

DATENQUALITÄT UND GOVERNANCE

Maximaler Mehrwert aus Daten im Zeitalter von AI

Schnellere Ergebnisse und robuste Governance mit
Datenprodukten

In der schnelllebigen Wirtschaft von heute stellen Daten eine wichtige strategische Ressource dar. Sie sind Dreh- und Angelpunkt für die Optimierung von Prozessen, fördern Innovationen und treiben Wachstum voran. Daten erfassen nicht nur Zahlen und Statistiken, sondern liefern auch Erkenntnisse, Trends und Muster, die den Weg zum Erfolg ebnen können. Vor dem Hintergrund von generativer AI und Machine Learning stehen Unternehmen heute vor einer zentralen Frage:

Wie können sie das volle Potenzial von Daten ausschöpfen, schneller Ergebnisse erzielen und gleichzeitig das Vertrauen in die Daten und ihre Governance stärken?

Geschwindigkeit ist dabei ein wichtiges Kriterium. Laut einer aktuellen IDC-Studie sagen 75 % der Entscheider, dass ihre Daten in wenigen Tagen ihren Wert verlieren.¹

Es reicht nicht, die riesigen Datenmengen in den Griff zu bekommen. Informationen müssen zur qualitativ hochwertigen, kuratierten und einfach zugänglichen Ressource werden und auf die Anforderungen der jeweiligen Geschäftsbereiche zugeschnitten sein. Doch wie jeder, der mit Daten arbeitet, weiß, ist dies leichter gesagt als getan. Die Verantwortlichen in den Bereichen Daten und Analyse stehen vor verschiedenen Herausforderungen, die die Umsetzung ihrer Dateninitiativen ausbremsen:

- **Dem projektorientierten Datenmanagement mangelt es an Skalierbarkeit.** Jedes Projektteam arbeitet innerhalb eigener Rahmenparameter wie Umfang, Ziele und Stakeholder. Dieser fragmentierte Ansatz erschwert die funktionsübergreifende Zusammenarbeit und verhindert, dass aus Daten Mehrwert geschaffen werden kann. Zudem birgt er vielfältige Risiken: mangelhafte Abstimmung zwischen den Projektbeteiligten, doppelter Arbeitsaufwand, lückenhafte Governance, ineffiziente Datennutzung, höhere Kosten und in letzter Konsequenz Verzögerungen oder verfehlte Unternehmensziele.

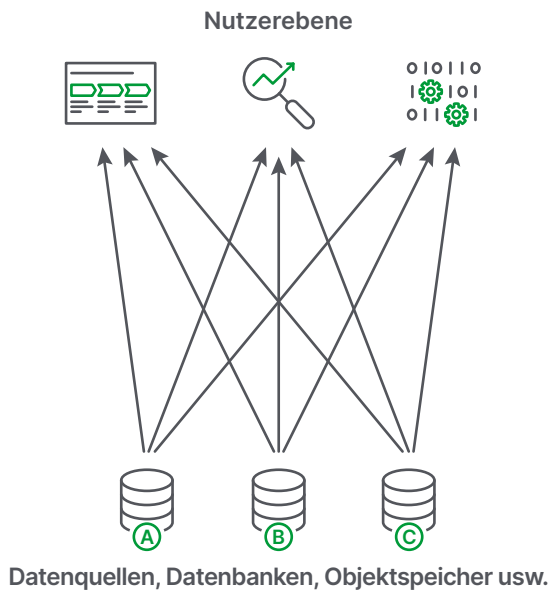
- Im gesamten Unternehmen **fehlt es systemübergreifend an konsistenter Datenqualität.** Dies untergräbt die Entscheidungsprozesse und erschüttert das Vertrauen in datengestützte Erkenntnisse.
- In der gesamten Daten- und Wertschöpfungskette **fehlen klare Zuständigkeiten und Verantwortlichkeiten** zwischen Datenproduzenten und -nutzern. Versuche, aus den Daten Mehrwert zu schaffen, wie es das Business erfordert, werden im Keim erstickt.
- **Fragmentierte Lösungen sowie Ad-hoc- und Do-it-yourself-Ansätze** im Datenmanagement verlängern die Amortisationszeit und ihre Anpassung und Pflege verursachen ohne qualifizierte Mitarbeiter hohe Kosten.

Angesichts dieser Herausforderungen wächst die Kluft zwischen Datenproduzenten und Datennutzern. Der Fokus der Produzenten liegt auf der IT. Für sie sind Datenplattformen wichtiger als die Auseinandersetzung mit den Anforderungen der Nutzer oder der Datenverwendung. Die Datennutzer in den Fachbereichen haben Schwierigkeiten, ihre Vorstellungen zu vermitteln und misstrauen den bereitgestellten Daten. **Damit Unternehmen diese wachsende Kluft schließen können, benötigen sie ein strategisches, gezieltes Konzept für das Datenmanagement.**

¹ IDC-Blog, „Navigating the Planes of Enterprise Intelligence Architecture“, Juni 2023.

Datenmanagement im Wandel

Bislang haben Unternehmen zwei unterschiedliche Ansätze verfolgt, um aus ihren Daten Mehrwert zu erzielen: den **fragmentierten** und den **zentralisierten** Ansatz.



