tdwi
WISSEN

BIG DATA EFFIZIENT NUTZEN

E-Book

Hybrides Datenmanagement mit der IBM Cloud Private for Data

tdwi.eu

Herausgeber
SIGS DATACOM GmbH
Lindlaustraße 2c
53842 Troisdorf
info@sigs-datacom.de
www.sigs-datacom.de

TDWI E-Book in Kooperation mit IBM

Inhalt

Vorwort

Innovative Konzepte wie Smart City, Autonomes Fahren oder Industrie 4.0 sind auf Gedeih und Verderb sowohl auf die performante und intelligente Verarbeitung von Daten als auch auf den reibungslosen Datenaustausch zwischen allen involvierten Parteien angewiesen.

Viele Parteien – z. B. Kommunen und Behörden, IT-Hersteller, Maschinenlieferanten oder -betreiber und natürlich die Anwenderunternehmen – sind daher permanent auf der Suche nach weiteren Daten, um ihre eigenen Bestände sinnvoll zu ergänzen. Sie bieten ihre eigenen Daten auf entsprechenden Datenbörsen an oder vernetzen und integrieren sich mit dem „Internet of Things" (IoT) – entweder gegen gutes Geld oder im Austausch gegen andere Daten.

Denn Daten erheben, vernetzen und auf dieser Basis intelligente Lösungen entwickeln: Das ist der Kerngedanke von sowohl Industrie 4.0 als auch Smart City. Träger dieser Lösungen ist die Cloud. Doch „die Cloud" gibt es nicht. Vielmehr gibt es sehr viele sehr unterschiedliche Cloud-Services, die von diversen Anbietern stammen – von Global-Playern ebenso wie von absoluten Spezialisten oder von Maschinenlieferanten. Dazu kommen die bewährten Anwendungssysteme des Unternehmens, in Form maßgeschneiderter Webservices aus der Private Cloud.

Um in diesem äußerst heterogenen, schnelllebigen Umfeld für eine einerseits performante, andererseits auch sichere und zuverlässige Analyse und Verarbeitung all dieser unterschiedlichen Daten zu sorgen, hat IBM die „Cloud Private for Data" entwickelt. Sie bildet auf Basis bewährter IBM-Technologien kombiniert mit verbreiteten Open-Source-Modulen eine offene Anwendungsplattform für hybrides Datenmanagement – und bringt dabei Private-, Public- und Multi-Cloud-Umgebungen geschickt unter einen Hut. So lassen sich die Vorteile der Public Cloud nutzen, ohne die Datensicherheit und Compliance zu gefährden.

Quelle: Shutterstock

1 Einleitung

Wissen entsteht durch die Gewinnung, Verarbeitung, Verknüpfung und Sammlung von Informationen. Informationen sind Daten mit einer Bedeutung für uns, die wir über die Interpretation dieser Daten herstellen.

Weil im Zuge der Digitalisierung die Industriegesellschaft zu einer Wissens- oder zumindest zu einer Informationsgesellschaft transformiert wird, gewinnen die Daten immer mehr an Bedeutung. Nicht nur personenbezogene Daten, sondern auch ganz schlichte Massendaten, die z. B. den Verkehrsfluss beschreiben, das Wetter oder die Arbeit von Maschinen.

Weil Daten allerorten sind, weil wir alles und jeden durch Daten beschreiben können und weil wir über immer bessere und kostengünstigere Technologien verfügen, diese Daten überall und jederzeit lückenlos zu erfassen, zu speichern, intelligent zu verarbeiten und zu korrelieren, werden Daten gerne als das neue Öl bezeichnet – also als der Rohstoff schlechthin im „Internet of Things". Das IoT wiederum ist das Fundament, auf dem der Megatrend „Digitalisierung" aufsetzt.

Letztendlich trägt also die effiziente Beschaffung, Verarbeitung und Analyse der Daten maßgeblich dazu bei, dass die Digitalisierungsprojekte zum Erfolg geführt werden können. So gewinnt ein Unternehmen schneller und effizienter solche Informationen, die einen Wissensvorsprung vor der Konkurrenz bedeuten. Und damit einen klaren Vorteil im Wettbewerb.

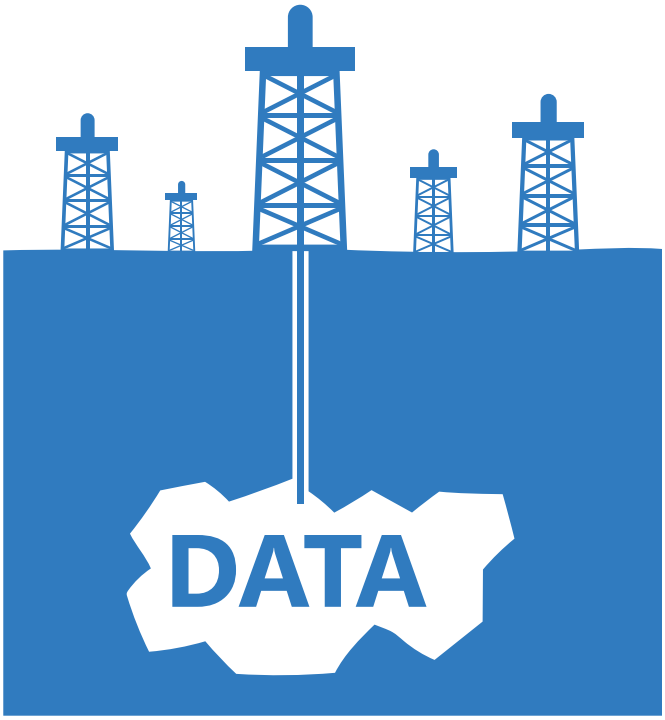
Quelle: Shutterstock

1 Einleitung

1.1 Massendaten – der Schmierstoff für das Zeitalter Industrie 4.0

Mit enormer Dynamik schreitet die Digitalisierung voran. Deutlich wird das an der Integration von IT in Produktklassiker wie das Auto oder die Waschmaschine. Aber auch in Gebäuden, in der Produktion oder in der Logistik geht schon heute nichts mehr ohne IT.

Gerade die Industrie treibt diesen Wandel voran – wird aber auch von ihm getrieben. Die digitale Transformation erfasst sämtliche Stufen der industriellen Wertschöpfung. Das beginnt in der Logistik und geht über die Produktion bis hin zur Dienstleistung. Der industrielle Kern Deutschlands und Europas durchläuft eine Phase grundlegender Veränderung.



Quelle: Shutterstock

Daten, ein Motor der Digitalisierung

Ein Motor dieser Veränderung sind die Daten, die laufend erfasst und immer häufiger auf Vorrat gespeichert werden. Sie werden in Echtzeit zur Steuerung und Kontrolle von Geschäftsprozessen genutzt, aber auch später für Dokumentation, Qualitätsmanagement, Berichtswesen, Entscheidungsfindung und viele andere Aufgaben.

War der Rohstoff Öl im 19. Jahrhundert der Treibstoff für die Entwicklung der Industriegesellschaft, sollen Daten der Treibstoff für die entstehende Informationsgesellschaft sein, lautet eine vielzitierte These. Doch der Vergleich hinkt. Öl ist ein realer Stoff mit physikalischen Eigenschaften wie Gewicht, Temperatur oder Volumen, während Daten „virtuell" sind. Das heißt zum Beispiel: Daten können – völlig anders als das Öl – relativ einfach vermehrt und geändert, transportiert oder vernichtet werden.

Die Anziehungskraft von Big Data

Bei „Big Data" ist das allerdings leichter gesagt als getan – und alles andere als trivial. Massendaten haben ein großes „virtuelles" Gewicht und entwickeln damit eine große Trägheit, lassen sich also nicht so einfach verschieben. Außerdem wächst bei „Big Data" wie in der realen Physik mit der Masse die „Anziehungskraft", so dass die Anwendungen häufig dort laufen müssen, wo die Daten liegen, um beispielsweise die Netzwerke zu entlasten oder um die Verarbeitung zu beschleunigen. Deshalb ist es eine gute Idee, die „Schwerkraft" der Daten dadurch aufzuheben, dass Daten nur noch ein einziges Mal erstellt werden und dann zugreifbar und analysierbar sind, wo auch immer sie gespeichert werden.

Gemeinsam ist dem Öl der Industriegesellschaft und den Daten der Informationsgesellschaft, dass beide einen Wert haben. Das gilt nicht nur für Stammdaten (etwa zu Produkten, Kunden oder Lieferanten) und für personenbezogene Daten, deren Nutzen offenkundig ist, sondern zunehmend auch für Massendaten, die von Sensoren auf dem Shopfloor den Fertigungsfortschritt und den Maschinenzustand dokumentieren. Genauso gilt es auch für Massendaten der Logistik, die Ort und Zustand von Waren während Lagerung oder Transport beschreiben. Hier sind der Fantasie keine Grenzen gesetzt – hier entstehen in den Think Tanks großer Unternehmen oder bei innovativen Start-ups laufend neue Geschäftsideen und Verbesserungsvorschläge für bestehende Geschäftsmodelle.

1 Einleitung

1.2 Daten als strategische Ressource für Geschäftsinnovation

Die Digitalisierung verändert die Statik herkömmlicher Branchen und die Mechanik traditioneller Geschäftsmodelle, diagnostiziert beispielsweise das Fraunhofer-Institut. Um ihre Wettbewerbsfähigkeit zu sichern, müssen Unternehmen demnach das Innovationspotenzial der Digitalisierung nutzen.

Dabei kommt es laut Fraunhofer-Institut auch auf effiziente Datendienste an, z. B. für die Anonymisierung von Daten, für die Integration oder für die Einstellung von „Verfallsterminen" bei der Datennutzung. Erst dann könne die datengetriebene Innovation gelingen, die in vier Facetten ausgeprägt werden könne:

- **Produktinnovation:** In der pharmazeutischen Industrie beispielsweise erlaubt die Nutzung von Gesundheitsdaten wirksamere, individuellere Medikamente und Behandlungskonzepte. Dafür arbeiten diverse Akteure zusammen, etwa Pharmakonzerne, Krankenkassen, Ärzte und Patienten – und tauschen Daten aus. Gleichzeitig sollte aber der Patient zu jeder Zeit Souverän über seine eigenen Daten bleiben und allein entscheiden, was damit geschieht.
- Service-Innovation: Modernes Verkehrsmanagement nutzt bei der Routennavigation von Fahrzeugen längst nicht mehr nur Kartenmaterial, sondern dynamisch verschiedene Datenquellen, etwa von Verkehrsleitzentralen oder vom Wetterdienst.
- Prozess-Innovation: Der Einzelhandel vermeidet „Out-of-Stock"-Situationen im Supermarkt, indem Warenflüsse zu jeder Zeit mit Datenflüssen gekoppelt werden. Daten zur Ware (Lokation, Zustand usw.) sind allen Partnern in der Supply Chain zu jeder Zeit verfügbar, so dass der Einzelhandel, die Lieferanten und die Logistikdienstleister ihre Lieferkette gemeinschaftlich steuern und überwachen können. Daten werden hier anders als im Gesundheitswesen zum Gemeinschaftsgut, von dem alle Wertschöpfungspartner profitieren.
- **Organisationsinnovation:** Bei der Fertigung von Kleinserien müssen Stammdaten der Produkte, Auftragsdaten, Transportangaben usw. gemeinschaftlich von Herstellern, Zulieferern und Logistikdienstleistern bewirtschaftet werden.

Abbildung 1: *Auch im Internet of Things muss die Souveränität über die Daten gewährleistet bleiben. (Quelle: Fraunhofer-Institut)*

1 Einleitung

1.3 Die Cloud als Medium für die Datenjäger und -sammler

Innovative Konzepte wie Smart City, Autonomes Fahren oder Industrie 4.0 sind auf Gedeih und Verderb auf den reibungslosen Austausch von Daten zwischen allen involvierten Parteien angewiesen. Deshalb werden Daten wertvoll: weil viele Parteien – z. B. Kommunen, IT-Hersteller, Maschinenlieferanten oder -betreiber – die Datenschätze gern nutzen möchten. Das kann für die Produktentwicklung sinnvoll sein, aber beispielsweise auch beim „Anlernen" von KI-Systemen für die automatische (und damit schnelle, fehlerfreie) Reaktion auf Ereignisse. Die Beispiele sind mannigfaltig.

Daten erheben, vernetzen und auf dieser Basis intelligente Lösungen entwickeln: Das ist der Kerngedanke sowohl von Industrie 4.0 als auch Smart City. Die Gedanken kreisen dabei um das Wohl der Bürger oder der Belegschaften bzw. Kunden, aber auch um Geschäfte mit Apps z. B. zur Parkplatzsuche oder Hochwasserwarnung.

Als ideales Medium für Speicherung und Austausch der Daten könnte die Public Cloud gelten, gäbe es nicht gut begründete Sicherheitsbedenken. Die Public Cloud ist günstig, skalierbar, flexibel und unkompliziert – vorausgesetzt, es kümmert sich ein kompetenter Provider um Betrieb, Management und Wartung. Spielt das Sicherheitsrisiko bei der IoT-Anwendung keine oder nur eine untergeordnete Rolle, kann die Public Cloud guten Gewissens als Plattform genutzt werden.

Hohes Sicherheitsniveau

Die Private Cloud mit ihrem hohen Sicherheitsniveau bietet sich dann als Plattform an, wenn viele mobile Mitarbeiter, sensible Geräte oder externe Partner auf die Daten zugreifen. Da die meisten Unternehmen für unterschiedliche Zwecke verschiedene Private Clouds nutzen, sei es aus dem eigenen Rechenzentrum, sei es von unterschiedlichen Providern, entstehen schnell sogenannte Multi Clouds (als Kombination mehrerer Public und/oder Private Clouds) oder auch Hybrid Clouds, die Public und Private Cloud unter einen Hut bringen.

Dann wird es sehr schnell sehr komplex. Nicht nur, weil das Management von Multi und/oder Hybrid Clouds alles andere als einfach ist, sondern auch, weil es beim Datenmanagement hapert. Und dann fehlt der Schmierstoff für das Zeitalter Industrie 4.0.

