VERSICHERER DER ZUKUNFT – MIT DATA SCIENCE ZUM ERFOLG

Erforderliche Erweiterungen der IT-Infrastruktur – Artikel 5

163 Zettabyte – laut einer Prognose des Analystenhauses IDC entspricht dies der weltweit generierten Datenmenge im Jahr 2025, was die zehnfache Menge der in 2016 generierten Daten ist.[1] Bereits die heute erhobenen Datenmengen stellen die bestehende IT-Infrastruktur einer Assekuranz vor neue Herausforderungen.

In unserer Artikelserie Versicherer der Zukunft – mit Data Science zum Erfolg sind wir auf die sich durch Data Science bietenden Möglichkeiten eingegangen. Um die sich daraus ergebenden Chancen nutzen zu können, gilt es, eine zukunftsorientierte Infrastruktur zu schaffen und bestehende Ressourcen zu erweitern.

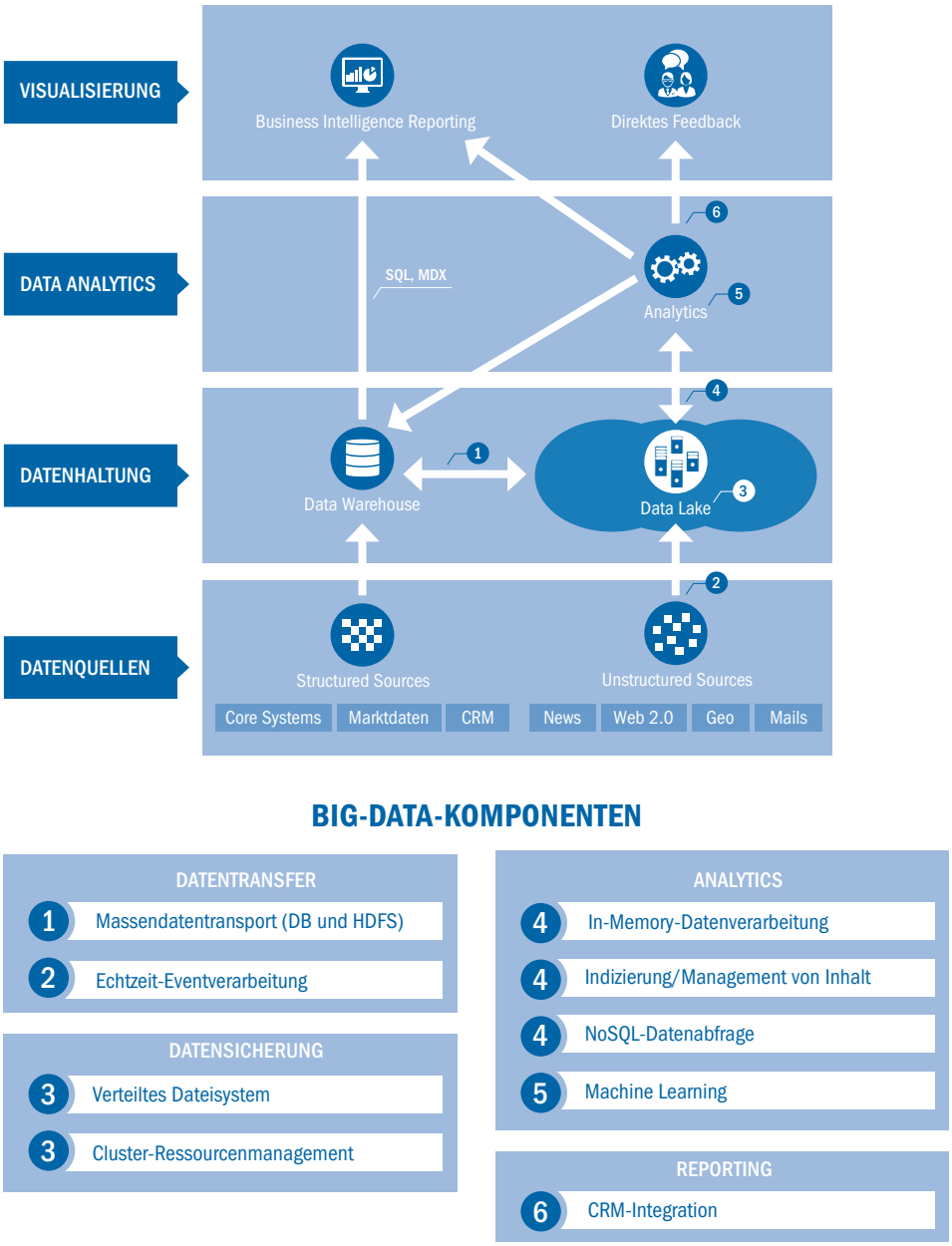

Abbildung 1: Data Science – IT-Komponenten und Aufbau

[1] IDC: Data Age 2025: The Evolution of Data to Life-Critical, April 2017.

Aufgrund der stetig wachsenden Datenmengen müssen Datenbanken skalierbar sein, Anwendungen brauchen mehr Rechenleistung, und neue Technologien zur Datenverarbeitung werden benötigt. Die Vielfalt der Herausforderungen unterscheiden wir im Folgenden nach drei Themengebieten:

- Datenhaushalt
- Datenanalyse & Verarbeitung
- Ergebnisaufbereitung

Datenhaushalt

Um Daten aus unterschiedlichen Quellsystemen, Dateiformaten und Strukturen für Data-Science-Anwendungen verfügbar zu machen, gilt es, diese zur Weiterverarbeitung an einem zentralen Ort vorzuhalten. Dies betrifft einerseits interne Kundendaten oder Prozessdaten, andererseits aber auch externe Daten, welche den bestehenden Datenhaushalt erweitern. Externe Daten können unter anderem kostenpflichtig von spezialisierten Anbietern erworben werden oder sind frei verfügbar. Diese Daten unterteilen sich in:

- Sensordaten (Satelliten, Standorte …)
