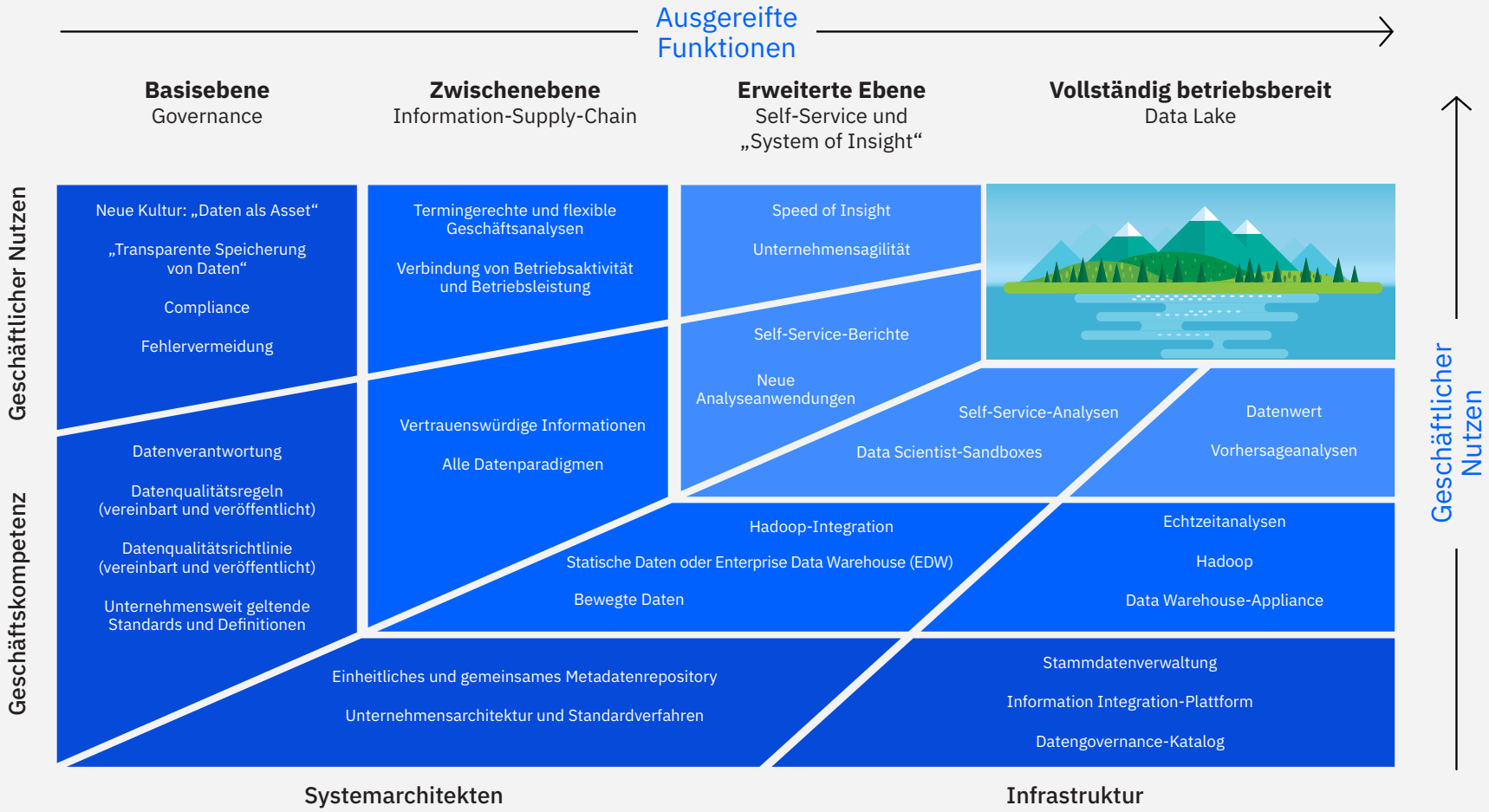
Wie Unternehmen mit dem kontrollierten Data Lake Erkenntnisse gewinnen

