

Data Lake: Zentrale Datenaufbewahrung für den technischen Einkauf

Ein Data Lake ist ein zentraler, skalierbarer Speicher, in dem Unternehmen Daten in ihrem ursprünglichen, rohen Format aus unterschiedlichsten Quellen zusammenführen. Für den technischen Einkauf bedeutet das: Stammdaten, Bestellhistorien, Lieferanteninformationen und Maschinendaten liegen an einem Ort und stehen für Analysen bereit. Damit wird der Data Lake zur technischen Grundlage für datengetriebene Entscheidungen in Beschaffung und Supply Chain.

Das Wichtigste in Kürze

- Ein Data Lake speichert strukturierte, semi-strukturierte und unstrukturierte Daten meist im Originalformat (Schema-on-Read).
- Im Unterschied zum Data Warehouse erfolgt keine vorab definierte Strukturierung, sondern eine bei der Auswertung.
- Für den Einkauf liefert er die Datenbasis für Lieferantenrisiko-Analysen, Should-Costing und Demand Planning.
- Typische Quellen sind ERP-Systeme, Einkaufsplattformen, Lieferantenportale, Sensordaten und Dokumente.
- Ohne Data Governance drohen Data Swamps – ungenutzte, unzuverlässige Datensammlungen.
- In der Praxis fallen laufende Speicher- und Betriebskosten häufig deutlich geringer aus als bei klassischen Data Warehouses.

Definition: Data Lake

Ein Data Lake ist eine zentrale Speicherarchitektur, die große Datenmengen beliebiger Formate aufnimmt – von ERP-Tabellen bis zu PDF-Zeichnungen und Maschinendaten. Die Daten werden zunächst unstrukturiert abgelegt und erst zum Zeitpunkt der Auswertung aufbereitet (Schema-on-Read). Damit unterscheidet sich der Ansatz grundlegend vom klassischen Data Warehouse, das Daten vor der Speicherung einem festen Schema unterwirft.

Abgrenzung zum Data Warehouse

Das Data Warehouse liefert kuratierte, modellierte Daten für standardisierte Berichte, etwa zur Skontoquote oder zur Backorder Rate. Der Data Lake hingegen speichert zuerst alles und erlaubt flexible, auch explorative Analysen. In vielen Unternehmen koexistieren beide Ansätze: Der Data Lake dient als Rohdatenquelle, aus der kuratierte Bereiche in Warehouse oder Data Mart überführt werden.

- Data Warehouse: Schema-on-Write, hochqualitative Stammdaten, Reporting
- Data Lake: Schema-on-Read, alle Formate, flexible Analysen und KI
- Data Lakehouse: hybrides Konzept, das beide Prinzipien verbindet

Begriffsherkunft und Grundidee

Die Metapher des Sees beschreibt den Ansatz: Daten strömen aus vielen Quellen ungefiltert

zusammen und bleiben in ihrer ursprünglichen Form erhalten. Erst wer Wasser entnimmt, entscheidet über Aufbereitung und Verwendung. Diese Idee entstand mit dem Aufkommen kostengünstiger, verteilten Speicher- und Rechenarchitekturen, die previously undenkbbare Datenmengen wirtschaftlich handhabbar machen.

Aufbau und Architektur eines Data Lakes

Ein Data Lake besteht typischerweise aus einer billigen, skalierbaren Objektspeicher-Ebene, einer Metadaten- und Katalogschicht sowie Werkzeugen für Ingestion, Verarbeitung und Analyse. Moderne Architekturen trennen Speicher und Rechenleistung, sodass Kosten und Performance unabhängig voneinander skaliert werden können. Offene Dateiformate haben sich als De-facto-Standard etabliert.

Datenquellen im Einkauf

Für Beschaffungsdaten speisen häufig ERP-Systeme (Bestellungen, GR/IR Clearing, Materialstammdaten), E-Kataloge, Lieferantenportale, Ausschreibungsmanagement-Tools, Frachtdaten und Wartungssysteme den Data Lake. Hinzu kommen unstrukturierte Inhalte wie Zeichnungen, Lastenhefte, Zertifikate und E-Mail-Verkehr. Auch externe Daten – Rohstoffindizes, Marktberichte, Risikofeeds – erweitern das Bild für Risikoanalysen und Should-Costing.