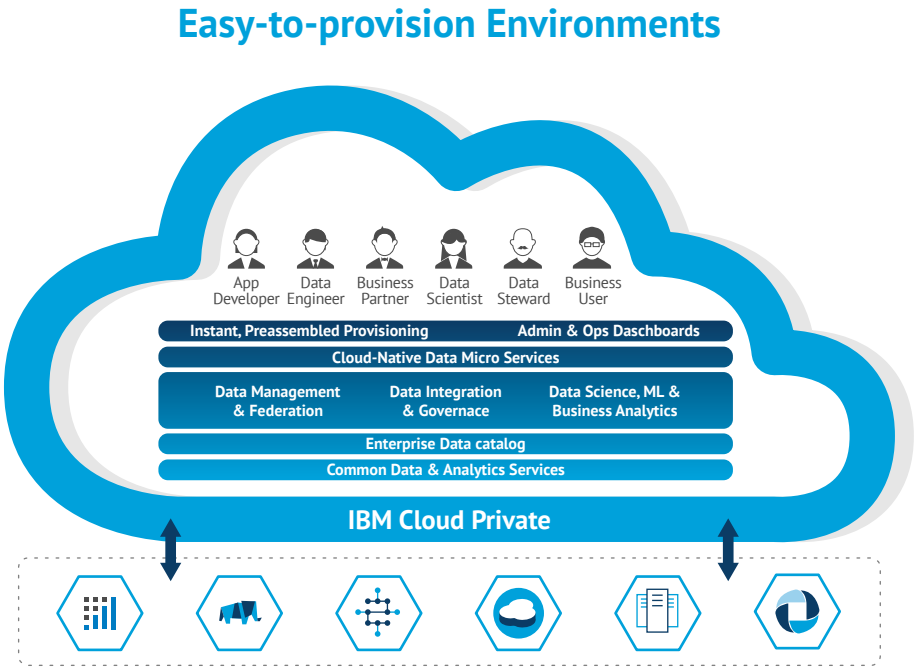


Abbildung 2: *Die IBM Cloud Private for Data vereinfacht die Bereitstellung aller Daten im Unternehmen. (Quelle: IBM)*

1 Einleitung

Um in diesem äußerst heterogenen Umfeld für einerseits performante, andererseits auch sichere und zuverlässige Analyse bzw. Verarbeitung aller Daten in der Cloud zu sorgen, hat IBM die „Cloud Private for Data" (siehe Kapitel 3) entwickelt. Sie bildet auf Basis bewährter IBM- und Open-Source-Technologie eine offene Anwendungsplattform für hybrides Datenmanagement (siehe Kapitel 4) – und bringt dabei Private-, Public- und Multi-Cloud-Umgebungen geschickt und sicher unter einen Hut.

So lassen sich die Vorteile der Public Cloud nutzen, ohne sich Probleme bei Datenschutz oder Compliance einzuhandeln. Außerdem kann der IT-Chef dank der Flexibilität und Skalierbarkeit der IBM Cloud Private for Data rasch und gezielt auf neue Anforderungen im Unternehmen (etwa bei Änderungen des Geschäftsmodells) reagieren. Neue Anforderungen von Kunden und Lieferanten lassen sich damit ebenso einfacher umsetzen wie neue Gesetze, die im Zuge der Digitalisierungsdebatte auf EU- und Bundesebene zu erwarten sind. Das hat sich schon bei Inkrafttreten der EU-DSGVO gezeigt, die – Stichwort „Recht auf Vergessenwerden" – völlig neue Spielregeln bei der Datenhaltung fordert. So wird etwa die ePrivacy-Verordnung weitere Veränderungen bringen.

1.4 Data Warehouse oder Data Lake als Datenspeicher?

Hinzu kommt: Viele Unternehmen haben heute schon Probleme, ihre Daten aus lange bestehenden „Silos" (für Abteilungen, Marken, Märkte usw.) zu erheben, um ihre zentralen Data Lakes bzw. Data Warehouses damit zu füllen, all diese zentralisierten Daten dann vernünftig zu verwalten und neue Analysefunktionen dafür zu entwickeln.

Ein Klassiker der gut strukturierten Datenhaltung ist das „Data Warehouse", ein für Analysezwecke optimiertes zentrales Repository, das Daten aus vielen heterogenen Quellen sinnvoll zusammenführt und verdichtet. Diese Idee, eine „Single Source of Truth" zu schaffen, stammt bereits aus den 80er-Jahren und hat sich bestens bewährt. In Data-Warehouse-Projekten wird über Jahre mit wachsender Erfahrung im Umgang mit den Daten geklärt, welche Daten überhaupt zu sammeln sind, wie sie am besten verdichtet und bereinigt werden und welche Kennzahlen und Resultate mit in die Auswertungen einfließen sollen.

Anwendungsgebiete für das Data Warehouse

Es gibt heute sehr viele Anwendungsgebiete für das Data Warehouse, etwa für die langfristige Daten-Dokumentation oder für das klassische Berichtswesen. Neue Anforderungen in den Industriezweigen schaffen noch weitere Einsatzszenarien. Weil sich zum Beispiel im Handel das Kaufverhalten der Verbraucher sehr rasch verändert, bietet sich das Data Warehouse als „Single Source of Truth" an. So kann beispielsweise „Predictive Analytics" stets im Vergleich mit vorliegenden Datenprofilen genutzt werden, um Abweichungen im Kaufverhalten bzw. neue Verhaltensmuster zu erkennen.

Auch in der Produktion ist ein Data Warehouse sinnvoll, etwa wenn zur Validierung der Qualität von Produkten oder Prozessen historische Daten genutzt werden. Empfehlenswert ist es deshalb nach wie vor, ein solides Data Warehouse aufzubauen, die Daten zielgerichtet zu säubern und aufzubereiten und so eine stabile Basis für neue Technologien und Tools zu schaffen, etwa auf Basis von KI. Wird das Data Warehouse zu groß oder zu komplex für performante Analysen in einer bestimmten Anwendung oder Abteilung, werden Teile davon in einen sogenannten „Data Mart" ausgelagert.

In der Regel werden in der Praxis die Datenbestände von Data Warehouse und Data Mart langfristig vorgehalten. Bei vielen heutigen Big-Data-Projekten ist das völlig anders, weil da zu Projektbeginn noch nicht einmal das gewünschte Ergebnis bekannt ist. Oft entsteht die Fragestellung auch spontan, z. B.

1 Einleitung

aufgrund neuer Marktentwicklungen oder konkreter Kundenwünsche. Weil man also nicht weiß, aus welchen Daten die benötigten Informationen gewonnen werden sollen, sammelt man einfach möglichst alle (Roh-)Daten, derer man habhaft werden kann – ohne sie in irgendeiner Art und Weise zu konsolidieren. Das wird dadurch erleichtert, dass Datenspeicher so schnell, hochskalierbar und kostengünstig geworden sind – vor allem auch Datenspeicher in der Cloud.

„Auf Verdacht" unterschiedlichste Daten speichern

So entstehen „Data Lakes", die vor zwanzig Jahren noch unbezahlbar teuer gewesen wären. Darin werden „auf Verdacht" unterschiedlichste Daten in ihrem natürlichen Format an zentraler Stelle verwaltet, wobei das Datenhaltungssystem (meistens NoSQL oder Hadoop) die Kollokation dieser Daten in verschiedenen Formaten (in der Regel Dateien) erleichtert. Ein Data Lake ist somit die Summe aller „Datentöpfe" (Repositories); das können ebenso Archive auf Hadoop-Basis sein wie ein Object Storage oder ein Data Warehouse.

Die Idee besteht heute nicht mehr darin, alle Daten des Unternehmens über eine einzige zentrale Datenbank zugänglich zu machen. Vielmehr werden simple „Rohdaten" und „veredelte" – sprich verdichtete und bereinigte – Daten im Data Lake für verschiedene Aufgaben wie Reporting, Visualisierung, Analytik und maschinelles Lernen zusammengeführt. Das können strukturierte Daten aus Datenbanken sein, semistrukturierte Daten (CSV, Logs, XML, JSON), unstrukturierte Daten (E-Mails, Dokumente, PDFs) oder sogar binäre Daten (Bilder, Audio, Video).

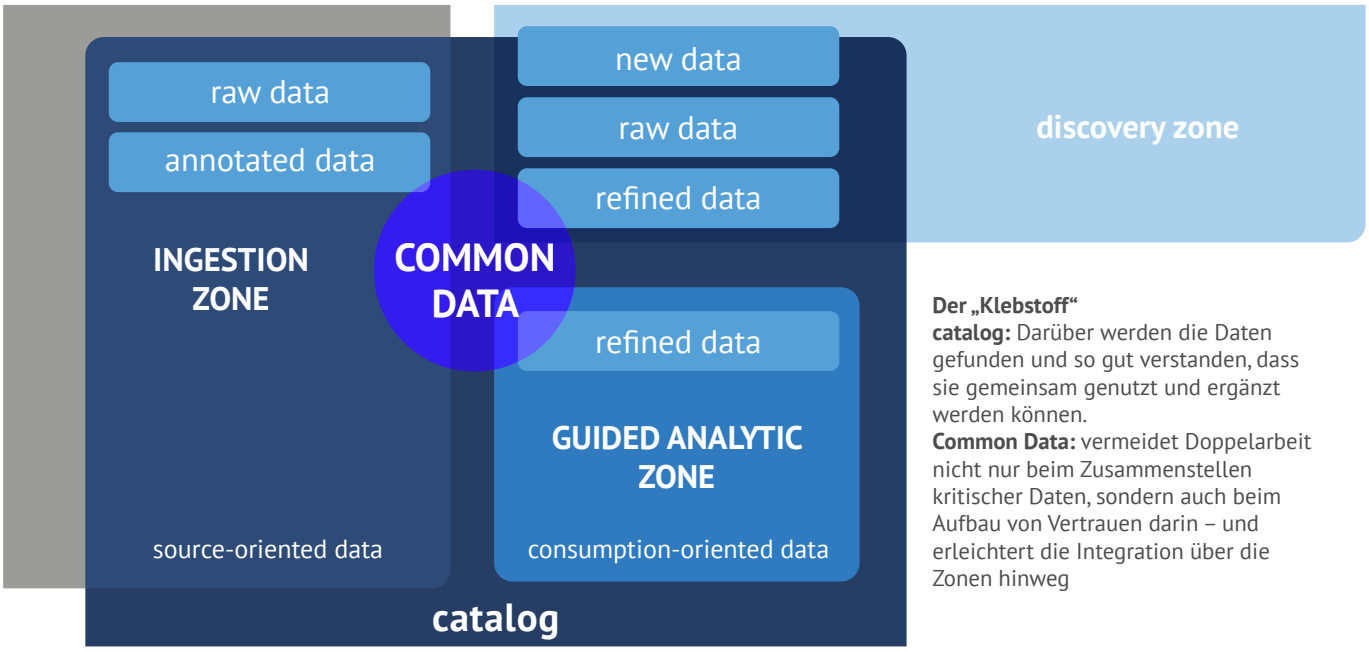


Abbildung 3: *Mit Hilfe des „Data Lake" können unterschiedlichste Nutzergruppen mit unterschiedlichen Datenbeständen arbeiten. (Quelle: IBM)*

1 Einleitung

1.5 Wenn es beim Datenmanagement hapert

Hapert es dann beim zentralen Datenmanagement, geht der Überblick über den Datenbestand verloren. Dann mutiert der „Data Lake" schnell zum „Data Swamp". In diesem Datensumpf (oder auch „Datenfriedhof") sind wichtige Daten für die Benutzer nur schwer zu finden und damit praktisch wertlos.

Abhängig von den Anwendungsanforderungen werden in verschiedenen Unternehmensbereichen unterschiedliche Datenhaltungssysteme verwendet. In ihnen werden jeweils die neuesten Daten untergebracht, die durch Kundentransaktionen, Benutzer-Tweets, Sensor-Inputs, Maschinendaten usw. generiert werden.

Mit Hilfe des „Data Lake" können all diese Daten sauber von diversen User-Gruppen für ihre Zwecke genutzt werden, ohne dass die Compliance gefährdet wäre.

Komplexe Daten-Ökosysteme

Die Auswertung von Daten in solch komplexen Umgebungen ist eine große Herausforderung. Dafür sind heute in der Regel „Data Scientists" oder Statistiker verantwortlich, die für Analysen auf Basis des Datenbestandes gerne Scale-out-Verarbeitungssysteme wie Hadoop oder Spark nutzen.

Es gibt aber durchaus einige immanente Probleme beim Zugriff auf Datenseen, wie etwa die Alterung von Daten oder den Datenschutz. Es kann sein, dass in der Datenflut die jüngsten (und damit interessantesten) Daten übersehen werden (z. B. weil sie noch gar nicht im „Data Lake" angekommen sind) oder dass ungültige, weil veraltete Daten fälschlicherweise in die Analyse einbezogen werden. Deshalb ist es so wichtig, dass bei der Implementierung des Data Lake auch die notwendigen Datenmanagement- und Governance-Mechanismen durchgängig implementiert werden.

2 Der Lebenszyklus von Daten

Auch Daten haben wie die Anwendungssysteme einen Lebenszyklus, der je nach Verwendungszweck von der Beschaffung über die Nutzung bis zum Löschen sehr unterschiedlich sein kann. In der Analytik beispielsweise wird die Kette von Aktivitäten normalerweise folgendermaßen beschrieben: Datengenerierung, Datenerfassung, Datenaggregation, Datenkontextualisierung und KI.

2.1 Datengenerierung, Datenanalyse und Betrieb

Um alle Phasen im Lebenszyklus der Daten zu beherrschen, ist ein konsistentes Datenmanagement erforderlich. Das gilt vor allem für die drei Kernaktivitäten Datengenerierung, Datenanalyse und Betrieb datenorientierter Geschäftsmodelle. Denn erhebt ein Unternehmen Daten, und zwar völlig unabhängig von Anwendungen wie Big Data, Machine Learning oder Sensoren, müssen diese Daten rechtskonform verarbeitet werden. Ohne unternehmensweites Datenmanagement ist diese Rechtskonformität kaum zu gewährleisten.

Bei personenbezogenen Daten etwa gibt die seit dem 25. Mai 2018 geltende Datenschutzgrundverordnung gemeinsam mit der ebenfalls neuen E-Privacy-Verordnung den Rahmen vor.

Das heißt beispielsweise auch, dass jede Person über ihre Daten Auskunft erhalten muss – und diese Daten auf Wunsch auch wieder gelöscht werden müssen –, mit entsprechenden Konsequenzen für das Datenmanagement.

Zu den drei Kernaktivitäten im Datenmanagement gehört die Datengenerierung. Darunter versteht man die Beschaffung und Sammlung von Daten, inklusive der Sensordaten im IoT-Umfeld. Das können Echtzeit-Verkehrsinformationen für die Tourenplanung sein, aber auch Maschinendaten aus der Fabrik.

Die anschließende Datenanalyse bezieht sich auf die Kombination und Analyse verschiedener Datenquellen mit dem Ziel, Muster oder Zusammenhänge zu finden und Algorithmen anzuwenden, um Personalisierung, Prognosen oder ortsbasierte Dienste bereitzustellen.

Datenorientierte Geschäftsmodelle sind die höchste Stufe der Wertschöpfung im Datenlebenszyklus – und heute vielfach erst in der Entwicklung. In solchen Geschäftsmodellen liefern die Daten nicht nur neue Erkenntnisse und Hilfestellungen bei der Entscheidungsfindung, sondern bilden auch die Grundlage innovativer Services für die Kunden oder Lieferanten.

