Bild: InputUX, stock.adobe.com

Architektonischer Tapetenwechsel: Datenschatz sucht neues Zuhause

Vom ETL-Altbau ins Lakehouse

Ein Beitrag von Florian Kretzschmar, Hannes Reinhardt und Mattis Hartwig

Die fortschreitende Digitalisierung von Geschäftsprozessen bringt eine große Vielfalt an nutzbaren Datenquellen mit sich. Egal welche Branche, Unternehmen nutzen eine breite Palette an datenerzeugenden und -speichernden Systemen. Eng daran geknüpft ist auch das nahezu exponentielle Wachstum der verfügbaren Datenmengen. Vorhandene Datenintegrationsstrategien geraten durch die damit einhergehenden Anforderungen zunehmend an ihre Grenzen. Viele Unternehmen erkennen diesen Modernisierungsdruck und machen sich auf die Suche nach neuen Lösungen. Dieser Artikel zeigt anhand eines Praxisberichts, welche Chancen und Herausforderungen der Umzug auf eine Data-Lakehouse-Plattform mit sich bringt.

Zweifelsfrei stellen Daten eine wertvolle und vielseitig einsetzbare Ressource dar, deren Ausschöpfung einen echten Mehrwert beiträgt – sowohl in einzelnen Geschäftsbereichen als auch zur allgemeinen Wettbewerbsfähigkeit. So nimmt die Konzeption und Umsetzung einer gelungenen Datenintegrationsstrategie eine wichtige Rolle in der Ausrichtung moderner Unternehmen ein. Nicht zuletzt eröffnen die jüngsten Fortschritte im Bereich der generativen KI neue Formen der Interaktion mit unternehmensinternem Wissen.

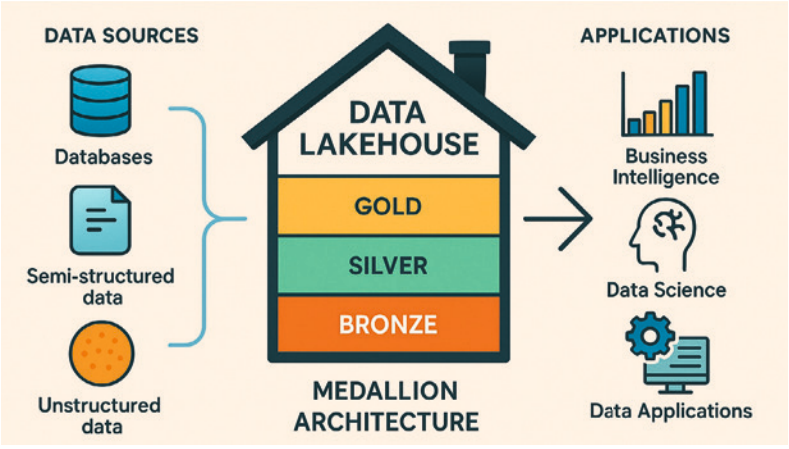


Abb. 1: Data Lakehouse und Medaillon-Architektur (Quelle: GPT-4o ImageGen)

Komplexere Anforderungen …

Mit dem rasant wachsenden Datenvolumen sowie der breiten Diversität an Quellen und Anwendungsfällen geht eine ganze Reihe an neuen technischen Herausforderungen einher. Dazu gehört die Notwendigkeit, alle erdenklichen Datenformate effizient zu lesen, zu speichern und weiterzuverarbeiten – und das möglichst schnell und effizient. Gleichzeitig steigen sowohl der Bedarf an Hardware-Ressourcen als auch die Ansprüche an deren Skalierbarkeit enorm. Ein weiterer Faktor ist die Unterstützung zusätzlicher Programmiersprachen in ETL-Tools, etwa um Machine- Learning-Algorithmen (ML) direkt in Daten-Pipelines zu integrieren oder flexibler in der Implementierung von Transformationen zu sein.