- Geschäftsprozessdaten (Behördendaten, Transaktionsdaten …)
- Individuelle Aktivitätsdaten (Social-Media-Daten, Nachrichten …)

Abhängig von der jeweiligen Datenquelle weisen diese Daten häufig auch eine unterschiedliche Struktur auf. Sie liegen entweder strukturiert, semi-strukturiert oder unstrukturiert vor. Während strukturierte Daten in einem fest vorgegebenen Format vorliegen, folgen semi-strukturierte Daten nur einer gewissen Grundstruktur. Ein im Zusammenhang mit semi-strukturierten Daten häufiger genanntes Beispiel ist die E-Mail, welche unter anderem aus einer Anrede, einem Textbaustein sowie einer Grußformel besteht. Bei unstrukturierten Daten ist lediglich das Dateiformat bekannt.

Aufgrund der Menge sowie Komplexität der von Versicherern gesammelten Daten werden diese auch als Big Data bezeichnet. Ursprünglich in den 90er-Jahren formuliert beschreibt Big Data das Phänomen von Daten, welche generell sehr groß sind und die Speicherkapazität des Hauptspeichers, lokaler Datenträger und sogar externer Festplatten ausreizen.[2] Die heute gängigste Definition von Big Data basiert auf den sogenannten drei Vs: Velocity, Variety und Volume, zu Deutsch: Geschwindigkeit, Vielfalt und Volumen.

Traditionelle relationale Datenbankmanagementsysteme (RDBMS), wie sie in einem klassischen Data Warehouse zum Einsatz kommen, geraten hierbei schnell an ihre Grenzen, da sie meist vertikal skalierbar sind (CPU, RAM, SSD) und nur eine begrenzte Menge von Daten speichern. Hinzu kommt, dass häufige Transaktionen großer Datenmengen viel Zeit benötigen. Um das Problem von Big Data zu meistern, werden skalierbare Datenbankmanagementsysteme benötigt. Not-only-SQL(NoSQL)-Datenbanken bieten eine Möglichkeit, die Datenmenge verteilt auf Servern abzuspeichern, was eine einfache horizontale Skalierung des Speichers ermöglicht und einen schnellen Zugriff erlaubt.