Ingestion und Verarbeitung

Beim Ingestion unterscheidet man Batch-Verfahren für periodische Ladeläufe und Streaming für Echtzeitdaten, etwa Maschinen- oder Logistikdaten. Auswertungen erfolgen häufig in einer ELT-Logik: Extraktion, Laden, anschließende Transformation im Zielsystem. Die Verarbeitungsschicht bereitet Rohdaten in Zonen auf – von Raw über Cleaned bis Curated – und macht sie so für unterschiedliche Nutzergruppen verzehrbar.

- Raw-Zone: unveränderte Originaldaten, revisionsfähig archiviert
- Cleansed-Zone: bereinigte, deduplizierte Daten mit Metadaten
- Curated-Zone: fachlich modellierte Daten für Einkauf und Controlling

Metadaten und Data Catalog

Ein zentraler Katalog dokumentiert Herkunft, Format, Besitzer und Qualität jedes Datensatzes. Ohne diese Schicht verfällt der Data Lake zum Data Swamp: Nutzer finden Daten nicht, vertrauen ihnen nicht oder doppelten Aufwände. Herkunftsverfolgung (Lineage) zeigt, aus welchen Quellen eine Kennzahl resultiert – wichtig etwa, wenn Durchlaufzeit BANF-PO oder Komplexitätsreduktion-Potenziale nachvollziehbar ausgewiesen werden sollen.

Einsatzszenarien im technischen Einkauf

Der Data Lake entfaltet im Beschaffungskontext Wert, weil er erstmals geteilter Einkaufs-, Qualitäts-, Logistik- und Lieferantendaten zusammenführt. Darauf aufbauend lassen sich Auswertungen automatisieren, die früher manuelle Excel-Arbeit bedeuteten – von Spend-Analysen über Lieferbewertungen bis zu Prognosen.

Spend-Analyse und Should-Costing

Mit konsolidierten Bestelldaten lassen sich Ausgaben nach Warengruppen, Lieferanten und Werken durchdringen. Kombiniert mit Stücklisten, Materialpreisen und Prozessdaten entstehen Should-Costing-Modelle, die faire Zielpreise für Bauteile herleiten. Auch Funktionskostenrechnungen

und Gesamtkostenrechnungen profitieren, weil Mengen-, Kosten- und Verbrauchsdaten aus einer Quelle stammen.

Lieferantenmanagement und Risikofrüherkennung

Im Lieferantenmanagement speist der Data Lake Lieferantenbewertungen aus Qualität, Liefertreue und Finanzkennzahlen. Verknüpft mit externen Risikoquellen unterstützt er die systematische Risikoanalyse beim Lieferantenrisiko – etwa wenn Single Sourcing-Konstellationen oder kritische Vorprodukte identifiziert werden sollen. Dual Sourcing-Entscheidungen lassen sich datenbasiert vorbereiten, statt auf Bauchgefühl zu setzen.

Disposition und Bedarfsplanung

Verbrauchs-, Bestands- und Auftragsdaten aus dem Data Lake verbessern Demand Planning und Material Requirements Planning (MRP). Historische Backorder-Daten helfen, die Backorder Rate als Frühindikator zu überwachen, während Bestands- und Lieferdaten Aussagen zur Sicherheitskapazität kritischer Teile ermöglichen. Available to Promise (ATP)-Auskünfte gegenüber dem Vertrieb werden belastbarer, wenn sie auf konsolidierten Echtzeitdaten beruhen.

Wichtige KPIs und Kennzahlen rund um den Data Lake

Der Erfolg eines Data Lakes lässt sich messen – sowohl hinsichtlich Datenqualität als auch hinsichtlich des Geschäftsnutzens. In der Praxis bewähren sich Kennzahlen, die Nutzung, Qualität und Zeitersparnis abbilden.

Datenqualität und Operation

Typische Richtwerte: Die Datenkomplettität sollte häufig über 98 % liegen, damit Auswertungen vertrauenswürdig sind. Die Datenaktualität wird über die Latenz gemessen – von Minuten im Streaming bis zu Tagesläufen im Batch. Anteil automatisierter Pipelines und Fehlerquote der Ingestion zeigen, wie stabil die Plattform läuft.

Nutzung und Geschäftsnutzen

Zahl aktiver Nutzer, Anzahl kuratierter Datensätze und Time-to-Insight – die Zeit von der Fragestellung bis zum Ergebnis – machen den Nutzen greifbar. Auf Einkaufsseite zählen abgeleitete Werte: Verbesserung der Skontoquote durch rechtzeitige Zahlungsplanung, sinkende Durchlaufzeit BANF-PO durch Prozessanalyse oder höhere Zielerreichung im Ausschreibungsmanagement.

