Qlik

WHITEPAPER

DATENQUALITÄT UND GOVERNANCE

AI-fähige Daten: Auf diese sechs Kriterien kommt es an

So schaffen Sie eine vertrauenswürdige Datengrundlage für AI

Inhalt

Zusammenfassung

- In diesem Dokument beschreiben wir sechs Kriterien, anhand derer Sie sicherstellen können, dass Ihre Daten fit für den Einsatz mit Artificial Intelligence (AI) sind.
- Diese lauten wie folgt: Daten müssen vielfältig, aktuell, genau, sicher, auffindbar und einfach von Maschinen nutzbar sein.
- Das Dokument stellt ferner den AI Trust Score vor, mit dessen Hilfe Sie feststellen können, wie gut Ihre Daten diese Kriterien erfüllen.

Einführung

Artificial Intelligence soll Bereichen wie Gesundheitswesen, Fertigung und Kundenservice erhebliches Optimierungspotenzial bieten und die Erfahrungen von Kunden und Beschäftigten gleichermaßen verbessern. Bereits heute nutzen Datenexperten AI-Technologien wie Machine Learning (ML), um mathematische Vorhersagen zu erstellen, neue Erkenntnisse zu gewinnen und fundierte Entscheidungen zu treffen. Hinzu kommt, dass aufkommende AI-Technologien wie Generative AI (GenAI) verblüffend realistische Inhalte generieren können, die das Potenzial haben, die Produktivität praktisch überall im Unternehmen zu steigern.

Laut Gartner werden 90 % der globalen Unternehmen bis 2025 mit GenAI arbeiten. Zudem werden bis 2026 mehr als 80 % aller Unternehmen GenAI-fähige Anwendungen in der Produktion einsetzen.[1]

Doch ohne verlässliche Daten kann AI nicht erfolgreich sein. Mithilfe der sechs in diesem Paper beschriebenen Kriterien stellen Sie sicher, dass Ihre Daten AI-fähig sind.

[1] Gartner: „Analysts to Discuss Generative AI Trends and Technologies", Oktober 2023.

Die sechs Kriterien für AI-fähige Daten

Es wäre absurd zu glauben, man müsste AI-Initiativen nur mit Daten überhäufen und dann passiert ein Wunder. Doch genau so wird in der Praxis häufig vorgegangen. Dieser Ansatz mag bei den ersten AI-Projekten noch erfolgreich gewesen sein, doch mit zunehmender Projektreife verbringen Data Scientists immer mehr Zeit mit der Korrektur und Nachbereitung der Daten.

Ferner müssen die für AI verwendeten Daten hochwertig und genau für diese intelligenten Anwendungen aufbereitet sein. Wer genaue und vollständige Daten haben möchte, muss viel Zeit in die manuelle Bereinigung und Optimierung der Daten investieren, denn es gilt sie so zu organisieren, dass sie für Maschinen problemlos lesbar sind. Außerdem benötigen diese Daten oft weitere Informationen (wie Definitionen und Label), um die semantische Bedeutung für automatisiertes Lernen anzureichern und die AI dabei zu unterstützen, Aufgaben effektiv auszuführen.

Daher gilt: Je früher Daten für nachgelagerte AI-Prozesse aufbereitet werden können, desto besser. Fertige AI-fähige Daten sind das bereits gewaschene und zerkleinerte Gemüse für eine Köchin. Statt sich wie sonst durch die Gemüselieferung zu arbeiten, spart sie Zeit und Aufwand. So ist sichergestellt, dass das Gericht prompt serviert werden kann. Das Diagramm unten zeigt sechs wichtige Kriterien, die erfüllt sein müssen, damit Daten für den AI-Einsatz bereit und geeignet sind.

Auf den folgenden Seiten werden die einzelnen Kriterien ausführlich erläutert.