Die wichtigsten Bausteine für den effizienten Zugang zu vertrauenswürdigen Daten

IBM **Cloud**

Der Mehrwert kontrollierter Data Lakes

Data Lakes sind ideal für Unternehmen, in deren Strategie Daten oberste Priorität haben. Aber auch Sicherheit ist ein wichtiger Aspekt, wenn mehrere Teams Unternehmensdaten gemeinsam nutzen. Die Lösung für diese Herausforderungen ist ein kontrollierter Data Lake, der strukturierte wie unstrukturierte Daten in vertrauenswürdiger, sicherer und kontrollierter Weise verfügbar macht. Unternehmen, die Nutzen aus Daten über Kunden, Mitarbeiter, Transaktionen und andere Assets ziehen, profitieren von kontrollierten Data Lakes, da sie Informationen schnell identifizieren, analysieren, gemeinsam nutzen und erfolgreich umsetzen können.



Die Architektur eines kontrollierten Data Lakes

Die Architektur eines kontrollierten Data Lakes umfasst wichtige Designentscheidungen wie etwa die Speicherung von Daten in drei zentralen Bereichen. Data Lake-Repositorys bieten eine Plattform, um Daten zu speichern und Analysen so nah wie möglich am Speicherort der Daten auszuführen. Hinzu kommen Data Lake-Services, die das Suchen, Abrufen, Aufbereiten, Transformieren, Verarbeiten und Verschieben der Daten in und aus den Datenspeicher-Repositorys ermöglichen. Aber auch das Informationsmanagement und die Governance-Fabric tragen dazu bei, Daten im Data Lake zu kontrollieren und zu verwalten.

Zur Prüfung und Verbesserung der Datenqualität werden Governance-Funktionen eingesetzt, die Daten zudem vor Missbrauch schützen. Dadurch wird sichergestellt, dass Daten gemäß der Phase ihres Lebenszyklus ordnungsgemäß aktualisiert, aufbewahrt und schließlich gelöscht werden.

Governance, Organisation und Datenqualität spielen eine wichtige Rolle bei der Verwaltung eines Data Lakes. Während der Data Lake flexiblen Zugang zu Daten bietet, benötigen Sie ein Kontrollsystem, das umfassende Sicherheit, Schutz und kontinuierlichen Nutzen sicherstellt. Ein kontrollierter Data Lake setzt sich aus drei Ebenen zusammen:

