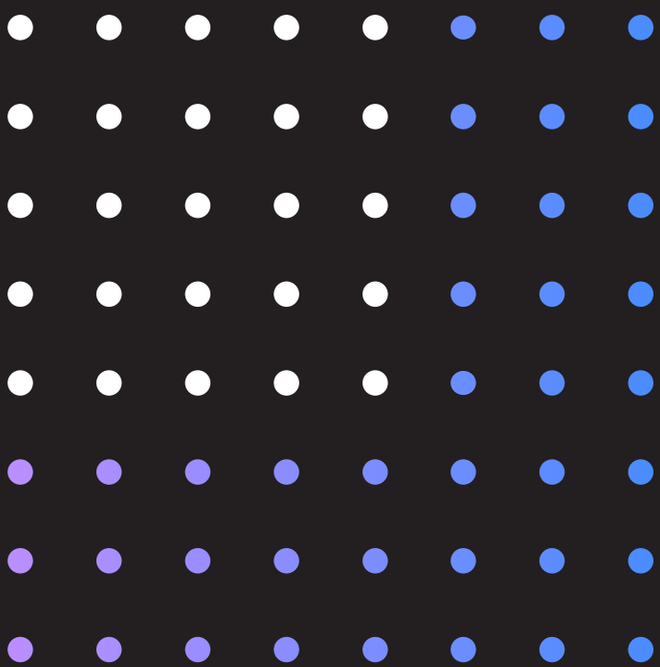
Bereitstellung von geschäftsfähigen Daten mit intelligenter Datenkatalogisierung und Data Lake Governance

IBM Watson Knowledge Catalog liefert eine durch maschinelles Lernen gestützte Data Governance-Plattform für alle Data Lake-Herausforderungen



IBM

Inhalt

03

Data Lake-Herausforderungen mit einem DataOps-Ansatz bewältigen

03

Herausforderungen bei der Verwendung von Data Lakes in Unternehmen

05

IBM Watson Knowledge Catalog

06

Zentrale Wissensressource und Zugriffsstelle

08

Vier Vorteile eines kontrollierten Data Lakes für KI

09

Fazit

Wichtige Erkenntnisse

- Nur wenige Unternehmen ziehen den erwarteten Nutzen aus den Data Lakes, die sie zum Speichern und Analysieren ihrer Daten für vertrauenswürdige Einblicke erstellt haben.
- DataOps lösen die Herausforderungen, denen Unternehmen gegenüberstehen, wenn sie Daten aufrufen, vorbereiten, integrieren und für die Benutzer verfügbar machen, während sie gleichzeitig Unternehmensrichtlinien und behördliche Vorschriften einhalten müssen.
- Zu den häufigen Herausforderungen bei Data Lakes gehören die Schwierigkeit und die Kosten des Imports neuer Datenquellen in den Data Lake, die Unfähigkeit, interne und externe Datensätze zu integrieren, das mangelnde Vertrauen in die Datenverwaltung, der fehlende Zugang zu Self-Service-Tools für die Datenaufbereitung und die mangelnde Kompetenz, die Daten im DataLake zu finden und zu verstehen.
- Eine Data Governance-Plattform für Unternehmen mit Funktionen für Katalogisierung, Datenqualität und Daten-Discovery kann ein fehlgeschlagenes Data Lake-Projekt in einen echten geschäftlichen Nutzen verwandeln.
- IBM Watson® Knowledge Catalog basiert auf IBM Cloud Pak™ for Data und liefert einen auf maschinelles Lernen (ML) gestützten Katalog für Daten-Discovery, Datenkatalogisierung, Datenqualität und Data Governance.Sie hilft den Datennutzern, Datenbestände, Datensätze und analytische Modelle schnell zu entdecken, zu kuratieren, zu kategorisieren und gemeinsam zu nutzen.
- Wenn Organisationen kein tiefes Verständnis ihrer Daten haben, wird es schwieriger, diesen Informationen zu vertrauen und sie mit allen Formen der künstlichen Intelligenz (KI), einschließlich ML und tiefem Lernen, zu nutzen.

Data Lake-Herausforderungen mit einem DataOps-Ansatz bewältigen

Vor zehn Jahren begann man, einen flexiblen, vielseitigen Ansatz für den Aufbau eines zentralen Datenspeichers zu finden, in dem alle Unternehmensdaten gespeichert werden können. Die Lösung war der Data Lake: eine Allzweck-Datenspeicherungsumgebung, in der praktisch jede Art von Daten gespeichert werden konnte. Damit konnten Geschäftsanalysten und Data Scientist die für jedes Dataset geeigneten Analyse-Engine und -werkzeug auf jeden Datensatz an seinem ursprünglichen Standort anzuwenden.

Diese Data Lakes wurden in der Regel mit Apache Hadoop und Hadoop Distributed File System (HDFS) zusammen mit Engines wie Apache Hive und Apache Spark erstellt. Als diese Data Lakes größer wurden, traten zunehmend Probleme zutage. Während die Technologie physisch skalierbar war, um riesige und vielfältige Sammlungen von strukturierten und unstrukturierten Daten zu erfassen, zu speichern und zu analysieren, wurde den praktischen Aspekten der Einbettung dieser Fähigkeiten in Geschäftsabläufe zu wenig Aufmerksamkeit geschenkt.

