DataOps & MLOps: Daten intelligenter nutzen

_Durch KI generiertes Bild. Prompts: elektronische Geräte, die in einen riesigen Trichter fallen, digitale Kunst.

Fachartikel

_Heiko Faller, Managing Director Enterprise Solutions und Mitglied der Geschäftsleitung
_Dr. Wilhelm Kleiminger, Head of Data Science
Ergon Informatik

Erschienen im SMART insights 2023 Magazin

ergon smart people – smart software®

Für viele Unternehmen hat sich Big Data als Flop erwiesen. Der Aufwand für die Datenschatzsuche ist gross, der Mehrwert eher klein. Hier setzen MLOps und DataOps an: Mit effizienten Prozessen schaffen sie aus Rohdaten wirkliche Wettbewerbsvorteile.

In den letzten Jahren haben viele Unternehmen grosse Datenmengen gesammelt. Die Ergebnisse sind jedoch grösstenteils ernüchternd. Ob im KMU, im Gesundheitswesen oder in der Finanzbranche, oft verursacht der vermeintliche Datenschatz hohe Kosten mit wenig echtem Ertrag. Das muss nicht sein: Richtig eingesetzt, schafft der Data-Centric-Ansatz einen deutlichen Mehrwert im Geschäftsalltag und verkürzt die Time-to-Market von neuen Produkten. «DataOps plus MLOps» heisst die Zauberformel, die mit dezidierten Pipelines und qualitativ hochwertigen Datensätzen kontinuierliche Wettbewerbsvorteile schafft.

DataOps und MLOps umfassen Prozesse, Best Practices und Technologien, mit denen sich Daten und Machine-Learning-Modelle schneller bereitstellen lassen – strukturiert, automatisiert und wiederverwendbar. Laut einer Studie von McKinsey können Unternehmen allein durch DataOps die Produktivität um 10 Prozent steigern, die Time-to-Market gar um 30 Prozent reduzieren. Statt aufwendig riesige Datenschätze auszuheben, erschafft eine DataOps-MLOps-Pipeline die Ausbeute quasi inhouse: Mit den richtigen Prozessen entsteht aus Rohdaten ein echter Wert.

So profitieren Unternehmen von Machine Learning

Mit Machine Learning (ML) finden Unternehmen Lösungen für Probleme, bei denen die traditionelle Software-Entwicklung scheitert. Im Unterschied zu herkömmlicher Software ist ein ML-Modell nicht ausprogrammiert, sondern erlernt seine Funktionalität mit beispielhaften Eingabe- und Ausgabedaten. Das ist besonders in Bereichen hilfreich, bei denen sich die Funktionalität nicht programmatisch beschreiben lässt. Ein Beispiel ist die automatische Spracherkennung: Die vielen Eigenheiten der menschlichen Sprache lassen sich nicht abschliessend in Quellcode abbilden. Heute gibt es zwar zahlreiche ML-Modelle aus unterschiedlichsten Bereichen, die sich mit wenig Aufwand auf einen konkreten Anwendungsfall anpassen lassen. Trotzdem ist der Sprung vom technischen Prototyp zum produktiven Einsatz noch gross. Dementsprechend wichtig sind die Datenaufbereitung, die Auswahl und das Training des ML-Modells.

Strukturiertes Vorgehen dank dedizierter Pipelines

DataOps und MLOps basieren auf dem Konzept von Pipelines. Dabei erfasst das System auf der einen Seite der Pipeline die rohen Daten und gibt auf der anderen Seite fertige Reports, Dashboards und Machine-Learning-Modelle aus. Innerhalb der Pipeline machen wohldefinierte Transformationsschritte die gewünschten Ergebnisse nachvollziehbar und reproduzierbar. In der Praxis bedeutet dies eine effiziente, transparente und beschleunigte Datenaufbereitung.

Ein Beispiel: Ein Unternehmen entwickelt ein eigenes Voice-User-Interface, um eine Applikation

mit Sprachbefehlen zu bedienen. Dazu passt das Unternehmen mithilfe einer MLOps-Pipeline ein bestehendes Sprachmodell mit domänenspezifischen Daten auf den eigenen Anwendungsfall an. Zunächst liest es dazu Trainingsdaten ein. Je nach Anforderung transformiert es diese Daten, beispielsweise von der Zeit- in die Frequenzdomäne. Anschliessend reichert es die Daten zusätzlich an. Es maskiert zum Beispiel bestimmte Frequenzen und entfernt so störende Nebengeräusche.

Mit diesen neu aufbereiteten Daten lässt sich ein bestehendes Sprachmodell auf den neuen Anwendungsfall anwenden. Ein solches Transfer Learning bringt auch mit wenig Trainingsmaterial gute Ergebnisse – ein klarer Effizienzgewinn.