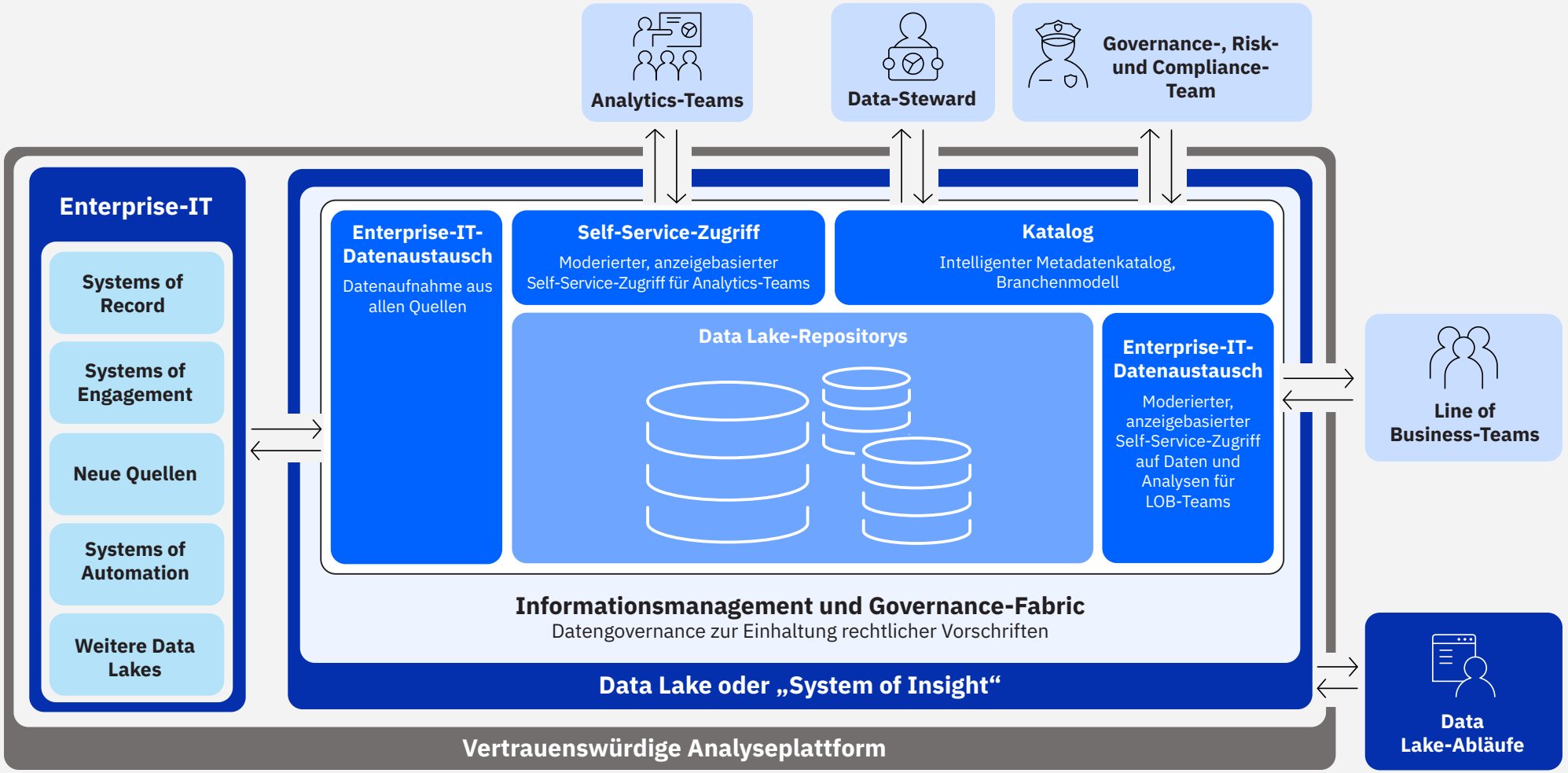
- Basisebene, die hauptsächlich auf Datengovernance aufbaut
- Zwischenebene, die die anfänglichen Data Lake-Repositorys um neue und zusätzliche Datentypen und Verhaltensweisen erweitert
- Erweiterte Ebene, die Self-Service-Analysen unterstützt

Je nach Arbeitsgruppe im Unternehmen haben diese Ebenen einen bestimmten Nutzen. Systemarchitekten profitieren von einer veröffentlichten Referenzarchitektur, die von einem einheitlichen und gemeinsamen Metadatenrepository unterstützt wird. Und Data Scientists verfügen über einen kontrollierten Bereich, in dem sie aktuelle Sandboxes einrichten können.

Die grundlegenden Vorteile eines Data Lakes liegen in der Governance. Die Governance fördert eine datenorientierte Kultur, in der Daten im Besitz der Fachanwender sind, die sich über Regeln und Richtlinien abstimmen. Dieser gemeinsame Prozess schafft ein gegenseitiges Verständnis und hilft, Unklarheiten zwischen Teams zu vermeiden. Diese gemeinsame Grundlage eröffnet aber auch Zugang zu vertrauenswürdigen Daten und liefert schnellere Erkenntnisse aus Analyseanwendungen. Der geschäftliche Nutzen verlagert sich vom Bewusstsein für Daten über deren Relevanz hin zu ständig verfügbaren flexiblen Analysen.

Ein modularer, skalierbarer Data Lake umfasst mehrere Elemente, die den Self-Service-Zugriff im ganzen Unternehmen fördern.

Die Architektur eines kontrollierten Data Lakes

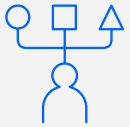


Die vier Typen von Datenkonsumenten

Die Konsumenten der Daten aus dem Data Lake unterscheiden sich erheblich. Die Unterschiede im Umgang mit Daten zu verstehen, ist ein wesentlicher Aspekt erfolgreicher Governance.