Bis 2022 werden über 80 % der Data Lake-Projekte keinen Nutzen liefern, da erfolgreiche Analysen und Data Science-Vorgänge wegen Schwierigkeiten beim Auffinden, Kategorisieren und Kuratieren von Daten erschwert werden.[1] Daher blieben Fragen wie die Folgenden oft unbeantwortet: „Welche Daten sollten im Data Lake abgelegt werden?“, „Wer wird ihn verwenden?“, „Wie erleichtern wir die Suche nach Daten?“, „Woher kommen diese Daten?“ und „Wie verhindern wir den Missbrauch von Daten?“ Diese kritischen Einschränkungen bei der Behandlung von Personen-, Prozess- und Technologiefragen führten effektiv zu erfolglosen Data Lake-Implementierungen.

Mittlerweile haben viele Unternehmen ihre Fehler erkannt und ihre Führungsteams für die Data Lake-Implementierung geändert und starten jetzt den zweiten, dritten oder sogar vierten Versuch einer erfolgreichen Data Lake-Implementierung – dieses Mal mit Schwerpunkt auf Datenvorgängen (Data Operations, DataOps).

Dieses Whitepaper gibt eine Einschätzung der allgemeinen Herausforderungen, denen sich Data Lakes gegenübersehen, und stellt neue Ansätze wie DataOps vor, die dazu beitragen können, sie aus einem Datensumpf in das Herzstück der geschäftsfähigen Datenpipeline eines Unternehmens zu verwandeln.

DataOps ist eine kooperative Datenmanagementmethode, die sich auf die Verbesserung der Kommunikation, Integration und Automatisierung von Data Lakes zwischen Datenmanagern und Datenkonsumenten innerhalb eines Unternehmens konzentriert.

Einführung in DataOps

DataOps kombiniert Best Practices von DevOps, Datenmanagement und Data Governance in einem gemeinsamen Framework, bei dem Datenflüsse von mehreren Beteiligten gemeinsam entwickelt und verwaltet werden. DataOps beseitigt die Herausforderungen, denen Unternehmen gegenüberstehen, wenn sie Daten effizient aufrufen, vorbereiten, integrieren und für die Nutzer verfügbar machen möchten, während sie gleichzeitig Unternehmensrichtlinien und behördliche Vorschriften einhalten. Diese effizienten Verfahren können in einer Geschäftseinheit, einem Analyseteam oder sogar in einem Betriebsprozess zu finden sein.

Für diese Methodik müssen die Personen-, Prozess- und Technologieprobleme gelöst werden, die den Unterschied zwischen erfolgreichen und erfolglosen Data Lake-Implementierungen ausmachen. Im Hinblick auf die Technologie unterstreicht DataOps die Bedeutung einer komplett integrierten End-to-End-Plattform für Datenaufnahme und -integration, Datenqualität, Data Governance und Datennutzung, um einen kontrollierten Data Lake zu erstellen. Regeln zur Validierung der Datenqualität sollten automatisch beim Integrationsprozess ausgeführt werden, um eine kontinuierliche Daten-Pipeline für das ganze Unternehmen aufrecht zu erhalten. Der Aufnahmeprozess sollte vollständig in den Datenkatalog integriert werden, der zum Kernstück der Pipeline wird. Datennutzer sollten auf die Datenqualitätswerte und Daten-Profiling-Ergebnisse aus dem Datenkatalog zugreifen und sich darauf verlassen können, dass das Unternehmen mit denselben Daten im Kontext arbeitet.

Das Wachstum der Daten übertrifft die Fähigkeit der Unternehmen, Nutzen aus ihnen zu ziehen. Auf die Frage nach den größten Herausforderungen bei der Nutzung von Insight-Systemen gaben Unternehmen folgende Antworten: 1) 40 % führen vorhandene Geschäftsprozesse mit Quelldaten für deren Analyse zusammen, und 2) 39 % beziehen, erfassen, verwalten und kontrollieren die Daten während des Wachstums.[2] Heutzutage geht es nicht nur um den Schutz der enormen Investitionen von Zeit und Ressourcen, die bereits in Data Lake-Technologien getätigt wurden –es gibt einfach keine Alternative. Von der Implementierung der KI bis hin zur Durchführung umfassender Analysen ist es von entscheidender Bedeutung, einen vollständigen Überblick über so viele Daten wie möglich zu haben, was bedeutet, dass Sie eine Architektur benötigen, die in der Lage ist, alle diese Daten an einem Ort zu speichern, zu analysieren und zu verwalten. In vielen Fällen ist ein kontrollierter Data Lake die einzige realistische Option, um diese Anforderungen zu erfüllen.

Die Unternehmen von heute können – und müssen – eine Möglichkeit finden, Wert aus ihrem Data Lake zu schöpfen. Dazu muss dieser eine einsatzbereite Daten-Pipeline für DataOps unterstützen.

