2 Architekturen und Technologien für Data Lakes

Carsten Dittmar • Peter Schulz

Lange Zeit galt das Data Warehouse als zentrales Architekturkonzept für dispositive Reporting- & Analysezwecke. Im Zuge der zunehmenden Digitalisierung und der damit einhergehenden Masse an zur Verfügung stehenden Datenmengen, aber auch des breiten Spektrums an potenziellen datenbasierten Use Cases hat mittlerweile der Data Lake dem klassischen Data Warehouse den Rang abgelaufen. In diesem Beitrag wird nach einer Einordnung das Konzept des Data Lake vorgestellt und anschließend werden gängige Architektur- und Technologiemuster der Praxis skizziert.

2.1 Historie der dispositiven Datenplattformen

Die richtige Architektur für die dispositive Datenverarbeitung war seit den 90er-Jahren mit dem Data Warehouse klar definiert. Mit dem Data-Warehouse-Konzept setzte sich die Idee einer separaten Datenbasis für dispositive Reporting- und Analysezwecke in der Praxis durch, die redundant Datenbestände aus den operativen Systemen speichert. Bisherige Architekturansätze, die zu Reporting- und Analysezwecken direkt auf die operativen Systeme und ihre Daten zugriffen, waren mit dem Data Warehouse obsolet. Die potenziellen Gefahren einer Inkonsistenz der redundanten dispositiven Datenbestände gegenüber den operativen Datenbeständen wurden durch einen gesteuerten Datenintegrationsprozess in das Data Warehouse sowie einer Beschränkung auf einen Lesezugriff auf das Data Warehouse beantwortet [Gluchowski et al. 2008, S. 128 ff.].

Die Standardarchitektur sieht idealtypisch ein singuläres (Enterprise) Data Warehouse vor, das aus den unterschiedlichen operativen Quellsystemen die relevanten Daten aufsammelt. In einem mehrschichtigen Datenintegrations- und Datenveredelungsprozess werden die Daten im Data Warehouse harmonisiert, integriert und persistiert. So soll aus einer Datenperspektive ein Single Point of Truth entstehen, aus dem anschließend für unterschiedliche Anwendungsfälle Datenextrakte – in der Regel in einer multidimensionalen Aufbereitung – in abgrenzbaren Data Marts gespeichert werden [Schnider et al. 2016, S. 6 ff.].

Zwei architekturelle Grundmuster haben sich zur Umsetzung der mehrschichtigen Data-Warehouse-Architektur bewährt. Der Hub-and-Spoke-Ansatz nach Inmon gilt nach wie vor als die Reinform des Data Warehouse, da in der Schicht des Core Data Warehouse eine integrierte und umfassende (Unternehmens-)Datensicht erstellt wird [Inmon 2005]. Als pragmatische Architekturalternative schreibt man gemeinhin dem Data Mart Bus nach Kimball Vorteile bei der schnellen Umsetzung und Erweiterbarkeit zu, da der Fokus einer Harmonisierung nur auf die strukturgebenden Stammdaten gelegt wird [Kimball & Ross 2002].

Auf den Datenschatz des Data Warehouse greift der Anwender selten direkt zu. Zumeist nutzt er Berichts- und Analysewerkzeuge (Business Intelligence), die den Zugriff auf Standardberichte erlauben, aber mittlerweile dem User auch die Möglichkeiten geben, im Self-Service eigene Datenanalysen auf Basis individueller Datenzusammenstellungen durchzuführen. Der überwiegende Fokus liegt im Data Warehouse auf der vergangenheitsorientierten Analyse von Kennzahlen entlang von konsolidierten Auswertungsstrukturen. Damit werden in der Regel Fragestellungen wie »Was ist passiert?« und »Warum ist es passiert?« adressiert.

Um eine gültige und konsolidierte Wahrheit zu allen strukturierten Daten in einem Unternehmen zu repräsentieren, stellt das Data Warehouse die Daten in vorab definierten Datenmodellen zur Verfügung. Bevor Daten integriert werden, ist demzufolge dieses Datenmodell zu entwickeln und zu implementieren. Der hohe Anspruch an korrekte und unternehmensweit harmonisierte Daten führt in der Regel dazu, dass es recht lange dauert, bis Daten aus einer neuen Datenquelle in dieser konsolidierten Sicht integriert sind, weil im Vorfeld viel Konzeptions- und Abstimmungsaufwand nötig wird.