… und deren Konsequenzen

Bestehende Dateninfrastrukturen wurden oft über viele Jahre hinweg aus verschiedenen, spezialisierten Systemen und Tools aufgebaut. Hierbei besteht das Risiko, einigen der vielfältigen Anforderungen nicht beziehungsweise nicht mehr vollends gerecht zu werden. Im Hinblick auf die Ausschöpfung der zahlreichen Potenziale kann dies zu kleinen bis hin zu unüberwindbaren Hindernissen führen [Arm21].

zum Inhalt

Diese Aspekte verdeutlichen, dass die Modernisierung ihrer Dateninfrastruktur kein optionales Vorhaben, sondern vielmehr eine wichtige Voraussetzung für den langfristigen Erfolg von Unternehmen ist. Dafür bedarf es einer Abkehr von traditionellen – bisher sicherlich bewährten – Ansätzen und die Adaption neuer, zukunftsfähigerer Konzepte.

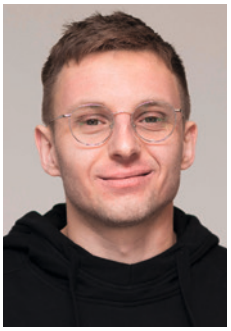
Die Idee des Data Lakehouse

Ein Data Lakehouse (DLH) ist eine moderne Datenarchitektur, die die Stärken von Data Lakes – Flexibilität und Skalierbarkeit – mit denen von Data Warehouses – unter anderem Datenverwaltungsfunktionen und Query-Performance – in einem einzigen System vereint [HaZ24; Sch24]. Konkret bedeutet das: Alle Daten, ob strukturiert (SQL-Datenbanken), semi-strukturiert (JSON, Parquet) oder unstrukturiert (Text, Bilder), liegen zentral im DLH. Durch diese Verschmelzung entstehen neue Möglichkeiten, BI-Analysen, Data Science und ML auf dem gesamten Datenbestand durchzuführen, ohne zwischen verschiedenen Plattformen wechseln zu müssen (siehe Abbildung 1).

Folgende Punkte lassen sich als **Schlüssel-Features** für ein Data Lakehouse anführen [Clo24]:

- **Offenes und kosteneffizientes Speichersystem:** DLHs basieren auf günstigem Cloud-Storage (zum Beispiel Objektspeicher wie S3) und offenen Dateiformaten wie Parquet. Dies senkt die Speicherkosten und vermeidet einen Vendor-Lock-in. Die Trennung von Speicher und Rechenleistung erlaubt zudem eine nahezu unbegrenzte Skalierbarkeit, da bei wachsender Datenmenge einfach weitere Ressourcen angebunden werden können.
- **ACID-Transaktionen und Datenqualität:** Eine transaktionale Metadatenschicht bietet ACID-Transaktionssicherheit und eine konsistente Datenversionierung direkt auf den Rohdaten im Data-Lake-Speicher. Gleichzeitig können Schema-Änderungen kontrolliert vorgenommen werden (Schema-Evolution), sodass die Datenqualität hoch bleibt.
- **Unterstützung vielfältiger Transformationen und Workflows:** Auf einer DLH-Plattform sind sowohl klassische SQL-Analysen als auch ML-Workflows ausführbar. Moderne Engines (zum Beispiel Apache Spark) ermöglichen auch auf riesigen Rohdatensätzen performante Ad-hoc-Abfragen. Gleichzeitig können Data Scientists direkt auf den Daten im Lakehouse arbeiten, ohne diese erst in andere Umgebungen übertragen zu müssen.
- **Einheitliches Daten- und Governance-Modell:** Da alle Daten in einem System liegen, lässt sich Data Governance zentral und einheitlich umsetzen – Sicherheitsrichtlinien, Zugriffsrechte und Datenkataloge gelten für den gesamten Datenbestand. Ein DLH etabliert somit eine einheitliche Datenquelle für das Unternehmen. Alle Fachbereiche greifen auf dieselben, stets aktuellen Daten zu, wodurch Redundanzen und Inkonsistenzen vermieden werden.

