures, in: Business Intelligence Journal, 19. Jg., 2015, 48-55.

Litzel, N. (2018) Was ist ein Data Lake?, https://www.bigdata-insider.de/was-ist-ein-data-lake-a-686778/, Abruf am 2019-01-14.

Roßnagel, A. (2013): Big Data – Small Privacy? – Konzeptionelle Herausforderungen für das Datenschutzrecht. In: Zeitschrift für Datenschutz 2013, S. 562-567.

Russom, P. (2017): Data Lake Management Innovations, January 23, 2017, https://upside.tdwi.org

Russom, P. (2019): Data Lakes Purposes, Practices, Patterns, and Platforms, Best Practices Report Q1 2017, https://info.talend.com/rs/talend/images/WP_EN_BD_TDWI_DataLakes.pdf, Abruf am 2019-05-21.

Schäfer, R.; Goetze, D. (2016): Integration von Data Lakes in BI-Landschaften, DW-Konferenz 2016, Regensdorf/Zürich, 2016-11-22.

Serra, J. (2019): Operational Data Store (ODS) Defined, http://www.jamesserra.com/archive/2015/02/operational-data-store-ods-defined/, Abruf am 2019-05-19.

Verykios, V.; Elmagarmid, K.; Bertino, E.; Saygin, Y.; Dasseni, E. (2004a): Association rule hiding, Knowledge and Data Engineering, IEEE Transactions on 16(4): 434-447.

Über unseren Sponsor

collibra®

Kontaktadresse

Collibra UK and Germany
Collibra UK
1 Fore Street
London, EC2Y 9DT – UK

(t) +44 203 695 6965
(f) +44 203 006 8844

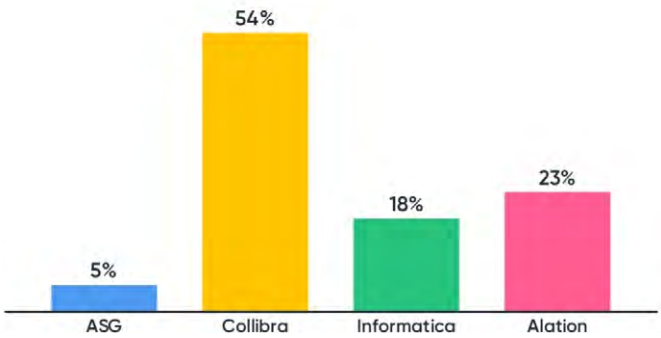
https://www.collibra.com

Das Unternehmen hinter der Platform

Collibra Data Governance Center / Data Catalog ist eine unternehmensweite Data Governance-Lösung, die Governance- und Stewardship-Funktionen für alle Data Citizens bietet.
Wir entwickeln, implementieren die Software und Betriebsmodelle und begleiten die Reise unserer Kunden zu einer echten Data Governance.

Collibra war Gewinner des BARC Speed Pitches (Data Catalog) per Zuschauervoting der TDWI München 2019 und erhielt als Preis zusätzliche Zeit, um dem Kernteam noch Tipps und Tricks mit auf den Weg zu geben.
Dabei beantworteten sie folgende Fragen: „Was empfehlen Sie Ihren Kunden, um Data Cataloging erfolgreich zu implementieren und zu betreiben? Welche praktischen Tipps können Sie ihnen geben?“

Wie viel Zeit vergeuden Sie, um die richtigen Daten zu finden?
Hören Sie auf, nach Daten zu suchen und beginnen Sie, deren Chancen zu nutzen. Mit einer angemessenen Daten Governance- und Kataloglösung erhalten Sie die erforderliche Grundlage, um den maximalen Wert aus Ihren Daten zu ziehen und diese für Analysen, für die Optimierung von Geschäftsprozessen und für die datengestützte Entscheidungsfindung zu nutzen.