Risiken, Governance und Kontrollen

Der größte Risikofaktor ist nicht die Technik, sondern fehlende Ordnung: Ohne klare Verantwortlichkeiten, Metadaten und Qualitätsregeln entsteht ein Data Swamp, dessen Inhalte niemand traut. Datenschutz, Zugriffskontrolle und Kostenkontrolle gehören von Beginn an zum Konzept.

Data Governance und Rollen

Eine Data Governance definiert Datenbesitzer, Qualitätskriterien, Berechtigungen und Lebenszyklen. Data Stewards pflegen Begriffe und Definitionen, etwa wann ein Lieferant als „kritisch“ gilt oder wie die Backorder Rate berechnet wird. Eine Kompetenzmatrix Einkauf hilft, gezielt Data-Literacy-Fähigkeiten im Team aufzubauen und Analysen nicht an einzelne Spezialisten zu binden.

Datenschutz, Sicherheit und Kosten

Personenbezogene Daten unterliegen dem Datenschutzrecht; konkrete Ausgestaltung sollte immer juristisch im Einzelfall geprüft werden. Technisch bewähren sich rollenbasierte Zugriffskonzepte, Verschlüsselung und pseudonymisierte Testdaten. Ohne Kostenkontrolle wachsen Speicher- und Rechenkosten unbemerkt – Lebenszyklusregeln, die alte Daten in künstlichere Speicherklassen verschieben, halten die Betriebskosten im Rahmen.

Datenqualität sicherstellen

Qualitätsregeln sollten automatisiert geprüft werden: Vollständigkeit von Bestellfeldern, Konsistenz von Lieferantenstammdaten, Duplikate in Materialnummern. Besonders Stammdaten sind der Schwachpunkt vieler Einkaufsanalysen; hier lohnt sich vorab ein Bereinigungsprojekt, bevor der Data Lake gefüllt wird. Ein Rollout Management mit Pilotbereichen verhindert, dass Fehler großflächig propagiert werden.

Aktuelle Entwicklungen

Der Data-Lake-Ansatz hat sich in den letzten Jahren deutlich weiterentwickelt. Offene Tabellenformate, Lakehouse-Architekturen und KI-gestützte Auswertung verschieben die Grenzen zwischen Data Lake und Data Warehouse.

Lakehouse und offene Formate

Lakehouse-Architekturen kombinieren die flexible Speicherung des Data Lakes mit Transaktions- und Qualitätsfunktionen des Data Warehouses. Offene Tabellenformate ermöglichen es, dass verschiedene Abfrage-Tools auf denselben Daten arbeiten, ohne Kopien zu erzeugen. Für mittelständische Fertiger senkt das Einstiegshürden, da Analysewerkzeuge auf bestehenden Speichern aufsetzen können.

KI und generative Assistenz im Einkauf

Große Sprachmodelle lassen sich auf kuratierten Data-Lake-Daten betreiben, um Anfragen in natürlicher Sprache zu beantworten – etwa „Welche Lieferanten lieferten im letzten Jahr verspätet?“ Auch bei der Feasibility Review Lieferant oder der Vorbereitung von Anfragen mit KI unterstützt eine gut gepflegte Datenbasis die Qualität der Ergebnisse. Voraussetzung bleibt saubere Governance; KI-Ergebnisse sind kritisch zu validieren.

Typische Kennzahlen

Datenkomplettität: > 98 %

typischer Zielwert für kuratierte Kernfelder wie Bestelldaten

Datenlatenz: Minuten bis 24 h

je nach Verfahren: Streaming nahe Echtzeit, Batch üblicherweise täglich

Time-to-Insight: Stunden bis wenige Tage

Zeit von der Fragestellung bis zum belastbaren Ergebnis

Fehlerquote Ingestion: < 1 %

Anteil fehlgeschlagener Ladeläufe als Stabilitätsindikator

Backorder Rate: < 5 %

üblicher Richtwert; verbessert sich mit besseren Dispositionsdaten

Praxisbeispiel

Ein mittelständischer Maschinenbauer mit rund 40.000 Teilestammsätzen konsolidierte ERP-Daten, Lieferantenportalinhalte und Qualitätsdaten in einem Data Lake. Zuvor lagen Auswertungen zur Liefertreue und zu Beschaffungskosten in verteilten Excel-Listen vor und waren wochenaktuell.