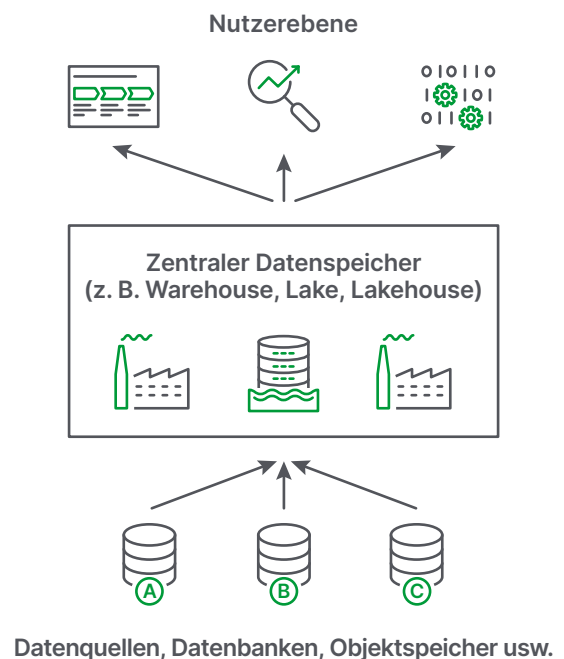
DER FRAGMENTIERTE ANSATZ

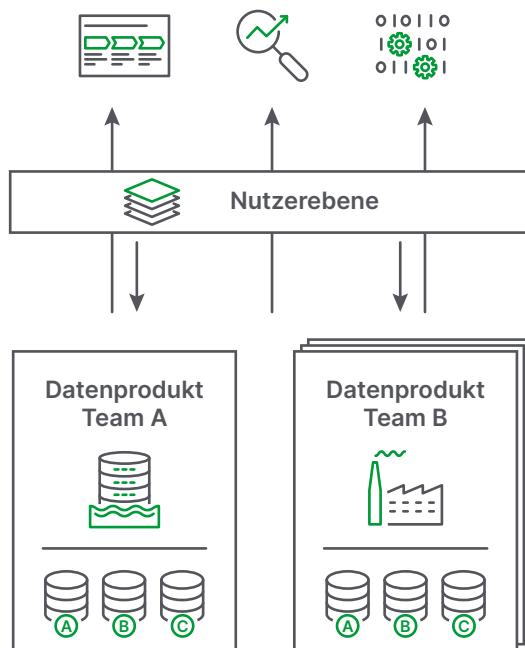
Bei einem fragmentierten Ansatz sammeln einzelne „Datenprojektteams“ Daten aus unterschiedlichen Quellsystemen und verwenden sie zur Umsetzung konkreter Anwendungsfälle. Dafür nutzen die Teams ihre eigenen Tools und Technologien. Sie greifen unabhängig voneinander auf die Daten zu, um diese für ihre jeweiligen Anforderungen zu extrahieren, zu integrieren und zu analysieren. Diese Methode ist zwar kurzfristig von Vorteil, in großen Unternehmen lohnt sie sich aber langfristig nicht. Denn sie führt zu mehr Datensilos, doppeltem Arbeitsaufwand, höheren Kosten und Verspätungen bei den Ergebnissen.

DER ZENTRALISIERTE ANSATZ

Bei einem zentralisierten Ansatz dagegen übernimmt ein zentral organisiertes Datenteam das Extrahieren, Bereinigen und Aggregieren umfangreicher Datenmengen. Dieses zentrale Team hat die Aufgabe, die Analyseanforderungen des Unternehmens umfassend abzudecken. Es erstellt komplexe Datenpipelines, um Daten aus verschiedenen Systemen abzurufen, zu bereinigen und in Cloud-basierten Data Lakes, Warehouses oder Lakehouses (wie Databricks Delta Lake, Snowflake, BigQuery, S3 usw.) zusammenzuführen.

Zentralisierung erscheint wie die ideale Lösung des Fragmentierungsproblems – und in gewissem Maße ist sie das auch. Sie hat jedoch einen gravierenden Nachteil: Da sie eine präzise Koordination zwischen den Abteilungen erfordert, um Änderungen oder Aktualisierungen vorzunehmen, schränkt sie die Flexibilität des Unternehmens erheblich ein. Weil der Fokus bei diesem Ansatz zudem auf generellen und unternehmensweiten Datenanforderungen liegt, übersieht das zentrale Team häufig die speziellen Bedürfnisse und Präferenzen der Stakeholder und Nutzer in den Fachbereichen. Das Ergebnis ist eine suboptimale Lösung, die den Erfordernissen des Unternehmens nicht gerecht wird.





DER FÖDERIERTE ANSATZ

Die beiden herkömmlichen Ansätze – der fragmentierte und der zentralisierte – haben ihre Grenzen und können die Bedürfnisse der Unternehmen hinsichtlich Datennutzung nur unzureichend erfüllen. Angesichts dieser Herausforderungen vollzieht sich derzeit ein Paradigmenwechsel. Bei dem zunehmend beliebteren **föderierten Ansatz** erfolgt die Bereitstellung von Daten strategisch mithilfe von Datenprodukten, während die technischen Funktionen der Plattform zentral verwaltet werden.

Auf diese Weise können mehrere Teams parallel arbeiten, wobei sich jedes auf die für seinen Bereich wichtigen Datenprodukte konzentriert. Dieser Ansatz führt durch die funktionsübergreifende Zusammenarbeit der Beteiligten aus den Fachbereichen und der Technik schneller zu Ergebnissen.

Der föderierte Ansatz bietet mehrere entscheidende Vorteile:



Flexibilität und Agilität

Durch ein schlankeres End-to-End-Datenmanagement können Unternehmen **zeitnah Mehrwert** aus ihren Daten generieren. Der föderierte Ansatz sorgt durch seine Produktorientierung für Agilität, weil die Bereitstellung und Wertschöpfung auf funktionsübergreifende Produktteams verteilt ist, während die technischen Funktionen der Plattform zentral abgewickelt werden. Dadurch können Unternehmen **schnell auf neue geschäftliche Anforderungen reagieren** und zugleich die Datenintegrität und Verantwortlichkeiten während des gesamten Prozesses sicherstellen.



Wirtschaftlichkeit

Das föderierte Modell basiert auf **Wiederverwendbarkeit** und ermöglicht Unternehmen, ihre Investitionen zu maximieren und Betriebskosten zu senken. Werden darüber hinaus Basiskomponenten zentral von verschiedenen Produktteams gemeinsam genutzt, **minimiert dies redundante Arbeitsschritte und verbessert die Pflege der Daten nachhaltig**.



Schnellere Umsetzung von Anwendungsfällen

Da funktionsübergreifende Teams einfacher zusammenarbeiten können, kommt es **schneller zu bereichsspezifischen Ergebnissen**, die auf die sich ändernden geschäftlichen Anforderungen und neue datengesteuerte Initiativen zugeschnitten sind. Durch mehrere parallele, teamübergreifende Entwicklungsstreams **ermöglicht der föderierte Ansatz Skalierbarkeit und eine individuelle, bedarfsgerechte Anpassung** der Anwendungsfälle.



Mehr Vertrauen

Das föderierte Modell **schafft Vertrauen in die datengesteuerte Entscheidungsfindung mit verbesserter Governance, Transparenz und Compliance**. Die Zuständigkeiten für die Datenbereiche sind klar definiert. Dies erleichtert die Abstimmung und klärt die Verantwortlichkeiten zwischen allen Interessengruppen im Bereich Daten und Analysen. Das Data Governance Framework des Unternehmens wird so zusätzlich gestärkt.