Datenorientierte Geschäftsmodelle

Bei Weitem nicht jedes Unternehmen braucht ein datenorientiertes Geschäftsmodell. Moderne Unternehmen werden in Zukunft aber zumindest datengetrieben arbeiten. Damit ist gemeint, dass in solchen Unternehmen alle Personen und Prozesse, die Daten für bessere Entscheidungen nutzen können, auf diese Daten auch direkt zugreifen können. Und zwar unabhängig davon, wo diese Daten sich befinden.

Datengetrieben zu sein bedeutet für ein Unternehmen also nicht, zu Beginn eines jeden Tages oder einer Woche einige vordefinierte Berichte zu erstellen. Es geht vielmehr darum, Entscheidungsträgern die Möglichkeit zu geben, spontan, eigenständig und unabhängig voneinander Daten zusammenzustellen und zu analysieren. Und das selbst dann, wenn Unternehmen mit sehr vielen unterschiedlichen oder sehr großen Datenbeständen arbeiten.

Wie erwähnt kann dabei auch für Massendaten der Datenschutz zum Thema werden. Neben Anonymisierung und Pseudonymisierung der personenbezogenen Daten rücken – gerade im Hinblick auf Big-Data-Prozesse – daher auch solche Ansätze in den Vordergrund, die den Datenschutz bereits durch die technische Gestaltung der Anwendungen sicherstellen (z.B. „Privacy by Design“).

2 Der Lebenszyklus von Daten

2.2 Der Lebenszyklus von KI-Daten

Ein typisches Beispiel für einen solchen Lebenszyklus von Daten findet sich im Bereich der „Künstlichen Intelligenz". Hier kann ein iterativer „Software Development Lifecycle" (SDLC) auf die Daten angewendet werden, um die Systeme für die Beschaffung und Bewertung von Datensätzen und für ihre Kontrolle zu verbessern, also für die beiden Kernaktivitäten Datengenerierung und Datenanalyse.

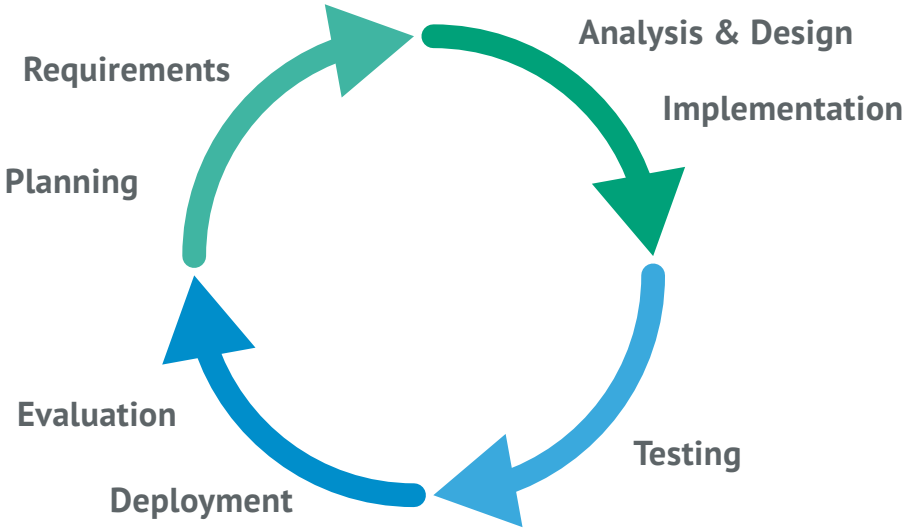


Abbildung 4: *Der Lebenszyklus von KI-Daten (Quelle: IBM)*

Effektives Sammeln, Vorbereiten und Verwenden von Daten für KI-Anwendungen wird am besten parallel zum Lebenszyklus der Softwareentwicklung realisiert. Im Geiste der jüngsten Entwicklungen in der agilen Entwicklung empfiehlt sich ein genau definierter, aber iterativer Ansatz zur Verwaltung der Daten für die KI-Anwendungen – anstelle eines sonst oft üblichen, starren „Wasserfall"-Ansatzes. Wird ein Projekt initiiert, tritt es (mit seinen Daten) in den Planungs- und Anforderungsquadranten ein – und durchläuft dann nach und nach alle Lebenszyklen (Abbildung 4).

Entwickler denken in diesen Phasen meistens über funktionale Spezifikationen, Datenbankschemata, Code-Schnittstellen- und Strukturdiagramme, Programmcode und Testfälle nach. Vorzugsweise werden Datenflussdiagramme zwar als Teil des Designs verwendet, aber allzu oft werden die eigentlichen Daten bisher nur fragmentarisch gemanagt, was insbesondere in der KI-Entwicklung schadet.

Analyse und Design

Beispielsweise sollten die Rohdatenquellen für die KI-Anwendung bereits zusammengestellt werden, während die Entwickler die Anwendung noch entwerfen. Wer mit dem Sammeln dieser Daten beginnt, lernt schnell, wie sie überprüft, erweitert, verwaltet und bewertet werden müssen. Formatgrenzen und Parameter werden frühzeitig festgelegt, z. B. die minimale und maximale Größe von Bildern oder die Länge von Audiodaten. Das erleichtert später die Inbetriebnahme enorm.

Gern übersehen wird beim Datendesign die Herkunft der Daten. Woher kommen sie? Wie gut lässt sich ihr Weg bis in die KI-Anwendung verfolgen? Wie sorgfältig wurden sie behandelt, bevor sie von der KI-Anwendung verarbeitet werden? Wie haben sie sich zuvor verändert, sei es durch Algorithmen oder Bearbeitung von Sachbearbeitern? All diese Fragen sollten beim Start von KI-Projekten geklärt werden.

Gibt es Anomalien, Verzerrungen oder andere Probleme, müssen die Daten eventuell vor der Verarbeitung „repariert" oder gar verworfen werden. In diesem Sinne ist der Datenlebenszyklus auch wichtig für kontinuierliche Systemverbesserungen. Wird die Anwendung „reifer", lernen Entwickler und „Data Scientists", welche Techniken und Prozesse effektiv bei der Datengenerierung sind und welche nicht.

2 Der Lebenszyklus von Daten

Den Datenfluss im Griff

Ein wichtiges, aber selten genutztes Design-Werkzeug ist das Datenflussdiagramm (DFD). Es wurde erstmals in den späten 1970er Jahren von Ed Yourdon und Larry Constantine als ein wichtiger Teil des Software-Engineering-Prozesses beschrieben. Entwickler in Bereichen wie Online-Sicherheit haben gelernt, wie wichtig ein Datenflussdiagramm z. B. für eine Online-Banking-Anwendung ist, wenn Daten vom Browser des Remote-Benutzers über Schichten mit zunehmenden Sicherheitsanforderungen in den Privatkunden-, Kreditkarten- und Back-Office-Abstimmungssystemen weitergeleitet werden. Eine ähnlich detaillierte Beschreibung des Datenflusses ist ein wichtiger Bestandteil der Datenaufbereitung für KI-Anwendungen.

Ein DFD für KI-Daten basiert in der Regel auf gemeinsamen Datenerfassungs- und Vorbereitungstechniken.

Ein solches Datenflussdiagramm handelt eigentlich von Daten, nicht von Prozessen, auch wenn die gerundeten Kästchen Prozesse darstellen. Diese Prozesse beschreiben aber nur, was mit den Daten zu tun ist. Die Pfeile deuten die Schritte an, die die Daten im Lebenszyklus durchlaufen. Das Ziel für die Daten ist der „Corpus", also die Datenbasis für die KI-Anwendung. Die in diesem Beispieldiagramm enthaltenen Datenverarbeitungsprozesse sind:

- **Datenerfassung** als der Prozess, bei dem Rohdaten gewonnen werden, die für die Aufnahme in den Corpus vorbereitet werden. Solche Rohdaten könnten z. B. durch Digitalisierung oder Extraktion aus dem Internet gewonnene Daten sein.
- Unter **„Data Wrangling"** versteht man das Konvertieren des Datenformats in eines, das als Input für die KI-Anwendung taugt, sowie das Verknüpfen der Rohdaten mit ihren Metadaten, einschließlich der Informationen über ihre Herkunft. Das Wrangling versucht auch, brauchbare Datenelemente zu identifizieren und Duplikate zu eliminieren.
- Bei der **Datenbereinigung** werden beschädigte, unvollständige oder ungenaue Daten korrigiert oder entfernt.
- Mit **Scoring & Inkorporation** sind statistische Analysen gemeint, mit denen die „Gesamtgesundheit" des resultierenden Corpus sichergestellt wird. Jedes Datenelement kann entsprechend seiner Eignung für den Corpus bewertet werden.

Ein beliebtes Modell für den Bewertungsschritt ist die „Explorative Datenanalyse" (EDA), mit vielen Visualisierungen der Beziehungen zwischen den Variablen über die Population hinweg. Eine gründliche EDA ist ein wichtiger Faktor für eine Verbesserung der Datenqualität.

Viele offene Fragen

Auch hier bleiben viele Fragen offen: Wie lassen sich Grundsätze wie „Privacy by Design" oder Datenminimierung zeitgemäß umsetzen? Genügen Maßnahmen wie Anonymisierung, Pseudonymisierung oder Verschlüsselung als Datensicherungsmaßnahme oder muss über weitergehende Ansätze nachgedacht werden, etwa über Datentreuhändermodelle der Industrie-4.0-Konsortien?

Brisante Themen wie Datenhoheit, die Neujustierung der Verantwortlichkeit bei autonomen Maschinen oder rechtliche Grenzen bei der Vernetzung in Wertschöpfungsketten werden aktuell durch die Politik adressiert.

Hier sind weitere rechtliche Vorgaben zu erwarten, die der IT-Chef dann wie zuletzt die Datenschutzgrundverordnung rasch und konsequent umsetzen muss. Ein hybrides Datenmanagement schafft hier auch die Flexibilität und Skalierbarkeit, um die Datenarchitektur immer à jour zu halten.

3 Hybrides Datenmanagement mit der IBM Cloud Private for Data

Mitte März hat IBM mit Blick auf datenorientierte Geschäftsmodelle und datengetriebene Unternehmen das neue Produkt „Cloud Private for Data“ lanciert. Dabei handelt es sich um eine Plattform für die effiziente Datenwirtschaft und maschinelle Lernprozesse bei der datengestützten Entscheidungsfindung. Die Grundidee: Alle Daten werden nur noch ein einziges Mal erstellt und sind dann immer und überall im Unternehmen nutzbar. Und all das auch im unternehmenseigenen Rechenzentrum so einfach, flexibel und skalierbar, wie es die User aus der Public Cloud gewohnt sind.

Diese Plattform wurde entwickelt, um Unternehmen dabei zu helfen, völlig neue Erkenntnisse aus ihren Daten zu gewinnen, aber auch, um ereignisgesteuerte Anwendungen zu erstellen und zu nutzen. Mit diesen neuen, ereignisgesteuerten Anwendungen können die Unternehmen künftig in Echtzeit die Datenströme von IoT-Sensoren, aus dem Online-Handel, von den unterschiedlichsten mobilen Geräten oder aber aus Social Media wie Facebook oder Twitter analysieren – „Big Data“ eben Und all das möglichst performant und absolut sicher.

In dieser vorgefertigten Lösung der Enterprise-Klasse sind viele Daten- und Analysedienste bereits sauber integriert und über eine einfache, kollaborative und aufgabengesteuerte Oberfläche nutzbar. Für den Einsatz IBM Cloud Private for Data ist keine spezielle Konfiguration erforderlich. Ganz wie ein gewiefter Datenwissenschaftler kann also auch „Otto Normalanwender“ die für seine Analyse relevanten Daten finden, auf ihrer Basis direkt Ad-hoc-Analysen durchführen, Modelle erstellen und diese Modelle in einer einzigen Umgebung in den Produktionsbetrieb überführen.

Auf diese Weise erhalten die Daten endlich ihren vollen Wert – dank des nahtlosen Zugriffs darauf, egal in welchem Rechenzentrum des Unternehmens und welcher Cloud sie sich auch befinden. Dieser Zugriff erfolgt auf Basis einer Cloud-nativen Datenarchitektur, die vollständig geschützt hinter der Firewall liegt.

Vorbereitet für den Einsatz mit KI-Systemen

Ganz nebenbei werden die Daten auch noch für den Einsatz mit KI-Systemen vorbereitet, sprich die Prozesse für die Datenaufbereitung sind bereits eingebaut. IBM Cloud Private for Data beschleunigt so als leistungsstarkes System die Erfassung, Organisation, Integration und Analyse von Daten aus dem gesamten Unternehmen – und damit auch die Einführung innovativer KI-Anwendungen. Und: Data Scientists, Datenadministratoren oder Anwendungsentwickler können dank rollenspezifischer Schnittstellen viel schneller und effizienter arbeiten.