- Automatisierte Spend-Analyse nach Warengruppen, Werken und Lieferanten statt monatlicher Manuellerfassung
- Frühwarn-Dashboard zum Lieferantenrisiko, verknüpft mit Liefertreue- und Qualitätsdaten
- Verbesserte Demand Planning durch Auswertung historischer Verbrauchs- und Backorder-Daten
- Reduktion der Durchlaufzeit BANF-PO um etwa ein Viertel durch Prozessengpass-Analyse auf den konsolidierten Daten

Fazit

Der Data Lake ist für den technischen Einkauf weniger ein Selbstzweck als die Voraussetzung für datenbasierte Entscheidungen: von Spend-Analyse über Lieferantenrisiko bis zur Bedarfsprognose. Entscheidend für den Erfolg sind nicht Speichertechnologien, sondern Governance, Datenqualität und klare Anwendungsfälle mit erkennbarem Nutzen. Wer mit einem überschaubaren Pilotbereich startet und kuratierte Daten schrittweise ausbaut, vermeidet Data Swamps und unnötige Kosten. In Kombination mit KI-gestützten Auswertungen wird der Data Lake zum Fundament einer modernen, proaktiven Beschaffung. Der Aufwand zahlt sich dort aus, wo Entscheidungen bisher auf verteilten, manuell gepflegten Daten beruhen.

Häufige Fragen

Was ist ein Data Lake einfach erklärt?

Ein Data Lake ist ein zentraler, sehr flexibler Speicher, in dem ein Unternehmen alle Daten in ihrem ursprünglichen Format sammelt – von Tabellen über Dokumente bis zu Sensordaten. Struktur und Aufbereitung erfolgen erst bei der Auswertung. Das erlaubt flexible Analysen, erfordert aber klare Ordnung, damit kein ungenutzter „Data Swamp“ entsteht.

Was ist der Unterschied zwischen Data Lake und Data Warehouse?

Ein Data Warehouse speichert strukturierte Daten nach einem vordefinierten Schema und dient standardisiertem Reporting. Ein Data Lake nimmt alle Datenformate unstrukturiert auf und erlaubt flexible, auch explorative Analysen. In der Praxis ergänzen sich beide: Der Data Lake liefert Rohdaten, das Warehouse kuratierte Berichtsdaten.

Was ist ein Data Swamp?

Ein Data Swamp ist ein Data Lake ohne Governance: Daten liegen unsortiert, ohne Metadaten, Verantwortliche oder Qualitätsprüfung. Nutzer finden und vertrauen den Inhalten nicht, wodurch der Speicher seinen Zweck verliert. Prävention setzt bei Datenkatalog, definierten Zonen und klaren Datenbesitzern an.

Welche Datenquellen kommen für den Einkauf in einen Data Lake?

Typisch sind ERP-Systeme mit Bestell- und Stammdaten, Lieferantenportale, Qualitätsmanagementsysteme, Fracht- und Logistikdaten sowie Dokumente wie Zeichnungen und Zertifikate. Externe Quellen wie Markt- und Risikodaten ergänzen das Bild. Je vollständiger die

Quellenabdeckung, desto aussagekräftiger werden Risikoanalysen und Should-Costing-Modelle.

Lohnt sich ein Data Lake für mittelständische Unternehmen?

Ja, wenn klare Anwendungsfälle bestehen und Datenquellen beherrschbar sind. Cloud-basierte, skalierbare Architekturen haben Einstiegshürden gesenkt, sodass auch kleinere Umsetzungen wirtschaftlich möglich sind. Entscheidend ist, mit wenigen kuratierten Auswertungen zu starten und erst dann zu skalieren – ein Vollausbau ohne konkrete Fragen führt häufig zu ungenutzten Kosten.

Wie stellt man die Datenqualität in einem Data Lake sicher?

Durch automatisierte Qualitätsprüfungen, definierte Datenbesitzer, einen Metadatenkatalog und gegliederte Zonen von Roh- bis kuratierten Daten. Vorab lohnt sich die Bereinigung kritischer Stammdaten. Regelmäßige Qualitätsreports machen Schwächen sichtbar, bevor sich Fehler in Auswertungen verbreiten.