Datenprodukte einführen

Ein föderierter Ansatz ist der optimale Ausgangspunkt für Datenprodukte. Für viele Unternehmen markiert er eine Evolution: von der Kontrolle zur Zusammenarbeit, vom Projekt zum Produkt und von Rigidität zu Agilität – verbunden mit der Möglichkeit, hohen Mehrwert zu schaffen. **Data Mesh** und **Data Fabric** stehen zwar für unterschiedliche Datenmanagementkonzepte, können aber gemeinsam eine starke Synergie bilden, um das volle Potenzial der Datenprodukte zu erschließen. Beim Data Mesh ist die Zuständigkeit für die Daten dezentralisiert und wird von den nutzenden Teams gesteuert. Das Datenprodukt ist der konkrete und wirkungsvollste Aspekt des Data-Mesh-Ansatzes. Das Data Fabric ist die zugrundeliegende Datenmanagementarchitektur, die eine optimale Bereitstellung von Datenprodukten für die Teams der Fachbereiche ermöglicht.²



Laut der CDAO Agenda Survey 2024 von Gartner haben 50 % der teilnehmenden Unternehmen bereits Datenprodukte eingeführt. Weitere 29 % planen eine Pilotphase oder die Einführung innerhalb des nächsten Jahres.³

Was genau ist ein Datenprodukt? Ein Datenprodukt liefert qualitativ hochwertige, kuratierte, gebrauchsfertige und AI-fähige Datensätze, auf die Mitarbeiter problemlos zugreifen und die sie für verschiedene Anwendungsfälle in einem Unternehmen nutzen können. Ein Datenprodukt kann man sich als eine Art Mikroservice für Daten vorstellen. Mikroservices stellen Anwendungsfunktionen über kleinere, kontextbezogene und modulare Einheiten bereit. Ebenso bietet ein Datenprodukt über Serviceschnittstellen Zugang zu fachbezogenen und relevanten Daten. Ein Datenprodukt erfüllt mehrere zentrale Kriterien:



Bereichsorientierung und Ownership

Jedes Datenprodukt muss einem bestimmten Bereich zugeordnet sein und es muss einen Datenverantwortlichen (Person oder Team) geben, der für den Erfolg des Produkts, beispielsweise Mehrwert schaffen oder Akzeptanz erhöhen, verantwortlich ist.



Suchbar und auffindbar

Datenprodukte werden in einem Datenproduktkatalog registriert und können von den Nutzern problemlos gesucht und gefunden werden. Das Produkt beschreibt sich selbst: welche Daten es beinhaltet, wie darauf zugegriffen werden kann, die Semantik sowie seine Herkunft und Qualität.



Produktgedanke

Bei Datenprodukten wird der Produktgedanke auf Daten übertragen, einschließlich Zuständigkeit, Versionierung und Kunden (in diesem Fall die Datennutzer). Die Produkte werden durch das Auswerten von Nutzungsmustern und auf Basis von Kundenfeedback optimiert. Sie haben einen Lebenszyklus genau wie echte Produkte.



Wiederverwendbar

Datenprodukte werden einmalig erstellt und dann für andere Anwendungsfälle wiederverwendet. Andere Datenprodukte werden aus bereits vorhandenen erstellt.



Standardisierung und Interoperabilität

Datenprodukte folgen standardisierten Formaten, Schemata und Metadatenkonventionen. Sie vereinfachen die Interoperabilität und die Integration in andere Komponenten des Datenökosystems.

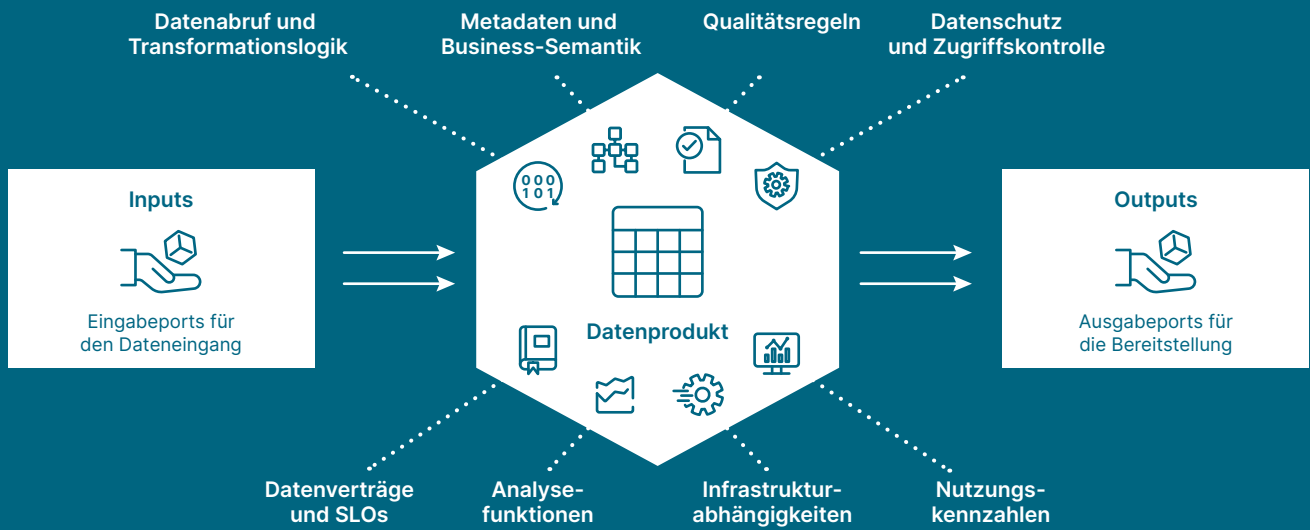


Zugänglich

Der Zugriff auf Datenprodukte erfolgt über eine Standardschnittstelle wie SQL oder eine REST-API, die bei der Erstellung der Produkte definiert und implementiert wird. Die Schnittstellen verbergen die Komplexität innerhalb des Datenprodukts.

² Gartner: „6 Lessons Data Leaders Can Learn from Adopters of Data Mesh“ von Robert Thanaraj et al. 4. März 2024. GARTNER ist eingetragenes Warenzeichen und Dienstleistungsmarke von Gartner, Inc. und/oder seinen Niederlassungen in den USA und international und wird hier mit Genehmigung verwendet. Alle Rechte vorbehalten.

³ Gartner: „Gartner Chief Data and Analytics Officer Agenda Survey for 2024“. 12. März 2024. Die Ergebnisse dieser Studie spiegeln keine allgemeingültigen Erkenntnisse oder den Markt als Ganzes wider, sondern repräsentieren lediglich die Stimmung der befragten Teilnehmer und Unternehmen. GARTNER ist eingetragenes Warenzeichen und Dienstleistungsmarke von Gartner, Inc. und/oder seinen Niederlassungen in den USA und international und wird hier mit Genehmigung verwendet. Alle Rechte vorbehalten.

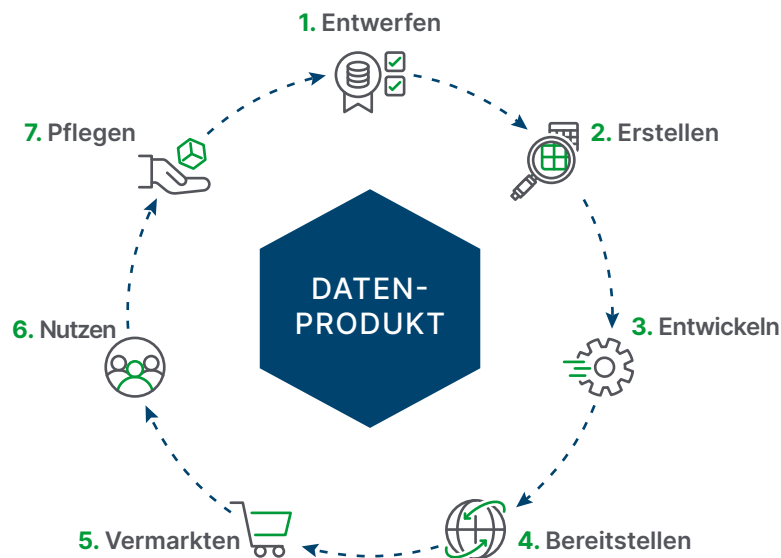


Ein Datenprodukt ist ein eigenständiges Konstrukt, in dem sich Daten, Transformationen, Code, Governance-Richtlinien, SLAs, Zugriffsmuster und Infrastrukturabhängigkeiten als Ganzes verbergen. Das obige Diagramm veranschaulicht die Zusammensetzung eines Datenprodukts.

