Datenanalyse für die Chefetage

Data Lakehouse wird strategisches Analytics-Instrument

In den Führungsetagen von Unternehmen haben die Entscheiderinnen und Entscheider längst erkannt, dass in vorhandenen Daten viel strategisches Potenzial brachliegt. Die Diskussion, auf welcher Dateninfrastruktur Vorstände unternehmensrelevante Entscheidungen treffen können, führt direkt zur Datenhaltung im Data Lakehouse – einer aktuellen Weiterentwicklung des klassischen Data Warehouse.

Bildquelle: Bigstockphoto

*Von Dr. Thomas Wörmann**

„Gerade die Versicherer, die stolz sind, viele Mathematiker zu beschäftigen, schöpfen zunehmend Mehrwert aus Daten – und das mithilfe von Machine Learning und Statistik“, erläutert Dr. Sarah Detzler, Competence Lead Data Science and Machine Learning bei SAP. Allerdings muss jedes Unternehmen seinen eigenen Weg finden, um etwa nutzenstiftende KI-Use-Cases sichtbar zu machen, Kunden und Mitarbeiter bei der Datenstrategie mitzunehmen und Know-how sowie Data-Science-Infrastrukturen aufzubauen – und vor allem Vorstände zu befähigen, auf dieser Grundlage strategisch relevante Entscheidungen zu treffen. Eine Frage ist, welchen Beitrag Business Analytics in der Zukunft bei der Bewertung langfristiger Daten- und auch Firmenstrategien liefern wird.

Bildquelle: SAP

Dr. Sarah Detzler, Competence Lead Data Science and Machine Learning bei SAP:

„Gegebenheiten ändern sich, es kommen neue Daten hinzu, und ein Modell kann plötzlich ein paar Prozentpunkte schlechter performen als am Anfang. Man muss hinterfragen, in welcher Geschwindigkeit neue Daten einfließen und wann sich gewisse Prozesse ändern. Auf dieser Grundlage kann man das Modell neu trainieren und auf die neue Datenlage anpassen.“

Daten, Wissen, Handlungsempfehlungen

Die dafür erforderlichen betriebswirtschaftlichen Zusammenhänge werden bereits seit Langem in multidimensionalen Datenmodellen vorgedacht, Bewertungskriterien in geeigneten Kennzahlen formalisiert und – das klassische Kon-

Dr. Thomas Wörmann ist Senior Data Analytics Consultant bei IKOR Informationsfabrik.

strukt – als Data Mart eines Data Warehouse in eine geschlossene Lösung überführt.

Das klassische Data Warehouse – mit seinem eingebauten Domänenwissen – stößt allerdings schnell an seine Grenzen: Der operative Aufwand, seine Transformationslogik zu erstellen und weiterzuentwickeln, ist enorm. Aus strategischer Sicht beschränkt die enge Kopplung der fachlichen Logik an den dimensionalen Modellansatz die analytische Breite des Modells. Problemstellungen, die über den bekannten Wirtschaftskontext hinausgehen, lassen sich in diesem Rahmen kaum sinnvoll analysieren. Fragen zu künftigen Szenarien sind nur unzureichend behandelbar – etwa, wenn das Marketing neue Kundensegmente oder ein erweitertes Produktspektrum anvisiert. Weder lassen sich große Mengen analytischer Daten effizient und kostengünstig speichern noch die Ergebnisse neuer analytischer Methoden erfolgreich einbinden.

Modellbasierte Beschreibungen versus Prozessbewertung des Data Warehouse

Mit dem neuen Paradigma von Data Science – vor allem dem Methodenspektrum der künstlichen Intelligenz und des Machine Learning – stehen völlig neue Werkzeuge zur datengestützten Wissensverarbeitung bereit. Diese sind gerade bei unvollständigem oder nur statistisch vorhandenem Wissen Erfolg versprechend. Viel Potenzial steckt in einer höheren Dunkelverarbeitungsquote und automatisierten Prozessen – etwa in den Versicherungsbereichen Schaden oder Antrag. Rund um neue Geschäftsmodelle sind oft KI-Lösungen wie telematik-, daten- und nutzungsbasierte Tarife im Einsatz. Wahrscheinlichkeiten wie Next Best Action liefern nicht nur dem Marketing einen schnellen Return on Investment. Insbesondere modellbasierte Beschreibungen weichen beim Data-Science-Ansatz erheblich von der kennzahlengetriebenen Prozessbewertung des klassischen Data Warehouse und seinen fachbezogenen Data Marts ab.