Herausforderungen bei der Verwendung von Data Lakes in Unternehmen

Gemeinsame Nutzung von Daten
Wenn ein Team innerhalb eines Unternehmens einen neuen Datensatz erwirbt oder erstellt, kennt es den Wert der Daten und die zugehörigen Anforderungen wahrscheinlich ganz genau. Wenn sie z. B. geschäftlich vertrauliche Informationen, persönlich identifizierbare Informationen (PII) oder Kundendaten enthalten, weiß das Team, wie diese Informationen verwendet werden sollten und wie sie nicht verwendet werden sollten, und es werden Vorkehrungen getroffen, um sicherzustellen, dass niemand im Team sie missbraucht.

Sie werden sich auch bewusst sein, dass andere potenzielle Nutzer der Daten außerhalb ihres Teams möglicherweise nicht das gleiche Verständnis für den Wert der Daten oder die mit einem Missbrauch verbundenen Risiken haben. Aufgrund dieser Risiken ist das Team natürlich sehr vorsichtig im Hinblick auf die Freigabe der Daten und deren Speicherung an einem Ort, der nicht in seiner Kontrolle liegt.

Das hat negative Folgen für Data Lakes. Wenn die Mitarbeiter den Data Lake lediglich als unkontrollierten Ablageort für Daten betrachten, vertrauen sie ihm ihre wertvollen Daten nur ungern an. Dann können andere Teile des Unternehmens nicht von diesen Daten profitieren. So scheitert der ganze Plan, den Data Lake als Self-Service-Repository für die gemeinsame Nutzung von Unternehmensdaten zu verwenden.

Integration von Daten
Selbst wenn ein Team der Integration seiner Daten in den Data Lake zustimmt, kann dieser Prozess sehr mühselig sein. Im ursprünglichen Konzept des Data Lakes sollten Daten im Rohformat importiert werden – ohne die komplexen ETL-Prozesse (Extraktion, Transformation und Laden) eines traditionellen Data Warehouse. In der Realität erfordern aber fast alle Datenquellen ein gewisses Maß an Vorverarbeitung, bevor sie für eine aussagekräftige Analyse herangezogen werden können.

Daher kann die Integration einer neuen Datenquelle in einen Data Lake oft mehrere Monate dauern. Und da viele dieser Daten bisher eher in kleinen operativen Silos als in Unternehmenssystemen gehalten wurden, gibt es möglicherweise Dutzende oder sogar Hunderte von Quellen, die insgesamt zu integrieren sind.

Das bedeutet, dass die für Geschäftsanalysten oder Data Scientists erforderlichen Informationen in vielen Fällen noch nicht zum Data Lake hinzugefügt wurden und vielleicht erst in mehreren Monaten oder sogar Jahren hinzugefügt wird. Auch das stellt ein beträchtliches Hindernis für die Einsatzbereitschaft dar.

Speichern von Daten
Die Kosten für die im Handel erhältlichen Speicher- und Rechenressourcen sind zwar in den letzten Jahren erheblich gesunken, Hadoop-Cluster sind aber nach wie vor nicht kostenlos. Die Speicherung enormer Datenmengen in einem Data Lake ist weitaus kostengünstiger als ihre Speicherung in einer leistungsstarken Data Warehouse-Appliance. Die Kosten können aber dennoch beträchtlich sein.

Darüber hinaus ist das Wert-Mengen-Verhältnis der Big Data in einem Data Lake im Vergleich zu Daten, die traditionell in Data Warehouses gespeichert werden, relativ gering. Möglicherweise müssen Sie einen sehr großen Heuhaufen lagern, um die Handvoll hochwertiger Nadeln darin zu finden.

Wenn Sie nicht wissen, welche Datasets wirklich nützlich und wertvoll für Ihre Data Scientists sind, investieren Sie unter Umständen erhebliche Summen in die Integration und Speicherung von Daten, die im Data Lake untergehen und nie genutzt werden.

Auffinden von Daten
Angenommen, Sie haben die wertvollsten zu speichernden Datensätze identifiziert, Ihre Stakeholder davon überzeugt, sie gemeinsam zu nutzen, und es ist Ihnen gelungen, sie

Herausforderungen bei der Verwendung von Data Lakes in Unternehmen

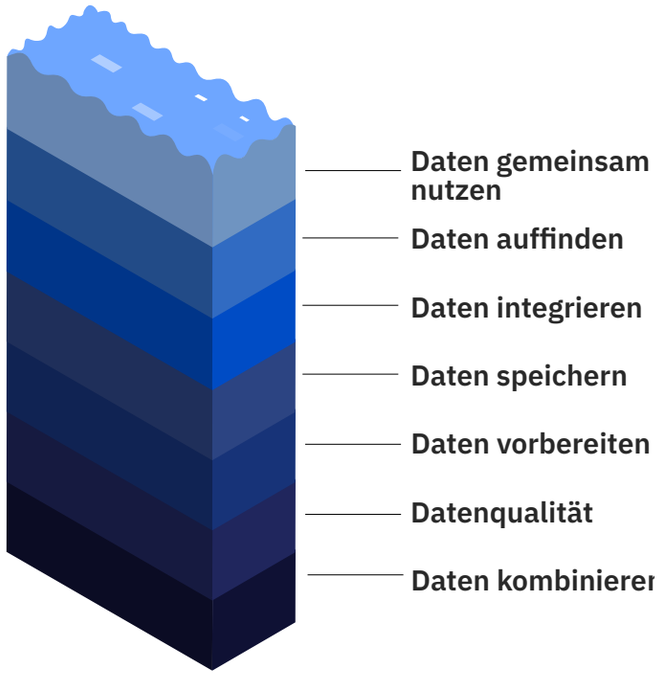


Abbildung 1: Unternehmen, die Data Lake-Technologien einsetzen, können auf eines oder mehrere dieser gemeinsamen Probleme stoßen.