Die genannten Features erleichtern die Datenintegration erheblich. So entfällt beispielsweise die Trennung zwischen Rohdatenspeicherung und Analyse: In traditionellen Architekturen mussten Daten erst aus dem Data Lake ins Data Warehouse übertragen werden – ein aufwendiger und langsamer Prozess. Zwar müssen Daten auch im Data Lakehouse zunächst aus den Quellsystemen geladen werden, doch



FLORIAN KRETZSCHMAR unterstützt die singularIT im Bereich Data Science und Data Engineering. Er verfügt über umfangreiche Programmierkenntnisse in Sprachen wie Python und R. Als wissenschaftlicher Mitarbeiter an der Technischen Universität Chemnitz hat er Module in den Bereichen Data Science, Machine Learning und Statistik geleitet. Des Weiteren hat Florian Kretzschmar bereits erste Publikationen zur Anwendung von Machine Learning veröffentlicht, die seine umfangreiche Expertise in diesem Bereich unterstreichen.
E-Mail: florian.kretzschmar@singular-it.de

HANNES REINHARDT konzipiert und entwickelt bei der singularIT individuelle Backend-Lösungen, die auf die Anforderungen der Kunden zugeschnitten sind. Dabei ist er aktiv in den gesamten Konzeptions- und Entwicklungsprozess eingebunden. Er orientiert sich an aktuellen Entwicklungen und gestaltet Datenbanken sowie REST-APIs nach Best Practices. Seine umfangreiche Erfahrung mit Datenbanken und der Datenbanksprache SQL befähigt ihn zudem zur Umsetzung moderner, cloudbasierter ELT-Prozesse zur Optimierung von Business Intelligence in Kundenprojekten.
E-Mail: hannes.reinhardt@singular-it.de

DR. MATTIS HARTWIG ist einer der beiden Geschäftsführer der singularIT. Er hat Wirtschaftsinformatik und Sustainability Management studiert und parallel die singularIT gegründet, die mittlerweile über 50 Softwareentwickler und Data Scientists beschäftigt und individuelle Softwarelösungen in den Bereichen Web, Cloud, Mobile, Data Analytics, Data Science und Künstliche Intelligenz entwickelt. In seiner Promotion im Bereich Künstliche Intelligenz hat er intensiv zu probabilistischen Modellen geforscht und seine Ergebnisse international publiziert. Neben seiner Tätigkeit als Geschäftsführer forscht er am Deutschen Forschungszentrum für Künstliche Intelligenz zu Themen rund um KI und übernimmt Lehrtätigkeiten an der Universität zu Lübeck.
E-Mail: mattis.hartwig@singular-it.de

zum Inhalt

Abb. 2: Traditionelle Lösungen sorgen zunehmend für Frust und Probleme (Quelle: GPT-4o ImageGen)

entfällt im Vergleich zu klassischen Architekturen die aufwendige Vorabtransformation. Die Daten werden in ihrem Rohformat gespeichert und können bedarfsorientiert transformiert werden (Extract-Load-Transform statt Extract-Transform-Load). Dadurch beschleunigt sich der Analysezugriff erheblich und neue Datenquellen lassen sich einfacher integrieren. Die Integration wird nicht nur **effizienter**, sondern auch **robuster** – weniger Schnittstellen bedeuten weniger Fehlerrisiken und geringeren Wartungsaufwand. Zudem sinkt die Total Cost of Ownership (TCO) [Dre24], da redundante Altsysteme wegfallen.

Die Versprechen des Data Lakehouse zeigen, warum viele Organisationen auf dieses Konzept setzen [Dre24]. Im nächsten Abschnitt wird exemplarisch erläutert, wie ein Unternehmen seine bestehenden ETL-Prozesse auf die Lakehouse-Plattform von Databricks migriert hat, um von diesen Vorteilen zu profitieren.

Warum ein Tapetenwechsel nötig war

Im Mittelpunkt steht ein Unternehmen, das innerhalb weniger Jahre zu einem der größten europäischen Player seiner Branche avancierte. Das starke Wachstum ging mit einer rasant zunehmenden Datenmenge sowie in Komplexität und Umfang steigenden Anforderungen an den BI-Bereich einher. Diese umfassten nicht nur die Erweiterung und Neuentwicklung von Standard-Reports, sondern auch die flexible und zügige Bearbeitung von Ad-hoc-Anfragen aus verschiedenen Fachbereichen.