3 Hybrides Datenmanagement mit der IBM Cloud Private for Data

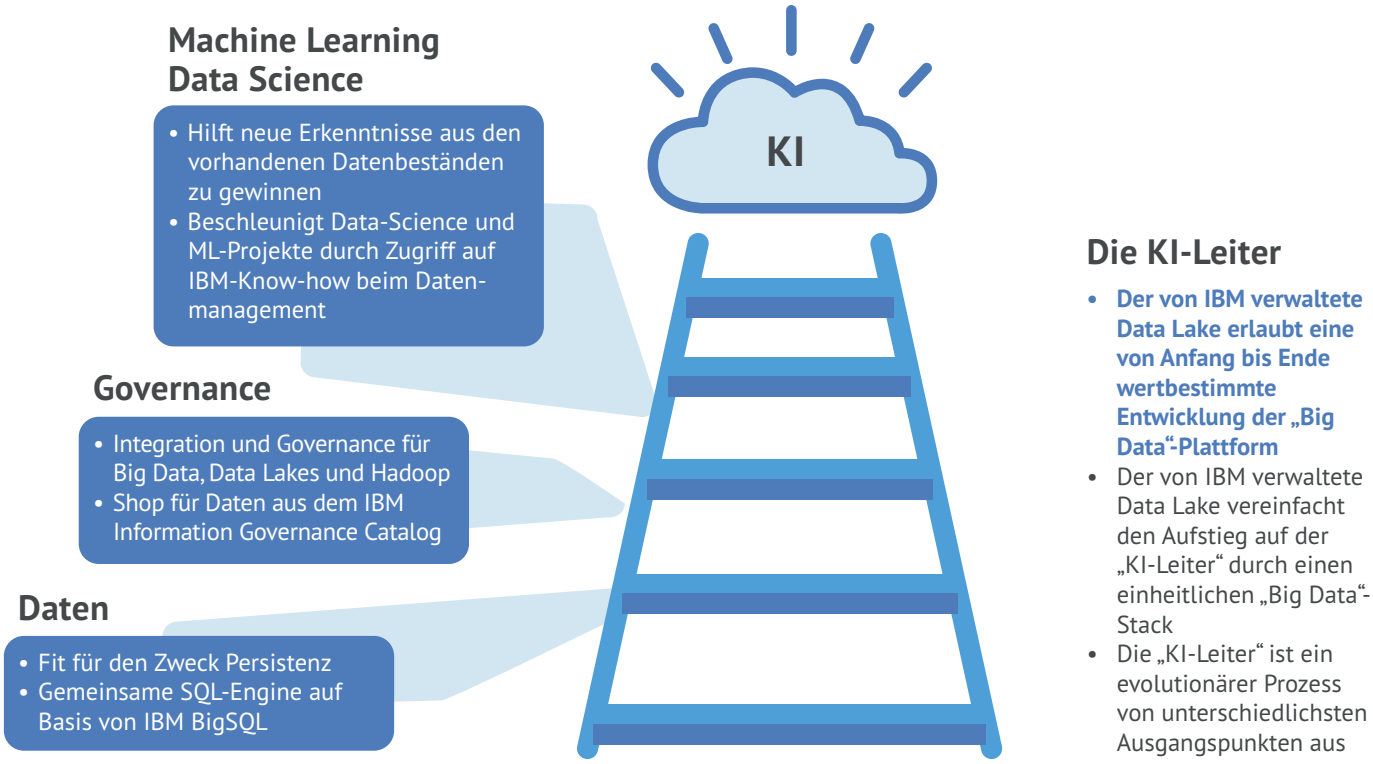


Abbildung 5: *IBM Cloud Private for Data bildet für den IT-Chef quasi die Trittleiter beim Aufstieg zum KI-Einsatz im Unternehmen. (Quelle: IBM)*

Ereignisgesteuerte Anwendungen können in der IBM Cloud Private for Data automatisch bestimmte Aktionen auslösen, was die Reaktionszeiten eines Unternehmens auf Kundenwünsche oder Marktereignisse enorm verkürzen kann. So wird die neue IBM Cloud Private for Data zur integrierten und einheitlichen Plattform für unterschiedlichste Anwendungen aus den Bereichen künstliche Intelligenz (KI), Data Science/Analytics, Datentechnik und Governance. Es lassen sich ML-Modelle und Dashboards erstellen, die ein Kunde dann über APIs sehr schnell bei der Erstellung spezifischer Apps für seine Zwecke nutzen kann.

Unternehmensweite Datenplattform

Vor allem aber erzeugt die IBM Cloud Private for Data eine unternehmensweite Datenplattform, die sämtliche heterogenen Datenquellen innerhalb und außerhalb des Unternehmens virtualisiert und in einheitlicher Form bereitstellt – eine entscheidende Voraussetzung dafür, dass datengesteuerte Geschäftsmodelle überhaupt realisierbar werden.

Die IBM Cloud Private for Data verbessert so die Datenarchitektur im Unternehmen (bzw. macht ihre Implementierung überhaupt erst praktikabel), automatisiert die Generierung von Metadaten durch Discovery-Prozesse, sorgt für Datenintegration und -virtualisierung und damit schließlich insgesamt für ein zentralisiertes und automatisiertes Datenmanagement. Fragen zu Compliance, Data Governance, Data Science oder auch KI „out of the box" lassen sich damit schnell klären. Auf diesem Fundament fällt dann auch der Aufbau neuer, datengesteuerter Geschäftsmodelle leichter.

3 Hybrides Datenmanagement mit der IBM Cloud Private for Data

3.1 Datenarchitektur

Die Daten eines Unternehmens sind in der Regel vollkommen heterogen und liegen in den unterschiedlichsten Metriken, Formaten und Datenspeichern vor. Diese Vielfalt der Datenarten wird noch verstärkt, falls die firmeneigenen Daten mit externen Daten ergänzt und validiert werden.

Neben der Art der Daten sind auch noch andere Metadaten wichtig, zum Beispiel ihre Herkunft oder der Zeitpunkt der Erstellung bzw. der letzten Änderung. Daten können aus unterschiedlichsten Quellen stammen; sie können durch manuelle Eingabe, Hardware (z. B. Maschinen, Kassen, Fahrzeuge) oder Software (z. B. ERP-Systeme, Suchmaschinen, E-Commerce-Plattformen) erzeugt werden. Und es gibt auch Daten, die von verschiedenen Sensoren eingesammelt worden sind.

Ein Maschinenbauer, der technische Daten seiner Geräte während ihres Einsatzes beim Kunden sammelt, um beispielsweise für vorbeugende Wartung den Verschleiß von Bauteilen vorherzusagen oder um seiner Entwicklungsabteilung Hinweise für die Verbesserung des Produktes zu liefern, ist ein exemplarischer Nutzer historischer Daten.

„Schwarmdaten" dagegen, die ein Verbund mehrerer Autos während der Fahrt austauscht (zum Beispiel für lokale Gefahrenwarnungen bei Staus, starkem Regen oder Glatteis) sind typische Echtzeitdaten. Ebenso Logistikdaten, die Sensoren in Paletten oder LKWs an die Zentrale funken, um den Standort zu melden oder auch Umgebungsparameter wie die Temperatur.

Ein spezieller Fall sind personenbezogene Daten, die zwar als Basis für Produktpersonalisierung oder personalisierte Kundenkommunikation offensichtlich sehr attraktiv sind, aber den Datenschutzbestimmungen unterliegen und daher eine dokumentierte und umkehrbare Zustimmung des Nutzers benötigen. Das hat wiederum spezielle Konsequenzen für das Datenmanagement, erfordert beispielsweise das dynamische Maskieren von Daten.

Diese Kategorie der Daten ist aber nicht scharf von den anderen Datenarten zu unterscheiden, weil technische Daten, z. B. Fahrzeugdaten, mit einer Person in Beziehung stehen und daher ebenfalls den Datenschutzbestimmungen unterliegen können. Dementsprechend komplex sind Aufgaben wie Pseudonymisierung oder Anonymisierung.

Auch deshalb ist eine stringente Datenarchitektur nützlich – weil sie solche Zusammenhänge sichtbar macht und zum Beispiel verhindert, dass über die technischen Daten seines Autos die bevorzugten Reiseziele oder der Musikgeschmack des Fahrers ermittelt werden könnte.

3 Hybrides Datenmanagement mit der IBM Cloud Private for Data

3.2 Datenintegration durch Datenvirtualisierung

Datenintegration und -synchronisation sorgen für Datenkonsistenz und weitgehende Redundanzfreiheit, verbessern also die Datenqualität und senken Speicherkosten. IBM Cloud Private for Data bietet auf Basis der Technologie Queryplex auch Features und Tools zur Abfrage von Datenquellen über mehrere Systeme und selbst über mehrere Clouds hinweg.

Auf diese Weise lassen sich die Daten der vorhandenen Systeme direkt für analytische Anwendungen bereitstellen. Die Analyse kann also zu den Daten gebracht werden – anstatt wie bisher üblich die Daten zu den Analysen zu bewegen. Außerdem entsteht ein Migrationspfad für IBM-Kunden mit der Möglichkeit, eine Vielzahl innovativer Technologien zu nutzen – entweder von IBM, von IBM-Partnern oder von Open-Source-Anbietern.

IBM Cloud Private for Data Experiences

Unter dem Namen IBM Cloud Private for Data Experiences wurde mittlerweile eine Testversion für Entwickler und neue Kunden lanciert, damit sie die Plattform schnell und unkompliziert ausprobieren können. Mit dieser Testversion können Entwickler und Data Scientists zum Beispiel Daten für maschinelle Lernmodelle und Big-Data-Analysen sammeln und aufbereiten.

Die Grundidee von Queryplex ist es, die Daten aus den unterschiedlichsten Quellsystemen nicht wie sonst vielfach üblich in ein zentrales System zu bringen, sondern die vielen vorhandenen Systeme einfach für analytische Abfragen nutzbar zu machen. Dieser Ansatz bietet sich etwa an, wenn die Daten schneller erzeugt

Quelle: Shutterstock

werden, als sie zu einem zentralen System (etwa wegen fehlender Netzwerkbandbreite) übertragen werden können, etwa in einem IoT-Szenario, das aus Hunderttausenden oder Millionen von Edge-Devices besteht, die Sensordaten sammeln und speichern. Häufig haben diese Systeme durchaus noch freie CPU-Zyklen, die für zusätzlichen analytischen Workload genutzt werden können, gleichzeitig aber häufig eine beschränkte Netzwerkbandbreite zur Cloud bzw. zu zentralen Systemen.

Daten, die von unterschiedlichen Werkzeugen erzeugt und/oder bearbeitet werden, nehmen oft auf die gleichen Fakten und Informationen Bezug. Es muss daher im Zuge der Datenintegration und -synchronisation sichergestellt werden, dass diese teilweise „überlappenden" Daten konsistent sind und bleiben. Dafür sorgt die IBM Private Cloud for Data, indem sie die Repositories (z. B. Data-Warehouse- oder Hadoop-Systeme) kontrolliert und zugänglich macht.

3 Hybrides Datenmanagement mit der IBM Cloud Private for Data

3.3 Datenmanagement, Governance, Data Science und KI „out of the box“

Damit bildet IBM Cloud Private for Data eine technisch ausgereifte Plattform, auf der sogar innovative Anwendungen aus dem Bereich der KI aufsetzen können, weil das Sammeln, Organisieren und Analysieren von Daten radikal vereinfacht wird.

Datenarchitektur für den KI-Einsatz

Denn gerade für den KI-Einsatz müssen Unternehmen im Vorfeld die erwähnte Datenarchitektur aufbauen. Die IBM Cloud Private for Data bietet als Fundament dafür eine sehr hohe Sicherheit mit eingebauten Funktionen für Maschinelles Lernen, Governance und Analytics. Hinzu kommen Mechanismen, die eine Integration in Cloud-Rechenzentren und Private-Cloud-Umgebungen vereinfachen.

Zunächst wird IBM Cloud Private for Data aber auf der IBM Cloud Private-Plattform gestartet, Diese Anwendungsplattform kann dank der Open-Source-Container-Software Kubernetes innerhalb von Minuten bereitgestellt werden. Sie bildet eine wirklich integrierte Umgebung für die Entwicklung von Analysen, KI- und ML-Anwendungen – und zwar unter der Einhaltung von Compliance-Vorgaben dank durchgängiger Governance-Mechanismen. Darüber lassen sich Maschinen in der Fabrik automatisch steuern, aber auch intelligente Automatismen und Apps zur Unterstützung der Geschäftsprozesse entwickeln.

Eingebaut ist zum Beispiel der IBM Information Governance Catalog. Er gibt Unternehmen die Möglichkeit, ein starkes Datengovernance- und Stewardship-Programm aufzubauen und aufrechtzuerhalten, durch das Daten zu verlässlichen Informationen werden. Diese verlässlichen Informationen lassen sich wiederum in verschiedensten Projekten nutzen, etwa Big-Data-Analysen, Stammdatenverwaltung (MDM), Lebenszyklusmanagement sowie Sicherheits- und Datenschutzinitiativen. Außerdem kann ein Unternehmen mit Hilfe des Governance Catalogs seine IT besser an den Geschäftszielen ausrichten.

Starke Daten-Governance und -Stewardship

Mit dem IBM Information Governance Catalog wird nicht nur das Vertrauen in die Daten gestärkt, sondern auch das Potenzial integrierter Metadaten genutzt. Dabei hilft der Governance Catalog zusätzlich durch den Aufbau eines gemeinsamen Vokabulars. Das etabliert im Unternehmen ein gemeinsames Verständnis und beschleunigt die Umsetzung verlässlicher, konsistenter Maßnahmen. Die Governance betrifft dabei nicht nur die Daten selbst, sondern auch alle Ergebnisse der Datenverarbeitung. Das können Modelle für Machine Learning oder Big-Data-Analysen sein, aber auch Skripts, Reports oder Auswertungen. Die Ergebnisse einer Big-Data-Analyse beispielsweise sind besser nachvollziehbar, weil dokumentiert ist, welcher Algorithmus aktiv war.

Auf diese Weise unterstützt der Governance Catalog auch den Aufbau einer Informationsarchitektur, die eine Grundlage der datengestützten Ökonomie bildet. Diese Architektur macht deutlich, woher die Daten stammen und wie Daten im Unternehmen miteinander verknüpft sind. Das erleichtert es sehr, auf dem aktuellen Stand der sich rasant ändernden Informationslage zu bleiben und rechtzeitig Einfluss auf die Geschäftsprozesse zu nehmen. Last but not least lassen sich konsistente Governance-Richtlinien erstellen, die definieren, wie Informationen strukturiert, gespeichert, umgewandelt und übertragen werden sollen.

IBM Cloud Private for Data kann nicht nur On-Premises auf der IBM Cloud Private ausgeführt werden, sondern wird künftig auf praktisch allen Clouds verfügbar sein – und zwar mit branchenspezifischen Lösungen für Finanzdienstleistungen, Gesundheitswesen, Fertigung und mehr. Zum Beispiel wird IBM Cloud Private for Data durch die Partnerschaft mit Red Hat über OpenShift auch die Technologien von HortonWorks einbinden, die ebenfalls für OpenShift zertifiziert sind.

3 Hybrides Datenmanagement mit der IBM Cloud Private for Data

Operationalisierung von Erkenntnissen aus Datenanalysen

So adressiert IBM mit Cloud Private for Data zwei der größten Herausforderungen für IT-Chefs: Die Bereinigung und Gestaltung der Daten des Unternehmens sowie die Operationalisierung von Erkenntnissen aus den auf diesen Prozessen aufsetzenden Datenanalysen.

Daten können so schneller für maschinelles Lernen und KI-Projekte vorbereitet werden – und die Ergebnisse dieser Projekte sind dann sehr einfach unternehmensweit nutzbar. IBM Cloud Private for Data nutzt dazu wichtige Funktionen bewährter IBM-Produkte, wie z. B. Data Science Experience, Information Analyzer, Information Governance Catalog, Data Stage, Db2 oder Db2 Warehouse.

Das Zusammenwirken all dieser Funktionen ist so konzipiert, dass sie möglichst einfach und schnell Einblicke in die Geschäftsdaten erlauben und diese Daten in einer geschützten, kontrollierten Umgebung nutzbar sind. Mit anderen Worten: Die neue Lösung stellt eine mächtige Dateninfrastruktur-Schicht für KI und Data Science hinter der Firewall bereit.