in Ihren Datenbestand zu integrieren, dann müssen Sie es anderen Nutzern ermöglichen, sie zu finden, zu verstehen und richtig zu nutzen. Die Qualität der Daten im Data Lake stellt eine weitere Herausforderung dar. Sie sind nicht sicher, ob die Daten eine hohe oder niedrige Qualität haben, aber sie werden dennoch dem Data Lake zugeführt.

Bei den meisten Data Lakes ist das aber nur schwer zu erreichen. Daten werden oft ohne jeglichen Kontext gespeichert, was es für einen neuen Benutzer schwierig oder unmöglich macht, sie ohne Rücksprache mit dem ursprünglichen Eigentümer zu entschlüsseln. Die Terminologie ist oft so domänenspezifisch, dass eine Metrik, die in einem Bereich des Unternehmens verwendet wird, unter einem völlig anderen Namen bekannt sein kann - oder auf eine subtil andere Art und Weise - von einem anderen definiert wird. Das Potenzial für Verwirrung und Fehlinterpretationen kann so groß sein, dass viele Datensätze für einen Analysten, der nicht bereits mit ihnen vertraut ist, effektiv wertlos oder sogar gefährlich sind.

Kombinieren interner und externer Daten
Schließlich sollte selbst der größte Data Lake nicht dazu genutzt werden, jedes Dataset zu speichern, das potenziell von den Data Scientists eines Unternehmens verwendet werden könnte. Es ist beispielsweise nicht sinnvoll, ein vollständiges Replikat von Google Maps, Weather.com® oder Bloomberg in Ihren Data Lake zu importieren, nur weil einer Ihre Data Scientists raumbezogene Analysen durchführen bzw. Wetterdaten oder Aktienkurse in einen Algorithmus integrieren möchte.

Da der Data Lake nicht alle Daten enthält, die Ihre Geschäftsanalysten für die Analyse benötigen, müssen sie Zeit für die Suche nach diesen Daten in verschiedenen Anwendungen aufwenden. Da ein sehr großer Teil der nützlichen Analysen wahrscheinlich die Kombination von

internen und externen Datensätzen beinhaltet, erhöht dies erneut die Zugangsbarriere und verringert aus der Sicht des Benutzers den wahrgenommenen Wert des Data Lakes.

Vorbereiten von Daten
Die Datenvorbereitung ist aus vielen Gründen eine Herausforderung – vom Wissen, wo die Daten zu finden sind, bis zu ihrer Formatierung. Die Vorbereitung von Daten für die Analyse ist die ineffizienteste und zeitaufwendigste Aufgabe für die Datennutzer. Datennutzer verbringen den Großteil ihrer Zeit damit, Informationen zu finden, zu bereinigen und zu formatieren. Daher haben sie weniger Zeit für die Datenanalyse und -modellierung und das Ableiten von Erkenntnissen, die geschäftliche Vorteile bringen sollen.

Der eingeschränkte Zugang zu den verwalteten Datensätzen hat auch zu einer übermäßigen Abhängigkeit von der IT während der Vorbereitungsphase geführt. Dieser eingeschränkte Zugang verdeutlicht den Bedarf an besseren Self-Service-Funktionen und Datenbearbeitungsfähigkeiten im Unternehmen, um dieses Hindernis aus dem Weg zu räumen.

Datenqualität
Durch die Ablage von Daten in einem Data Lake können diese Daten unbrauchbar werden. Da keine Regeln zur Datenqualität oder -validierung auf Daten angewendet werden, bevor diese einem Data Lake zugeführt werden, ist die Vertrauenswürdigkeit und Brauchbarkeit der Daten nicht gewährleistet. Hohe Datenqualität ist ein wesentliches Merkmal für die Vertrauenswürdigkeit von Daten bei der Entscheidungsfindung. Daten stellen eine wertvolle Ressource dar, die während der gesamten Nutzung in einem Unternehmen verwaltet werden muss. Da Informationsquellen stets zunehmen und immer vielfältiger werden, während Initiativen zur Einhaltung behördlicher Vorschriften zielgerichteter werden, ist es unerlässlich, Informationen aus diesen unterschiedlichen Quellen auf konsistente, vertrauenswürdige und wiederverwendbare Arten zu integrieren und aufzurufen.

Ein ganzheitlicher Ansatz für den Aufbau kontrollierter Data Lakes

Die meisten Data Lakes nutzen Apache Hadoop und das zugehörige umfangreiche Ökosystem aus Open-Source-Projekten für ihre Datenspeicherschichten und Analyse-Engines. Es ist dabei keine Überraschung, dass die Open-Source-Community um Hadoop die Probleme im Zusammenhang mit aktuellen Data Lake-Implementierungen erkannt hat. Dementsprechend wurden in letzter Zeit viele Projekte eingeleitet, um die verschiedenen Probleme individuell zu beseitigen. Gleichermaßen sind einige proprietäre Tools auf dem Markt erhältlich, die dieselben Probleme lösen sollen.