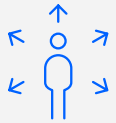
Analytics-Teams

- Data Scientists, die Daten verwalten und Modelle entwickeln
- Entwickler von Analyseanwendungen, die Modelle in Anwendungen umsetzen
- Anwendungsentwickler, die Analyseanwendungen in betriebliche Systeme einbinden



Data-Steward

- Optimiert die Datenqualität und bereitet ETL-Prozesse vor
- Katalogisiert Daten und verwaltet Metadaten
- Bringt Datenschutz und Privatsphäre in Einklang



Governance-, Risk- und Compliance-Team

- Datengovernance-Spezialisten, die die Datengovernance- und Sicherheitsrichtlinien entwickeln
- Setzt Datenschutzkontrollen in allen Prozessen durch
- Legt Anforderungen zur Aufbewahrung, Archivierung und Löschung von Daten fest und stellt die Einhaltung von Richtlinien und Verordnungen sicher

Line-of-Business Teams

- Leiter einzelner Geschäftsbereiche wie CMOs, CFOs oder CHROs
- Chief Data Officers (CDOs), die eine aufstrebende Rolle als Datenverantwortliche haben
- LOB-Führungskräfte, die Systeme zur Erreichung bestimmter Geschäftsergebnisse oder relevanter Erkenntnisse implementieren

Die Bausteine eines kontrollierten Data Lakes

Ein kontrollierter Data Lake ist eine Referenzarchitektur, die unabhängig von einer bestimmten Technologie ist und mit Governance und Informationsmanagement kombiniert wird. Der Data Lake ist kein käufliches Produkt wie Hadoop oder Enterprise Data Warehouse (EDW) und kann diese auch nicht ersetzen. Ein kontrollierter Data Lake ist eine lokale oder cloudbasierte Lösung für Unternehmen, die Daten in den Mittelpunkt ihrer Betriebsabläufe rücken. Die Bausteine eines kontrollierten Data Lakes umfassen:

Ein unternehmensweiter IT-Daten-austausch umfasst das Extrahieren, Analysieren, Aufbereiten, Transformieren und den Austausch von Daten zwischen Data Lakes und Enterprise-IT-Systemen sowie die Integration isolierter Daten in den Data Lake. Weitere Merkmale sind die fortlaufende Bereinigung und Qualitätssicherung der Daten.

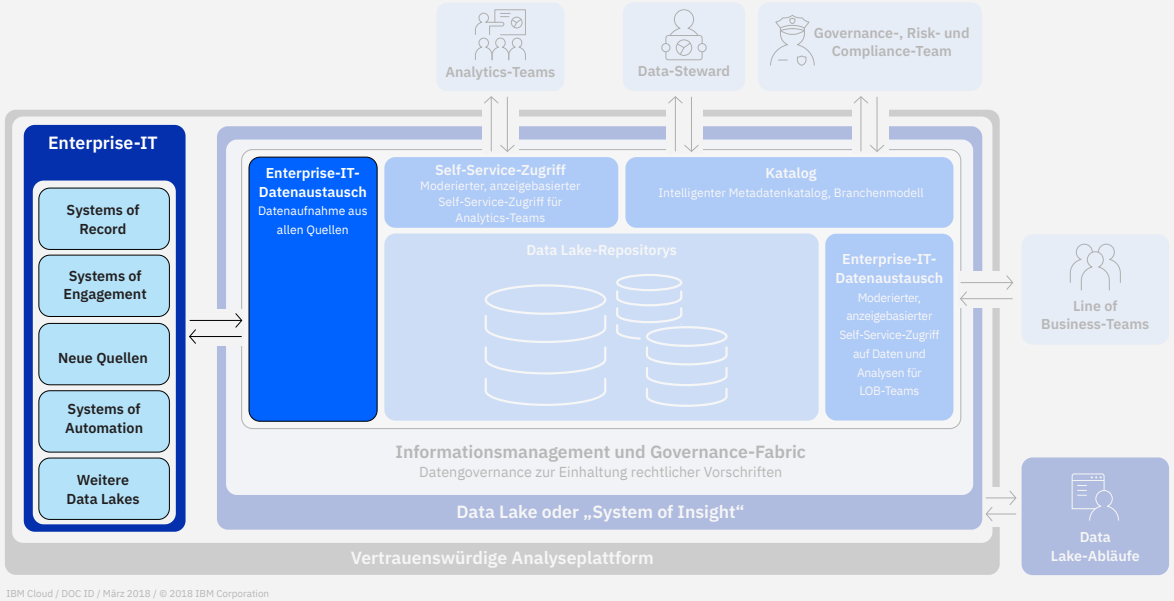
Katalog -Services beschreiben die Daten im Data Lake, z. B. was sie bedeuten, wie sie klassifiziert und wie sie in einem geeigneten Governance-System kontrolliert werden.