Auf diese Weise lässt sich relativ einfach und sauber ein hybrides Datenmanagement realisieren, um jede Art von Daten zugreifbar und analysierbar zu machen – unabhängig davon, wo diese Daten sich befinden. Unternehmen gewinnen damit die Freiheit, ihre Datenquellen im laufenden Betrieb problemlos zu erweitern, weil sie beispielsweise in einem Bereich eine neue Anwendung einführen oder eine weitere Datenquelle einbinden. Durch die einheitliche Governance und die Integration aller Daten entsteht eine virtuelle „Single Source of Truth“, und zwar mit der nötigen Flexibilität und Skalierbarkeit, so dass das Datenmanagement einfach mit dem Unternehmen und seinen Anforderungen mitwachsen kann.

4 Das technische Fundament

IBM Cloud Private for Data ist eine hochintegrierte Sammlung von Microservices für Daten und Analysen, die auf einer Cloud-nativen Architektur basieren. Sie bietet unterschiedlichste Funktionen, die jeweils speziell auf „Data Scientists", Geschäftsanwender, Dateningenieure, IT-Manager, Datenadministratoren und Anwendungsentwickler zugeschnitten sind. Diese können damit ihre Daten finden, organisieren und analysieren. Ziel ist es, privaten Rechenzentren der Unternehmen beim Datenmanagement die Flexibilität und Elastizität von öffentlichen Cloud-Infrastrukturen zu verschaffen und dabei außerdem KI-, Data-Science- und Entwicklungsplattformen zu integrieren.

IBM Cloud Private for Data bietet für jede dieser Rollen wichtige Funktionen, wie hybrides Datenmanagement, einheitliche Governance und Integration oder Business Analytics und KI. Die zweckmäßigen, rollenorientierten Funktionen sind für das Tagesgeschäft der jeweiligen Nutzergruppe maßgeschneidert und auf eine integrierte, sinnvolle Weise implementiert. Genutzt wird dazu das aktuelle Deep-Analytics-Portfolio von IBM, das Funktionen und Services für knifflige Herausforderungen bei Datenmanagement und Analyse beisteuert.

Weil die Plattform zusätzlich über APIs nahtlos erweitert werden kann, ist IBM jederzeit in der Lage, weitere Funktionen über sein Partner-Ökosystem bereitzustellen – und auch Kunden können unternehmensspezifische Anwendungen über die APIs integrieren. Partner-Lösungen wie MongoDB, DataMeer oder HortonWorks werden in der Form integriert, dass sie auch im Browser-Menü zu finden sind und ihre Metadaten in den Governance-Catalog aufgenommen werden.

Partnerschaften erweitern den Datenhorizont

Mit DataMeer hat sich IBM zusammengeschlossen, um eine „Datenpipeline" für IBM Cloud Private for Data zu schaffen – mit umfassenden Funktionen für die Datenvorbereitung, Exploration und Pipeline-Operationalisierung. DataMeer soll als integraler Bestandteil der Lösung eine nahtlose Kunden- und Benutzererfahrung bieten, um die Datenaufbereitung innerhalb der Plattform zu optimieren und die Daten schneller für maschinelles Lernen und KI-Projekte vorzubereiten. Die Zusammenarbeit adressiert zwei der größten Herausforderungen für Data Scientists – die Bereinigung und Gestaltung der Daten sowie die Operationalisierung der Erkenntnisse, die bei ihrer Verarbeitung gewonnen werden.

Durch die Zusammenarbeit mit MongoDB werden zum Beispiel NoSQL-Datenbanken als Teil von IBM Cloud Private for Data nutzbar. NoSQL-Datenbanken eröffnen Entwicklern auch neue Optionen für die Datenvirtualisierung und die Nutzung unterschiedlicher Serverplattformen. Durch ihre Skalierbarkeit und Flexibilität erweitern NoSQL-Datenbanken die Möglichkeiten vertrauenswürdiger Benutzer dadurch, dass diese JSON-Daten auch mithilfe der Tools von MongoDB erkunden und nutzen können. Dabei adressiert das dokumentenbasierte Datenmodell von MongoDB mehrere Probleme, die das relationale Datenmodell nicht adressieren soll:

- große Mengen sich schnell ändernder strukturierter, halbstrukturierter und unstrukturierter Daten
- agile Sprints, schnelle Schema-Iterationen und häufige Code-Pushs
- API-gesteuerte, objektorientierte Programmierung
- geographisch verteilte Scale-Out-Architektur statt teurer monolithischer Architektur.

Gemeinsam mit HortonWorks und Red Hat hat IBM eine Open-Hybrid-Architecture-Initiative angekündigt. Die Kooperation hat das Ziel, ein gemeinsames Bereitstellungsmodell zu entwickeln, mit dem Unternehmen ihre Big-Data-Workloads auf hybride Weise über lokale, Multi-Cloud- und Edge-Architekturen hinweg verarbeiten können. Hortonworks bringt seine offenen Datenplattformen auf Basis von Apache Hadoop, Apache NiFi und Apache Spark ein, Red Hat die Anwendungsplattform OpenShift. In einer ersten Phase optimieren die Unternehmen das Zusammenspiel der Hortonworks-Technologien Data Platform, DataFlow und DataPlane mit IBM Cloud Private for Data. Auf diese Weise sollen Anwender Big-Data-Workloads in Containern entwickeln und bereitstellen können. Dies macht es Kunden letztlich einfacher, Daten-Applikationen über hybride Cloud-Implementierungen hinweg zu verwalten.

4 Das technische Fundament

4.1 Linux, Virtualisierung, Docker und Kubernetes

Linux, Virtualisierung, Docker und Kubernetes bilden das technische Fundament für die IBM Cloud Private for Data, basiert sie doch auf diesen offenen Standards und der Container-Technologie. Das hat Folgen für das gesamte darin eingebundene Softwareportfolio von IBM, das mit Containern neu entwickelt werden soll. Unter anderem gilt das für WebSphere, MQ Series und Db2.

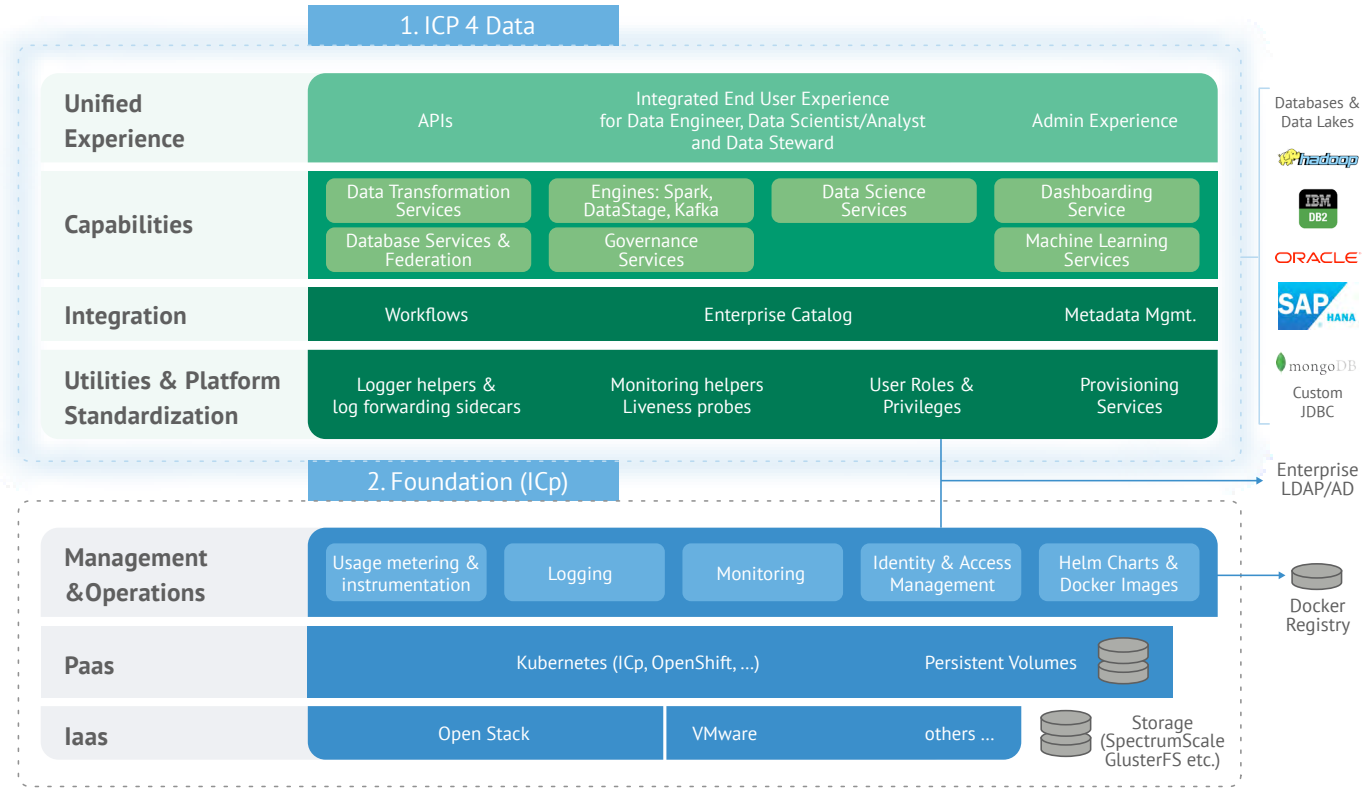


Abbildung 6: *Die technische Architektur der IBM Cloud Private for Data (Quelle: IBM)*

Container-Technologien sind eine sichere und zuverlässige Methode, mit der Anwendungen über mehrere IT-Footprints hinweg übertragen werden können – etwa von bestehenden Rechenzentren in die öffentliche Cloud und umgekehrt. Parallel zur Containerisierung der eigenen Software entwickelt IBM deshalb gemeinsam mit der Open Source Company Red Hat ein Portfolio an On-Premise-, Cloud-nativen und Hybrid-Cloud-Infrastrukturlösungen, das Unternehmen einen klaren Weg für die Einführung von Hybrid Cloud Computing eröffnen soll.

Auf diese Weise lassen sich dann die bisherigen Technologieinvestitionen maximieren, weil vorhandene Anwendungen und Daten sich einfacher in die Hybrid Cloud verschieben lassen, wenn IBM Cloud Private und Red Hat OpenShift als Grundlage dienen. OpenShift ist eine Container-Anwendungsplattform, die auf Docker und Kubernetes basiert. Unabhängig von der vorhandenen Anwendungsarchitektur ermöglicht es OpenShift, einfach, schnell und in fast jeder Infrastruktur, öffentlich oder privat, neue Lösungen zu entwickeln und bereitzustellen.

4 Das technische Fundament

4.2 Verwaltung von Container-Anwendungen mit Kubernetes

Kubernetes (oft einfach auch nur „K8“ genannt) ist ein Open-Source-System zur Automatisierung der Bereitstellung, Skalierung und Verwaltung dieser Container-Anwendungen. Es ist als Plattform für automatisierte Roll-outs, Skalierungen sowie Betrieb und Wartung von Anwendungscontainern auf verteilten Hosts konzipiert, unterstützt verschiedene Container-Tools (inklusive Docker) und arbeitet auf den wichtigsten Cloud-Plattformen. Das heißt konkret: nicht nur auf der IBM Cloud und OpenShift, sondern auch auf z. B. Microsofts Azure, Amazons AWS oder Oracles OCI.

Kubernetes ist eine Umgebung, die Deployment und Management von Applikationen sogar vollständig übernehmen kann. Sie ist heute schon weitestgehend standardisiert, so dass sich damit zustandslose Applikationen überall problemlos installieren lassen. Spannend wird es, sobald Daten mit ins Spiel kommen; das heißt: Eine Anwendung kann in unterschiedlichen Zuständen sein, die durch die Daten definiert werden.

Ein entscheidender Aspekt ist dann, wie diese Daten vorgehalten und zur Verfügung gestellt werden. Ratsam ist die Nutzung einer Hybrid-Cloud-Plattform, bei der „private“ Daten in jedem Fall im eigenen Unternehmen verbleiben und dennoch – je nach Bedarf – gemeinsam mit anderen Ressourcen, etwa aus einer Public Cloud, verarbeitet werden können.

4.3 Integration mit anderen IT-Systemen

Die Unterstützung von Docker durch IBM Cloud Private for Data ist insofern wichtig, als dass diese Open-Source-Software eine Isolierung von Anwendungen durch Containervirtualisierung erlaubt. Das vereinfacht die Bereitstellung von Anwendungen in der Cloud, weil sich Container, die alle nötigen Anwendungs-Pakete enthalten, ganz simpel als Dateien transportieren und installieren lassen. So lassen sich im Container ganz einfach neue Anwendungen bereitstellen, die mit der IBM Cloud Private for Data arbeiten. Die Container gewährleisten dabei die saubere Verwaltung aller wichtigen Ressourcen einer Anwendung: Code, Laufzeitmodul, Systemwerkzeuge und Systembibliotheken.

Programmierer können dank dieser Zentralisierung neue Anwendungen schneller entwickeln und bestehende Anwendungen einfacher modernisieren. Die resultierenden Anwendungen lassen sich dann mit Hilfe von Kubernetes auf der IBM Cloud Private for Data automatisiert bereitstellen und betreiben, während sie dabei gleichzeitig andere Cloud-Dienste von IBM (wie Watson, IoT und Blockchain) auf der Container-Platform OpenShift nutzen. Dazu erweitert IBM seine Cloud Private for Data und seine Middleware-Angebote zu Red Hat Certified Containers.

Außerdem plant IBM auch neue „Watson Data Kits“, um die Entwicklung von KI-Anwendungen zu beschleunigen. Diese Kits sollen dann Unternehmen branchenübergreifend mit vorbereiteten, maschinenlesbaren, branchenspezifischen Daten versorgen, was eine Skalierung von KI im gesamten Unternehmen deutlich vereinfacht. Diese Kits werden zunächst die Reise-, Transport- und Lebensmittelindustrie mit Input versorgen, etwa mit Details zu Reisezielen oder Rezepten.

Unternehmen profitieren nach der Containerisierung von der Schnelligkeit und Einfachheit des IBM Cloud Private Self-Service-Katalogs, von der performanten Bereitstellungs-Engine und vom übergreifenden Betriebsmanagement per OpenShift über alle Footprints der Hybrid-Cloud (einschließlich der IBM Public Cloud). Natürlich kann die IBM Cloud Private for Data nach wie vor auch ohne OpenShift als Virtualisierungsschicht genutzt werden.

4 Das technische Fundament

4.4 Compliance und Governance

Die Idee hinter dem Produkt IBM Cloud Private for Data basiert auf zwei strategischen Aspekten. Erstens wird eine integrierte Plattform geschaffen, über die der Zugriff auf alle Unternehmensdaten einfach und gleichzeitig kontrolliert erfolgen kann. Zweitens basiert diese integrierte Plattform auf Microservices, die ohne Anpassungen und Konfigurationen „out of the box" direkt zusammenarbeiten.