- 1 Datensatz:** Er ist der Mittelpunkt des Produkts. Hierbei kann es sich um zusammengehörige Tabellen, Dateien, Graphen oder Vektordarstellungen handeln. Ein Datenprodukt kann mehrere Datenebenen enthalten: einen Rohdatensatz, der die Daten so darstellt, wie sie aus einem Quellsystem übernommen wurden, einen kuratierten Datensatz, der bereinigte und integrierte Daten umfasst, sowie einen gebrauchsfertigen, speziell für eine einfache Nutzung angelegten Datensatz.
- 2 Inputs:** Sie stehen für die Eingabeports oder die Eingangsdatenquelle. Das sind beispielsweise Tabellen in einer SAP-Anwendung, die eingelesen werden müssen, Tabellen in Snowflake, die als Quelle verwendet werden sollen, ein anderes Datenprodukt, aus dem ein neues abgeleitet wird, oder eine Message Queue zum Empfang von Echtzeitdatenfeeds.
- 3 Outputs:** Diese Mechanismen sorgen für die Bereitstellung der Daten. Ein Datenprodukt kann eine Vielzahl von standardisierten Schnittstellen wie SQL, REST-API, Vektorsuche, Nachrichten und Dateien bereitstellen. Die Ausgabeports können auch je nach Art des Diensts variieren, den das Datenprodukt bietet. Ein Output kann beispielsweise Zugriff auf historische Daten ermöglichen, während ein anderer den jeweils neuesten Daten-Snapshot bereitstellt.
- 4 Datenabruf und Transformationslogik:** Hierbei geht es um den Code und die Logik für die Datenpipeline zur Datenübernahme – wie die Übertragung der Daten aus einer Quelldatenbank zu einer Cloud-Datenplattform oder das Einlesen von Tabellen – und die anschließende Anwendung der Transformationslogik. Mit ihr werden die Daten aus dem Quell- in das Zieldatenformat gebracht, das die Form von Tabellen, Ansichten oder Vektoreinbettungen haben kann.
- 5 Metadaten und Business-Semantik:** Dies sind die technischen (Tabellennamen, Dateinamen, Spaltendefinitionen) und Business-Metadaten (Bereich, Pro-
- duktnamen, Art des Produkts, Zweck, Zuständigkeit, Tag usw.), die das Datenprodukt verständlicher machen.
- 6 Datenqualitätsregeln:** Sie definieren die verschiedenen Datenqualitätsprüfungen, die das Datenprodukt durchlaufen muss. Jede Regel verfügt über Konfigurationsinformationen unter anderem zum Ausführungszeitpunkt, zum Geltungsbereich der Regel und ob diese auf alle Daten oder nur einen Teil angewendet werden soll. Qualitätsregeln werden nach einem festgelegten Zeitplan ausgeführt. Dies sorgt für genaue Daten und schafft Vertrauen.
- 7 Richtlinien für Datenschutz und Zugriffskontrolle:** Sie steuern den Zugriff auf den Datensatz innerhalb des Datenprodukts. Enthält der Datensatz beispielsweise personenbezogene Daten, kann eine Datenschutzrichtlinie diese für alle Benutzer maskieren oder nur bestimmten Benutzergruppen den Zugriff gestatten.
- 8 Datenverträge und SLOs:** Sie definieren die Sollwerte für das Serviceziel des Datenprodukts. Ein SLO könnte beispielsweise festlegen, dass die Zuverlässigkeit der Datenqualität mehr als 95 % beträgt oder dass die Daten alle fünf Minuten aktualisiert werden.
- 9 Analysefunktionen:** Sie enthalten die zusätzliche benutzerdefinierte Business-Logik, um die Funktion wiederverwendbar zu machen. Eine Funktion eines Datenprodukts für Bestellungen könnte „Delinquent Orders“ sein und die komplexe Anwendungslogik enthalten, mit der festgestellt wird, welche Bestellungen überfällig sind. Analysefunktionen können über einen Ausgabeport wie eine API zugänglich gemacht werden.
- 10 Infrastrukturabhängigkeiten:** Dies sind Informationen zu den Infrastrukturanforderungen der mit dem Datenprodukt verbundenen Datensätze. Beispielsweise werden die Datensätze auf einer Datenplattform (wie Snowflake oder Databricks) gespeichert, wobei der dazugehörige Code (z. B. die Transformationslogik) zur Ausführung eine Rechenumgebung (z. B. SQL) benötigt.

Der Lebenszyklus von Datenprodukten

Datenprodukte sind nicht statisch. Wie physische Produkte müssen sie kontinuierlich verwaltet und permanent kontrolliert werden. Für jedes Datenprodukt ist ein auf den jeweiligen Bereich spezialisiertes Datenteam zuständig. Seine Aufgabe ist es, Mehrwert zu schaffen, die Akzeptanz zu steigern und den Lebenszyklus des Produkts zu unterstützen.



Die wichtigsten Phasen im Lebenszyklus von Datenprodukten sind:

- 1 Entwerfen:** Der Prozess beginnt mit der Identifizierung der jeweiligen fachbereichsspezifischen Anforderungen. Die Nutzer, ihre Anwendungsfälle und die Art, wie sie normalerweise Daten einsetzen, muss dabei natürlich auch berücksichtigt werden.
- 2 Erstellen:** Der nächste Schritt ist die Gestaltung der Lösung. Er beinhaltet das Bestimmen der erforderlichen Datensätze und ihrer Quellen sowie die Definition der Merkmale des Datenprodukts nach Funktionalität, Interaktivität, Darstellung usw.
- 3 Entwickeln:** Nach der Erstellung folgt der Aufbau der technischen Infrastruktur und der Funktionalitäten. Dazu werden die Rohdatensätze aus ihren Quellen erfasst und integriert, die Qualität sichergestellt und die Daten transformiert.
- 4 Bereitstellen:** In dieser Phase werden die Daten „zugänglich gemacht“. Das beinhaltet die Produktdokumentation, die Festlegung der Zuständigkeiten, die Integration in bestehende Unternehmenssysteme oder Workflows für einen nahtlosen Zugriff sowie die Aktivierung des Datenprodukts bzw. seine Bereitstellung für die Nutzung.
- 5 Vermarkten:** Nach der Aktivierung wird es Zeit, das Datenprodukt bekannt zu machen und seine Verwendung zu fördern. Hierzu müssen die Nutzer geschult und Kommunikationskanäle zum Einholen von Feedback eingerichtet werden.
- 6 Nutzen:** Nach dem erfolgreichen Launch beginnen die Nutzer, mit dem Datenprodukt zu interagieren: Sie suchen, finden, erkennen und erhalten Zugang. In dieser Phase liegt der Fokus auf der Überwachung der Anwendung und dem Einholen von Feedback. In der Regel werden Daten darüber erfasst, wie Nutzer mit dem Datenprodukt interagieren, wie gut es funktioniert und ob es die Anforderungen der Geschäftsbereiche erfüllt.
- 7 Pflegen:** Datenprodukte sind nicht statisch. Damit sie auch künftig einen Mehrwert bieten, müssen sie optimiert werden. Basierend auf dem Feedback der Nutzer und den sich ändernden Anforderungen des Business fließen die Vorschläge für neue Dimensionen, Features und Funktionalitäten dann in die nächste Version des Datenprodukts ein.