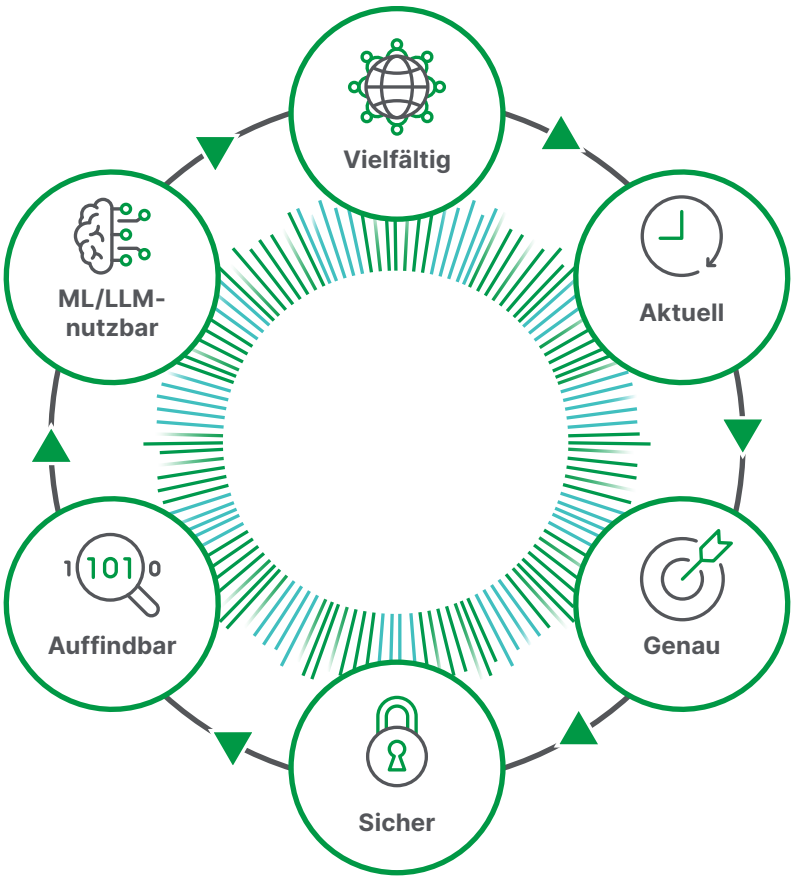

Abbildung 1: Sechs Kriterien für AI-fähige Daten

1 Daten müssen vielfältig sein.

Verzerrungen in AI-Systemen – auch bekannt als „Machine Learning Bias“ oder „algorithmische Verzerrung“ – treten auf, wenn die Ergebnisse von AI-Anwendungen menschliche Voreingenommenheit widerspiegeln, wie beispielsweise soziale Ungleichheiten. Dazu kommt es, wenn der Entwicklungsprozess des Algorithmus vorurteilsbehaftete Annahmen enthält oder, was noch häufiger vorkommt, die Trainingsdaten verzerrt sind. Bei der Prüfung der Kreditwürdigkeit kann ein Algorithmus daher einen Kredit ablehnen, wenn er nur wenige finanzielle Merkmale nutzt.

Das erste Kriterium ist daher, dass AI-Modelle viele unterschiedliche Daten benötigen. Dies erhöht die Datenvielfalt, reduziert Verzerrungen und sorgt dafür, dass AI-Anwendungen ausgewogene Entscheidungen treffen.

Datenvielfalt bedeutet, dass Ihre AI-Modelle nicht auf eng gefassten und isolierten Datensätzen aufsetzen dürfen. Greifen Sie stattdessen auf eine Vielzahl von Datenquellen zurück, die verschiedene, für das Problemfeld relevante Muster, Perspektiven, Variationen und Szenarien umfassen. Dies können durchaus gut strukturierte Cloud- oder On-Premises-Daten sein. Die Daten können aber auch aus Mainframes, Datenbanken, SAP-Systemen oder Software-as-a-Service-Anwendungen stammen. Oder es handelt sich um unstrukturierte Daten, die in Form von Dateien oder Dokumenten auf einem Unternehmenslaufwerk gespeichert sind.

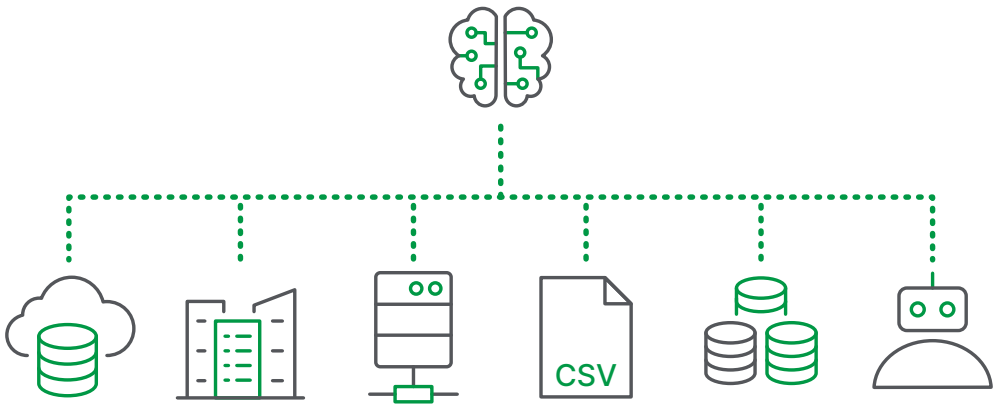

Abbildung 2: Datenvielfalt

Für die Integration in Ihre ML- und GenAI-Anwendungen müssen unbedingt vielfältige Daten in unterschiedlichen Formaten erfasst werden.

2 Daten müssen aktuell sein.

Zwar profitieren ML- und GenAI-Anwendungen von vielfältigen Daten, doch deren Aktualität ist ebenso wichtig. So wie die Wettervorhersage von gestern keinerlei Wert für einen heutigen Ausflug hat, liefern auch AI-Modelle, die auf veralteten Informationen basieren, falsche oder irrelevante Ergebnisse. Darüber hinaus können AI-Modelle mit aktuellen Daten Trends besser im Blick behalten, sich an veränderte Umstände anpassen und bessere Ergebnisse liefern. Das zweite Kriterium für AI-fähige Daten ist daher Aktualität.

Wenn Sie die neuesten Daten für ihre AI-Initiativen nutzen wollen, dann sollten Sie latenzarme Echtzeit-Datenpipelines einrichten und implementieren. Bei der Bereitstellung aktueller Daten aus relationalen Datenbanksystemen kommt oft **Change Data Capture (CDC)** zum Einsatz. Für Daten von IoT-Geräten, die mit geringer Latenz verarbeitet werden müssen, bietet sich dagegen **Stream Capture** an. Sobald die Daten erfasst sind, werden die Ziel-Repositories aktualisiert und **Änderungen fortlaufend** in Quasi-Echtzeit vorgenommen.