Bestrebungen, eine dispositive Datensenke neben den operativen Systemen obsolet werden zu lassen und alle operativen und auch dispositiven Anfragen mit demselben System beantworten zu können, hat insbesondere das SanssouciDB-Vorhaben des Hasso-Plattner-Instituts und der Stanford University vorangetrieben [Plattner & Zeier 2012]. Das daraus resultierende kommerzielle Produkt ist die SAP HANA-Datenbank, die in der aktuellen Version der Enterprise-Resource-Planning-(ERP-)Lösung des Herstellers eingesetzt wird. Diese Lösung stellt für einige Data-Warehouse-Fragestellungen eine tatsächliche Alternative dar, kann Daten außerhalb der operativen Applikation aber nur schwer integrieren.

2.2 Das Data-Lake-Konzept

Im neuen Jahrtausend stieg mit dem Aufkommen neuer Datenquellen wie Social-Media-Daten oder IoT-Daten und dem enormen Anstieg des Datenvolumens durch die Digitalisierung vieler Prozesse auch der Bedarf, diese neuen Daten ebenfalls in einer zusammenführenden Datenplattform zur Verfügung zu stellen. Viele dieser Daten liegen jedoch in semistrukturierter oder unstrukturierter Form vor. Mit der steigenden Relevanz dieser Datenquellen wurde die Idee des Data

Lake geboren. Die Idee wird gemeinhin James Dixon zugeschrieben, der in einem Blogpost von 2010 erstmalig das Bild eines Data Lake prägt [Dixon 2010]. Der Data Lake stellt alle Quelldaten – interne und externe, strukturierte und unstrukturierte – auch in ihrer nicht aufbereiteten Form als Rohdaten zur Verfügung. Somit stehen die Daten möglichst unmittelbar nach der Datenerzeugung schnell und unverfälscht in einem Data Lake bereit. Dadurch werden Einblicke zur Echtzeit ermöglicht, die auf Basis von Vorhersage- und Szenariomodellen die Fragestellungen »Was wird wahrscheinlich passieren?« und »Was kann unternommen werden, damit es passieren wird?« beantworten können.

Die Speicherung von Rohdaten ohne jegliche Datenveredelung auf feinster Granularität oblag in einer klassischen Data-Warehouse-Architektur eher der Datenschicht der Staging Area. Daher wird häufig der Data Lake als Fortentwicklung dieser Schicht gesehen.

Der effiziente Umgang mit großen polystrukturierten Datenmengen, eine schnelle (oft nahezu in Echtzeit) Verarbeitung von Datenströmen und die Beherrschung komplexer Analysen für neue Data-Science- und Machine-Learning-Anwendungen stehen beim Data Lake zulasten der Harmonisierung und Integration der Daten im Vordergrund. Die Struktur der Daten steht damit zugunsten einer schnellen und vollständigen Integration in den Data Lake nicht schon bei der Speicherung, sondern erst im Rahmen der nachgelagerten Analyse im Fokus. Somit ist das Ziel eines Data Lake die Schaffung von flexiblen Strukturen zur Bändigung der komplexen Integration der Vielzahl von Datenquellen.

Zumeist ist zum Zeitpunkt der Datenspeicherung noch gar nicht festgelegt, welche Analysen mit den Daten durchgeführt werden sollen. Der Data Lake bildet also das Eldorado für den Data Scientist, der explorative Analysen wie Cluster-/Assoziationsanalysen, Simulationen und Vorhersagen über komplexe Algorithmen durchführen möchte. In der folgenden Tabelle 2–1 sind wesentliche Charakteristika des Data Warehouse und des Data Lake vergleichend zusammengefasst.