Es entsteht eine zuverlässige, skalierbare und wartbare Plattform für vorkonfektionierte Lösungen und zusätzliche Services, die sich dann sehr einfach up- und downgraden lassen. Dabei wird jede Art von Daten virtualisiert und damit zugreifbar gemacht, wo immer sie sich auch befinden. Das befreit die User von der zeitraubenden Arbeit, in den sich ständig ändernden Datenquellen die richtigen Daten zu suchen. Sie finden die Daten über den Governance Catalog, vorausgesetzt, die automatisierte Discovery ist aktiviert und die manuelle Zuweisung der Metadaten gemäß Business-Vokabular ist erfolgt.

Vertrauenswürdige Analytik-Grundlage

Außerdem entsteht direkt eine vertrauenswürdige Analytik-Grundlage, da IBM Cloud Private for Data alle Daten kontrolliert virtualisiert und direkt über eine zentrale Bedienoberfläche zugreifbar macht. Das verbessert die Agilität des Unternehmens spürbar und macht die Daten auf ganz neue Art und Weise nutzbar, zum Beispiel für das Qualitätsmanagement oder für datengesteuerte Geschäftsmodelle.

Dabei ist immer die nötige Governance gegeben. Fehlerhafte oder unvollständige Datenbestände werden angezeigt und lassen sich im Zuge der Überprüfung manuell vervollständigen bzw. korrigieren, was die Datenqualität insgesamt deutlich verbessert. Es lassen sich Richtlinien und Datenkataloge definieren und umsetzen, die den Zugriff auf Datenbestände und analytische Modelle kontrollieren.

In den Datenkatalogen werden auch die Metadaten für die Datenbestände gespeichert, was das Auffinden erleichtert und Mehrwert bei der Verarbeitung der Daten bringen kann. Aussagekräftige Dashboards lassen sich vom User ohne Codierung und ohne die Notwendigkeit mächtiger Reportingwerkzeuge erstellen. Last but not least ist ein „Data Asset Lifecycle Management" (DALM) vorgesehen, mit dem sich regelrechte Daten-Pipelines zur automatisierten Versorgung bestimmter Anwendungen realisieren lassen. Außerdem dokumentiert und kontrolliert DALM die „Datenlieferkette", was die Nachvollziehbarkeit von Datenanalysen verbessert und das Verständnis der Analyseergebnisse erleichtert.

4 Das technische Fundament

4.5 Teamarbeit und Datenschutz

IBM Cloud Private for Data macht es allen Teams und Abteilungen einfacher, Daten intelligenter als bisher zu analysieren und ganz neue Erkenntnisse zu gewinnen – sei es beispielsweise über das Kundenverhalten, sei es über die Produktionsprozesse.

User, aber auch Datenwissenschaftler, Dateningenieure und Entwickler können mit rollenspezifischen Bedienoberflächen schneller arbeiten, ohne dass Datenschutz und Datensicherheit gefährdet würden. Dabei gibt es eine einheitliche, übergreifende Bedienerschnittstelle, die personalisiert werden kann, aber auch kollaborativ gut nutzbar ist. Das entlastet die Mitarbeiter, die dann mehr als bisher strategische Aufgaben wahrnehmen können und nicht länger Zeit mit dem Auffinden und Validieren von Daten vergeuden.

Kontrolle des Lebenszyklus der Enterprise Daten

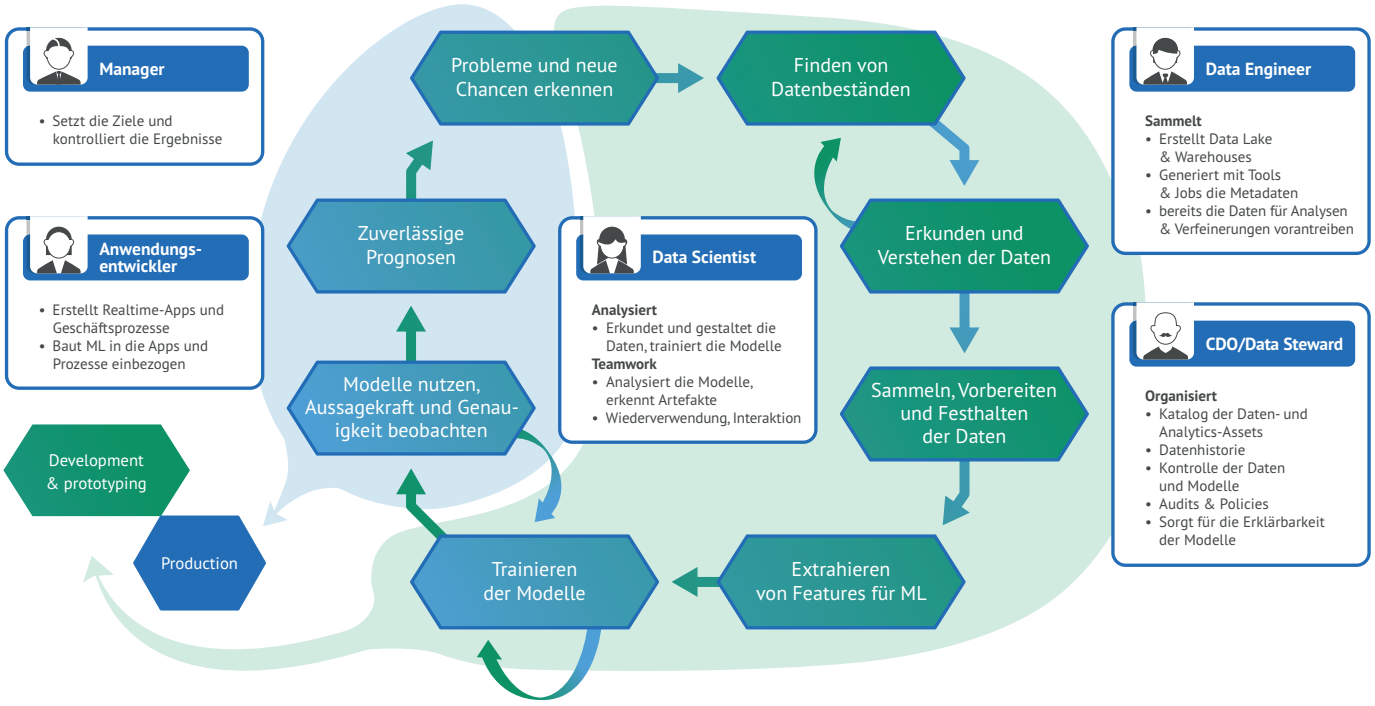


Abbildung 7: *Der Daten-Lebenszyklus wird mit den nötigen Sicherheitsmaßnahmen unterstützt, so dass alle „Stake-Holder" mit den Daten arbeiten können.(Quelle: IBM)*

Personalisiert und kollaborativ gut nutzbar

IBM Cloud Private liefert wie beschrieben den Container, das Microservices-Management und die Laufzeit-Umgebung für IBM Cloud Private for Data, was sowohl Management als auch User-Experience auf Basis offener Tools und Standards vereinheitlicht. Darauf aufbauend implementiert IBM Cloud Private for Data auch die Informations-Governance, also die Orchestrierung von Personen, Prozessen und Technologien bei der Nutzung von Daten als Unternehmensressourcen. Und die Visualisierung verbessert auch das Verständnis der Daten.

4 Das technische Fundament

Governance – Geschäftspolitik und Geschäftsregeln
Die Richtlinien der Informations-Governance definieren die Grundsätze im Umgang mit Informationsressourcen. Jede Richtlinie sollte nicht nur relevant für die Unternehmensziele sein, sondern auch einfach genug für die Benutzer, damit sie auch nachhaltig gelebt wird und nicht in Aktenordnern verstaubt. Entsprechende Regeln legen dann Kriterien fest, anhand derer bestimmt wird, ob Daten richtlinienkonform genutzt werden.

Mit Services für automatisierte Discovery und Informations-Governance unterstützt IBM Cloud Private for Data die Administratoren, Business-Analysten oder Daten-Ingenieure beim Aufspüren und Katalogisieren vorhandener Datenbestände, beispielsweise für die Nutzung in KI-Anwendungen. Dies erleichtert es außerdem enorm, jedes einzelne Datum während seines kompletten Lebenszyklus im Sinne der Geschäftsziele zu überwachen und gleichzeitig die Anforderungen der unterschiedlichen Geschäftsprozesse zu berücksichtigen.

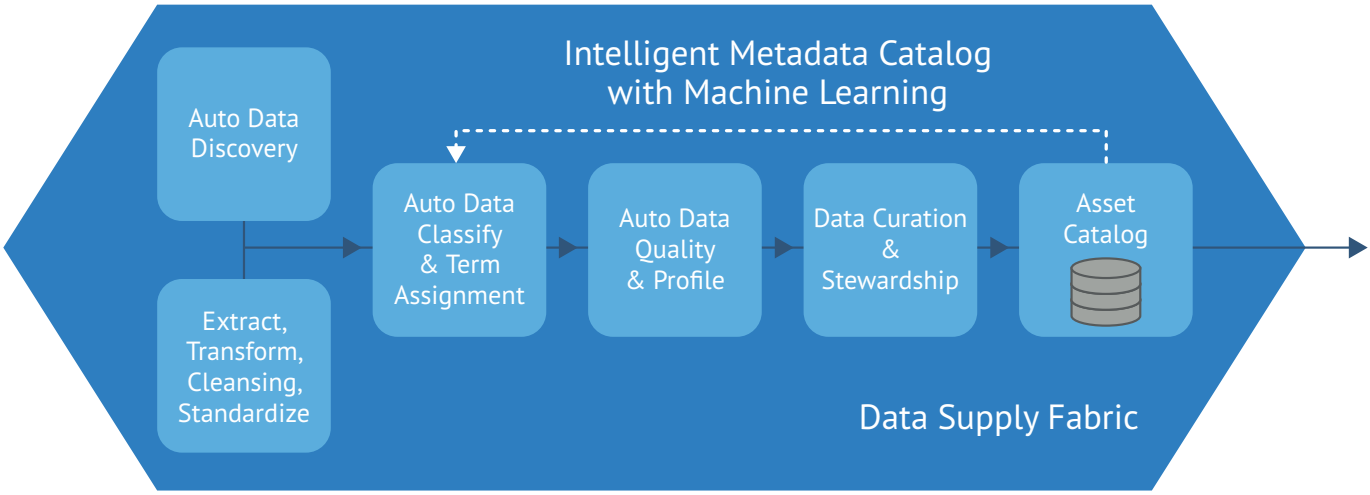


Abbildung 8: *Automated Discovery erweitert den Datenbestand kontinuierlich. (Quelle: IBM)*

Außerdem wird der Prozess des Katalogisierens, Analysierens und Klassifizierens von Daten aus neuen Datenquellen automatisiert, indem die Datensätze registriert und die entsprechenden Metadaten erzeugt und dem Datenkatalog hinzugefügt werden. Eigenschaften wie Datenklasse, Datentyp oder -format werden ebenfalls automatisch identifiziert und mit dem Prozentsatz der Wahrscheinlichkeit vorgeschlagen. Eine Qualitätsanalyse der Datensätze ermittelt prozentual die Datenklassenverletzungen.

Auf diese Weise wird die Suche nach Daten einfach: Wer in der Symbolleiste nach Daten-Assets sucht, kann in den Ergebnissen auch auf das Beziehungsdiagramm zugreifen und darüber direkt alle Beziehungen des gesuchten Assets erkunden. Die Metadaten zu jedem Dateifeld und zu jeder Tabellenspalte einer Datenbank können von jedem Benutzer bewertet werden, der eine Zugriffsberechtigung darauf hat. Die durchschnittliche Bewertung beeinflusst wiederum das Ranking bei der Anzeige des Datensatzes in den Suchergebnissen.

4 Das technische Fundament

Daten transformieren
Mit IBM Cloud Private for Data können Transformationsjobs erstellt, bearbeitet, geladen und ausgeführt werden. Dazu sind Funktionen wie die automatische Metadatenweitergabe integriert. Das macht speziell ETL-Arbeiten sehr komfortabel und effizient, zumal jede bekannte Verbindung zu Datenquellen für ETL-Jobs direkt verfügbar ist.

Wie erwähnt, werden im Rahmen der Informations-Governance durch die IBM Cloud Private for Data auch die Rollen aller an der Informationsverarbeitung beteiligten Personen festgelegt. Der „Data Steward" (Datenadministrator) ist für die Katalog- und Datenqualität im Unternehmen verantwortlich. Er definiert nicht nur den Datenkatalog und das gemeinsame Vokabular (Geschäftsglossar, für das allgemeine Verständnis der Daten), sondern verantwortet z. B. auch die Einhaltung der Governance-Regeln und Richtlinien.

Daten-Ingenieure wiederum bereiten die Datenbestände für eine Verarbeitung bzw. Analyse vor, zum Beispiel durch eine Datenauswahl oder durch das Erstellen virtualisierter Ansichten der Daten. Daten-Ingenieure schaffen Datenstrukturen, sorgen für die Automatisierung der Neuaufnahme von Metadaten, bewerten und verbessern die Datenqualität und erstellen neue Regeln.