So könnten Sie sich also dazu verleiten lassen, die Probleme mit Ihrem Data Lake nacheinander einzeln zu lösen. Wenn die Anzahl der Datasets für die Verwaltung zu hoch wird, fügen Sie ein Katalogisierungstool hinzu. Wenn Benutzer sich beschweren, dass sie die erforderlichen Daten nicht finden, stellen Sie ein Frontend mit einer Suchfunktion voran. Wenn die Data Stewards nicht mehr verfolgen können, woher Daten kommen und wer sie verwendet, stellen Sie Datenherkunftstools und ein Data-Governance-Framework bereit.

Das hört sich einfach an, in der Praxis ist dieser stückchenweise Ansatz aber meist mit wesentlich höherer Komplexität und verringerter Verwaltbarkeit verbunden, besonders wenn der Umfang des Data Lakes zunimmt. So wie das Hinzufügen neuer Datenquellen zu einem Data Lake Ihre ETL-Anforderungen komplexer macht, vergrößert das Hinzufügen neuer Tools auch die Komplexität der nicht funktionsbezogenen Anforderungen des Data Lakes.

Anstatt dass Sie Daten mit einer integrierten End-to-End-Plattform integrieren, Qualitätsvorgänge daran ausführen und diese für die effektive Verwendung durch Geschäftsanalysen katalogisieren, werden Sie bei individuellen Tools in der Regel feststellen, dass jedes Tool Fehler auf eigene Weise verwaltet und eigene Protokollierungsansätze nutzt. Daher können Fehlerbehebung und Problemlösung extrem viel Zeit in Anspruch nehmen.

Ein weiterer, noch bedeutender Nachteil des Nach-und-Nach-Ansatzes zeigt sich, wenn Sie die häufigen Data Lake-Probleme aus einer weniger technischen, sondern eher konzeptbezogenen Sichtweise betrachten. Sie müssen sich bewusst sein, dass Skalierbarkeit, Auffindbarkeit, Integration, Datenqualität und Governance keine separaten Probleme sind: Sie sind untrennbar miteinander verknüpft. Um sie zu lösen, ist ein ganzheitlicher Ansatz erforderlich.

> Skalierbarkeit, Auffindbarkeit, Integration, Datenqualität und Governance sind keine separaten Probleme: Sie sind untrennbar miteinander verknüpft. Um sie zu lösen, ist ein ganzheitlicher Datenmanagementansatz erforderlich.

IBM Watson Knowledge Catalog Daten-Discovery, Datenkatalogisierung und Datenqualität

Der IBM Watson Knowledge Catalog auf Basis von IBM Cloud Pak for Data hilft Datennutzern, Datenbestände, Datensätze, analytische Modelle und deren Beziehungen zu anderen Mitgliedern der Organisation schnell zu entdecken, zu kuratieren, zu kategorisieren und gemeinsam zu nutzen. Data Governance-Teams können damit Geschäftsglossare, Richtlinien und Regeln definieren und erhalten erweiterte Workflows für die Data Governance. Der Katalog dient als zentrale Wissensressource für Datenentwickler, Data Stewards, Data Scientists und Geschäftsanalysten, um Self-Service-Zugriff auf Daten zu erhalten, denen sie vertrauen können.

Lösungen wie IBM Watson Knowledge Catalog gestützt auf IBM Cloud Pak for Data können alle erforderlichen Funktionen bereitstellen, um die wesentlichen Probleme der modernen Data Lakes in einer einzelnen, umfassenden Plattform zu lösen. Der Katalog trägt dazu bei, die Grundursache dieser miteinander verbundenen Probleme anzugehen: das weit verbreitete Versagen der Data Lakes, wirksame Werkzeuge zur Erfassung, Speicherung und Verwaltung von Metadaten und zur Verfolgung der Datenabstammung bereitzustellen.

In vielerlei Hinsicht hängt der Nutzen eines Data Lakes von den darin enthaltenen Metadaten genauso ab wie von den Daten selbst. Ohne Metadaten, die Informationen zur Herkunft, zum Ersteller, zum Inhalt und zu den berechtigten Benutzern eines Datasets sowie zu dessen Nutzung angeben, werden die Daten selbst nahezu wertlos. Selbst wenn Benutzer sie finden können, wissen sie nicht, was sie bedeuten oder wie sie sie verwenden. Zudem vertrauen sie diesen Daten einfach nicht.

Watson Knowledge Catalog
Vertrauenswürdige und aussagekräftige Daten

Daten organisieren

Wissen

Daten müssen vollständig, anwendbar und überall zugänglich sein. Erkennen, klassifizieren und verstehen Sie alle Datentypen.

Daten kontrollieren

Vertrauen

Daten müssen sicher, bereinigt und einfach auffindbar sein, um vertrauenswürdigen Self-Service-Zugriff zu unterstützen. Bestimmen Sie die Herkunft und Qualität der Daten.

