Bild: Shutterstock

Data Lake und Data Warehouse – Konkurrenz oder Ergänzung?

Datenseen in der BI-Landschaft

Ein Beitrag von
Thomas Weiler

Business Intelligence ist nicht tot, befindet sich aber im Wandel. Wesentlichen Anteil daran hat die verstärkte Nutzung sogenannter Data Lakes. Sie spielen in aktuellen Informationsarchitekturen eine immer größere Rolle und werden häufig im Umfeld von Big-Data-Initiativen genutzt. Allerdings herrscht immer noch Unsicherheit, ob Data Lakes in Konkurrenz zu BI-Systemen stehen oder diese ergänzen. Dieser Artikel positioniert einen Data Lake in seiner ursprünglichen Form und zeigt, wie er sich als integrative Komponente einer Big-Data-Strategie in die unternehmensweite Informationsinfrastruktur einfügt.

Den Begriff „Data Lake" prägte der ehemalige CTO von Pentaho, James Dixon, im Jahr 2010 (vgl. [Dix10]). Er bezeichnet sinngemäß alle Daten als Data Lake, die in einem analytischen System vorhanden sind, aber nicht unmittelbar in Data Marts analysiert werden. Seine Annahme war, dass auch in diesen Daten wertvolle Informationen enthalten seien.

Formal ist ein Data Lake ein dezidierter, technologisch unabhängiger Speicherort für quellsystemnahe, anforderungsgetriebene strukturierte und unstrukturierte Daten unter Obacht von Governance-Prozessen, deren Nutzung primär auf Data Science abzielt. Er umfasst atomare und unabhängige Daten in anwendungsneutraler Form, beschrieben durch ein Metadatensystem.

Diese Definition stellt das Zielbild eines Data Lake dar. In der Praxis werden im Bereich Governance und Metadaten häufig aufgrund fehlenden Verständnisses, technologischer Unterstützung oder schlichtweg Kosten-Nutzen-Überlegungen bewusst Kompromisse eingegangen.

Wo werden Data Lakes genutzt?

„Wofür brauche ich denn eigentlich einen Data Lake? Ich habe doch schon ein Data Warehouse." Das mag für viele Anwendungen gelten. Ein Data Lake sollte aber in Betracht gezogen werden, sobald eines oder mehrere der folgenden Kriterien zutreffen:

- Es müssen große Mengen von Daten und Dateien verarbeitet und gespeichert werden
- Data Science ist als Anwendungsgebiet gefordert
- Es müssen polystrukturierte Daten gespeichert oder strukturierte und unstrukturierte Daten verbunden werden
- Der Zugriff auf Daten soll plattformübergreifend virtualisiert werden

Aus Anwendungssicht dient ein Data Lake, im Gegensatz zu vordefinierten Analysemustern in DWH-Data-Marts, zur kontextfreien Analyse heterogener Daten. Getrieben wird das Data-Lake-Konzept vor allem von technologischer Weiterentwicklung: Konkrete Anwendungsfälle, für die ein Data Lake im Gegensatz zum DWH geeignet ist, sind beispielsweise Streaming Analytics, die Analyse medialer Daten oder sozialer Netzwerke, Analytik im sensorischen Bereich, etwa der Predictive Maintenance, Image Recognition oder Sentiment-Analysen – kurzum: Analysen, die auf großen Datenmengen und/oder unterschiedlich strukturierten Daten basieren.

Data Lake versus Data Warehouse / BI

Aus den zuvor beschriebenen Eigenschaften des Data Lake ergeben sich zwangsläufig bestimmte Architekturvarianten (vgl. [Lin18]). Bevor diese auf den Ebenen Systemarchitektur, Prozess- und Datenarchitektur sowie der technischen Architektur betrachtet werden, geht es zunächst um die Abgrenzung von Data Lakes gegenüber Data Warehouse / BI.

Vielfach ist zu hören, dass ein Data Lake nichts anderes als die Wiedergeburt des bekannten Data-Warehouse-Konzeptes sei, nur mit noch mehr Daten, neuen Datentypen/-arten und Techniken. Richtig ist, dass beide Konzepte im Grunde Daten-Repositories darstellen, die sich auf den ersten Blick ähneln. Tatsächlich verfolgen beide Ansätze ähnliche Ziele, haben aber komplementäre Eigenschaften und Methoden zur Zielerreichung, die in Tabelle 1 zusammengefasst sind.