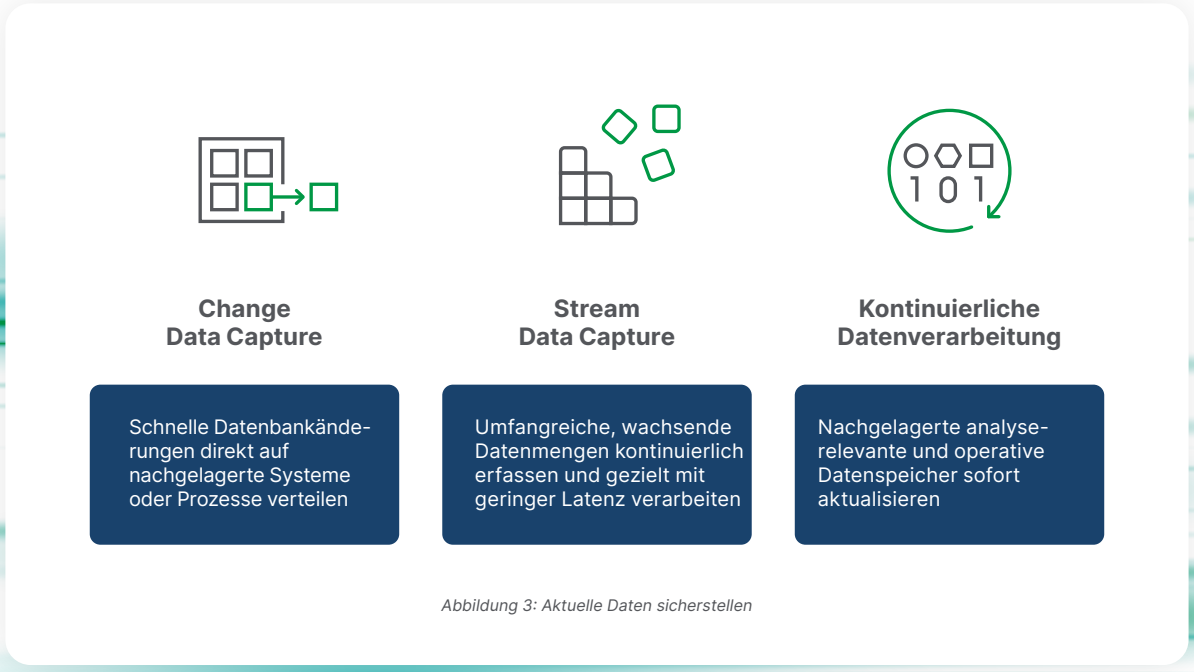

Abbildung 3: Aktuelle Daten sicherstellen

Wichtig: Aktuelle Daten ermöglichen stimmige und fundierte Vorhersagen.

3 Daten müssen genau sein.

Der Erfolg jeder ML- oder GenAI-Initiative hängt von einer entscheidenden Komponente ab: korrekten Daten. Denn AI-Modelle funktionieren wie besonders aufnahmefähige Schwämme, die Informationen aufsaugen, um zu lernen und Aufgaben zu erfüllen. Sind die Informationen falsch, ist es, als würde der Schwamm schmutziges Wasser aufsaugen. Die Folge: verzerrte Ergebnisse, unsinnige Einstellungen und letztendlich ein mangelhaft funktionierendes AI-System. Das dritte Kriterium ist daher die Datengenauigkeit. Sie ist eine grundlegende Voraussetzung für die Entwicklung verlässlicher und vertrauenswürdiger AI-Anwendungen.

Datengenauigkeit umfasst drei Aspekte. Erstens muss ein **Quelldatenprofil** erstellt werden, um die Eigenschaften, Vollständigkeit, Verteilung, Redundanz und Form nachzuvollziehen. Dieses Profiling wird auch als explorative Datenanalyse (kurz EDA) bezeichnet.

Zweitens müssen Sie **Korrekturmaßnahmen festlegen**. Dazu werden Regeln für die Datenqualität erstellt, implementiert und kontinuierlich hinsichtlich ihrer Wirksamkeit überwacht. Hier sollten Sie Ihre Data Stewards einbinden, um Sie beim Deduplizieren und Zusammenführen der Daten zu unterstützen. Alternativ kann AI durch maschinengestützte Vorschläge zur Verbesserung der Datenqualität beitragen sowie den Prozess automatisieren und beschleunigen.

Beim letzten Aspekt geht es um Datenherkunft und Auswirkungsanalyse. Mit Tools für Data Engineers und Data Scientists werden die Auswirkungen potenzieller Datenänderungen aufgezeigt und der Ursprung der Daten zurückverfolgt. Ziel ist, versehentliche Änderungen an den von AI-Modellen verwendeten Daten zu verhindern.