Data Warehouse: Datenbasis für Systems of Record	**Data Lake: Datenbasis für Systems of Innovation**
■ Stellt 80 % der Analysen mit 20 % der Daten bereit ■ Optimiert für wiederholbare Prozesse ■ Unterstützt Vielzahl von unternehmensinternen Informationsbedarfen ■ Fokus auf vergangenheitsorientierte Auswertungen ■ Schema-on-Write mit harmonisiertem Datenmodell	■ Originäre Erweiterung der Staging Area des DWH ■ Speichert Rohdaten für Exploration und Analyse ■ Optimiert Daten unkompliziert für Analytics-Lösungen ■ Fokus auf unbekanntes Data Discovery und zukunftsorientierte Data Science & Artificial Intelligence ■ Schema-on-Read mit Echtzeit-Rohdatenbewirtschaftung

Tab. 2–1 *Charakteristika von Data Warehouse und Data Lake im Vergleich*

Damit ist auch klar, dass ein Data Lake ein Data Warehouse nicht ersetzt, sondern ergänzt. Beide Architekturkonzepte haben ihre Relevanz und bedienen zueinander unterschiedliche Use Cases [Gartner 2020].

Verschiedene übergreifende Gesamtarchitekturen sind denkbar. Sofern der Data Lake als übergreifender Staging Layer fungiert, folgt in einer sequenziellen Architekturabfolge das Data Warehouse hinter dem Data Lake. In der Regel stehen jedoch beide Systeme parallel und isoliert nebeneinander. Häufig wird im letzten Falle über Virtualisierungstechnologie für den Anwender ein virtueller Datenmarktplatz geschaffen, der beide Architekturkonzepte vermeintlich vereint [Leisten 2020].

Aktuell Verbreitung findet die Variante des Data-Mesh-Konzeptes, das neben seinen Architekturaspekten auch Governance und prozessuale Aspekte abdeckt [Dehghani 2020].

Der Reifegrad der Nutzung des Data-Lake-Konzeptes kann anhand der Dimensionen des verwendeten Datenhaushaltes und der Nutzung unterschieden werden [Gorelik 2019, S. 9 ff.]. Dieses Modell nimmt Bezug auf die ursprünglich von James Dixon aufgestellte Definition: »Wenn Sie sich Data Mart als einen Vorrat an Wasser in Flaschen vorstellen – gereinigt, verpackt und strukturiert für einen einfachen Verbrauch, ist der Data Lake ein großes Gewässer in einem natürlicheren Zustand. Der Inhalt des Datensees fließt von einer Quelle herein, um den See zu füllen, und verschiedene Benutzer des Sees können kommen, um zu untersuchen, einzutauchen oder Proben zu entnehmen« ([Dixon 2010], eigene Übersetzung aus dem Englischen). Die nachfolgende Abbildung 2–1 verdeutlicht die fünf unterschiedlichen Ausprägungen.

Ein Data Puddle ist die erste Adaptionsform des Data-Lake-Konzeptes. Er ist gekennzeichnet durch seinen eingeschränkten, nur den aktuellen Nutzungsfall abdeckenden Datenhaushalt und eine lokale Nutzung durch (IT-)Experten, die einen hohen Grad an manuellen Tätigkeiten zur Nutzung bedingt. Hier stehen Kostensenkung und höhere Performance im Vergleich zum Einsatz klassischer Technologien eines Data Warehouse im Vordergrund, ein Mehrwert aus der Nutzung gegenüber einem Data Warehouse wird jedoch nicht geschaffen.

Als Data-Pond-Konstrukt wird eine Vielzahl von nebeneinander isoliert bestehenden Data Puddles bezeichnet. Ein gängiges Beispiel ist die Kopie von mehreren Datenhaushalten aus Data Warehouses in separate Systeme, die auf typischen Data-Lake-Technologien basieren. Neue Erkenntnisse sind auch mit diesem Archetyp nur sehr eingeschränkt und umständlich zu gewinnen.