THOMAS WEILER ist BI-Experte bei mayato. Er verfügt über die Erfahrung aus zwei Jahrzehnten erfolgreicher Tätigkeit in Konzeption, Modellierung, Realisierung und Optimierung von Business-Intelligence-Systemen verschiedenster Hersteller.
E-Mail: thomas.weiler@mayato.com

Aus den angeführten – unterschiedlichen – Anwendungseigenschaften folgt, dass Data Lakes und Data Warehousing sich nicht gegenseitig ersetzen, sondern zwei unterschiedliche, sich ergänzende Architekturkonzepte darstellen. Zum Beispiel kann ein Data Lake auch dazu genutzt werden, Daten eines BI-Systems zu archivieren. Auch wenn diese Nutzungsart nicht die ursprüngliche Verwendung eines Data Lake ist, können technische und monetäre Aspekte ihn dafür prädestinieren. Umso mehr wird damit aber auch ein Katalogisierungssystem im Data Lake notwendig – hierzu gibt es mittlerweile Ansätze mit Werkzeugen aus beiden Systemwelten, für Big-Data-Systeme etwa Talend Data Catalog, Atlas oder Navigator. Eine weitere Interaktion kann das Stammdaten-Management darstellen, das in der Praxis vielfach im DWH bereits existiert und Daten des Data Lake, zum Beispiel Prozessdaten, Sensordaten oder Streaming-Daten, zur Auswertung anreichert.

Es existieren auch Anwendungsfälle wie Streaming-Analysen, Alerting Engines, Sensoranalysen oder Predictive Maintenance, die mit beiden Konzepten darstellbar sind. Welches Konzept im

Ziel / Eigenschaft	Data Warehouse / BI-System (RDBMS)	Data Lake (Big Data, NoSQL)
Daten	Daten sind vor der Einspeisung strukturiert und modelliert („schema-on-write").	Speichert alle Arten von Daten, strukturiert, semistrukturiert und unstrukturiert. Erst bei Verwendung der Daten werden diese typisiert („schema-on-read").
Prozesse	Überführung der Daten in Informationen durch Vereinheitlichung, Konsolidierung, Historisierung und Versionierung	Primäre Verarbeitung von Rohdaten und Fokussierung auf die „as-is"-Sicht, die sich in Versionen historisch abbilden lässt. Data Lakes können jedoch auch über mehrere Schichten/Zonen verfügen, in denen Daten qualitätsgesichert oder angereichert werden.
Speicherung	Ausgereifte Basistechnologie mit RDBMS und oder In-Memory-Technologie. Teilweise komplexe Infrastrukturen spezialisierter Anbieter für performante Analytik.	Überwiegend Verwendung von Big Data-Technologien, die Open-Source-Komponenten auf kostengünstiger Infrastruktur ermöglichen, aber meist neue Skills im Unternehmen erfordern
Nutzungsmuster	Performante Analytik in vordefinierten Analyseräumen (kontextsensitiv) Eher starre „Advanced Analytics" durch Modellerweiterung	Weniger performante Analytik in vordefinierten Analyseräumen Flexiblere „Advanced Analytics" durch Datenerweiterung Kontextfreie Analytik mit Latenz
Sicherheit	Sehr ausgereift aufgrund langjährigen Einsatzes und in vielfältigen Schichten des DWH abbildbar	Bei weitem nicht ausgereift und aktuell ein wichtiges Thema
Nutzer	Ziel des DWH-Konzepts: Nutzung der Daten durch Fachbereiche, Management und spezialisierte Analysten in strukturierten Analyseräumen	Data-Lake-Nutzer sind primär Data Scientists, die keine (Daten-) Grenzen wollen und kennen. Das kann nur mit völliger Strukturfreiheit erreicht werden.