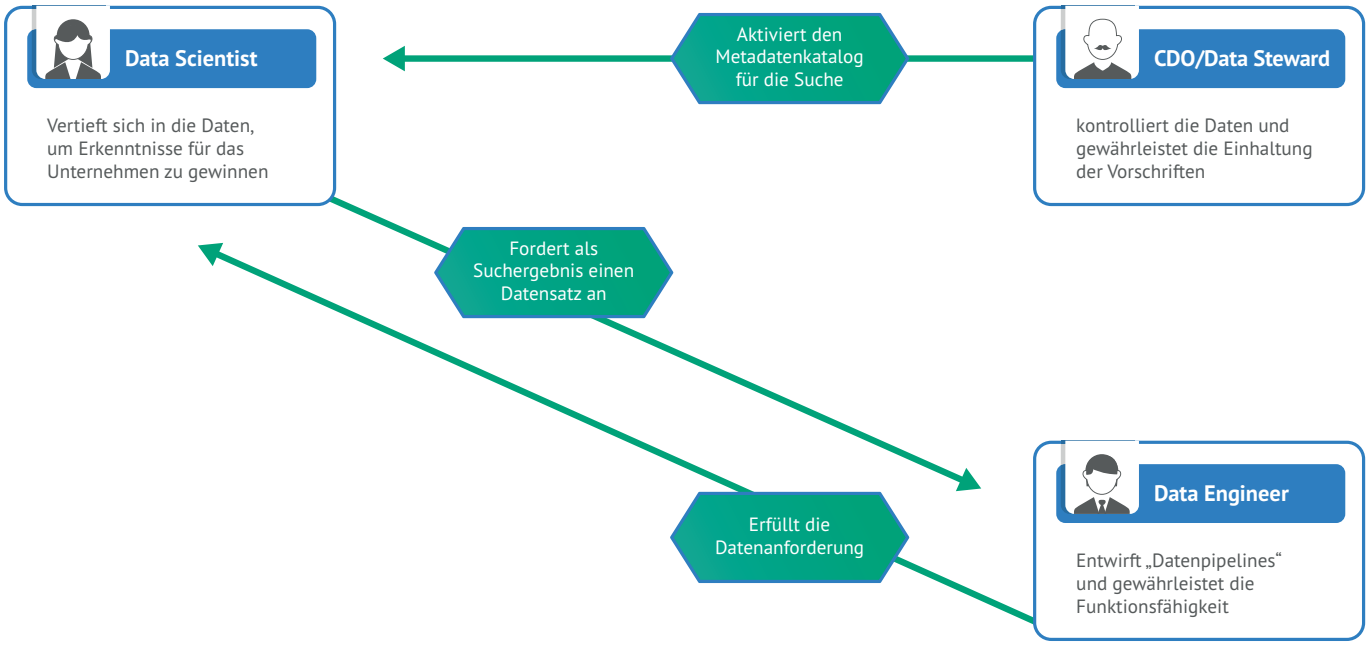


Abbildung 9: *Beispiel für die Zusammenarbeit der Datenarbeiter (Quelle: IBM)*

4 Das technische Fundament

Der „Data Scientist“ soll aus den vorhandenen Daten neue Erkenntnisse gewinnen, indem er relevante Daten aufspürt, untersucht und visualisiert, die gefundenen Daten veredelt und ergänzt oder auch Algorithmen entwickelt und verbessert, etwa zur Darstellung von Erkenntnissen, für maschinelle Lernmodelle oder für Skripte. Dabei können dem Data Scientist auch Richtlinien vorgegeben werden, um die Compliance zu wahren.

Business-Analysten wiederum erstellen Analyse-Dashboards zur Visualisierung von Daten, die sie mit Kollegen oder Vorgesetzten teilen. Administratoren schließlich verwalten die Datenzugänge, APIs und ML-Modelle, die Benutzer und ihre Berechtigungen. Sie stellen auch neue Modelle bereit, planen umfangreichere Batch-Scoring- und Bewertungsjobs, kümmern sich um benutzerdefinierte Webdienste und Notebooks und überwachen die Integrität der bereitgestellten Modelle.

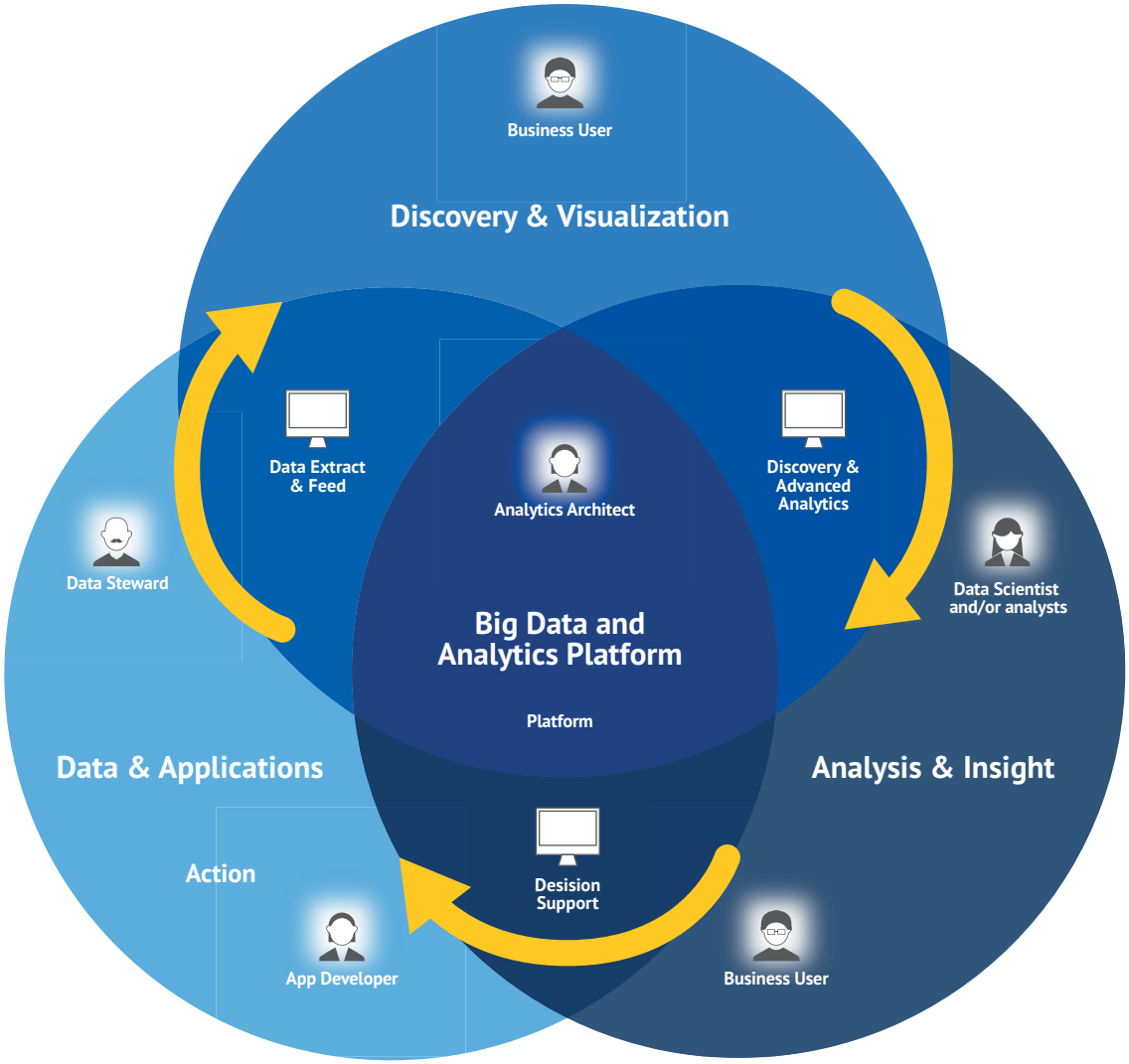


Abbildung 10: *Ein neues Berufsbild entsteht – der Datenarchitekt (Quelle: IBM)*

Die Jobs im Umgang mit den Daten sind heute schon sehr vielfältig. Dennoch entsteht gerade das neue Berufsbild des Datenarchitekten. Er ist eine Mischung aus Business-Analyst, Datenadministrator und „Data Scientist“, der sowohl den Aufbau der Datenplattform gestaltet als auch die notwendigen Tools und Anwendungen für Datenanalyse und Reporting bereitstellt und verwaltet. Auch der Datenarchitekt arbeitet direkt mit der IBM Cloud Private for Data. Und die Fachabteilungen gewinnen viel Flexibilität; sie können z. B. auch völlig neue Benutzerrollen beim Umgang mit den Daten schnell und unkompliziert implementieren, etwa für Kunden oder Lieferanten.

4 Das technische Fundament

4.6 Anwendungsbeispiele für unternehmensweites Datenmanagement

Dass sich Digitalisierung und Datenorientierung für Unternehmen lohnt, machen die Beispiele einiger Pioniere sehr schnell deutlich. Die südafrikanische Bankengruppe Nedbank beispielsweise hat bereits eine lange Tradition der Analyse interner, strukturierter Daten. Weil aber immer mehr Daten verfügbar werden und die analytischen Werkzeuge sich rasant weiterentwickeln, haben die Banker begonnen, interne und externe Daten zu nutzen, um innovative datengesteuerte Geschäftsmodelle zu schaffen. Darüber wollen sie neue Einnahmequellen erschließen.

Mit Unterstützung des Elite-Teams von IBM, das auf Technologien wie Watson oder IBM Analytics zurückgreifen kann, hat die Nedbank ganz neue Paradigmen entdeckt, wie sie Analysen handhaben könnte. Außerdem hat man die Art und Weise grundlegend geändert, wie neue Anwendungsprojekte angegangen werden, um datengetrieben zusätzliche Mehrwerte schaffen zu können.

Kognitive Lösung für die Mitarbeitervermittlung

Auch die japanische Zeitarbeitsfirma Forum Engineering setzt gezielt auf Analytics – und hat eine kognitive Lösung für die Mitarbeitervermittlung eingeführt. Diese Lösung analysiert automatisch sowohl strukturierte als auch unstrukturierte Daten aus internen Kandidatendateien, um Kunden möglichst optimale Kandidaten vorzuschlagen und ihnen auch die Gründe für diese Wahl zu erläutern. Dazu werden auch diejenigen Argumente mitgeliefert, die den vorgeschlagenen Kandidaten als besonders geeignet für die Anforderungen des Arbeitsplatzes erscheinen lassen.

In der Vergangenheit haben die Personalberater Tausende von Lebensläufen, Interviewnotizen und Kundenfeedbacks von Hand durchgesiebt, um Kandidaten anhand der Stellenanforderungen zu bewerten. Der Prozess war zeitaufwendig und oft ineffektiv. Daher musste das Unternehmen in der Regel mehrere Kandidaten vorstellen, bevor der passende gefunden war. Jetzt kann die kognitive Personalbesetzung in Sekundenschnelle umfangreiche Recherchen und Vergleiche durchführen, so dass die Personalspezialisten mehr Zeit für Kundenbeziehungen aufwenden können.

Das Resultat dieses Analytics-Projektes kann sich sehen lassen: Eine Verbesserung um 83 Prozent beim „Matching", was die Zahl der notwendigen Vorschläge bis zu einer Vermittlung drastisch reduziert. Diese Verbesserung steigert nicht nur die Kundenzufriedenheit, sondern stärkt auch den Ruf und die Glaubwürdigkeit von Forum Engineering in der Personalbranche. Das wiederum macht es einfacher, neue Kunden zu gewinnen. Last but not least spart der beschleunigte Platzierungsprozess auch Arbeitszeit und hilft der Agentur (dank der schnelleren Besetzung von Positionen als bei Wettbewerbern), Marktanteile zu gewinnen.

Analytics verbessert Marketingkampagnen

Ein anderes Beispiel: Um die Aufmerksamkeit der Verbraucher zu gewinnen, müssen Anzeigen relevant, zeitnah und überzeugend sein. Nur dann erhalten entsprechende Marketingkampagnen die nötige Aufmerksamkeit im allgemeinen „Hintergrundrauschen". Um das zu erreichen, verwendet die weltweit größte Medieninvestmentgesellschaft, GroupM, maschinelle Lernalgorithmen.

Diese Algorithmen verbessern das Targeting, das Timing und die Platzierung von Anzeigen. Ein Ergebnis des KI-Projektes: bis zu 50 Prozent höhere „Konversionsraten". Das ist ein durchschlagender Erfolg, denn mit diesen Raten wird die prozentuale Anzahl der Besucher einer Website oder eines Online-Shops beziffert, die zu zahlenden Kunden umgewandelt werden konnte.

4 Das technische Fundament

Der Grund für diesen Erfolg liegt auf der Hand. Einer der Schlüssel zu effektivem Marketing ist es ja, die richtigen Botschaften zur richtigen Zeit an die Zielgruppe zu liefern. Zum Beispiel wird eine Fastfood-Anzeige besser wirken, wenn sie ihre Leser kurz vor Arbeitsbeginn anspricht und dazu verleitet, sich beim Essen zu entspannen, anstatt den Abend in der Küche zu verbringen.

GroupM verwendet deshalb fundierte Analysen des Kundenverhaltens, um zu entscheiden, wann und wo Anzeigen platziert werden sollen. Aufgrund des wachsenden Umfangs der zu analysierenden Werbedatensätze und der zunehmenden Geschwindigkeit, mit der Entscheidungen in der Welt der Online-Werbung fallen müssen, ist es wichtig, diese Analysen so automatisiert wie möglich durchzuführen.

Den Energieverbrauch von Gebäuden schneller und genauer vorhersagen

Letztes Beispiel: BlocPower, ein kleines Start-up mit Sitz in Brooklyn/USA, vermarktet und finanziert ein Portfolio erneuerbarer Energien und Energieeffizienztechnologien für Kirchen, Schulen, kleine Unternehmen und gemeinnützige Organisationen in unterversorgten Innenstädten der USA. Als eines der ersten Unternehmen weltweit hat BlocPower IBM Data Science Experience eingeführt, um Informationen über die Gebäude von Kunden zu analysieren, denn der Energieverbrauch ineffizienter Gebäude erhöht die Betriebskosten eklatant.

Um Gebäudeeigentümer als neue Kunden zu werben, schickte BlocPower früher Ingenieure, die das Gebäude besichtigten und herausfinden sollten, wo und wie Energie verschwendet wird und welche Sicherheitsrisiken es gibt. Dabei galt es, zwei Herausforderungen zu meistern: Erstens war der Datenerfassungsprozess sehr arbeitsintensiv und konnte daher nicht skaliert werden, um den schnell wachsenden Kundenstamm zu unterstützen. Zweitens gab es keinen effizienten Weg, die gesammelten Daten zu nutzen.

Heute ist das anders. Mit IBM Data Science Experience wurden Schlüsselvariablen ermittelt, die den Energieverbrauch beeinflussen. So kann BlocPower den Energieverbrauch eines Gebäudes schneller und genauer vorhersagen. Basierend auf den Daten, die bereits erhoben sind, lassen sich zudem bessere Projektionen machen – zum Beispiel auf Basis einer Stichprobe von 300 Gebäuden, um den Energieverbrauch eines konkreten Gebäudes in New York City zu schätzen. BlocPowers Datenwissenschaftler sind bei der Erforschung komplexer Datensätze rund zehn Prozent effektiver geworden – und die Entwicklungsdauer statistischer Modelle wurde von Wochen auf Tage verkürzt.

Das sind nur einige wenige Beispiele für die Erfolge von Pionieren der Digitalisierung und die gezielte Nutzung von Daten. Solche Beispiele überzeugen immer mehr Unternehmer, denen IBM Cloud Private for Data die Umsetzung ihrer Ideen nun viel einfacher macht.

5 Standards in Arbeit

Die Open-Source-Standards Linux, Docker und Kubernetes werden laufend weiterentwickelt. All diese Weiterentwicklungen berücksichtigt IBM gemäß ihrer Cloud Computing Reference Architecture (CCRA), in die auch andere wichtige Standardisierungsbestrebungen einfließen. CCRA soll das reibungslose Zusammenspiel der Cloud Services untereinander und ihre plattformübergreifende Arbeit mit den Daten langfristig sicherstellen. Deshalb fließen in CCRA auch die Konzepte von ISO und anderen Gremien ein.