Die analytische Breite für maschinelle Wissensbildung ist meist erheblich größer: Das erforderliche Domänenwissen entsteht im Rahmen der Modellbildung zusammen mit der verwendeten Datenbasis. Das dafür notwendige explorative Vorgehen erfordert es allerdings, unterschiedliche Datenmengen einfach und schnell bereitzustellen.

Die mit Machine Learning entwickelten Modelle sind dadurch eng an die zugrundeliegenden Datenstrukturen gekoppelt, als sogenanntes Datenprodukt – einer komplexen Kombination aus Datengrundlage, Machine-Learning-Methodik und fachlichem Analysewissen. Diese Kapselung als Datenstruktur mit eigenem Lebenszyklus ist in einem klassischen Data Warehouse – selbstredend mit relationaler Datenhaltung, Schwerpunkt auf Datenqualität und -konsistenz sowie mangelhafter Skalierbarkeit – nur schwer darstellbar; Abhilfe schafft eine ubiquitäre Datenbevorratung per Data Lake.

Der Data Lake wird heute von allen Cloud-Anbietern als eigenständige Infrastruktur angeboten – ein aus dem Hadoop-Framework abgeleiteter Ansatz. Er ermöglicht eine offene, skalierbare und kostengünstige Datenhaltung und unterstützt den Einsatz von Machine Learning mit der präventiven Speicherung möglichst aller verfügbaren Daten. Mit einer der iterativen Arbeitsweise angepassten Schema-on-Read-Architektur werden die Daten zunächst ohne fachliche Transformation in die Datenplattform geladen.

Alle KI-Ergebnisse sollten API-fähig sein

Dann funktionieren auch Machine-Learning-Plattformen, wie bei der Zurich Gruppe Deutschland. Der Versicherer hat eine cloudbasierte State-of-the-Art-KI-Landschaft aufgebaut – auf Basis eines Hyperscalers: „Wir ziehen die Plattform hoch und bauen passende Git-Repositories für Datenbewirtschaftung und MLOps unserer Modelle. Zusätzlich ist jede KI-Anwendung in einer Function oder einem Container gekapselt. Dadurch können wir bis zu drei unterschiedliche Versionen parallel vorhalten. Zudem haben wir klare Namenskonventionen aufgesetzt und mit unserem Delivery Center in Barcelona Servicelevel definiert. Dieses analysiert zum Beispiel, ob unsere Container und Anwendungen lebendig sind“, erläutert Dr. Michael Zimmer, Chief Data Officer bei der Zurich Gruppe Deutschland. „Unsere Berechtigungskonzepte sind datenschutzkonform. Alle abnehmenden Systeme werden über Schnittstellen versorgt; alle KI-Ergebnisse sind also API-fähig. Die Daten halten wir in einem Data Lake vor.“ Bei Dateninhalt und Datenstruktur ist der Data Lake aber mehr oder weniger genau. Sollen die Daten in einem fachlichen Kontext genutzt werden, ist eine allen Daten gemeinsame Strukturgebung erforderlich.

Data Lake mit zentraler Zugriffsschicht

Diese Struktur wird mit dem Ansatz des Delta Lake – als konzeptionelle Erweiterung des Data Lake – erstellt und verwaltet. Der Delta Lake enthält zusätzlich eine zentrale, metadatengesteuerte Zugriffsschicht. Damit vereint er bekannte Methoden und Werkzeuge der Datenaufbereitung mit den neuen Tech-

Open Source spielt wichtige Rolle für Data Analytics

Quelloffene Standards sind auch bei der Arbeit mit Datenplattformen und in der Data Science entscheidend für den Erfolg. Nach der anfänglichen Ernüchterung beim Einsatz des in seiner Komplexität unterschätzten Hadoop-Framework entsteht jetzt eine Vielzahl cloudbasierter Platform-as-a-Service-Infrastrukturen auf Basis des Spark-Framework. Im maschinellen Lernen und bei künstlichen neuronalen Netzen findet eine zunehmende Konzentration auf wenige Open-Source-Standards statt, die meist aus Eigenentwicklungen der großen Player an Open-Source-Organisationen wie die Apache Foundation übergeben werden.