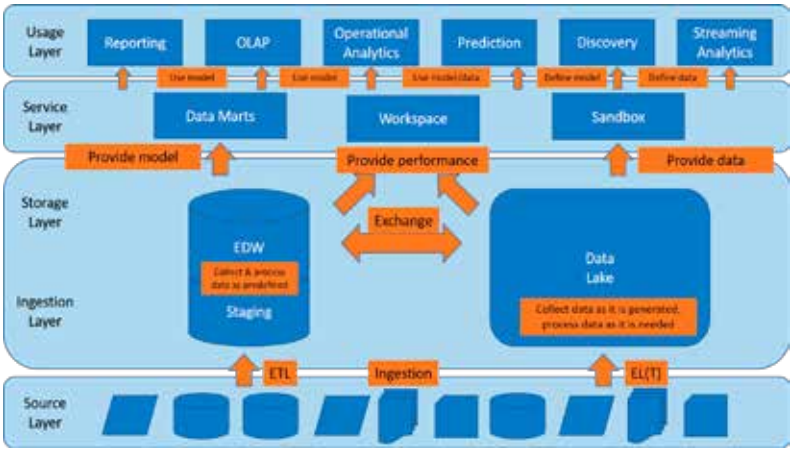
Einzelfall tragfähiger ist, richtet sich nach dem Umfang fachlicher Regeln sowie Skill-Kosten-Nutzen-Aspekten.

Eines ist dabei klar: Der Reifegrad des Data Warehousing ist im Vergleich zu dem des Data Lake sehr hoch. Unternehmen sind heute in der Lage, mit den richtigen Werkzeugen quasi ein „DWH von der Stange“ zu implementieren, nicht zuletzt durch „Data Warehouse Automation“-Werkzeuge und standardisierte Modellierungstechniken wie zum Beispiel Data Vault. Gleichzeitig reicht das Anwendungsspektrum des DWH nicht mehr aus. Mit den drei Vs des Big Data – Volume, Variety und Velocity – haben mittlerweile Datenarten in die Analytik Einzug gehalten, die mit technischen DWH-Komponenten nicht mehr effizient bearbeitbar sind. Technische Erweiterungen sind zwar möglich, aber kostspielig und aufwendig im Vergleich zu Data-Lake-Mechanismen, auch wenn diese bei weitem noch nicht so standardisiert und automatisiert sind wie im DWH. Daher ist das notwendige Know-how beim Aufbau eines Data Lake ein kritischer Aspekt für dessen Erfolg.

Systemarchitektur

Eine gängige Systemarchitektur einer Data-Lake-Lösung inklusive BI-System ist aus den vorangegangenen Ausführungen direkt ableitbar. Sie besteht neben den übergreifenden Funktionalitäten wie Datenqualität, Governance und Metadaten aus der Datenquellen-, Datenhaltungs-, Service- und Nutzungsschicht. Ein wesentlicher Kernpunkt ist hierbei das Zusammenspiel zwischen der Data-Warehousing- und der Data-Lake-Datenhaltung. Sie tauschen Daten untereinander aus und werden auf ähnliche Art und Weise von den „Abnehmern“ konsumiert. Das heißt, aus Systemsicht entsteht aus den beiden Konzepten ein neuer – logischer – Datenpool, den man als „analytischen Unternehmensdatensee“ bezeichnen könnte und der einen wirklichen „Single Point of Truth“ (SPOT) darstellt. Denn in diesem Datensee finden sich in einer (logischen/virtuellen) Instanz auch die Originaldaten, die im Gegensatz zu den konsolidierten und integrierten Daten in einem Data Warehouse die tatsächliche, unveränderte Wahrheit darstellen.

Abb. 1: Data-Lake-Prozess- und Datenarchitektur



In einer Data-Lake-Systemarchitektur werden partiell neue Begrifflichkeiten eingeführt und dabei teilweise auf die bekannte Data-Warehouse-Terminologie zurückgegriffen. Daher seien die wichtigsten Komponenten kurz erklärt.

- Die Schichtenarchitektur orientiert sich sehr stark an der klassischen BI-Systemarchitektur. Allerdings spricht man bei der „Staging Area“ oder „Landing Zone“ nun vom „Ingestion Layer“. Dies begründet sich in der Tatsache, dass der Data Lake nur über eine zentrale Datenschicht verfügt (im Gegensatz zur Trennung im EDW in Staging Area und Core Data Warehouse).
- Im Service Layer finden sich die bekannte OLAP-Struktur, der Workspace beziehungsweise eine Sandbox. Die Sandbox steht dabei synonym für eine Kombination aus Self-Service-Funktionalitäten und beschränktem, aber wahlfreiem Datenzugriff. Insbesondere für analytische Prozesse im Bereich des Data Mining oder der Prediction müssen Modelle flexibel gestaltet werden. Dabei kann der Datenzugriff der eingesetzten Analysewerkzeuge logisch durch entsprechende Schnittstellendefinitionen und -komponenten oder auch persistent durch weitere Datenspeicher erfolgen. Steht jedoch der wahlfreie Datenzugriff im Vordergrund und soll dieser möglichst performant erfolgen, kommt der Workspace zum Tragen. Dieser ist eine zusätzliche Datenschicht für Performance, zum Beispiel durch Einsatz von In-Memory-Komponenten wie In-Memory-Datenbanken oder -Engines.
- Der Usage Layer stellt die möglichen Analyseverfahren gleichberechtigt dar. Dabei ist nicht zwingend vorgegeben, welche Service-Layer-Komponenten genutzt werden. Hier ist jede Kombination möglich und im Einzelfall zu definieren.