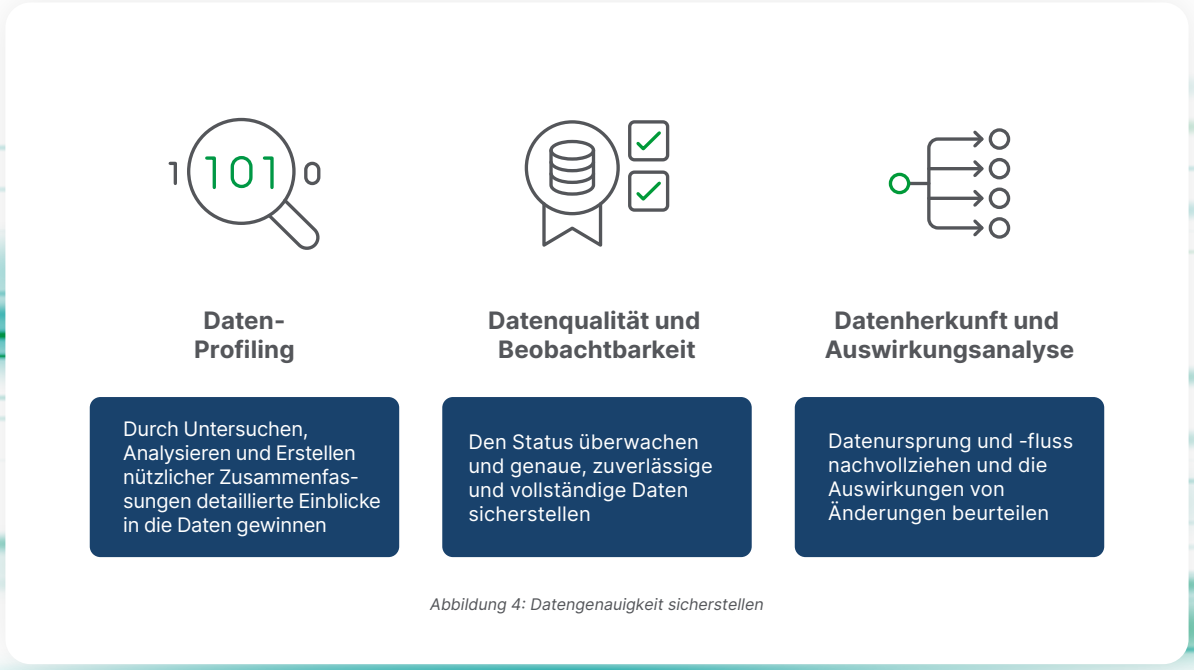

Abbildung 4: Datengenauigkeit sicherstellen

Mit hochwertigen und fehlerfreien Daten sorgen Sie dafür, dass Modelle relevante Muster und Beziehungen erkennen. Die Folge ist eine höhere Präzision bei Entscheidungen, Generierung und Prognosen.

4 Daten müssen sicher sein.

AI-Systeme nutzen oft sensible Daten, beispielsweise personenbezogene Informationen, Finanzunterlagen oder proprietäre Unternehmensdaten. Der Umgang mit diesen Daten erfordert ein hohes Maß an Verantwortungsbewusstsein. Fehlender Schutz für Daten in AI-Anwendungen lässt sich mit einer offenstehenden Tresortür vergleichen. Böswillige Akteure könnten sensible Informationen stehlen, Trainingsdaten manipulieren, um Ergebnisse zu verfälschen, oder GenAI-Systeme sogar komplett zum Erliegen bringen. Der Schutz der Daten ist unerlässlich für die Vertraulichkeit personenbezogener Informationen, die Integrität des Modells und die verantwortungsbewusste Entwicklung leistungsstarker AI-Anwendungen. Datensicherheit ist daher das vierte Kriterium für die AI-Fähigkeit von Daten.

Mit den folgenden drei Maßnahmen können Sie den Schutz der Daten umfassend automatisieren, da dies manuell nahezu unmöglich ist. Die **Datenklassifizierung** erkennt, kategorisiert und markiert Daten, die in den nächsten Schritt einfließen. Der **Datenschutz** definiert Richtlinien wie Maskierung, Tokenisierung und Verschlüsselung, um Daten zu verschleiern. Im Rahmen der **Datensicherheit** werden die Regeln zur Zugriffssteuerung festgelegt, also wer Zugang zu den Daten erhält. Die drei Konzepte arbeiten wie folgt zusammen: Zuerst werden Schutzstufen definiert und die Daten mit Kennzeichnungen wie *sensibel*, *vertraulich* oder *eingeschränkt* markiert. Im nächsten Schritt wird eine Schutzrichtlinie angewendet, um eingeschränkte Daten zu maskieren. Am Ende kommt dann eine Regel zur Zugriffssteuerung zum Einsatz, die die Zugriffsrechte beschränkt.

Abbildung 5: Zentrale Attribute der Datensicherheit

Mit diesen drei Maßnahmen schützen Sie Ihre Daten und tragen entscheidend dazu bei, das Vertrauen in Ihr AI-System insgesamt zu stärken und seine Zuverlässigkeit zu bewahren.

5 Daten müssen auffindbar sein.

Bei den bisher genannten Kriterien ging es in erster Linie darum, die richtigen Daten im richtigen Format rechtzeitig für die richtigen Personen, Systeme oder AI-Anwendungen bereitzustellen. Doch mit dem Sammeln von Daten ist es nicht getan. AI-fähige Daten müssen auch im System auffindbar und problemlos zugänglich sein. Stellen Sie sich eine Bibliothek vor, in der alle Bücher weggeschlossen sind: Das Wissen ist vorhanden, lässt sich jedoch nicht nutzen. Auffindbare Daten erschließen das wahre Potenzial von ML und GenAI. Mit ihrer Hilfe finden diese Workloads die nötigen Informationen, um zu lernen, sich anzupassen und wegweisende Ergebnisse zu erzielen. Daher ist die Auffindbarkeit das fünfte Kriterium für AI-fähige Daten.

Sorgfältig gepflegte Metadaten spielen eine wichtige Rolle bei der Auffindbarkeit von Daten. Neben den technischen Metadaten im Zusammenhang mit AI-Datensätzen müssen auch geschäftliche Metadaten und semantische Typisierungen definiert werden. Durch eine **semantische Typisierung** erhalten automatisierte Systeme eine zusätzliche Bedeutungsebene und betriebswirtschaftliche Begriffe liefern weiteren Kontext, um das menschliche Verständnis zu erleichtern. Ein bewährtes Verfahren ist die Erstellung eines **Unternehmensglossars**, in dem betriebswirtschaftliche Begriffe den technischen Elementen in den Datensets zuordnet werden. So wird sichergestellt, dass die Konzepte für alle transparent und verständlich sind. Mit AI-gestützten Erweiterungen können Sie zudem automatisch Dokumentationen erstellen und betriebswirtschaftliche Termini aus dem Glossar hinzufügen. Abschließend werden alle Metadaten indiziert und über einen **Datenkatalog** durchsuchbar gemacht.