Daten demokratisieren

Nutzen

Fähigkeit zu Self-Service-Discovery und automatisierter Entscheidungsfindung zum Vorantreiben des Unternehmens. Stellen Sie allen erforderlichen Benutzern eine Ansicht aller benötigten Informationen sowie Zugriff darauf bereit.

Abbildung 2: IBM Watson Knowledge Catalog liefert eine Vielzahl von Funktionen für Daten-Discovery, Datenkatalogisierung und Data Governance.

Zentrale Wissensressource und Zugriffsstelle

IBM Watson Knowledge Catalog wird von IBM Cloud Pak for Data unterstützt und löst diese Probleme durch Priorisierung der Metadaten. Sein Kernstück ist eine leistungsstarke Katalogisierungs-Engine, die alle Datasets und Analyseressourcen, auf die Ihr Unternehmen zugreifen kann, indiziert – unabhängig vom Speicherort der Daten (ob in Ihrem Data Lake, Data Warehouse oder Transaktionssystem oder sogar einer Reihe von Kalkulationstabellen). Dabei spielt es keine Rolle, ob die Daten strukturiert oder unstrukturiert sind und ob sie lokal gespeichert oder in der Cloud gehostet werden. Darüber hinaus kann der Katalog auch externe Datasets und Datenquellen einbeziehen, wie proprietäre Datenservices, die Ihr Unternehmen abonniert, oder offene Daten-APIs.

Der Datenkatalog liefert nicht nur eine zentrale Wissensressource für all Ihre Datasets, sondern auch eine zentrale Zugriffsstelle. Dank KI-gestützter Such- und Vorschlagsfunktionen können Geschäftsanalysten, Data Scientists, Datenqualitätsingenieure und Data Governance-Teams Assets einfacher finden. Dabei geben die verfügbaren Metadaten den Benutzern Informationen dazu, was sie gefunden haben und ob es für sie nützlich ist.

Eingebettete Self-Service-Funktionen für die Datenvorbereitung verkürzen die Zeit für die Transformation von Daten für die produktive Verwendung in Analysen und KI-Anwendungen. Geschäftsanalysten und Data Scientists müssen keine Zeit mit dem Vorbereiten und Analysieren der Daten verschwenden. Die Integration in eine unternehmensweite Datenvorbereitungslösung wie IBM® InfoSphere® Advanced Data Preparation hilft sicherzustellen, dass die verwalteten Datensätze, die über die Katalogoberfläche erstellt werden, denjenigen mit dem meisten Kontext zur Verfügung stehen, um Geschäftseinblicke und -aktionen für Geschäftsanwender zu fördern. Diese Integration unterstützt die Zusammenarbeit in der ganzen Daten-Pipeline.

Skalierbarkeit, Auffindbarkeit, Integration, Datenqualität und Governance sind keine separaten Probleme: Sie sind untrennbar miteinander verknüpft. Um sie zu lösen, ist ein ganzheitlicher Datenmanagementansatz erforderlich.

Dieser Katalog hilft zudem Data Stewards, die dem Chief Data Officer (CDO) unterstellt sind, indem Datasets getaggt und klassifiziert werden. Außerdem wird ihre Herkunft und Verwendung automatisch verfolgt, und mit dem integrierten Geschäftsglossar kann die Geschäftsterminologie für alle Daten standardisiert werden. So können Data Stewards leichter nachvollziehen, was jedes Dataset enthält, wo die sensiblen oder personenbezogenen Daten sind und wer zum Zugriff darauf berechtigt sein sollte.

Ein einzelner Katalog für mehrere Datenquellen innerhalb und außerhalb des Unternehmens