NoSQL-Datenbanken können jedoch nicht als alleinige Lösung des Problems gesehen werden, sondern vielmehr als Ergänzung zu einem Data Warehouse, da sie im Vergleich zu RDBMS nur einen begrenzten Umfang an Funktionalitäten bieten. Um zusätzlich unstrukturierte Daten speichern zu können und einen einheitlichen Zugriff auf alle Daten zu gewährleisten, gilt es, einen Data Lake aufzubauen. Bekannte Beispiele für einen Data Lake sind unter anderem das Hadoop Distributed File System (HDFS), Azure Data Lake oder Amazon S3 Services.

Ein weiteres Problem in diesem Zusammenhang ist die Tatsache, dass innerhalb des Unternehmens sehr oft Daten aktuell nicht an einem zentralen Ort gespeichert werden. So werden Daten von Tochtergesellschaften oft dezentral bei den jeweiligen Entitäten gespeichert, oder auch innerhalb einer Gesellschaft verwalten Organisationseinheiten ihre Daten teils eigenständig, wobei die Gründe hierfür verschieden sind. Hinzu kommt, dass ein Großteil der Kundendaten nicht bei dem Versicherungskonzern selbst gespeichert wird, sondern bei den Versicherungsmaklern. Diese Silostrukturen erlauben keine effiziente Nutzung der Daten. Um dies zu ermöglichen, ist ein Vorhalten in einem Data Lake oder Data Warehouse jedoch unabdingbar.

[2] Cox, Ellsworth: Application-controlled demand paging for out-of-core visualization, IEEE 8th conference on Visualization, 1997.

Datenverarbeitung und Analyse

Aufgrund der großen Datenmengen werden spezielle Tools und Frameworks benötigt, um die Daten für Analysen aufbereiten zu können. Diese erlauben es, Extract-, Transform-, Load(ETL)-Prozesse parallel auf einem Rechnerverbund auszuführen, wodurch eine schnelle Ergebnisbereitstellung erfolgen kann. Die aufbereiteten Daten ermöglichen es in einem nächsten Schritt, bereits erste einfache Analysen durchzuführen sowie diese zu visualisieren. Das Entwickeln von Machine-Learning-Modellen oder aufwendigen Analysen setzt jedoch den Einsatz von komplexeren Algorithmen voraus, welche auf den vorgehaltenen Daten ausgeführt werden müssen. Entsprechende Cluster Computing Frameworks, wie Apache Spark, erlauben das effiziente Ausführen. Die zwei für die Implementierung am häufigsten verwendeten Programmiersprachen im Bereich Data Science sind R und Python. Beide bieten verschiedene Möglichkeiten, statistische Analysen durchzuführen. Durch die Anbindung verschiedenster Maschine-Learning-Bibliotheken, wie keras oder sciki-learn, können in kürzester Zeit fortgeschrittene und zudem äußerst performante Algorithmen zum Einsatz kommen.

Ergebnisbereitstellung

Die implementierten Modelle können auf verschiedene Weisen genutzt werden. Eine Möglichkeit ist das direkte Ableiten von Handlungsempfehlungen und die Visualisierung von Zusammenhängen, eine andere Möglichkeit ist der Einsatz von Machine-Learning-Modellen zur Auswertung von Livedaten.

Datenvisualisierung

Die aus Analysen resultierenden Ergebnisse sind oft nicht in einer derartigen Form, dass sie für sich selbst sprechen. Es gilt, diese weiter aufzubereiten, sodass Implikationen und Sachverhalte transparent und verständlich sind. Da Ergebnisse häufig ein Teil von wöchentlichen oder monatlichen Reports sind, ist hierfür ein Standard notwendig. Anwendungen wie Tableau und QlikView ermöglichen eine interaktive Reporterstellung, welche diese Eigenschaft aufweist. Zudem bieten sie eine Schnittstelle zu Programmiersprachen wie R, was eine einfache Integration in eine bestehende Systemlandschaft fördert.