Bildquelle: Dataiku

Philipp Schützbach, Sales Engineer bei Dataiku:

„Wer ein Machine-Learning-Modell produktiv setzt, muss den Lebenszyklus des Modells aktiv managen – also wissen, wann und wie ein Modell neu trainiert werden muss, und er sollte verstehen, warum sich das Modell so verhält, wie es sich verhält.“

nologien: So wird SQL, als präferiertes Werkzeug des Data Engineer bei der Implementierung der Prozesslogik ebenso unterstützt wie eine Data-Frame-Schnittstelle als bevorzugte Datenstruktur im Data-Science-Kontext.
Eine gemeinsame semantische Datensicht mit unterschiedlichen Repräsentationen senkt die Redundanz in der Prozesslogik. Sie erleichtert zudem die Einbindung des technischen Know-hows. Aktuelle Entwicklungen des Delta Lake versprechen nicht nur hochperformante Zugriffe über SQL, sondern auch Zusagen bei Transaktions- und Integritätssicherheit im Rahmen des ACID-Modells.

Zukunft gehört offenen Transaktionsplattformen

Der Delta Lake erlaubt es, ubiquitäre analytische Plattformen in Richtung eines Data Lakehouse weiterzudenken. Dahinter verbirgt sich das Konzept einer organisationsübergreifenden Sicht auf den Datenbestand ohne technologische Einschränkungen. Organisationen nebst Führung können ihr Augenmerk wieder auf die fachliche Datenintegration in einem gemeinsamen Informationsmodell legen. Die Beteiligten sehen Daten in ihrer bevorzugten Repräsentationsform. Aufwendige Transformationsprozesse wie Data Governance, Datenqualität oder Single Point of Truth, werden da, wo sie notwendig sind, zentral bereitgestellt und gepflegt. Genau das ist in der schnelllebigen VUCA-Welt unerlässlich: „Gegebenheiten ändern sich, es kommen neue Daten hinzu, und ein Modell kann plötzlich ein paar Prozentpunkte schlechter performen als am Anfang. Man muss hinterfragen, in welcher Geschwindigkeit neue Daten einfließen und wann sich gewisse Prozesse ändern. Auf dieser Grundlage kann man das Modell neu trainieren und auf die neue Datenlage anpassen“, erläutert Detzler. „Ein Machine-Learning-Modell ist kein Selbstläufer“, bekräftigt auch Philipp Schützbach, Sales Engineer beim KI-Hersteller Dataiku. „Wer ein solches produktiv setzt, muss zum einen den Lebenszyklus des Modells aktiv managen, sprich: wann und wie ein Modell neu trainiert werden muss. Zum anderen sollte er oder sie auch verstehen, warum sich das Modell so verhält, wie es sich verhält.“


Bildquelle: Zurich Gruppe Deutschland

Dr. Michael Zimmer, Chief Data Officer bei der Zurich Gruppe Deutschland:

„Wir ziehen die KI-Plattform hoch und bauen passende Git-Repositories für Datenbewirtschaftung und MLOps unserer Modelle. Zusätzlich ist jede KI-Anwendung in einer Function oder einem Container gekapselt. Dadurch können wir bis zu drei unterschiedliche Versionen parallel vorhalten.“

Über den Use-Case-Mehrwert einzelner KI-Anwendungen hinausgehen

Wissen aus Daten zu ziehen und Vorstände zu befähigen, wissensbasierte Entscheidungen zu treffen, bleibt bis auf Ausnahmen auch in der Zukunft aufwendig. Eine rein technologiezentrierte Sichtweise führt allerdings nicht zum Ziel. Solange es keine wirklichen Alternativen zu der geschäftsprozessorientierten Datenanalyse gibt, wird das Konzept des Data Warehouse weiterhin Bestand haben.
Dennoch steht das Data Lakehouse mit seiner gemeinsamen semantischen Zugriffsschicht für einen Evolutionsschritt in Richtung einer umfassend verfügbaren Analyseplattform. Sie hilft dabei, Business Analytics im Rahmen neuer Geschäftsmodelle strategisch neu zu bewerten. Die eigentliche Revolution und Evolution findet in den Köpfen der Anwender statt – Top-down, Bottom-up und für alle Stakeholder. Versicherungen scheinen hier gut aufgestellt zu sein: „Sie treffen Vorhersagen, sind zahlenaffiner und weiter in Predictive Analytics als andere Industrien“, begründet Schützbach. „Der Branchenfokus liegt häufig lediglich auf dem Mehrwert eines spezifischen Use Case und nicht auf dem Mehrwert, den eine Plattform generieren kann. Der Plattformgedanke ist noch nicht gereift – das Bewusstsein, dass Kollaboration, Wiederverwendbarkeit und Integration die Businessbereiche fördern und Unternehmen dabei helfen, Hunderte produktive Use Cases zu skalieren.“ *(cr)* @