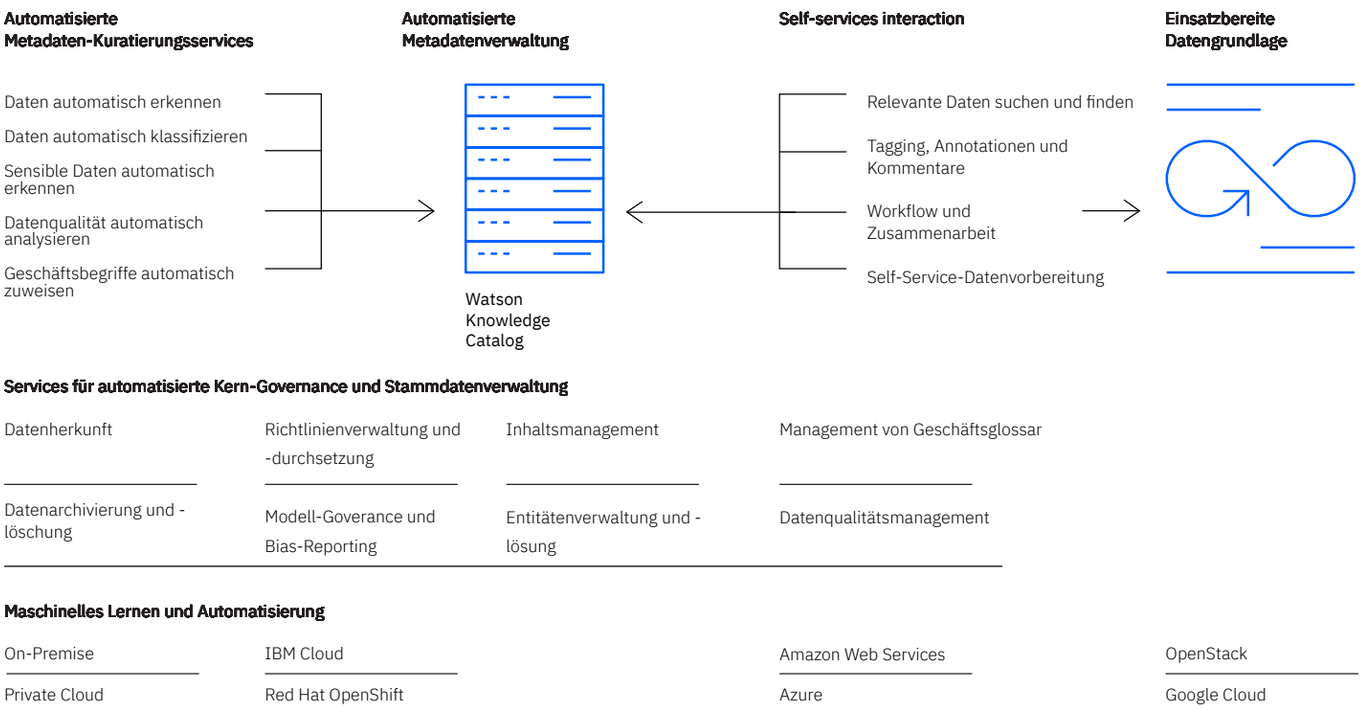


Abbildung 3: Mit dem intelligenten Metadatenindex von IBM Watson Knowledge Catalog können sowohl strukturierte als auch unstrukturierte Daten auch in Originalsystemen schnell von Benutzern gefunden und für intelligente Analysen herangezogen werden.

IBM Watson Knowledge Catalog macht Metadaten zur Priorität, sodass Sie eine zentrale Wissensressource und Zugriffsstelle für alle Datasets erhalten, auf die Ihr Unternehmen zugreifen kann.

Integrierte intelligente Daten-Discovery

Um die Auffindbarkeit von Daten noch zu verbessern, ermöglicht der Katalog den Benutzern das Taggen und Kommentieren von Datasets und Analyseressourcen. Dank dieser angereicherten Metadaten und des zusätzlichen Kontexts können Mitarbeiter gesuchte Elemente einfacher finden. Die Lösung umfasst zudem integrierte Daten-Discovery-Algorithmen, die den Inhalt der einzelnen Datasets mit ML automatisch klassifizieren. Da die Lösung häufig verwendete Feldtypen wie Namen, Adressen, Postleitzahlen und Sozialversicherungsnummern identifiziert, müssen Autoren die Daten nicht mehr manuell ergänzen. Durch die Integration von Automatisierung und ML werden das Kuratieren der Daten und das Metadatenmanagement automatisiert. Mit ihren integrierten Datenqualitätsfunktionen ermöglicht die Lösung umfangreiche Regeln für Daten-Profiling, Datenqualität und Validierung.

Automatisierte Datenvorgänge bieten eine kuratierte Daten-Pipeline mit Datenqualität und Data Governance. So sorgen Sie für einen kontinuierlichen Fluss hochwertiger kontrollierter Daten in den Data Lake.

Gleichermaßen stellt ein intelligentes Metadatenmodell Ihrer Assets eine einzigartige Möglichkeit dar, die Einhaltung von Gesetzen wie der Datenschutz-Grundverordnung (DSGVO) und dem California Consumer Privacy Act (CCPA) durchzusetzen.

IBM Watson Knowledge Catalog wird von IBM Cloud Pak for Data unterstützt und liefert vertrauenswürdige, hochwertige, einsatzbereite Daten für praktisch alle Datennutzer.

Alle Komponenten der Lösung wurden als Microservices konzipiert, mit einem einzelnen Satz aus Designprinzipien und einem gemeinsamen Ansatz für nicht funktionsbezogene Anforderungen, wie Skalierbarkeit, Fehlermanagement, Sicherheit und Protokollierung.

IBM Watson Knowledge Catalog bietet eine ML-Plattform für die Unternehmens-Governance – und ermöglicht so skalierbare KI.

Während ein stückweiser Do-it-yourself-Ansatz mit großer Wahrscheinlichkeit zu verwirrenden Fehlern und Leistungsengpässen führt, bietet IBM Watson Knowledge Catalog eine ML-Plattform für die Unternehmens-Governance – und ermöglicht so skalierbare KI.

IBM Watson Knowledge Catalog ist in drei Varianten erhältlich:

- Als Software-as-a-Service-Lösung (SaaS) in der IBM Cloud™
- In IBM Cloud Pak for Data
- Als Integration in IBM Watson Studio

Lösungen wie der IBM Watson Knowledge Catalog können den Nutzen realisieren, der mit Data Lake-Initiativen ursprünglich angestrebt wurde. Watson Knowledge Catalog mit intelligenten Katalogisierungs- und Governance-Funktionen trägt zum Aufbau eines vertrauenswürdigen und kontrollierten Data Lakes für KI bei.