Prozess- und Datenarchitektur

Der wesentliche Unterschied einer Data-Lake-Prozess- und Datenarchitektur gegenüber dem Warehousing ist die Bereitstellung von Daten in Echtzeit. Diese Daten unterliegen keiner Modellierung. Daher wird der Ingestion-Prozess auch als EL(T) bezeichnet. Im Gegensatz zum DWH-ETL-Prozess finden keine Transformationen statt, das heißt keine Integration, keine Konsolidierung. Eine Transformation findet, wenn überhaupt, erst bei der Nutzung der Daten statt. Dabei ist nicht determiniert, ob dies unter der Hoheit des Data Lake oder des nutzenden Analysewerkzeugs geschieht.

Weiterhin kooperieren DWH (EDW) und Data Lake. Beispielsweise können Stammdaten aus dem DWH mit Bewegungsdaten/Streams im Data Lake zur Sensoranalyse kombiniert oder Ergebnisse aufwendiger Prediction-Analysen zur einfachen Nutzung in das DWH zurückgeschrieben werden. Der Umfang und die Tiefe der Kooperation sind vielfältig. Es ist sogar denkbar, die klassische Staging Area im DWH durch den Data Lake zu ersetzen.

Die Architektur in Abbildung 1 unterscheidet in modellbasierte Datenbereitstellung aus dem Bereich des klassischen Data Warehousing und reine Datenbereitstellung im Data Lake. In der Architektur bilden die beiden Prozesse „Ingestion“ und „Provision“ die Kernfunktionalitäten. Daher ist es wichtig, die wesentlichen Kriterien zu diesen Prozessen zur Ableitung der technischen Architektur zu kennen.

Für den Prozess „Ingestion“ ist dabei zu untersuchen:

- Welche Datenarten und -typen müssen verarbeitet werden (Variety)?
- Mit welcher Frequenz müssen die Daten bereitgestellt werden (Velocity)?
- Welche Latenzzeiten sind möglich (Batch, Micro-Batch, Stream)?
- Mit welchen Datenmengen ist zu rechnen (Volume)?
- Welche Auswirkungen hat die Datenqualität (Veracity)?
- Bestehen zeitliche Abhängigkeiten (Timeline Consistence)?
- Welche Formen des Datenaustauschs und der Änderungserkennung liegen vor (Messaging, Push/Pull, CDC)?

Für „Provision“ sind wichtige Eckpunkte:

- Welche Ausprägungen haben die drei Vs (Volume, Variety, Velocity)?
- Welche Abfrage- und Entlademuster sind bekannt?
- Ist SQL-Unterstützung erforderlich?
- Sind Datenverbindungen notwendig (Joins)?
- Wie wahrscheinlich sind Schemaänderungen?
- Werden historische Abfragen benötigt?
- Welche Sicherheits- und Schutzmechanismen werden erwartet?
- Welche Datenspeicherungsstrategien liegen vor (Retention Policy)?

Technische Architektur

Die technische Architektur eines Data Lake bietet einen großen Spielraum. Es existiert eine Vielzahl von möglichen technischen Produkten aus dem Big-Data-Umfeld. Aus diesem Grund ist es unmöglich, „die eine“ technische Referenzarchitektur zu definieren. In Abbildung 2 wird daher auszugsweise auf gängige Beispielkomponenten zurückgegriffen.

Grundsätzlich können beliebige Komponenten der technischen Architektur cloudbasiert oder vor Ort genutzt werden, auch in Mischformen. Welche Aufteilung im Einzelfall sinnvoll ist, muss basierend auf der Unternehmensstrategie, dem vorhandenen Know-how und der Infrastruktur geklärt werden.