Ein Data Lake unterscheidet sich von einem Data Pond in zwei wesentlichen Faktoren: Erstens ermöglicht er Self-Service-Nutzung durch Anwender ohne IT-Beteiligung und zweitens enthält er Daten, die aktuell noch gar nicht genutzt werden, aber perspektivisch interessant werden können. Die breite Nutzung im Unternehmen wird maßgeblich dadurch erreicht, dass Daten für die Nutzung aufbereitet und vor allem durch (einfach zugängliche) Metadaten beschrieben sind. Mit dieser Aufstellung können von einer Vielzahl unterschiedlicher Nutzergruppen Antworten auf ihre individuellen Fragestellungen gewonnen werden.

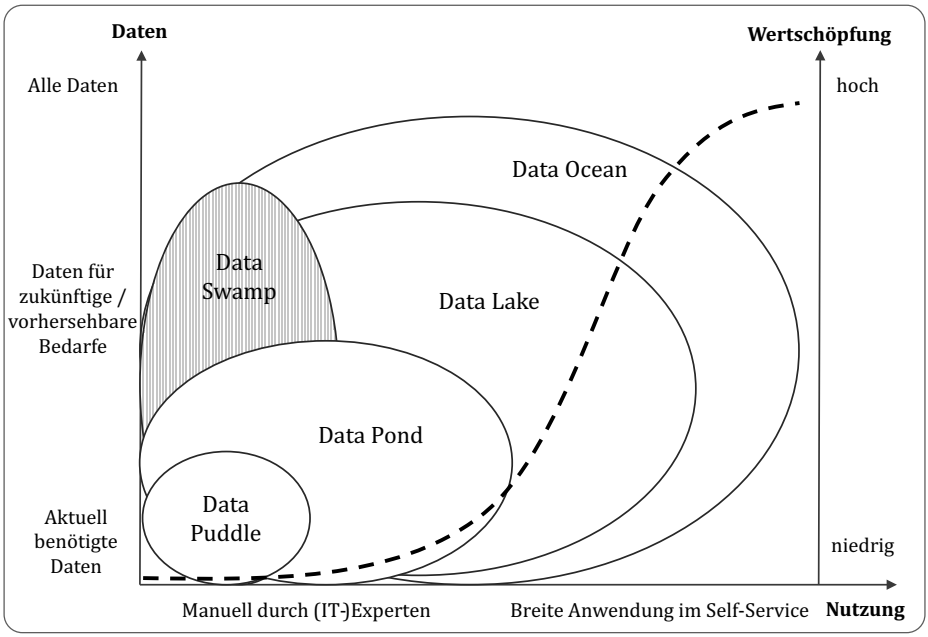

Abb. 2–1 *Reifegrad der Nutzung des Data-Lake-Konzeptes und resultierende Wertschöpfung (gestrichelte Linie) (eigene Darstellung in Anlehnung an [Gorelik 2019])*

Ein Data Ocean gilt als ultimative Antwort auf datengetriebene Entscheidungen, basierend auf allen (fachlichen) Daten eines Unternehmens und mit einem einfachen, verständlichen Zugang für alle Mitarbeiter. Die resultierende Wertschöpfung kann jedoch gegenüber einem gut positionierten Data Lake nur noch marginal erhöht werden.

Eine Sonderform stellt der berüchtigte Data Swamp dar. Er ist eine Ansammlung von verschiedenen Daten, die jedoch überhaupt nicht oder wenig organisiert und aufbereitet sind. Weiterhin fehlen Metadaten, was eine Nutzung durch eine breitere Anwenderbasis verhindert. In der Praxis wird in einem solchen Fall die Behandlung mit dem Ziel der Umwandlung in einen Data Lake durch Maßnahmen wie Aufbereitung der Daten und Zuordnung von Metadaten versucht.

2.3 Architektur eines Data Lake

Die idealtypische Architektur eines Data Lake durchlief eine Evolution: Glaubte man zunächst, eine einzige, große Plattform für alle Bedürfnisse im eigenen Rechenzentrum zu bauen sei optimal, hat sich hier ein Trend zur Nutzung multipler, aber orchestrierter Data Lakes und auch die Nutzung von Cloud-Angeboten durchgesetzt.

Ein Katalog an Architekturprinzipien für einen Data Lake ist in der nachfolgenden Tabelle 2–2 dargestellt.