Vier Vorteile eines kontrollierten Data Lakes für KI

1. Vertrauen in Daten durch Qualitäts- und Governance-Funktionen aufbauen

- Dank Datenqualitätsfunktionen können Sie die Qualität Ihrer Daten verbessern und hochwertige Daten in Ihrem Data Lake zur Verfügung stellen.
- Governance-Richtlinien werden automatisch festgelegt und durchgesetzt. Wenn Sie ein Dataset finden, wissen Sie also, ob und wie Sie es verwenden dürfen.
- Sie können Ihre Daten kuratieren, indem Benutzer Bewertungen, Kommentare und andere Informationen hinzufügen, anhand derer andere ermitteln können, ob ein Dataset für sie von Nutzen ist oder nicht.

2. Datennutzer unterstützen

- Ihre Geschäftsbereichsteams geben ihre Daten gerne weiter, weil sie darauf vertrauen, dass sie ordnungsgemäß verwaltet und vor Missbrauch geschützt werden.
- Sie können die Zusammenarbeit fördern und Daten durch dynamische Datenrichtlinien und deren Durchsetzung in vertrauenswürdige Unternehmenswerte verwandeln.
- Ihre Daten werden mit der Zeit immer leichter auffindbar und wiederverwendbar, da Benutzer relevante Tags und Metadaten hinzufügen, anhand derer andere Nutzen aus ihnen ziehen können.
- Sie erhalten über eine zentrale Oberfläche Zugriff auf alle Datasets Ihres Unternehmens, unabhängig von deren Speicherort.

3. Zeit sparen

- Die automatische Daten-Discovery reduziert den Zeit- und Arbeitsaufwand für das Hinzufügen von Metadaten für neue Datasets.
- Dank automatischer Datenkuratierung und Metadatenverwaltung können Sie Metadaten schneller erkennen und Begriffe schneller zuweisen. Außerdem wird die Zeit für die Erstellung des Geschäftsglossars damit verkürzt.
- Mit einfachen und intuitiven Self-Service-Tools für die Datenvorbereitung verbringen Ihre Datenbenutzer weniger Zeit mit der Datenvorbereitung und mehr Zeit mit der Entdeckung von Erkenntnissen.
- So können Ihre Data Scientists und Geschäftsanalysten bessere Analysen in kürzerer Zeit bereitstellen.
- Dank der intelligenten KI-gestützten Suche finden Sie die benötigten Daten in Sekundenschnelle und müssen nicht mehr wochenlang warten, bis ein anderes Team diese weitergibt.

4. Zunehmende Daten und Kosten verwalten

- Sie können die Speicherkosten optimieren, indem Sie die Kosten für die Aufnahme von geringwertigen Datensätzen in den Data Lake vermeiden.
- Außerdem können Sie alle externen Datensätze einsehen, die Ihr Unternehmen abonniert, und bezahlen somit nicht für mehr Abonnements als nötig.
- Sie können die Aufnahme neuer Datenquellen in den Data Lake nach der Nachfrage der Benutzer nach den Daten priorisieren und so die wertvollsten Quellen zuerst integrieren.

Wert aus Ihren Daten schöpfen

Ob Sie im Team des Chief Digital Officer, in der IT-Abteilung oder als Data Scientist oder Analyst für einen Geschäftsbereich arbeiten – Sie und Ihre Kollegen verfolgen ein gemeinsames Ziel. Wenn Sie einen Data Lake aufbauen können, der wirklich hält, was er verspricht, könnten Sie nicht nur Ihre eigene Arbeit viel einfacher und produktiver machen. Darüber hinaus können Sie wesentlich zu einem Wettbewerbsvorteil Ihres Unternehmens beitragen, den derzeit nur wenige Organisationen erreichen.

Wenn Sie mit einem sauberen Data Lake arbeiten, während die Konkurrenz noch im Sumpf steckt, eröffnen Sie sich Möglichkeiten, von denen die anderen nur träumen können. Diejenigen, die den Wert von zuvor nicht genutzten Daten als erstes ausschöpfen, erreichen einen echten First Mover Advantage.

Fazit

Sie möchten wissen, wo Ihre Daten aufbewahrt werden, wer sie verwendet und welchen Nutzen sie für Analysen für Ihr Unternehmen bringen.

Datenkataloge sind entscheidend für DataOps-Initiativen, da sie das automatisierte Metadatenmanagement durch Integration von Data Governance, Datenqualität und aktiver Richtlinienverwaltung unterstützen.

IBM Watson Knowledge Catalog mit intelligenten Katalogisierungs- und Governance-Funktionen trägt zum Aufbau eines vertrauenswürdigen und kontrollierten Data Lakes für KI bei. Der Katalog bettet Datenintegration, Datenqualität und Data Governance in Ihre Data Lake-Umgebung ein, damit Sie einsatzbereite Daten für DataOps und eine zentrale Wissensressource erhalten.

Weitere Informationen

Weitere Informationen erhalten Sie unter:
ibm.com/cloud/watson-knowledge-catalog

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
Germany
ibm.com/de

IBM Österreich
Obere Donaustrasse 95
1020 Wien
ibm.com/at

IBM Schweiz
Vulkanstrasse 106
8010 Zürich
ibm.com/ch

1. Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders–Gartner, Sept 2019
2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018

ASW12449-DEDE-03