Governanceunterstützt die Datenkontrolle im Data Lake und wendet geeignete Richtlinien, Sicherheits-, Qualitäts- und Schutzmaßnahmen auf die Daten im Data Lake an.

Self-Service-Zugriff umfasst den bedarfs-gesteuerten Zugriff auf den Data Lake in drei Servicebereichen: Nutzer von Analytics erhalten Zugriff auf gespeicherte Rohdaten, LOB-Teams verfügen über normalisierte Daten in vereinfachten Datenstrukturen, und Governance-, Risk- und Compliance-Teams nutzen kontrollierte Daten für ihre Audits.

Datenaufnahme aus verschiedenen Quellen

Aufnahme beschreibt die Extraktion, Transformation, Qualitätssicherung und den Austausch von Daten zwischen dem Data Lake, unternehmensweiten IT-Systemen und anderen vorhandenen Data Lakes. Ein Großteil der Daten im Data Lake stammt aus unternehmenseigenen IT-Systemen. Diese Datentypen können strukturiert, teilstrukturiert oder unstrukturiert sein. Zu den Datenquellen zählen Systeme, die für Geschäftsprozesse, Websiteprotokollierung oder die Überwachung von Aktivitäten zuständig sind. IBM bietet neben der Skalierbarkeit von Datenvolumen auch die vielseitige Transformation und Replikation von Daten.



Sie machen alles richtig, wenn...
- Daten ohne Unterbrechung in den Data Lake fließen
- Analysedaten transformiert, standardisiert und aufbereitet sind
- Speicherkosten auch bei steigendem Datenvolumen sinken
- Explorative Analysen in Sandboxes ausgeführt werden

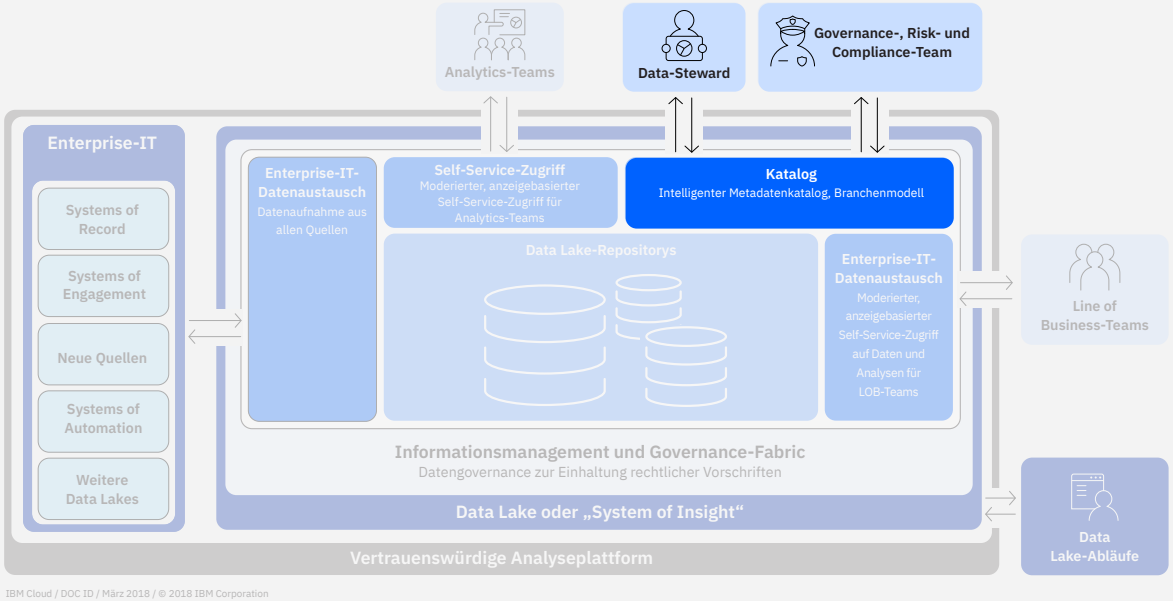
Es gibt Verbesserungspotenzial, wenn...
- Die Aktualität von Daten bei steigendem Volumen schwankt
- Unstrukturierte Informationsassets nicht verwertbar sind
- Speicherkosten zu hoch sind
- Die Datenbereinigung zu kompliziert ist, was höhere Kosten bei der Datenverarbeitung verursacht