Obwohl bereits eine Vielzahl an Datenquellen genutzt wurde – etwa eine SQL-Datenbank, die die Backends verschiedener Software-Lösungen beinhaltet, oder ein Microsoft Sharepoint –, waren die darauf aufbauenden ETL-Prozesse heterogen und ineffizient. Dabei wurde ein Data Warehouse per Pentaho Data Integration gespeist, während der Großteil der Transformationen direkt in PowerBI mit einer Mischung aus DAX-, PowerQuery- und SQL-Abfragen implementiert war. Die im Rahmen der ursprünglich eingesetzten Architektur verfügbaren Rechenkapazitäten erwiesen sich aufgrund der mittlerweile stark gestiegenen Menge an zu verarbeitenden Daten immer öfter als unzureichend. Die daraus resultierenden langen Ladezeiten bei der Aktualisierung von BI-Reports sorgten nicht nur für eine suboptimale User-Experience für die Endverbraucher, sondern erschwerten auch die Arbeit der Entwickler und Analysten beträchtlich.

Für die Verbesserung der bestehenden Datenmodelle und die Umsetzung neuer Anforderungen gab es unter diesen Bedingungen wenig bis keinen Spielraum. Gleichzeitig wäre eine Erweiterung der vorhandenen Rechnerinfrastruktur mit einer komplizierten Bedarfskalkulation, dem Kauf neuer Hardware und zusätzlichem Personalaufwand auf Seiten der IT-Abteilung verbunden, ohne dass ein späteres Wiederaufkommen der eben geschilderten Probleme dauerhaft hätte ausgeschlossen werden können. Hieraus entstand schließlich die Notwendigkeit und der Wunsch nach einem neuen Lösungsansatz. Nach einer eingehenden Vorabrecherche entschied sich das Unternehmen schließlich für die Migration zu einer modernen Lakehouse-Architektur auf der Plattform Databricks (DBX).

Die Einrichtung des neuen Zuhauses …

Das Projekt startete mit einer Onboarding- und Einarbeitungsphase, die sich stark an bewährten Best Practices der agilen Software-Entwicklung orientierte. Neben dem gegenseitigen Kennenlernen wurden klare Rollen innerhalb des Projektteams definiert und sowohl technische als auch fachliche Ansprechpartner benannt. In diesem Rahmen wurde auch gezielt der individuelle Einarbeitungsbedarf identifiziert, insbesondere im Hinblick auf theoretische Grundlagen moderner Datenarchitekturen und die den Daten zugrunde liegenden Business-Logiken. Parallel dazu machten sich die Teammitglieder mit der Funktionsweise und den Werkzeugen (siehe Tabelle 1) der Lakehouse-Plattform vertraut, um ein technologisches Verständnis für die neue Arbeitsumgebung zu erlangen.

Die nächste Etappe stand im Zeichen der Konzeption und Datenmodellierung. Konkret verstän-

zum Inhalt

Technologie/Tool	Im Projekt	Alternativen
Speicher-/Tabellenformat	Delta Lake	Apache Hudi, Apache Iceberg
Execution Engine	Apache Spark	Apache Flink, Apache Hadoop
ELT	DBX DLT Pipelines	dbt, SQLMesh
Orchestrierung	DBX Workflows	Apache Airflow, Dagster
Qualitätssicherung	DBX DLT Expectations	Great Expectations

Tab. 1: Im Projekt verwendete Tools und mögliche Open-Source-Alternativen

digte man sich auf eine Medaillon-Architektur, innerhalb derer die Daten schichtweise nach ihrem Reifegrad organisiert sind. Die Bronze-Schicht fungiert dabei als Landezone: Hier werden Rohdaten abgelegt, anschließend bereinigt und in ein einheitliches Format gebracht. Die Gold-Schicht wurde klassisch in Form von Data Marts für die einzelnen Geschäftsbereiche (Finance, Operations, Service etc.) realisiert, um (materialisierte) Sichten für das Berichtswesen bereitzustellen. Komplettiert wird die Architektur durch die Silber-Schicht, in der alle informationsanreichernden Transformationen stattfinden und die als Speicherort der daraus resultierenden Tabellen dient.

Die dreistufige Hierarchie der Datenbankobjekte in Databricks erleichterte die Implementierung des Architekturkonzepts samt zugehöriger Namenskonventionen. Dabei entspricht jede Schicht einem Katalog, während die darin liegenden Schemata entweder die Quellsysteme (Bronze, Silber-Basis) oder die fachlichen Domänen (angereicherte Silber-, Gold-Schicht) referenzieren. Somit entstand eine übersichtliche, leicht zu navigierende Datenlandschaft. Ferner orientierte sich auch die Strukturierung und Orchestrierung der ELT-Prozesse an der Medaillon-Architektur: Hier wurden für jedes Ziel-Schema sogenannte Master-Jobs erstellt, innerhalb derer dann verschiedene Teilprozesse gesteuert werden konnten. Abgerundet wurde das Set-up durch die Einrichtung von getrennten Workspaces für Testing/Entwicklung und Produktion sowie die Implementierung von Continuous-Integration- und Continuous-Delivery-Prozessen (CI/CD).