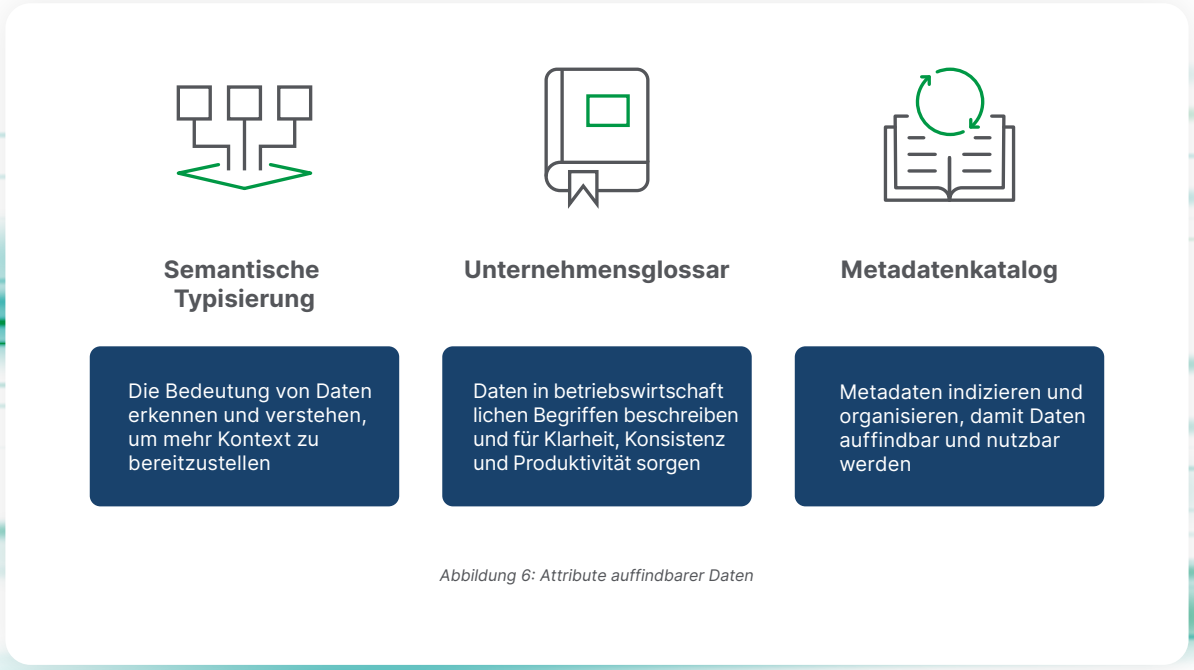

Abbildung 6: Attribute auffindbarer Daten

Dieses Konzept stellt sicher, dass die Daten sofort auffindbar, nutzbar, praxistauglich und für die jeweilige AI-Aufgabe wichtig sind.

6 Daten müssen für ML oder LLMs einfach nutzbar sein.

Wie wir bereits festgestellt haben, sind ML- und GenAI-Anwendungen leistungsstarke Tools. Allerdings steht und fällt ihr Potenzial mit ihrer Fähigkeit, aufbereitete Daten gut nutzen zu können. Anders als Menschen, die in der Lage sind, handschriftliche Notizen zu entziffern oder sich in unübersichtlichen Spreadsheets zurechtfinden, erfordern solche Technologien bestimmte Formate. Stellen Sie sich einen wählerischen Gast vor: Wenn er nicht isst, was Sie ihm servieren, bleibt er hungrig. Ähnlich werden auch AI-Initiativen fehlschlagen, wenn die Daten nicht im für ML-Experimente oder LLM-Anwendungen passenden Format vorliegen. Durch einfach verwertbare Daten erschließen Sie sich das Potenzial dieser AI-Systeme. Informationen werden ohne Probleme aufgenommen und in intelligente Maßnahmen für kreative Ergebnisse übersetzt. Das letzte Kriterium für AI-fähige Daten lautet demnach, für einfach nutzbare Daten zu sorgen.

Daten für Machine Learning nutzbar machen

Datentransformation ist das unsichtbare Rückgrat hinter ML-nutzbaren Daten. Algorithmen wie lineare Regression stehen zwar im Mittelpunkt, doch die Qualität und Beschaffenheit der Daten, mit denen sie trainiert werden, sind ebenso wichtig. Der Aufwand, der in die Bereinigung, Organisation und Nutzbarmachung der Daten für ML-Modelle investiert wird, macht sich bezahlt. Mit entsprechend aufbereiteten Daten können Modelle effektiv lernen. Das Ergebnis sind korrekte Prognosen, zuverlässige Ergebnisse und letztendlich ein insgesamt erfolgreiches ML-Projekt.

Allerdings bestimmt die zugrunde liegende ML-Infrastruktur, in welchem Format die Trainingsdaten vorliegen müssen. Herkömmliche ML-Systeme sind datenträgerbasiert. Viele Workflows von Data Scientists haben das Ziel, Best Practices und manuelle Programmierverfahren für den Umgang mit großen Dateimengen zu entwickeln. In letzter Zeit verwenden Lakehouse-basierte ML-Systeme verstärkt einen datenbankähnlichen Feature Store und bei den Data Scientists setzt sich zunehmend SQL als vorrangige Sprache für Workflows durch. So entstehen wohlgeformte und hochwertige tabellarische Datenstrukturen, die sich ideal für ML-Systeme eignen.