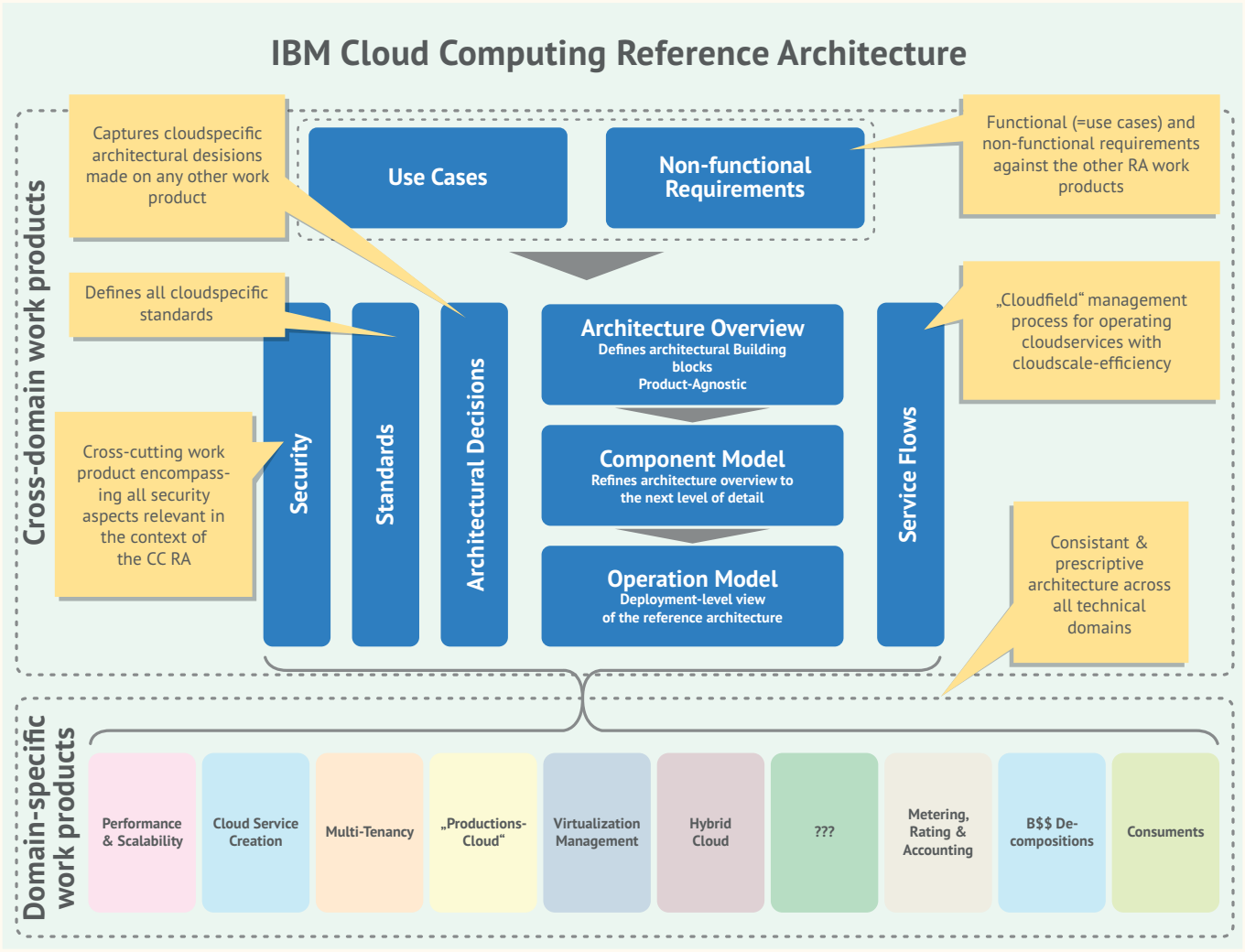


Abbildung 11: *Die IBM Cloud Computing Reference Architecture (CCRA) (Quelle: IBM)*

Eine der klarsten Definitionen des grundsätzlichen Aufbaus von Cloud-Infrastrukturen stammt vom „National Institute of Standards and Technology" (NIST). Auch andere Standards wie ITIL und COBIT spielen hier eine entscheidende Rolle, denn eine Voraussetzung für funktionierende Cloud-Strategien sind klare Compliance-Richtlinien, exakt definierte Services und ein vollständiger Servicekatalog.

Auch die nationalen und internationalen Normungsgremien liefern wichtigen Input für die Entwicklungsarbeit von IBM. Beiträge dazu leisten u. a. auch das Deutsche Institut für Normung mit der DIN SPEC 92222 als Referenzmodell für die industrielle „Cloud Federation" oder die International Organization for Standardization mit der Norm ISO/IEC 19944, die den Datenfluss über Geräte und Cloud-Dienstleistungen festlegt.

6 Ausblick

Eines ist klar: Datensammeln ist längst kein Privileg von Behörden und Großkonzernen mehr, sondern wird auch ein Thema für den Mittelstand. Bei den Einsatzszenarien sind der Fantasie kaum Grenzen gesetzt; sie reichen von der vorausschauenden Maschinenwartung in der Fabrik über die Verbesserung des Kundenservice bis hin zur personalisierten Werbung.

Dabei werden nicht nur unternehmenseigene Daten genutzt. Es entstehen bereits erste Börsen wie die 2015 in Frankreich gegründete Dawex oder das belgische Unternehmen Databroker, die aktiv mit Daten handeln. Hier können Unternehmen IoT-Daten verkaufen und sich etwas hinzuverdienen – sie können aber auch Zeit sparen bei der Datenbeschaffung und Daten zukaufen.

Genutzt werden die gekauften Daten zur praxisnahen Kalibrierung der Simulations- und Prognose-Modelle, zum „Anlernen" smarter Assistenten oder zur Konfiguration von Machine-to-Machine-Lösungen (M2M). Die gemeinsame Nutzung der IoT-Daten ist eine Win-Win-Lösung, denn davon profitieren alle – die Betreiber der Maschinen, deren Hersteller sowie die Software- bzw. Netzwerk- und Cloud-Provider. Die Marktforscher sehen regelrechte IoT-Ökosysteme rund um das innovative Geschäftsmodell „Data as a Service" (DaaS) entstehen, in denen Produzenten und Konsumenten die IoT-Daten gemeinsam nutzen.

„Sharing" der IoT-Daten

Dieses „Sharing" der IoT-Daten verstärkt den Mehrwert von Szenarien der Industrie 4.0 – vorausgesetzt, es basiert auf einem sicheren, kontrollierten Datenaustausch. Dann lassen sich die IoT-Daten nicht nur monetarisieren, sondern sie können allen Teilnehmern des Ökosystems langfristig wirtschaftliche Vorteile bringen. Initiatoren dieser Ökosysteme können Maschinenhersteller oder Vertreter der Sensorikbranche sein, aber auch Größen der IT-Branche.

Meistens haben diese Initiatoren bereits eine eigene „Datenpolitik" definiert. IBM beispielsweise schreibt darin bezogen auf die Datenverarbeitung mit dem cloud-basierten KI-System Watson: „To reap the societal benefits of cognitive, we will first need to trust it. We have created a system of best practices that guide the safe and ethical management of Watson [...]; a system that includes contracts and disclosures that help foster full transparency; a strategy that reflects our compliance with existing legislation and policy; and a framework that protects privacy and personal data."

Zum Thema „Data Ownership" heißt es bei IBM, dass der Kunde keinerlei Rechte an seinen Daten abtreten müsse; man biete vielmehr etliche Optionen für den fairen Umgang mit diesen Daten, bis hin zu ihrer Isolierung und zur Wahrung der Vertraulichkeit bei sämtlichen daraus gewonnenen Erkenntnissen. IBM werde diese Daten und Erkenntnisse nicht ohne Zustimmung des Kunden mit anderen Unternehmen teilen.

„Selbstverständlich gehören IBM die mit Watson verarbeiteten Daten nicht – und sie werden auch nicht gespeichert", heißt es weiter in dieser „Data Policy". Wer aber Watson in der IBM Cloud nutze, könne seine eigenen Datensätze mit anderen kombinieren, die entweder IBM gehören oder aber lizensiert bzw. „Open Data" sind.

6 Ausblick

Wachsende Bedeutung von Daten
Angesichts dieser Bedeutung von Daten stehen Unternehmen auf der ganzen Welt vor der Herausforderung, ihre Datenarchitekturen zu modernisieren, um mit den wachsenden Anforderungen an Daten und Analysen Schritt halten zu können. Dabei leistet die IBM Cloud Private for Data wertvolle Hilfestellung, indem sie die Fähigkeit der Unternehmen verbessert, datenorientiert zu arbeiten, und so die Kosten senkt, die Produktivität steigert und auch bessere Reaktionen auf Kundenwünsche oder Marktrends erlaubt.

Die gut integrierte Sammlung von Mikroservices, die auf einer Cloud-nativen Architektur basiert, liefert auch den Zugriff auf zweckgerichtete Daten und dedizierte Governance- und Analysefunktionen, die jeweils speziell auf Datenwissenschaftler, Geschäftsanwender, Dateningenieure, IT-Manager, Datenverwalter oder Anwendungsentwickler zugeschnitten sind.

Das fördert die Zusammenarbeit und Kommunikation innerhalb und außerhalb des Unternehmens, beseitigt Barrieren durch überkommenes Abteilungsdenken bzw. sperrige Datensilos – und schafft letztendlich gute Voraussetzungen für eine KI-getriebene Anwendungsentwicklung.

Auf diese Weise reduziert IBM nachhaltig die Probleme, die aus der „Anziehungskraft" der Daten resultieren. Denn die IBM Cloud Private for Data-Plattform versetzt Unternehmen in die Lage, Analysen überall dort durchzuführen, wo Daten bereits vorhanden sind – in einem IoT-Gerät, in einer privaten Cloud oder in Public Clouds wie Amazon Web Services, Microsoft Azure oder IBM Cloud. Und das, ohne die Daten zuvor in einen zentralen „Data Lake" verschieben zu müssen. So wird IBM Cloud Private for Data auch zum Sprungbrett für den Fall, dass ein breiterer (öffentlicher) Cloud-Einsatz geplant wird.

7 Literatur

„Was die Hybrid-Cloud zusammenhält“, TDWI e-Book, März 2018
https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IMW14951DEDE

“Reshaping Business With Artificial Intelligence”, MIT Sloan in Zusammenarbeit mit Boston Consulting, 6. September 2017
http://sloanreview.mit.edu/offers-ai2017/

“Breach Level Index 2017 Full Report”; Gemalto,
https://www6.gemalto.com/breach-level-index-2017-full-report?utm_campaign=breach-level-index&utm_medium=blog&utm_source=pardot

“The Journey Continues: From Data Lake to Data-Driven Organization”, IBM Redguide, 19. Februar 2018
http://www.redbooks.ibm.com/abstracts/redp5486.html?Open

“Governing and Managing Big Data for Analytics and Decision Makers”, IBM Redguide, 26. August 2014
http://www.redbooks.ibm.com/redpieces/abstracts/redp5120.html

“Designing and Operating a Data Reservoir”, IBM Redbook, 26. Mai 2015
http://www.redbooks.ibm.com/abstracts/sg248274.html?Open

“IBM Decision Optimization and Data Science”, IBM Redbooks Analytics Support Web Doc, 6. Dezember 2017
http://www.redbooks.ibm.com/abstracts/tips1357.html?Open

“Managing IBM Db2 Analytics Accelerator by using IBM Data Server Manager”, IBM Redbooks Analytics Support Web Doc, 6. Dezember 2017
http://www.redbooks.ibm.com/abstracts/tips1356.html?Open

“IBM Spectrum Scale: Big Data and Analytics Solution Brief”, IBM Redguide, 23. Januar 2018
http://www.redbooks.ibm.com/abstracts/redp5397.html?Open

“Hortonworks Data Platform with IBM Spectrum Scale: Reference Guide for Building an Integrated Solution”, IBM Redpaper, Juli 2017
http://www.redbooks.ibm.com/abstracts/redp5448.html?Open

“Building a data reservoir to use big data with confidence”, Mandy Chessell, Distinguished Engineer, IBM Analytics Group, 12. Oktober 2015
http://www.ibmbigdatahub.com/blog/building-data-reservoir-use-big-data-confidence

“From Data Lakes to Data Cakes”, Andrew White, Gartner Group, 14. Februar 2017
https://blogs.gartner.com/andrew_white/2017/02/14/from-data-lakes-to-data-cakes/

„Data-driven Business Models in Connected Cars, Smart Mobility & Beyond“, BVDW & Accenture, April 2018
https://www.bvdw.org/fileadmin/user_upload/20180509_bvdw_accenture_studie_datadrivenbusinessmodels.pdf

“IBM Industry Model support for a data lake architecture”, IBM, 2016,
https://public.dhe.ibm.com/common/ssi/ecm/im/en/imw14877usen/IMW14877USEN.PDF

“ITIL and cloud series: NIST and IBM Cloud Reference Architecture, what and how to”, 21, Oktober 2011
https://www.ibm.com/blogs/cloud-computing/2011/10/21/itil-and-cloud-series-nist-and-ibm-cloud-reference-architecture-what-and-how-to/

“GroupM Nordic: Embracing data science to increase marketing campaign conversion rates by up to 50 percent”, 2016
https://www.ibm.com/case-studies/GroupM_Nordic

“Forum Engineering Inc. Cognitive staffing solution removes subjectivity from the matching process, leveling perceptual bias”
https://www.ibm.com/case-studies/e447008m75681e05

“BlocPower uses data science to give insights to building owners to invest in energy upgrades and lower costs”

“Reference Architecture Model for the Industrial Data Space”, Fraunhofer 2017
https://www.fraunhofer.de/content/dam/zv/de/Forschungsfelder/industrial-data-space/Industrial-Data-Space_Reference-Architecture-Model-2017.pdf

“Query many data sources as one: IBM Queryplex for data analytics”, Sam Lightstone, IBM, März 2017
https://www.ibm.com/blogs/bluemix/2017/03/query-many-data-sources-one-ibm-queryplex-data-analytics/

“Big-brained data, Part 2: Apply the software development lifecycle to the data that feeds AI applications”, Uche Ogbuji, Oktober 2017
https://www.ibm.com/developerworks/library/cc-cognitive-big-brained-data-pt2/index.html

“Adding MongoDB to the IBM enterprise database ecosystem”, Mike Connor, IBM, Juni 2018
https://www.ibmbigdatahub.com/blog/adding-mongodb-ibm-enterprise-database-ecosystem

“IBM Defies Data Gravity”, Mitch Wagner , Light Reading, September 2018
https://www.lightreading.com/enterprise-cloud/iot-and-edge/ibm-defies-data-gravity/d/d-id/746071

E-Book

tdwi.eu