Datenqualität statt -quantität
Anwendungsspezifische Daten zu sammeln, ist aufwendig und teuer. Darum ist die Datenqualität meist wichtiger als die Quantität. Wichtig sind hier Metainformationen, die die Rahmenbedingungen bei der Datenaufnahme dauerhaft nachvollziehbar machen. So kann man diese im Training eines Modells berücksichtigen. Für ein Sprachmodell zum Beispiel wirken sich Umgebungsvariablen wie das Wetter, das genutzte Gerät oder das Geschlecht der Nutzer:innen auf die Genauigkeit eines Modells aus.

DataOps bietet die nötigen Konzepte und Tools, um Daten aus unterschiedlichen Quellen zusammenzuführen und vorzubereiten. Eine DataOps-Pipeline kann Datensätze aus verschiedenen Quellen einlesen, filtern und in ein einheitliches Format bringen. Zudem ermöglicht sie das Visualisieren von verschiedenen Statistiken, zum Beispiel der Worthäufigkeit oder der Geschlechterzusammensetzung der Sprecher:innen. So hilft eine DataOps-Pipeline auch bereits beim Sammeln von eigenen Datensätzen, jederzeit den Überblick zu behalten.

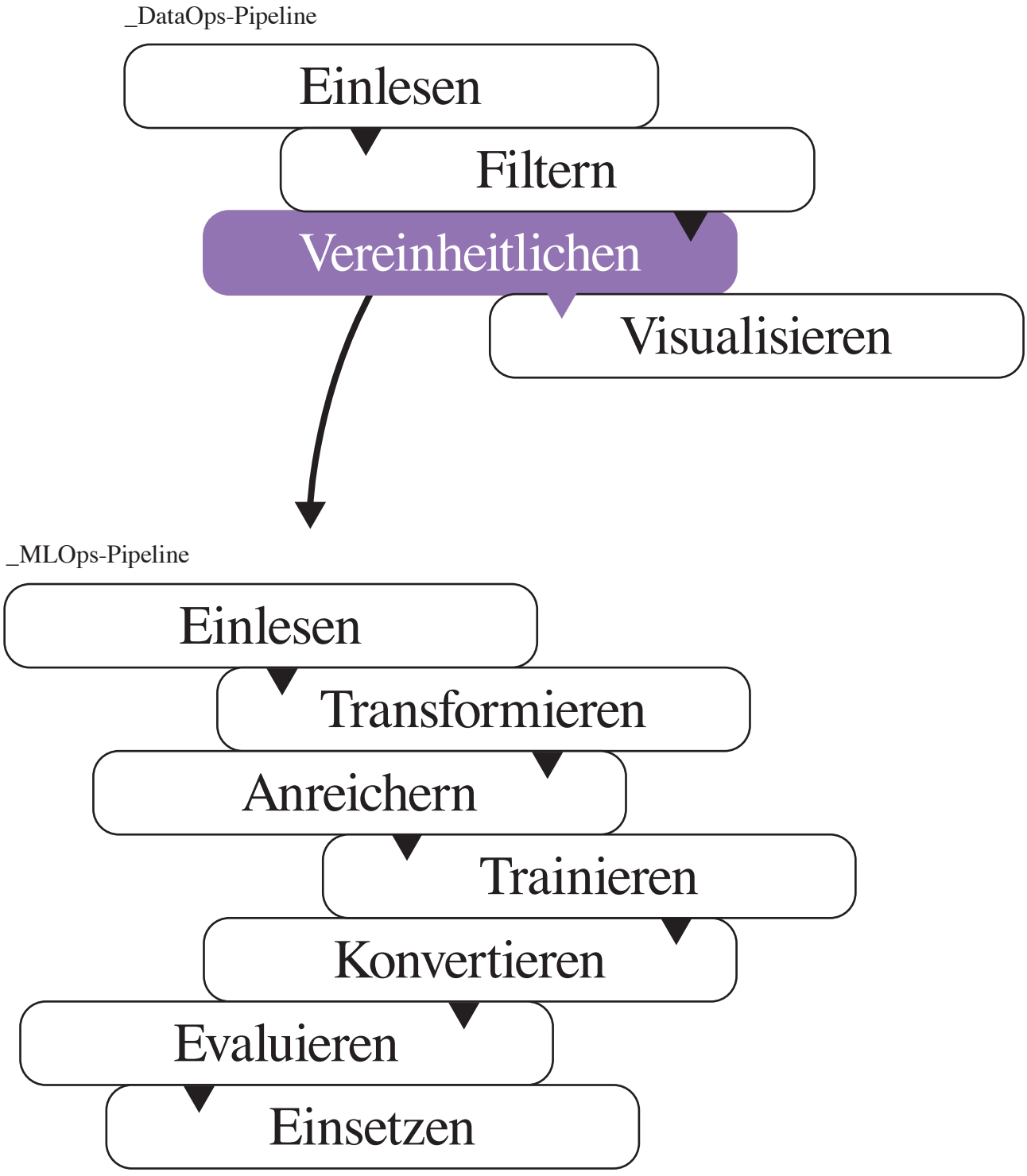
Daten und Modelle müssen verwaltet werden
Gemäss Robert C. Martin, Autor von «Clean Code», kann die Wahrheit in der traditionellen Software-Entwicklung nur im Code gefunden werden. Bei ML-Modellen hingegen ist die «Wahrheit» verteilt: in dem gewählten ML-Algorithmus, den genutzten Trainingsdaten und in allfälligen Konfigurationsparametern. Um Resultate zu verstehen und zu reproduzieren, müssen die einzelnen Eingabe- und Ausgabeartefakte sauber versioniert und verwaltet sein. Es muss immer nachvollziehbar sein, welche Datensätze im Training zum Einsatz gekommen sind – und welche nicht. Nur so kann das Modell auf veränderte Anforderungen eingehen, wie zusätzliches Vokabular oder neue Daten durch Umwelteinflüsse, zum Beispiel Hintergrundgeräusche.