Katalogisierung

Die Katalogisierung dient zur Kennzeichnung der Daten im Data Lake, damit Sie den Überblick über Ihre Informationsassets behalten. Über Katalogschnittstellen erhalten Data Lake-Benutzer Informationen über die Klassifizierung, Herkunft und Governance der Daten. IBM bietet die folgenden Katalogisierungsfunktionen für Data Lakes:

- Erfassung unstrukturierter Informationsassets im Katalog
- Integration offener Ökosysteme in praktisch jedes Informationsasset
  - Ein einheitlicher Unternehmenskatalog für praktisch alle Informationsassets
  - Unterstützung branchenspezifischer Daten und Geschäftsbegriffe
  - Bewertung und Social Tagging bei der Verarbeitung von Metadaten

Daten aus der Governance-Pipeline müssen verständlich sein, damit Geschäftsanwender technische Daten sinnvoll nutzen können. Beispielsweise kann eine neunstellige Ziffernfolge in den USA eine Sozialversicherungsnummer, eine Mitarbeiter-ID oder beides sein. Durch die Klassifizierung und Zuordnung von Geschäftsbegriffen erhalten Unternehmen aussagekräftige technische Daten. Die Automatisierung ist ein wichtiges Merkmal, damit der Prozess an das Volumen und die Vielfalt der Daten im Data Lake angepasst werden kann. Kuratierungsworkflows, Qualitätsprüfung und Datenkontrollen stellen anschließend sicher, dass Daten bereit für die Katalogisierung sind. Auf diese Weise werden Daten unternehmensweit verfügbar gemacht.



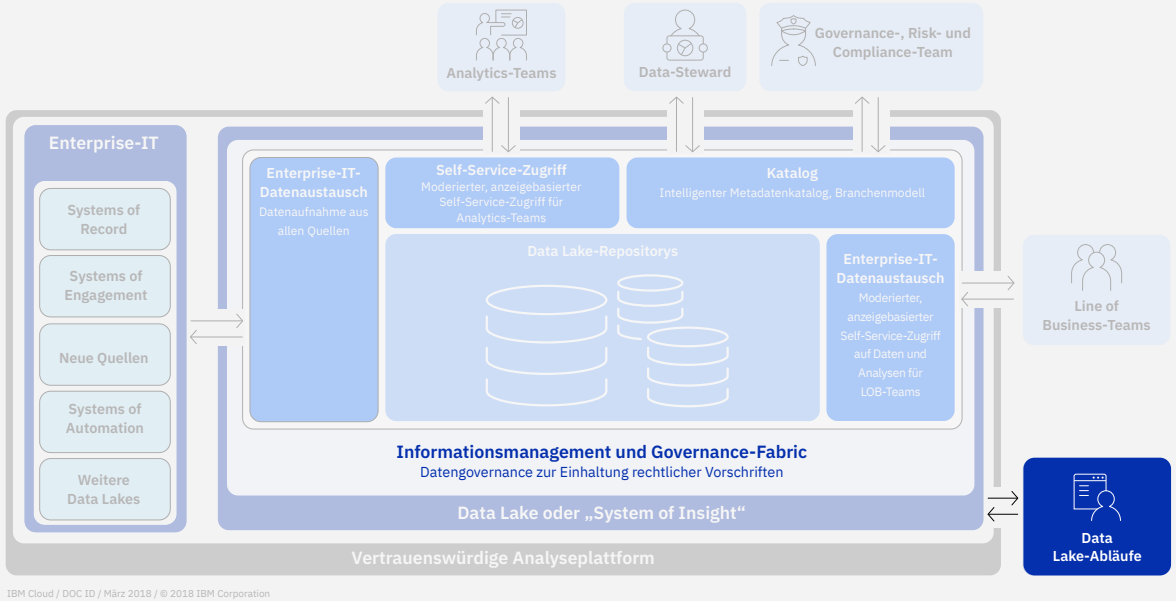
Sie machen alles richtig, wenn...
- Ergebnisse schneller verfügbar sind und mehr Zeit für die Analyse bleibt
- Wissensbestände im Kontext relevanter Daten präsentiert werden
- Die Datenabstammung transparent und Daten vertrauenswürdiger sind
- Informationsassets mehr Datenkonsumenten zur Verfügung stehen
- Compliance-Vorschriften für Daten eingehalten werden