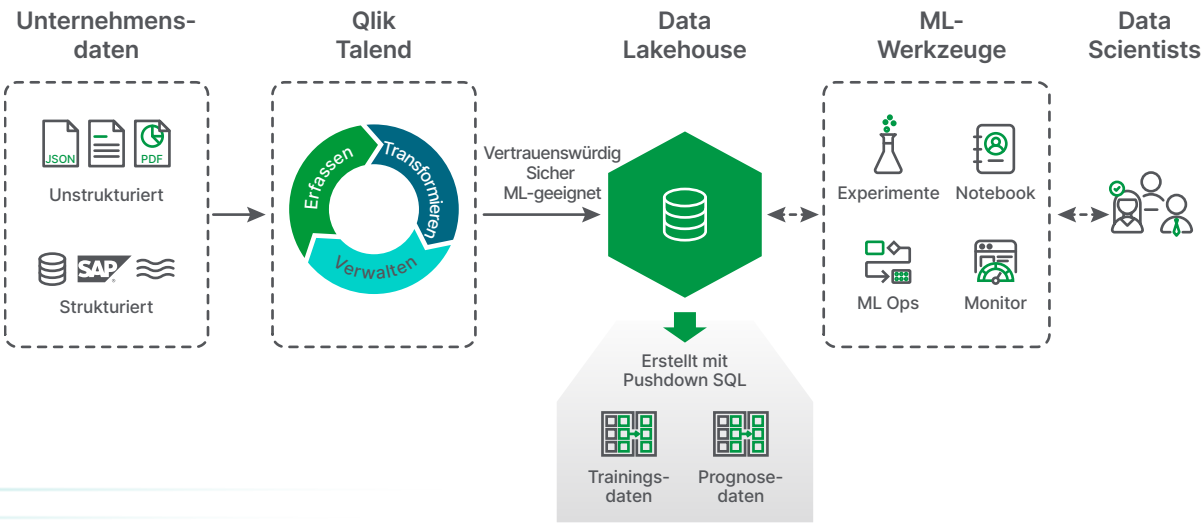

Abbildung 7: Daten für ML nutzbar machen

Daten für generative AI nutzbar machen

Large Language Models (LLMs) wie GPT-4 von OpenAI, Claude von Anthropic sowie LaMDA und Gemini von Google AI wurden bereits mit enormen Textdatenmengen trainiert. Sie sind das Fundament von GenAI. Das GPT-3-Modell von OpenAI wurde mit schätzungsweise 45 TB an Daten trainiert – einer Datenmenge, die mehr als 300 Mrd. Token entspricht. Trotz dieser Informationsfülle sind LLMs nicht in der Lage, konkrete Fragen zu Ihrer Geschäftstätigkeit zu beantworten, da sie nicht auf die Daten Ihres Unternehmens zugreifen können. Diese Modelle müssen Sie um Ihre eigenen Informationen ergänzen – erst dann erhalten Sie korrekte, relevante und vertrauenswürdige AI-Anwendungen.

Die Technik zur Integration Ihrer Unternehmensdaten in eine LLM-basierte Anwendung wird als Retrieval-Augmented Generation (RAG) bezeichnet. Hierbei werden in der Regel Textinformationen aus unstrukturierten, dateibasierten Quellen wie Präsentationen, E-Mail-Archiven, Textdokumenten, PDF-Dateien, Protokollen usw. verwendet. Die Texte werden in überschaubare Abschnitte zerlegt und in eine numerische Darstellung umgewandelt, die vom LLM für das sogenannte Embedding eingesetzt wird. Die entstehenden Einbettungen werden danach in einer Vektordatenbank wie Chroma, Pinecone oder Weviate gespeichert. Interessant ist, dass viele klassische Datenbankanbieter wie PostgreSQL, Redis und SingleStoreDB ebenfalls Vektoren unterstützen. Zudem bieten neuerdings auch Cloud-Plattformen wie Databricks, Snowflake und Google BigQuery Unterstützung für Vektoren.