Unternehmen, die diese Lebenszyklusphasen berücksichtigen, können Datenprodukte erstellen, die sinnvolle Erkenntnisse liefern, datengestützte Entscheidungen fördern und zum Erreichen der Geschäftsziele beitragen.

Was ist ein Datenprodukt NICHT?

In der Welt der Daten werden viele Begriffe nicht immer einheitlich verwendet. „Datenprodukt“ wird häufig mit Begriffen wie „Datensatz“ oder „Datenkatalog“ gleichgesetzt, obwohl es sich um unterschiedliche Konzepte handelt. In diesem Abschnitt werden einige der grundlegenden Unterschiede zwischen diesen Konzepten erläutert.



Ein Datensatz ist ein grundlegender Bestandteil vieler Datenprodukte, aber man muss sich klarmachen, dass er nur eine Komponente ist. Ein Datenprodukt ist weit mehr als nur der Rohdatensatz.

DATENSÄTZE: DAS AUSGANGSMATERIAL

Bei einem Datensatz handelt es sich in der Regel um eine Reihe organisierter (strukturierter oder unstrukturierter) Datenpunkte. Hierzu zählen beispielsweise Rohdaten wie Kundeninformationen, Finanzunterlagen oder Sensorwerte. Betrachten Sie Datensätze als einen Rohstoff, wie einen Stapel gefällter Baumstämme. Das Wertschöpfungspotenzial ist da, aber zuerst muss es aufbereitet werden. Der wichtigste Unterschied zwischen einem Datensatz und einem Datenprodukt ist der beim Datensatz fehlende bereichsspezifische Businesskontext.

DATENPRODUKTE: DAS ENDPRODUKT

Datenprodukte bestehen aus sorgfältig kuratierten und verarbeiteten Datensätzen. Diese werden, wie beschrieben, oft mit zusätzlichen Artefakten kombiniert, um Mehrwert zu generieren. Um bei der Baumstamm-Metapher zu bleiben: Datenprodukte sind die Möbel, die aus dem Holz (den Datensätzen) angefertigt wurden.

So unterscheiden sich Datenprodukte von Datensätzen:

- **Kontext- und Bereichsbezogenheit:** Datenprodukte sind auf konkrete Anforderungen der Geschäftsbereiche und Nutzer zugeschnitten. Sie gehen über die Rohdaten hinaus und enthalten kontextbezogene Informationen wie Definitionen, Herkunft und Qualitätsprüfungen.
- **Benutzerfreundlichkeit und Zugänglichkeit:** Datenprodukte werden erstellt, um einen Mehrwert zu bieten und die Nutzung zu erleichtern. Sie sind so aufbereitet und zusammengestellt, dass sie auch von technisch nicht so versierten Anwendern problemlos genutzt werden können.
- **Verwertbare Erkenntnisse:** Datenprodukte präsentieren Daten nicht nur, sondern bieten auch operative, analytische, prädiktive und präskriptive Einblicke. Diese werden über Visualisierungen/Dashboards, vorgefertigte Analysemodelle oder Hooks zur Integration mit anderen Tools via SQL oder APIs bereitgestellt.
- **Ownership und Verantwortlichkeit:** Bei Datenprodukten gibt es klare Zuständigkeiten. Sie regeln, wer die Verantwortung für verschiedene Aspekte ihrer Entwicklung, Nutzung und Akzeptanz im Unternehmen trägt. Wer für ein Datenprodukt zuständig ist, richtet sich nach dem Business-Kontext. Der Owner kann die Person sein, die das Datenprodukt erstellt, der Nutzer des Fachbereichsteams, der für die entsprechenden Anwendungsfälle verantwortlich ist, oder eine neu entstehende Funktion: der sogenannte Data Product Manager.



Datenprodukte verwandeln Datensätze in wertvolle, nutzbare Ressourcen, die zu informierten Entscheidungen und zum Wachstum des Unternehmens beitragen. Nach Einschätzung der Harvard Business Review „können Unternehmen, die Daten wie ein Produkt behandeln, den Zeitaufwand für deren Implementierung in neue Anwendungsfälle um bis zu 90 % reduzieren, ihre Kosten (für Technologie, Entwicklung und Wartung) um bis zu 30 % senken und die Risiken und Belastung durch Data Governance verringern“.⁴

⁴ „A Better Way to Put Your Data to Work“, Harvard Business Review, Juli/August 2022.

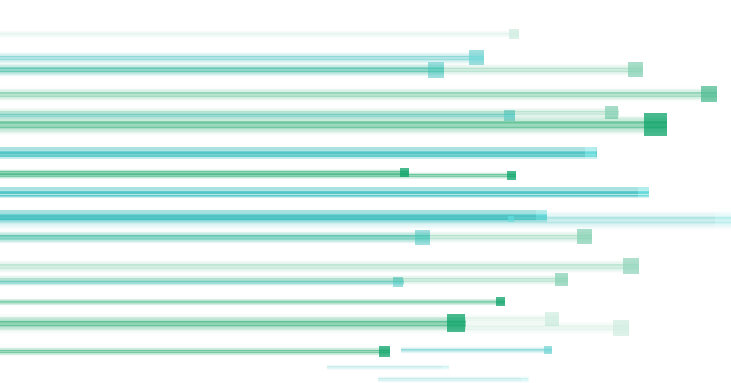
DER DATENPRODUKTKATALOG: DER HUB FÜR ENDPRODUKTE

Ein Datenproduktkatalog ist ein umfassendes Repository für Datenprodukte und die zugehörigen Metadaten (Beschreibungen, Tags, Versionen, Labels usw.) und sorgt dafür, dass diese leichter auffindbar und verständlich werden. Der Datenproduktkatalog unterstützt Datenproduzenten und -nutzer und bietet ihnen maßgeschneiderte Lösungen für ihre Anforderungen. Für Data Product Manager stellt der Katalog Tools bereit, mit deren Hilfe sie den gesamten Lebenszyklus eines Datenprodukts – von der Erstellung über die Aktivierung bis hin zum End-of-Life – konfigurieren, aktivieren und verwalten können. Nutzern dient der Katalog als Marktplatz, der ein einkaufs-ähnliches Erlebnis ermöglicht und die Verwendung fördert.