Ergebnisse aus BARC Speed Pitch@ TDWI München 2019

Über unseren Sponsor

HITACHI
Inspire the Next

Kontaktadresse

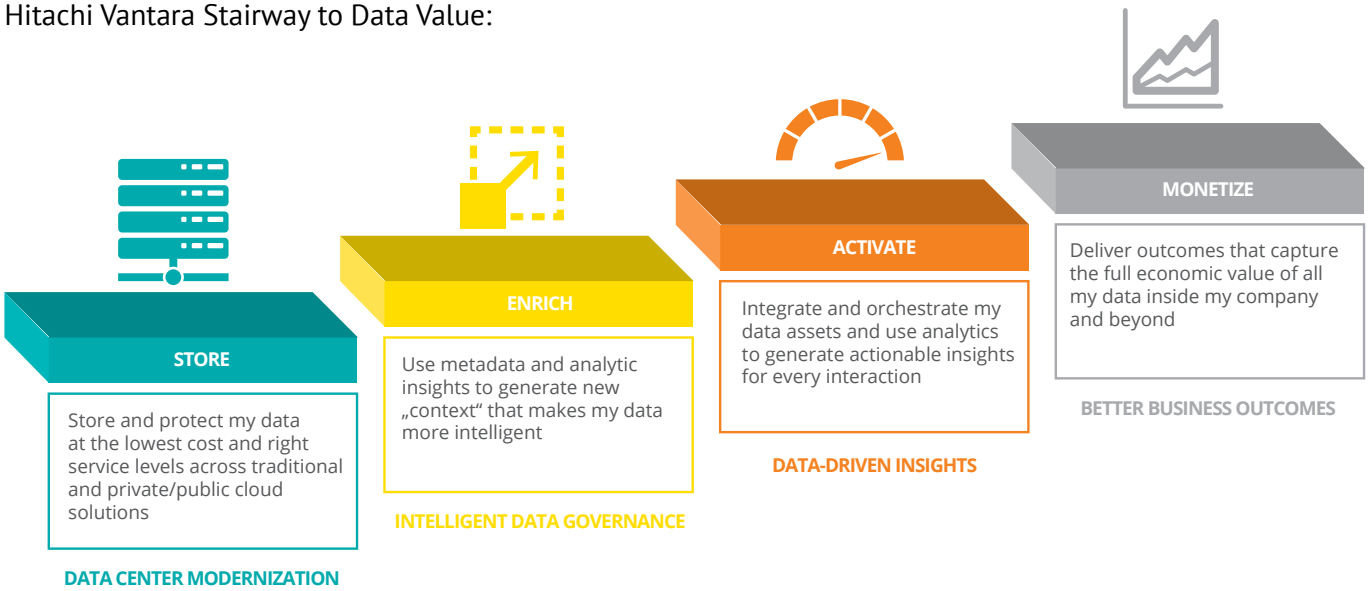
Hitachi Vantara GmbH
Im Steingrund 10
63303 Dreieich

Tel: +49-6103-804-1000
Fax: +49-6103-804-1111

E-Mail Info.de@hitachivantara.com
URL www.hitachivantara.com

Hitachi Vantara hilft datenorientierten Marktführern, den Wert ihrer Daten zu entdecken und zu nutzen, um intelligente Innovationen hervorzubringen und Ergebnisse zu erzielen, die für Wirtschaft und Gesellschaft von Bedeutung sind. Wir kombinieren Technologie, geistiges Eigentum und Branchenwissen, um Lösungen zum Datenmanagement, zur Datenintegration und -analyse zu liefern, mit denen Unternehmen das Kundenerlebnis verbessern, sich neue Einnahmequellen erschließen und Betriebskosten senken können. Nur Hitachi Vantara erhöht Ihren Innovationsvorsprung durch umfassendes Wissen in IT (Information Technology), OT (Operational Technology) und vielfältigen Fachgebieten. Gemeinsam mit anderen Organisationen arbeiten wir weltweit daran, aus Daten sinnvolle Ergebnisse zu gewinnen. Hitachi Vantara ist eine hundertprozentige Tochtergesellschaft der Hitachi Ltd. (Hauptsitz Tokyo, Japan) und hat weltweit mehr als 10.000 Unternehmenskunden (85% der FORTUNE Global 100) und gut 7.000 Mitarbeiter, 2000 Partner-Unternehmen und ist in 100 Ländern direkt präsent. Hitachi Ltd. wurde 1910 gegründet und ist weltweit führend bei Patentanmeldungen im Bereich der Big Data Analytics Technologien.

Hitachi Vantara Stairway to Data Value:



Über unseren Sponsor

Kontaktadresse

Salvatorplatz 3
Munich 80333
Germany

Phone: +44 (0)2032044300

E-Mail Dach_sales@splunk.com
URL https://www.splunk.com/de_de

Splunk Inc. (NASDAQ: SPLK) hilft Unternehmen, Fragen zu stellen, Antworten zu erhalten, Maßnahmen zu ergreifen und Geschäftsergebnisse aus ihren Daten zu gewinnen. Unternehmen nutzen marktführende Splunk-Lösungen mit Machine Learning für Monitoring-, Untersuchungs- und Reaktionsvorgänge im Zusammenhang mit sämtlichen Arten von Geschäfts-, IT-, Sicherheits- und IoT-Daten. Schließen Sie sich den Millionen passionierter Nutzer an und testen Sie Splunk gleich heute kostenlos.

Über unseren Sponsor

Kontaktadresse

Talend Germany GmbH
Baunscheidtstraße 17
53113 Bonn

https://de.talend.com/

Mit Integrationslösungen von Talend können datengetriebene Organisationen auf Anhieb Mehrwert aus all ihren Daten ziehen. Talend nimmt Integrationsbemühungen die Komplexität und stattet IT-Abteilungen so aus, dass sie schneller auf Geschäftsanforderungen reagieren können – und das zu vorhersehbaren Kosten. Talends skalierbare, zukunftssichere Lösungen basieren auf Open-Source-Technologie und decken alle bestehenden und sich entwickelnden Anforderungen an Integration ab. Talend hat private Investoren und die Hauptniederlassung befindet sich in Redwood City, Kalifornien.

Mehr Informationen finden Sie unter https://de.talend.com/

Auf Twitter finden Sie uns unter @Talend.

E-Book

tdwi.eu