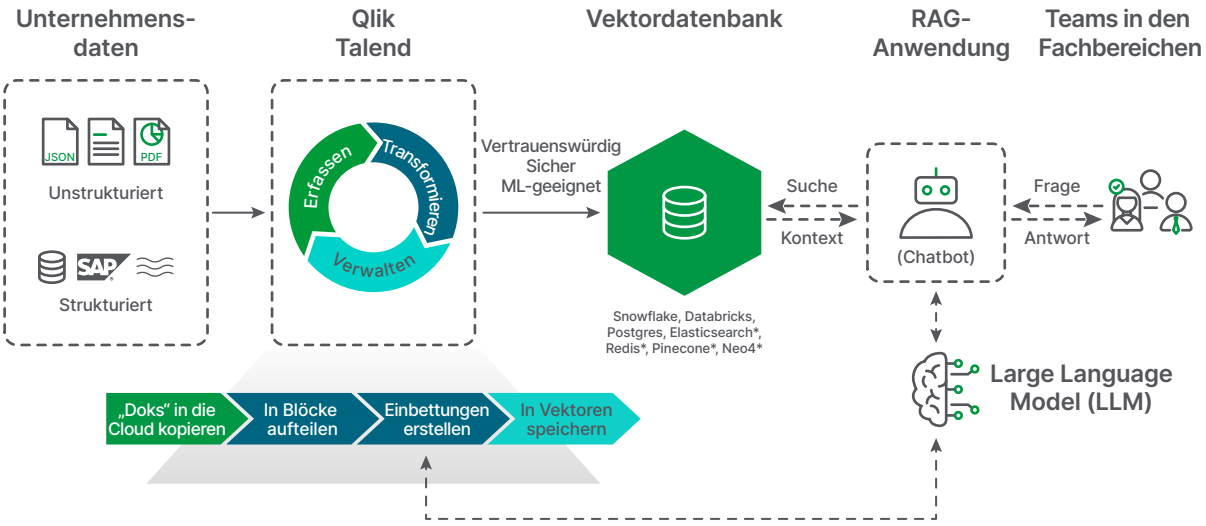

Abbildung 8: Daten für GenAI nutzbar machen

Ob Ihre Quelldaten strukturiert sind oder nicht: Das Konzept von Qlik stellt sicher, dass Ihre GenAI-, RAG- oder LLM-basierten Anwendungen hochwertige Daten problemlos nutzen können.

Der AI Trust Score

Nachdem wir nun die sechs Kriterien für die Verfügbarkeit und Eignung von Daten definiert haben, bleibt die Frage: Lassen sich diese Kriterien auch zusammenführen und problemlos im Alltag umsetzen? Wie schnell lässt sich feststellen, ob Daten AI-fähig sind? Eine Möglichkeit ist der AI Trust Score von Qlik.

Dieser unabhängige und einfach verständliche Indikator ordnet jedem Kriterium eine eigene Dimension zu und aggregiert anschließend die einzelnen Werte. Am Ergebnis können Sie die AI-Fähigkeit Ihrer Daten schnell und zuverlässig ablesen. Da sich Unternehmensdaten ständig ändern, wird auch der AI Trust Score regelmäßig geprüft und neu kalibriert, damit Datentrends bei der AI-Fähigkeit berücksichtigt werden.

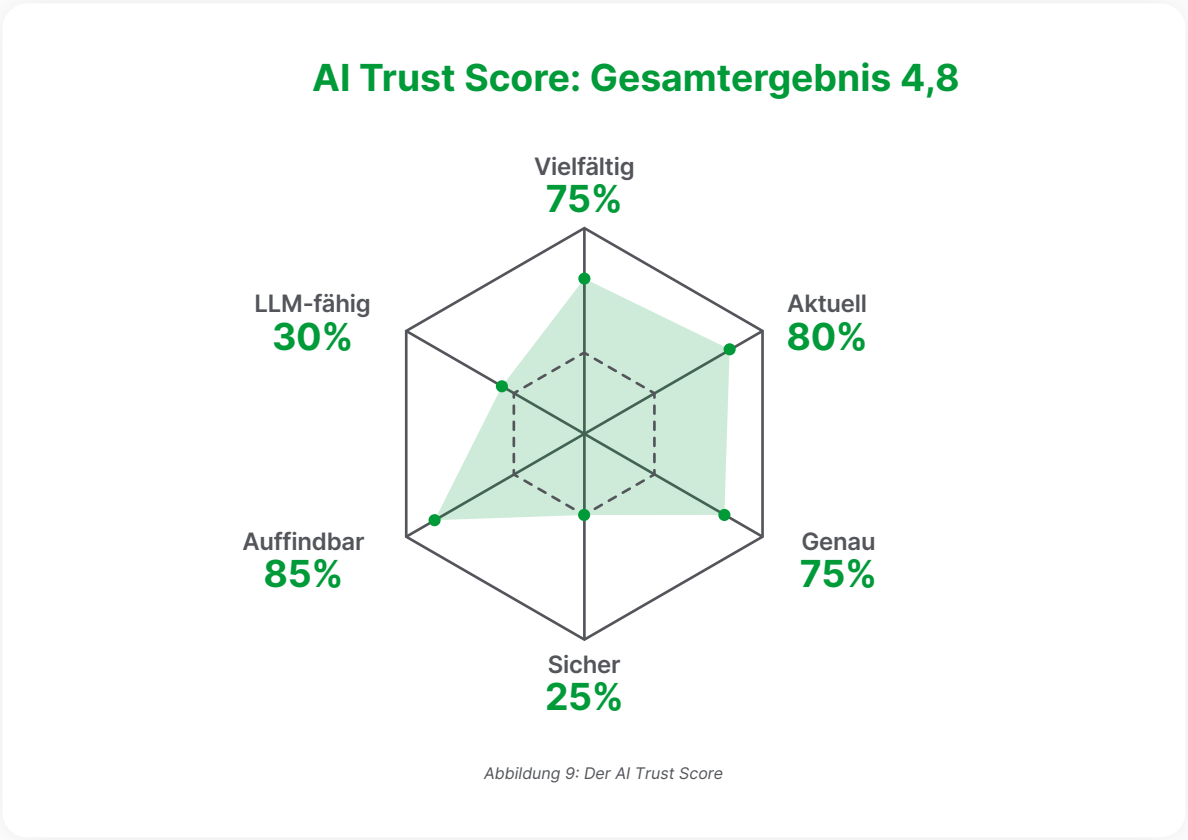

Abbildung 9: Der AI Trust Score

Ihr AI Trust Score aggregiert verschiedene Kennzahlen zu einem leicht verständlichen Wert, der die AI-Fähigkeit Ihrer Daten beschreibt.