Ein Datenproduktkatalog unterscheidet sich wie folgt von einem Datenkatalog:

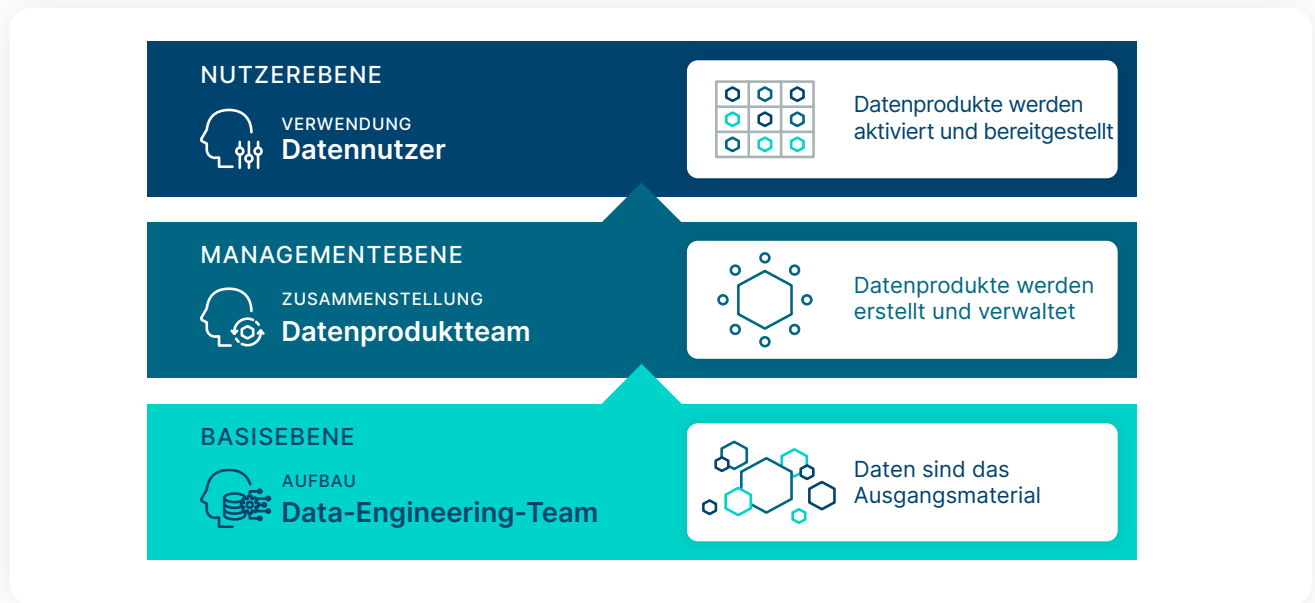
FUNKTION	DATENKATALOG	DATENPRODUKTKATALOG
Aufgabe	Data Governance stärken	Mehrwert aus Daten erhöhen
Umfang	Unternehmensorientiert	Fachbereichsorientiert
Managementmodell	Zentralisiert	Föderiert
Zweck	Auffindbarkeit, Compliance, Transparenz	Benutzerfreundlichkeit, Vertrauen, Verwertbarkeit
Inhalt	Metadaten verschiedenster Art, insbesondere technische, operative und geschäftliche Metadaten zu jedem Datenbestand	Vertrauenswürdige Datenprodukte, Metadaten zum Lebenszyklus von Datenprodukten, Dokumentation und Zuständigkeit
Einträge	Mehrere Hunderttausend	Mehrere Hundert
Zielgruppe	In erster Linie Data Governance	Datenproduzenten und -nutzer

Beim Datenmanagement müssen nicht nur Datenplattformen entwickelt, sondern auch deren effektive Nutzung sichergestellt werden. Allzu häufig investieren Unternehmen beträchtliche Beträge – nicht selten in Millionenhöhe – in den Aufbau von Dateninfrastrukturen, nur um auf ein ernstes Problem zu stoßen: **Beim Datenaustausch zwischen Produzenten und Nutzern hapert es.** Wieso? Viele Technologieplattformen liefern keine zuverlässigen Ergebnisse und sind nicht auf die konkreten Geschäftsziele der Nutzer ausgerichtet. **Daraus folgt: Sie benötigen in Ihrem Unternehmen beide Kataloge: einen Datenkatalog und einen Datenproduktkatalog.** Stellen Sie sich die Kataloge wie verschiedene Gänge in Ihrem Supermarkt vor: einen Gang mit Grundnahrungsmitteln wie Mehl, Butter und Zucker, den anderen mit Fertigprodukten wie Kuchen und Gebäck.



Datenproduktkatalog: Ebenen und Akteure

Der Datenproduktkatalog ist der wichtigste Hub bei der Verwaltung von Datenprodukten im Unternehmen. Er ist in drei Ebenen unterteilt, die jeweils auf eine bestimmte Stakeholdergruppe zugeschnitten sind. Ob Data Engineers, Data Stewards, Analysten oder neu entstehende Rollen wie die des Data Product Managers: Jede Gruppe interagiert auf ganz spezielle Weise mit diesen Ebenen, je nach Ziel und Anforderung. Das Ebenenkonzept trennt die unterschiedlichen Nutzeranforderungen voneinander und bietet das richtige Maß an Abstraktion, um die Einführung und Nutzung zu erleichtern.



DIE BASISEBENE

Die Basisebene ist das Fundament des Datenproduktkatalogs. Sie legt die technische Infrastruktur fest und stellt Funktionen zur Unterstützung der nachfolgenden Ebenen bereit.

Die wichtigsten Funktionen dieser Ebene sind:

- **Data Engineering:** Daten aus verschiedenen Quellen, insbesondere aus Dateien, Cloud Data Warehouses, Data Lakes und Echtzeit-Datenströmen übertragen und verarbeiten.
- **Metadatenmanagement:** Wertvolle Informationen zu den Daten, wie Datenherkunft, Data Dictionary, Schemata usw. zusammentragen und organisieren.
- **Sicherheit und Zugriffskontrolle:** Datensicherheitskontrollen durchführen, um die Daten zu schützen und zu gewährleisten, dass nur autorisierte Nutzer darauf zugreifen können.
- **Überwachung von Datenqualität und Datenherkunft:** Qualität kontrollieren, um die Richtigkeit der Daten sicherzustellen, und durch nachvollziehbare Daten die Zuverlässigkeit stärken.