… und dessen Vorzüge

Im Rahmen des Entwicklungsstarts begann das Projekt-Team zunächst mit der Migration der bestehenden ETL-Prozesse und Data Marts in die neu geschaffene Lakehouse-Umgebung. Während die Anbindung der verschiedenen Datenquellen größtenteils reibungslos verlief, stellte das Refactoring der in unterschiedlichen Sprachen implementierten Transformationen eine anspruchsvolle Herausforderung dar, die zusätzlich zur raschen Adaption der neuen Tools auch ein Verständnis der zu ersetzenden Systeme verlangte.

Die sich daran anschließende Entwicklung einheitlicher Verfahren zur Sicherung der Datenqualität markierte den letzten Schritt auf dem Weg zum Go-Live. In den darauffolgenden Monaten konnte das Team den Produktiv-Betrieb aufnehmen und auf verschiedenen Ebenen von der neuen Lösung profitieren:

- **Spaß an der täglichen Arbeit:** Die Kombination aus einheitlicher Architektur, sicherem Deployment-Prozess und leistungsfähigen Tools hat die Arbeit von Entwicklern und Analysten deutlich vereinfacht und angenehmer gemacht.
- **Erhöhte Reaktionsfähigkeit und Flexibilität:** Ad-hoc-Anfragen und Anforderungen, die zuvor aufgrund technischer Engpässe im Backlog verharrten, konnten nun zeitnah umgesetzt werden. Durch die flexible Anbindung neuer Datenquellen – zum Beispiel über APIs oder Partner-Konnektoren – konnten zahlreiche Chancen zur Erweiterung bestehender Analysen und zur Entwicklung innovativer Datenmodelle wahrgenommen werden.
- **Self-Service & Aktualität:** Die Einführung von Self-Service-Funktionalitäten und die Möglichkeit zur untertägigen Aktualisierung bestimmter Berichte haben den Nutzwert für End-User erheblich gesteigert – ein Szenario, das mit der vorherigen Lösung nicht realisierbar gewesen wäre.
- **Konzentration auf Entwicklung:** Durch die Nutzung von serverlosem Compute mit intelligentem Auto-Scaling [Dbx25] entfallen sowohl die zeitaufwendige Konfiguration und Verwaltung von Rechenressourcen als auch die bisherigen Kapazitätsengpässe.
- **Technologische Zukunftsfähigkeit:** Die Wahl einer State-of-the-Art-Plattform ist ein strategischer Vorteil. Die umfangreichen Möglichkeiten für Transformation und Orchestrierung, granulare Zugriffskontrollen auf Datenobjekte und die native Unterstützung von ML-Tools schaffen eine solide Basis für künftige Entwicklungen und Projekte, wie etwa die Integration von generativer KI in Geschäftsprozesse.

Den Projektabschluss bildete die Vorbereitung der vollständigen Übergabe an die internen Mitarbeiter des Unternehmens. Hierfür wurde in Kollaboration aller Beteiligten eine umfangreiche Dokumentation ausgearbeitet. Wichtige Inhalte waren detaillierte Beschreibungen der implementierten ELT-Prozesse sowie Erläuterungen zu Namens- und Strukturkonventionen, aber auch theoretische Hintergründe und nützliche Tipps und Hinweise für die tägliche Arbeit.