Produktivnahme von Machine-Learning-Modellen

Im Fall von Machine-Learning-Modellen ist nicht nur das Ergebnis eines Modells von Interesse, sondern auch das trainierte Modell an sich. Dieses kann für die Bestimmung der Kreditwürdigkeit in einem Antragsprozess verwendet werden oder als Bestandteil eines aktuariellen Cockpits. In beiden Fällen gilt es, das fertige Modell so zur Verfügung zu stellen, dass es in einen bereits bestehenden Prozess integriert werden kann. Hierzu müssen die notwendigen Schnittstellen geschaffen werden. Die durch Spezialisten in den Programmiersprachen R oder Python implementierten Modelle sind in ein Front-End zu integrieren. Da sich die Programmiersprache der Modelle jedoch häufig von der des Front-End unterscheidet, müssen hierfür APIs in Betracht gezogen werden. Wird das Modell auf einem Server bereitgestellt, kann es beispielsweise über die REST API in eine bestehende Website und in einen Antragsprozess eingebettet werden.

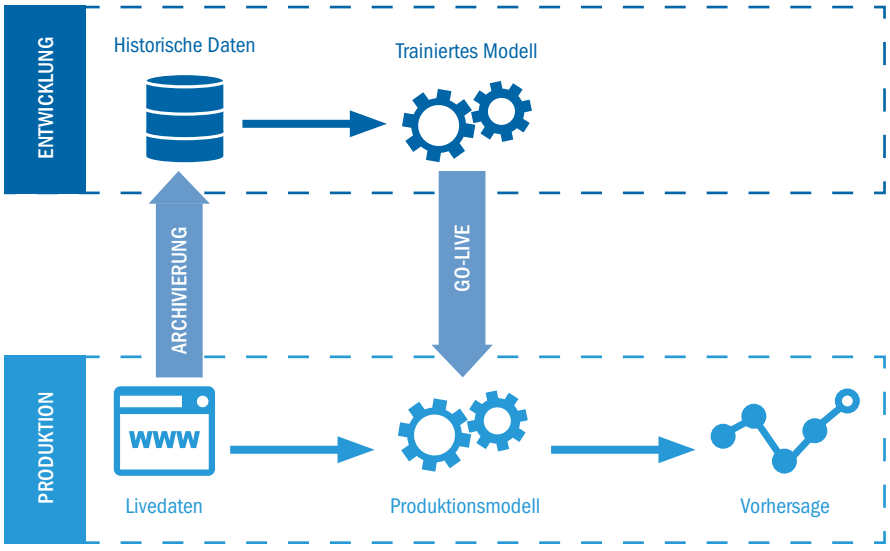

Abbildung 2: Zusammenspiel Modellentwicklung und Produktivnahme

Um die Möglichkeiten, welche sich durch Data Science bieten, optimal nutzen zu können, ist eine komplexe Infrastruktur zu schaffen. Anbieter von Infrastructure as a Service (IaaS), wie Amazon Web Services, Microsoft Azur oder Alibaba Cloud, bieten eine einfache Möglichkeit, eine entsprechende Infrastruktur aufzubauen und in die aktuelle Landschaft zu integrieren. Unter anderem werden Lösungen in den Bereichen Datenbankservices, Elastic Computing sowie Big Data & Analytics angeboten, aus welchen die notwendigen Anwendungen ausgewählt werden können. Vorteil dieser Services ist zum einen die rasche Verfügbarkeit der ausgewählten Anwendungen und zum anderen die hohe Skalierbarkeit. Speicherkapazitäten sowie Rechenleistungen können kurzfristig erhöht und weitere Anwendung in die Infrastruktur aufgenommen werden.

Bei vielen Versicherern ist die vorgestellte Infrastruktur in Grundzügen vorhanden, die Verwendung von IaaS, um diese zu erweitern, wird jedoch teils skeptisch betrachtet. Dies hängt unter anderem mit bestehenden Datenschutzrichtlinien zusammen sowie mit der Tatsache, dass Daten außerhalb des Konzernnetzwerks gespeichert werden und somit der Kontrolle des Unternehmens entzogen werden könnten. Es besteht jedoch nicht nur Skepsis gegenüber Cloud-Services, sondern auch generell gegenüber Data Science, was wir in unserem nächsten Artikel thematisieren.

Zu den Autoren:

Alexander Riesner
Manager

Office Wien
Praterstraße 31
1020 Wien

E-Mail alexander.riesner@zeb.at

Tobias Holler
Analyst

Office München
Theresienhöhe 13a
80339 München

E-Mail tobias.holler@zeb.de