Die Basisebene wird in erster Linie vom Datenplattformteam genutzt, das sich vor allem aus praxisorientierten und technischen Rollen wie Data Engineers, Data Stewards und Data Architects zusammensetzt. Das Ziel ist, die Funktionen dieser Ebene als Self-Service anzubieten, sodass auch Nutzer ohne technischen Hintergrund sie problemlos verwenden können. Die in der Basisebene benötigten Funktionen befinden sich in verschiedenen Tools für das Metadatenmanagement und für die Erfassung, Transformation, Qualität und Sicherung der Daten.

DIE MANAGEMENTEBENE

Die Datenprodukt-Managementebene ist ein neues Konzept. Sie bietet alle erforderlichen Funktionen, um die Anforderungen der ebenfalls neuen Rolle des Data Product Managers (oder Data Product Teams) zu erfüllen. Auf dieser Ebene finden Sie Funktionen zum Entwickeln, Verwalten und Pflegen vertrauenswürdiger, bereichsbezogener Datenprodukte. Die Ebene stützt sich auf die Dienste der Basisebene, um Datenbestände zu inventarisieren, zu übertragen, zu transformieren und zu sichern. Sie ist die Brücke zwischen den rohen Metadaten der Basisebene und den verwertbaren Erkenntnissen, die die Nutzer aus ihren Daten ziehen wollen.

Die wichtigsten Funktionen dieser Ebene sind:

- **Zusammenstellung von Datenprodukten:** Funktionen zum Aufbau eines Datenprodukts, zur Definition von Inputs und Outputs und zur Konfiguration der verschiedenen Artefakte.
- **Bereitstellung von Datenprodukten:** Versionsmanagement und der Workflow zur Aktivierung eines Datenprodukts, wenn dieses für den Einsatz durch die Nutzer bereit ist.
- **Pflege von Datenprodukten:** Nachverfolgung der Nutzung, Einholen von Feedback, Änderungen und Aktualisierungen sowie Erfassung aller Phasen des Datenproduktlebenszyklus, die in den vorherigen Abschnitten beschrieben wurden.

Die Datenprodukt-Managementebene ist primär für das Produktmanagementteam, konkret die Data Product Manager, gedacht, die den Lebenszyklus von Datenprodukten steuern, um die Unternehmensziele zu erreichen.

DIE NUTZEREbene (ODER AUCH MARKTPLATZEBENE)

Die Marktplatzebene befindet sich an der Spitze des Datenproduktkatalogs und ist der Teil des Systems, mit dem die Nutzer interagieren. Diese Ebene bietet Self-Service-Funktionen zum Untersuchen und Anwenden der Daten und fördert so eine datengesteuerte Unternehmenskultur.

Die wichtigsten Funktionen dieser Ebene sind:

- **Suchen und Auffinden:** Über intuitive Benutzeroberflächen können die Nutzer die zu ihren Anforderungen passenden Datenprodukte problemlos finden und untersuchen.
- **Informieren und Vertrauen:** Durch einen Blick auf Meta- und Beispieldaten erfahren die Nutzer mehr über das Datenprodukt. Beispielcode und Dokumentation unterstützen sie bei der Verwendung und der Einblick in die Qualität und Herkunft stärkt ihr Vertrauen in die Zuverlässigkeit der Daten.
- **Zugriff und Berechtigungen:** Auf dieser Ebene wird geregelt, wer auf welche Datenprodukte zugreifen darf und die Sicherheit der Daten und ihre Compliance mit den Data-Governance-Richtlinien wird gewährleistet.
- **Schnittstellen zur Datennutzung:** APIs, Visualisierungstools oder direkter Datenbankzugriff ermöglichen Nutzern die Anwendung der Datenprodukte. Je nach Produkt können verschiedene Schnittstellen verfügbar sein, die die unterschiedlichen Anforderungen und Fertigkeiten der Nutzer berücksichtigen.

Die Datennutzungsebene ist in erster Linie für Business-Analysten, Data Scientists, Anwender aus den Geschäftsbereichen und Entscheider im Unternehmen gedacht.

VORTEILE VON DATENPRODUKTEN FÜR DIE VERSCHIEDENEN AKTEURE IM DATENÖKOSYSTEM

Daten sind die wichtigste Ressource eines Unternehmens. Ihr volles Potenzial kann jedoch nur dann ausgeschöpft werden, wenn sie einfach zugänglich, leicht verständlich und effektiv nutzbar sind. Datenprodukte füllen diese Lücke. Sie verwandeln Rohdaten in einsatzbereite, kuratierte Daten, von denen alle Stakeholder im Unternehmen profitieren können.

Hier einige Vorteile von Datenprodukten für die wichtigsten Akteure im Datenökosystem:



Data-Engineering-Teams erhalten mit Datenprodukten eine gemeinsame Sprache in Form einer klar definierten Regelung, die die Kommunikation zwischen Datenproduzenten und -nutzern strukturiert und präzisiert. Dadurch können Datenproduzenten wiederverwendbare Datenprodukte schrittweise entwickeln und zur Senkung der Betriebskosten beitragen. Zudem muss das Data-Engineering-Team keine großen, zentralisierten und komplexen Datenpipelines mehr erstellen, sondern kann den Schwerpunkt stattdessen auf die Entwicklung kleinerer, unabhängiger Pipelines legen, die optimal auf die Implementierung der einzelnen Produkte zugeschnitten sind.



Für **Data-Governance-Teams** stellen Datenprodukte einen zentralen Prüfpunkt dar. Da die Produkte vertrauenswürdig und reguliert sind, wird es einfacher, Risiken zu minimieren. Das Governance-Team definiert zentral Richtlinien und Standards, die dann von den einzelnen Datenproduktteams passend umgesetzt werden.



Dem **Data Product Manager** erleichtern Datenprodukte die Abstimmung zwischen den Datenproduzenten und -nutzern. Dies sorgt für mehr Zusammenarbeit und Verantwortung im gesamten Datenökosystem. Für die Daten sind jetzt Ownership und Verantwortlichkeiten geregelt.



Datennutzer profitieren durch Datenprodukte von für ihren Fachbereich relevanten und kuratierten Datenbeständen, die ihnen alles Nötige bieten, um maximalen Mehrwert zu schaffen. Die Folge: weniger Nachbearbeitung und damit schnellere Ergebnisse.

Durch die gemeinsame Anwendung eines Datenproduktkatalogs können alle Beteiligten den Nutzen ihrer Datenbestände maximieren, datengestützte Entscheidungsprozesse vorantreiben und ihre Ziele erreichen.