Es gibt Verbesserungspotenzial, wenn...
- Die Suche und Kennzeichnung von Daten zu lange dauert
- Wissen undokumentiert vorliegt und für Kollegen nicht verständlich ist
- Der Zugriff auf Daten nicht transparent ist
- Compliance- und Governance-Anforderungen nicht erfüllt werden

Datengovernance und Informationsmanagement

Informationsintegration und Governance-Fabric unterstützen das System bei der effizienten Überwachung des Data Lakes, damit eingehende Informationen erkannt und Kontrollrichtlinien automatisch angewendet werden. Das Governance-Framework erleichtert die Dokumentation von Governance-Richtlinien und setzt Regeln für die Strukturierung, Speicherung, Transformation und Verschiebung von Informationen durch.

Die Anforderungen der Datengovernance werden im Katalog als Richtlinien, Regeln und Klassifizierungen dokumentiert. Die wesentlichen Vorteile von IBM bestehen darin, dass der Data Lake unstrukturierte Assets enthalten kann – ohne Kompromisse beim Volumen, der Vielseitigkeit und der Geschwindigkeit der Datenverarbeitung.



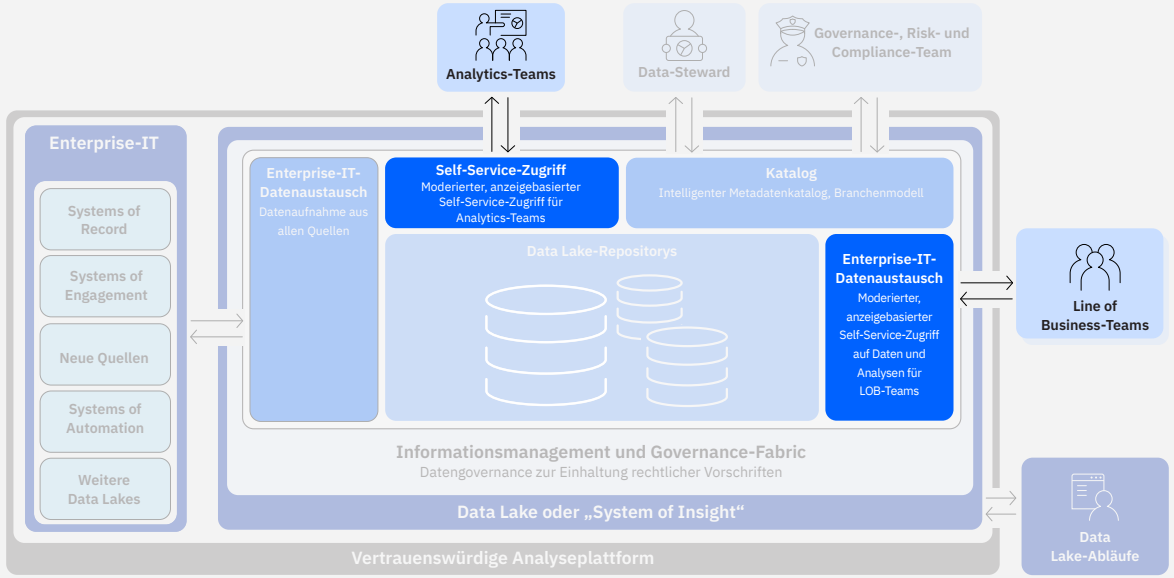
Sie machen alles richtig, wenn...
- Wachsende Datenvolumen kontrollierbar bleiben
- Gesetzliche Bestimmungen mithilfe branchenspezifischer Compliance-Tools eingehalten werden
- Stammdaten schnell eingepflegt werden
- Qualitätsdaten präzisere Erkenntnisse liefern
- Compliance-Prüfungen schnell erledigt sind
- Daten optimal geschützt sind