Qlik Talend: Datenbasis für AI

Echtzeitdaten, die besser durchdachte Entscheidungen, mehr operative Effizienz und Innovationen unterstützen, sind gefragt wie nie. Deshalb setzen erfolgreiche Unternehmen auf die marktführenden Datenintegrations- und Datenqualitätslösungen von Qlik Talend, um zuverlässige Daten für Data Warehouses, Data Lakes und andere Plattformen effizient bereitzustellen. Unsere umfassenden Angebote nutzen automatisierte Pipelines, intelligente Transformationen und zuverlässige Datensatzqualität. So bieten sie nicht nur die Flexibilität, die Datenprofis sich wünschen, sondern auch die Governance und Compliance, die Unternehmen erwarten.

Ob Sie Warehouses oder Data Lakes für aussagekräftige Analysen erstellen, durch Modernisierung operativer Dateninfrastrukturen die Effizienz steigern oder Multi-Cloud-Daten für AI-Initiativen nutzen wollen: Qlik Talend unterstützt Sie dabei.

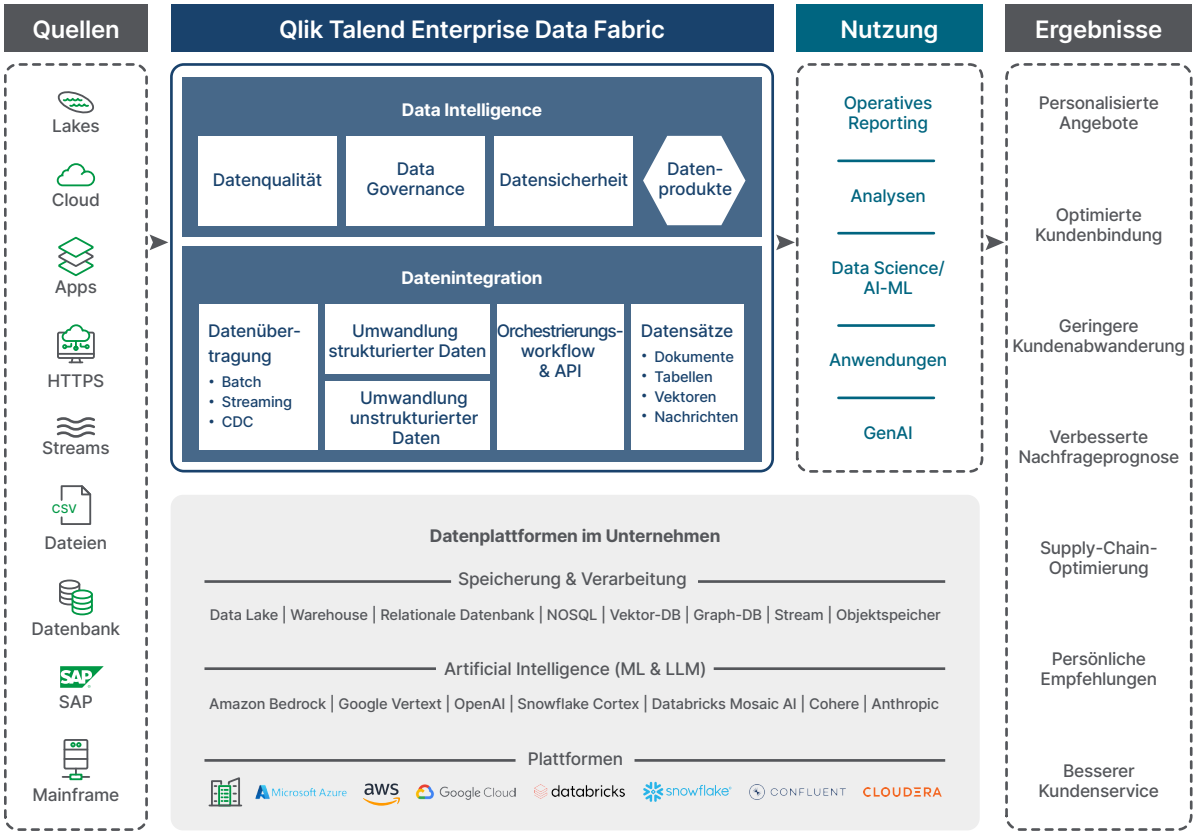

Abbildung 10: Qlik Talend Enterprise Data Fabric für AI und Analysen

Fazit

Trotz der Transformationskraft von Machine Learning und des gewaltigen Wachstumspotenzials von GenAI hängt der Erfolg jeder AI-Implementierung immer noch von geeigneten Daten ab. Wir haben in diesem Whitepaper sechs wichtige Kriterien für die Schaffung einer robusten und vertrauenswürdigen Datenbasis beschrieben, mit denen sich Ihr Unternehmen das volle Potenzial von AI erschließen kann.

Möchten Sie das AI-Potenzial Ihres Unternehmens voll ausschöpfen?

Starten Sie hier

Über Qlik

Qlik verwandelt komplexe Datenlandschaften in verwertbare Erkenntnisse, die zum Geschäftserfolg beitragen. Mehr als 40.000 Kunden weltweit nutzen unser Portfolio, das für moderne, businesstaugliche AI/ML und durchgängig hohe Datenqualität steht. Unsere Stärken sind Datenintegration und Data Governance und wir bieten umfassende Lösungen, die mit den unterschiedlichsten Datenquellen arbeiten. Intuitive Echtzeitanalysen von Qlik decken verborgene Muster auf und ermöglichen Teams, komplexe Herausforderungen zu meistern und neue Chancen zu nutzen. Unsere praxisnahen und skalierbaren AI/ML-Tools führen schneller zu besseren Entscheidungen. Als strategische Partner verbessern wir mit unserer plattformunabhängigen Technologie und unserem Know-how die Wettbewerbsfähigkeit unserer Kunden.

qlik.de

24-DQ-0001-01-WHM