„In der Hypothekenbranche beschränken sich Datenprodukte nicht auf Zahlen und Algorithmen, wir betrachten sie vielmehr als Transformationshelfer. Durch die Definition von kuratierten Self-Service-Datenprodukten wollen wir unseren Beratern ermöglichen, jedem Kreditnehmer die passende Hypothekenlösung anzubieten, während die Finanzteams Kosten und Risiken besser kalkulieren und kontrollieren können. Der Einsatz von Datenprodukten in unserem Unternehmen hat die Zukunft von Hypotheken für uns von Grund auf neu definiert.“

Julia Fryk

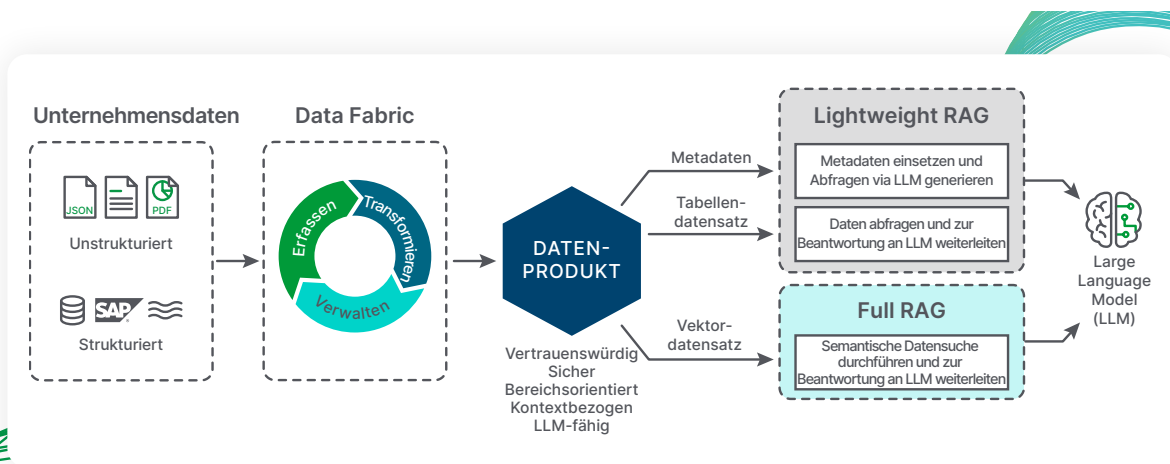
Principal Data Architect, Waterstone Mortgage

Mit Datenprodukten kontextbezogene Daten für LLMs bereitstellen

Generative AI bietet hervorragende Möglichkeiten für Produktivitätssteigerungen in allen Unternehmensbereichen. Im Mittelpunkt von GenAI stehen Basismodelle, beispielsweise LLMs wie GPT-3, Anthropic, Llama und BERT. Diese Modelle wurden mit einem umfangreichen Textdatenkorpus (Webseiten, Dokumente usw.) vor-trainiert. Basis-LLMs sind jedoch nicht in der Lage, interne, unternehmensspezifische Daten zu interpretieren. Deshalb müssen sie zunächst mit Unternehmensdaten unterfüttert werden.

Eine der effektivsten Techniken zur Bereitstellung von Unternehmensdaten für LLMs ist Retrieval Augmented Generation (RAG). Kontextbezogene Daten zu einem Prompt (einer Frage des Nutzers) werden **abgerufen**. Dann wird der Prompt mit diesen Daten **angereichert** und an das LLM übermittelt, um eine Antwort zu generieren. Es gibt zwei verschiedene RAG-Methoden: Lightweight RAG und Full RAG.

- Bei Lightweight RAG (für strukturierte Daten) übermittelt die RAG-Anwendung die Metadaten an das LLM, um eine SQL-Abfrage zu erstellen. Anschließend wird die Abfrage auf dem Tabellendatensatz ausgeführt und die Ergebnisse werden zur Erzeugung einer Antwort an das LLM weitergeleitet. Diese Methode kommt zum Einsatz, wenn eine eindeutige Antwort aus den Daten benötigt wird.
- Bei Full RAG (für strukturierte oder unstrukturierte Daten) werden die Daten dagegen zunächst vektorisiert. Danach führt die RAG-Anwendung eine semantische Suche innerhalb der vektorisierten Daten durch, um die Informationen abzurufen. Diese werden dann an das LLM übermittelt, das eine Antwort generiert.



Bei beiden Verfahren müssen für RAG hochwertige, bereinigte und kontextrelevante Daten zur Verfügung stehen. **Mithilfe von Datenprodukten werden kontextbezogene und bereichsorientierte Daten auf unterschiedliche Weise nutzbar.** Wie erwähnt, enthalten Datenprodukte umfassende Metadaten (geschäftliche, technische und bereichsorientierte) sowie hochgradig kuratierte Datensätze (Dateien, Tabellen oder vektorisierte Daten), die verwendet werden können. Aus diesem Grund sind Datenprodukte ideal für RAG-basierte Anwendungen.

Der Aufbau einer RAG-Anwendung beginnt mit dem Durchsuchen des Datenproduktkatalogs bzw. des Datenproduktmarktplatzes nach dem passenden Produkt für die gestellte Frage. Eine vertriebliche Frage lässt sich beispielsweise am besten mit einem „Opportunity“-Datenprodukt aus dem CRM beantworten, das auf den detaillierten, mit dem Datenprodukt verknüpften Metadaten basiert. Eine Frage zu einer Bestellung lässt sich wahrscheinlich am besten mit einem „Sales Order“-Datenprodukt aus SAP beantworten, das auf den detaillierten, mit diesem Datenprodukt verknüpften Metadaten beruht. Sobald das Datenprodukt gefunden wurde, kann die RAG-Anwendung den – tabellarischen oder vektorisierten – Datensatz des Datenprodukts basierend auf dem Anwendungsfall direkt abfragen.

Fazit

Weil Daten die Wirtschaft von heute immer stärker prägen, gewinnen Datenprodukte zunehmend an Bedeutung, um **maximalen Mehrwert aus Daten zu generieren**, schneller zu besseren Ergebnissen zu gelangen und zugleich **das Vertrauen in Daten und Governance zu stärken**. Qlik möchte Sie dabei unterstützen, eine verlässliche Datengrundlage für Self-Service-Funktionen aufzubauen. Während Ihre Datenteams in der Lage sind, Datenprodukte zu erstellen und zu verwalten, können Ihre Datennutzer diese Produkte im Handumdrehen durchsuchen, untersuchen und so einsetzen, dass sie zum Geschäftserfolg beitragen.

Eine letzte Anmerkung: Datenprodukte sind zwar leistungsstarke Tools, aber kein Allheilmittel. Ihr wahrer Vorteil liegt in der Identifizierung von Unternehmensanforderungen, in der Umsetzung einer gut durchdachten Strategie für föderierte Daten, in effektiver Governance und in einer teamübergreifenden Zusammenarbeit.

Möchten Sie das Potenzial Ihrer Daten voll ausschöpfen?

Mehr erfahren

„Unternehmen interessieren sich vermehrt für die Erstellung bereichsbezogener Datenprodukte, die mehr bieten als Datenintegration und -qualität. Die Datenqualitäts- und Governance-Angebote von Qlik verkörpern das Konzept der Datenprodukte und weisen dabei einen Reifegrad auf, der den Anforderungen der Kunden von heute Rechnung trägt.“

Stewart Bond

VP, Data Intelligence & Integration Software, IDC



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, businessstaugliche 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