Erkenntnisse & Fazit

Der Umzug in ein Data Lakehouse stellte ein anspruchsvolles und spannendes Projekt dar, das eine Vielzahl von Herausforderungen und Erkenntnissen mit sich brachte:

zum Inhalt

Abb. 3: Nach dem Umzug: ein glückliches Team, bereit für neue Aufgaben (Quelle: GPT-4o ImageGen)

- **Konzeption & Kompetenzen:** Für einige Aufgaben, wie die Konzeption von Datenarchitektur und ELT-Pipelines, oder die Implementierung von CI/CD ist bereits bestehendes technisches Know-how im Projektteam unverzichtbar. Gleichzeitig konnten alle Beteiligten von einem offenen Wissensaustausch profitieren und nachhaltige Kompetenzen aufbauen.
- **Planung & Kommunikation:** Die frühzeitige Implementierung von Best Practices aus der Softwareentwicklung, etwa eine agile Arbeitsweise und ein klares Anforderungsmanagement per Ticket-System, sowie eine enge, direkte Kommunikation zwischen allen Stakeholdern erwiesen sich als entscheidende Erfolgsfaktoren.
- **Datenqualität der Quellsysteme:** Einige der genutzten Datenquellen wiesen erhebliche Qualitätsprobleme auf, wie Dubletten, inkonsistente Formate oder unvollständige Datensätze. Zwar waren diese vor der Migration bereits bekannt, konnten aber mithilfe der neuen Architektur nun einheitlicher und frühestmöglich behandelt werden. Für die eingehende Untersuchung der Datenqualität und die Implementierung robuster Cleaning-Prozesse sollte ein hinreichendes Zeitbudget eingeplant werden.
- **Kostenkontrolle und Kosteneffizienz bei Cloud-Computing:** Durch die nahezu unbegrenzte Skalierbarkeit der Cloud gehören Engpässe bei Rechenressourcen theoretisch der Vergangenheit an. Praktisch bringt das meist damit einhergehende Pay-as-you-go-Preismodell jedoch Risiken mit sich. Ein präzises Monitoring der Compute- und Storage-Kosten muss selbst aufgebaut werden, um eine effektive Kostenkontrolle zu gewährleisten. Gleichzeitig spielt Kosteneffizienz eine entscheidende Rolle. Die Vorgabe einer kosteneffizienten Lösung konnte erfolgreich eingehalten werden. Wo zuvor leistungsstarke Laptops oft bei der Belastung durch Berichtsaktualisierungen an ihre Grenzen stießen, sorgt nun eine plattformseitig hochoptimierte Infrastruktur dafür, dass monatliche Betriebskosten im Bereich von dreistelligen bis unteren vierstelligen Beträgen bleiben. Gezielte Code-Optimierungen tragen zudem erheblich dazu bei, weitere Einsparungen zu erzielen.

Abschließend lässt sich festhalten, dass sich die Modernisierung der Datenintegrationsstrategie durch den Umstieg auf die Lakehouse-Plattform für das Unternehmen eindeutig gelohnt hat. Rückblickend wäre ein früherer Einsatz der Lösung ratsam gewesen, um das Aufkommen der geschilderten Probleme zu vermeiden. Vor allem Unternehmen, die sich noch im Aufbau einer Datenstrategie befinden, sollten frühzeitig auf ein Lakehouse setzen, um von der Skalierbarkeit, Flexibilität und Zukunftssicherheit der Lösung optimal zu profitieren.

Literatur

[Arm21] Armbrust, M. et al.: Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. Conference on Innovative Data Systems Research 2021. https://people.eecs.berkeley.edu/~matei/papers/2021/cidr_lakehouse.pdf, abgerufen am 23.4.2025

[Clo24] Cloudera Inc.: The Open Data Lakehouse. Whitepaper, 4.12.2024, https://www.cloudera.com/content/dam/www/marketing/resources/whitepapers/the-open-data-lakehouse.pdf?daqp=true#, abgerufen am 24.4.2025

[Dbx25] Databricks Inc.: Databricks Lakehouse-Plattform. 2025, https://www.databricks.com/de/glossary/serverless-computing, abgerufen am 23.4.2025

[Dre24] Dremio Corporation: State of the Data Lakehouse. Whitepaper, 2024, https://www.dremio.com/wp-content/uploads/2023/11/whitepaper-2024-state-of-the-data-lakehouse_report.pdf?, abgerufen am 23.4.2025

[HaZ24] Harby, A. / Zulkernine, F.: Data Lakehouse: A Survey and Experimental Study. SRRN, 2024. https://ssrn.com/abstract=4765588, DOI: https://dx.doi.org/10.2139/ssrn.4765588

[Sch24] Schneider, J. et al.: The Lakehouse: State of the Art on Concepts and Technologies. In: SN COMPUT. SCI. 5, 449 (2024)

zum Inhalt