Es gibt Verbesserungspotenzial, wenn...
- Wachsende Datenvolumen aus strukturierten und unstrukturierten Quellen nicht verwaltbar sind
- Die Datensuche zu lange dauert, was die Prüfbarkeit beeinträchtigt
- Compliance- und Governance-Anforderungen nicht erfüllt sind

Self-Service oder Berichterstellung

Mit Self-Service-Zugriff finden Sie über eine einfache Benutzeroberfläche relevante Informationen. Auf diese Weise können Entwickler selbstbestimmt arbeiten und verfügen über vertrauenswürdige Qualitätsdaten, um Analysemodelle in ihren Data-Science-Initiativen erfolgreich umzusetzen. Aber auch nicht-technische Benutzer sind in der Lage, Daten zu transformieren, bevor sie Modelle umsetzen und bereitstellen.

Der direkte Datenzugriff erleichtert IT-Entwicklern die Aufbereitung und Transformation von Daten. Darüber hinaus können Governance- und Compliance-Teams Daten einfacher kuratieren, um deren Prüfbarkeit sicherzustellen. Aber auch die Konsumenten dieser Lösungen profitieren, indem sie auf ihre Geschäftsanforderungen abgestimmte Berichte erstellen. Außerdem haben sie Zugriff auf sofort einsatzbereite Daten und können schnellere Entscheidungen treffen und bessere Erkenntnisse aus ihren Daten ziehen.



Sie machen alles richtig, wenn...

- Datennutzer Zugriff auf Kontextdaten haben
- Datenkonsumenten durch gemeinsames Wissen, Social Tagging und die qualitative Bewertung von Informationsassets über vertrauenswürdige Daten verfügen
- Daten für alle Datenkonsumenten unternehmensweit verfügbar sind
- Daten eine schnellere Wertschöpfung erzielen
- Das Innovationstempo steigt
- Datenexploration und Analysen agil und iterativ durchgeführt werden

Es gibt Verbesserungspotenzial, wenn...

- Die Suche und Aufbereitung von Daten länger als die Analyse dauert
- Unstrukturierte Assets nicht durchsucht oder verwertet werden
- Entscheidungen mangels vertrauenswürdiger Daten verzögert werden
- Das Innovationstempo sinkt

Warum IBM?

Laut einer von Radiant Advisors durchgeführten Studie nennen 72 % der befragten Führungskräfte Governance und Sicherheit als größte Herausforderungen, aber gleichzeitig auch als wichtigste Erfolgsfaktoren für ihr Unternehmen. Der erste Schritt besteht also darin, die Governance- und Informationsarchitektur als oberste Priorität zu behandeln. Dies fördert den offenen Austausch im Unternehmen und ermöglicht Datenbenutzern, ihre Datenanforderungen klar zu definieren. In einer Welt, in der minderwertige Daten zu minderwertigen Ergebnissen führen, zählt die Meinung jedes einzelnen Datenbenutzers.

Die Bereitstellung einer unternehmensweiten Plattform, die Datenintegration, qualitative Datenverarbeitung und Datengovernance vereint, ist entscheidend für den Erfolg Ihrer Analytics-Initiativen. Mit einer solchen Plattform können Unternehmen Daten reibungslos erfassen, hohe Qualität sicherstellen und kontrollierte Datenfeeds in Analyseprozesse einbinden. Mit einem kontrollierten Data Lake meistern Sie die Herausforderungen und schaffen die Grundlage dafür, dass vertrauenswürdige Daten in vielen Unternehmensbereichen verfügbar sind.

Kunden profitieren von der einzigartigen Breite und Tiefe der IBM Unified Governance and Integration-Plattform. Die Vorteile der Plattform sind: hohe Skalierbarkeit für große Datenmengen, branchenspezifische Beschleuniger, Auswertung strukturierter, unstrukturierter und teilstrukturierter Daten sowie höchste Kompetenz im Bereich maschinelles Lernen und künstliche Intelligenz. Damit bietet IBM seinen Kunden eine umfassende Lösung für die Umsetzung eines vertrauenswürdigen und kontrollierten Data Lakes.

Weitere Informationen finden Sie unter **ibm.com**/governed-data-lake.

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

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

IBM Schweiz
Vulkanstrasse 106
8010 Zürich
ibm.com/ch

Die IBM Homepage finden Sie unter:
ibm.com

03018403-DEDE-01