Dauerhaft in Produktivumgebungen eingebunden
Für die Integration eines Modells in die Produktivumgebung braucht es vielleicht weitere Anpassungen. Soll das Sprachmodell zum Beispiel nicht in der Cloud, sondern auf einem mobilen Endgerät ausgeführt werden, braucht es in der MLOps-Pipeline einen zusätzlichen Schritt für die Konvertierung des Modells. Dieses konvertierte Modell muss ausserdem noch einmal getestet werden. Dies stellt sicher, dass beim Konvertieren keine signifikanten Performance-Einbussen aufgetreten sind.

Der Feldbetrieb bringt Erfahrungen, die womöglich neue Veränderungen am Modell nach sich ziehen. Häufig zeigt sich erst im produktiven Einsatz, welche Randbedingungen noch zu wenig berücksichtigt sind. Um ein Modell kontinuierlich zu verbessern, müssen die Produktivdaten entsprechend verwendbar sein. Für ML-Modelle ist zudem ein eigener Release-Zyklus sinnvoll, der unabhängig vom Release-Zyklus der eigentlichen App läuft. Wenn sich eine Veränderung in den Daten abzeichnet, kann man so rechtzeitig reagieren.

Datenprojekte: «Start small and agile»
Unternehmen, die DataOps/MLOps nutzen wollen, starten am besten mit einem klaren, überschaubaren Business Case in einem kleinen Team. Dieses Team ermittelt, welche Schritte tatsächlich

_Strukturiertes Vorgehen
dank dedizierter Pipelines



«DataOps und MLOps machen Unternehmen im Zeitalter von KI wettbewerbsfähig.»
_Heiko Faller, MD Enterprise Solutions und Mitglied der Geschäftsleitung, Ergon

«DataOps und MLOps sind Prozesse, die ständig weiter verbessert werden müssen.»
_Dr. Wilhelm Kleiminger, Head of Data Science, Ergon

erforderlich sind, um eine produktive Lösung zu entwickeln. Ohne klares Konzept einen firmenweiten Ort mit strukturierten und unstrukturierten Daten anzulegen, ist kostspielig und oft nicht zielführend. Hier helfen Full Stack Data Scientists, auf das Wesentliche zu fokussieren. Sie unterstützen die notwendigen Schritte vom Requirements Engineering bis zum Produktivsetzen der Lösung. Wie bei der agilen Software-Entwicklung gilt es, schnell eine erste produktive Version in Betrieb zu nehmen. Damit lassen sich wertvolle Erfahrungen sammeln.

Ein solcher Ansatz ist leicht umsetzbar und belegt mit jedem weiteren gelungenen Projekt unmittelbar seinen Nutzen. Wenn Entwicklungsteams in verschiedenen Datensilos des Unternehmens arbeiten, ist es wichtig, dass sie sich austauschen. So entwickelt sich eine nachhaltig wirksame DataOps/MLOps-Kultur: In ihr geht es nicht um die gewählten Tools oder einen One-size-fits-all-Ansatz – sondern um klare Prozesse und klare Business Cases.

Besonderes Augenmerk gilt den Daten, die für die Abdeckung des Anwendungsfalls notwendig sind. Hier ist eine hohe Datenqualität gefragt. Dank strukturierten DataOps-/MLOps-Pipelines werden alle Prozessschritte, von der Datenerfassung bis zur produktiven Integration des ML-Modells, nachhaltig und nachvollziehbar durchgeführt. Das Vorgehen konvergiert und bietet somit mehr Erfolgspotenzial als eine Big-Data-Schatzsuche bei deutlich tieferen Kosten. />

DataOps und MLOps kurz erklärt

DataOps verbessert die Effizienz von Datenpipelines. DataOps-Werkzeuge und -Methoden überwachen die Datenqualität und machen Datenverarbeitungssysteme skalierbarer. MLOps, auch bekannt als DevOps für Machine Learning (ML), legt den Fokus auf das Entwickeln, das Bereitstellen und das Überwachen von ML-Modellen. MLOps-Methoden implementieren, überwachen und optimieren ML-Modelle in produktiven Systemen. Beide Ansätze fördern die Zusammenarbeit zwischen Datenwissenschaftler:innen und IT-Teams.