Fazit

„Vision trifft Realität“, das gilt auch im Data-Lake-Umfeld. **Eine Vision ist es, dass ein Data Lake die Probleme der Datensilos und Informationsvielfalt sofort löst.** Diese Aussage stimmt nur, wenn die Governance-Strukturen funktionieren. „Ansonsten wird aus dem Datensee ein Datensumpf, eine Ansammlung nicht verknüpfter Datenpools.“ (vgl. [ITR 14]) Unternehmen leiden dann immer noch unter Daten-Silos, „nur eben an einer Stelle“.

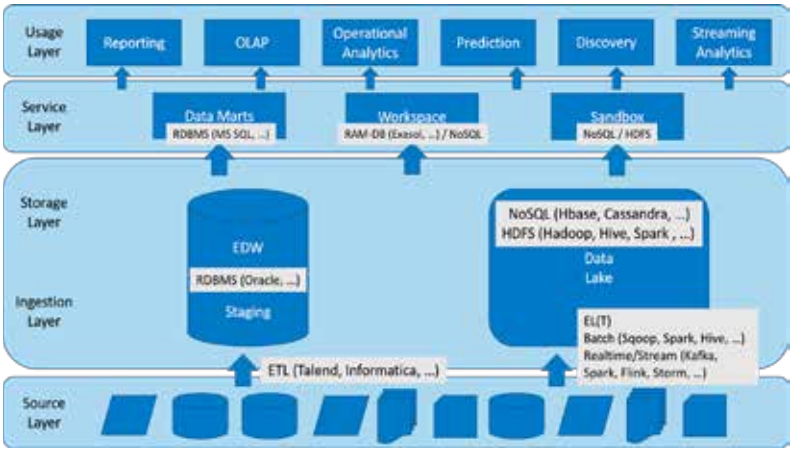


Abb. 2: Beispielhafte technische Data-Lake-Architektur

Ebenso ist es eine Vision, dass **ein Data Lake ein Produkt ist, das man kaufen kann und das immer auf Hadoop aufsetzt** (vgl. [IBM 18]. Ein Data Lake ist keine „Lösung von der Stange“, sondern eine Technologie-unabhängige Referenzarchitektur. Dass Data Lakes und Big Data beziehungsweise Hadoop fast immer zusammen genannt werden, liegt schlichtweg daran, dass Hadoop eine sehr effiziente technische Lösung darstellt und mittlerweile zu einem mächtigen Ökosystem gewachsen ist.

Die Vorstellung, dass ein Data Lake alle Daten aufnimmt, sie allen Anwendern zur Verfügung stellt und andere analytische Systeme überflüssig macht, stimmt so auch nicht. Ein Data Lake, der alle Daten aufnimmt, verkommt in kürzester Zeit zu einem unkontrollierten und unkatalogisierten zentralen Datenfriedhof. Hier wären wir wieder beim Ausgangspunkt: Die Datengrundlage hat sich verändert, die analytischen Prozesse sind komplexer geworden und gleichzeitig hat die technologische Entwicklung dieser Veränderung Rechnung getragen. Die klassische Analytik, wie sie im Data-Warehousing-Konzept umgesetzt ist, hat weiterhin Bestand. Nach wie vor sind klassische Bestands-, Produkt-, Transaktions- und Prozessanalysen wichtige Werkzeuge der Unternehmenssteuerung – und werden es auf absehbare Zeit auch bleiben. Insofern ergänzen sich die beiden Konzepte. Oder um es neudeutsch auszudrücken: „Stay with BI, improve with Data Lake.“

Literatur

[Alb17] Albrecht, J.: Data Lake Architektur. 31.1.2017, https://de.slideshare.net/JensAlbrecht2/data-lake-architektur-von-den-anforderungen-zur-technologie, abgerufen am 29.7.2019

[Dix10] Dixon, J.: James Dixon's Blog – Pentaho, Hadoop, and Data Lakes. 14.10.2010, https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/, abgerufen am 29.7.2019

[IBM18] IBM: Five myths about the data lake. 2018, www.ibm.com/downloads/cas/9ENMLD0L, abgerufen am 29.7.2019

[ITR14] IT-Rebellen: Gartner warnt vor Data Lakes. 4.8.2014, www.it-rebellen.de/2014/08/04/gartner-warnt-vor-data-lakes/, abgerufen am 29.7.2019

[Lin18] Linstedt, D.: Defining a Data Lake. 14.2.2018, www.youtube.com/watch?v=tDNjI1Yvqxw, abgerufen am 29.7.2019