Business News

DOAG Zeitschrift für die Anwender von Oracle Business- und BI-Lösungen

BS

Die Datenexplosion meistern

Sommer 2019 | ISSN 09 36-0360 | http://bs.doag.org

Das E-3 Magazin

Information und Bildungsarbeit von und für die SAP-Community

Überfordert?

Wir bieten Information und Bildungsarbeit von und für die SAP-Community

e-3.de | e3zine.com

Björn Bröhl
DOAG Vorstand Data Analytics

Liebe Mitglieder, liebe Leserinnen und Leser,
„Die Daten-Explosion meistern" lautete das Thema der Data Analytics, die am 26. Und 27. März 2019 im Phantasialand in Brühl stattfand. Einen „Knall" gab es zwar nicht, doch die Konferenzteilnehmer erlebten viele spannende Vorträge und Diskussionen.

In dieser Ausgabe der DOAG Business News greifen wir nun diese Thematik erneut auf. Mit welchen Herausforderungen sehen sich Unternehmen angesichts einer ständig wachsenden Menge von Daten konfrontiert und mit welchen Strategien können sie darauf reagieren?

Sie erwartet ein breiter Mix an Themen – von der Betrachtung der Vor- und Nachteile sowie möglicher Synergien zwischen Data Lake und klassischem Data Warehouse über einen weiterentwickelten multifunktionalen Data Lake bis zu einem Praxisbericht zum Einsatz agiler Methoden im Rahmen eines Data-Warehouse-Projekts.

Darüber hinaus finden Sie in diesem Heft eine Reihe interessanter Beiträge von Referenten der besagten Data-Analytics-Konferenz. Freuen Sie sich u.a. auf Artikel über die Einführung eines Data Hub, die Prozessvereinfachung durch ein Recommender-System sowie ein preisgekröntes Datenprojekt der Stadtverwaltung Kaiserslautern.

Viel Spaß mit der Lektüre dieser Ausgabe

Impressum

DOAG Business News wird von der DOAG Deutsche ORACLE-Anwendergruppe e.V. (Tempelhofer Weg 64, 12347 Berlin, www.doag.org), herausgegeben. Es ist das User-Magazin rund um die Applikations-Produkte der Oracle Corp., USA, im Raum Deutschland, Österreich und Schweiz. Es ist unabhängig von Oracle und vertritt weder direkt noch indirekt deren wirtschaftliche Interessen. Vielmehr vertritt es die Interessen der Anwender an den Themen rund um die Oracle-Produkte, fördert den Wissensaustausch zwischen den Lesern und informiert über neue Produkte und Technologien.

DOAG Business News wird verlegt von der DOAG Dienstleistungen GmbH, Tempelhofer Weg 64, 12347 Berlin, Deutschland, gesetzlich vertreten durch den Geschäftsführer Fried Saacke, deren Unternehmensgegenstand Vereinsmanagement, Veranstaltungsorganisation und Publishing ist.

Die DOAG Deutsche ORACLE-Anwendergruppe e.V. hält 100 Prozent der Stammeinlage der DOAG Dienstleistungen GmbH. Die DOAG Deutsche ORACLE-Anwendergruppe e.V. wird gesetzlich durch den Vorstand vertreten; Vorsitzender: Stefan Kinnen. Die DOAG Deutsche ORACLE-Anwendergruppe e.V. informiert kompetent über alle Oracle-Themen, setzt sich für die Interessen der Mitglieder ein und führt einen konstruktiv-kritischen Dialog mit Oracle.

Redaktion:
Sitz: DOAG Dienstleistungen GmbH
(Anschrift s.o.)
ViSdP: Mylène Diacquenod
Redaktionsleitung: Christian Luda
Weitere Redakteure: Lisa Damerow,
Marina Fischer, Sanela Lukavica,
Martin Meyer, Fried Saacke, Rolf Scheuch,
Dr. Frank Schönthaler

Druck:
adame Advertising and Media GmbH, Berlin,
www.adame.de

Fotonachweis:
Titel: © agsandrew | *https://de.fotolia.com*
S. 5: © Stuart Miles | *https://de.123rf.com*
S. 11: © Sebastien Decoret | *https://de.123rf.com*
S. 15: Quelle: Thyssenkrupp
S. 18: © ToheyVector | *https://de.fotolia.com*
S. 22: © mamanamsai | *https://de.123rf.com*
S. 26: © Kanda Euatham | *https://de.123rf.com*
S. 30: © rawpixel | *https://de.123rf.com*

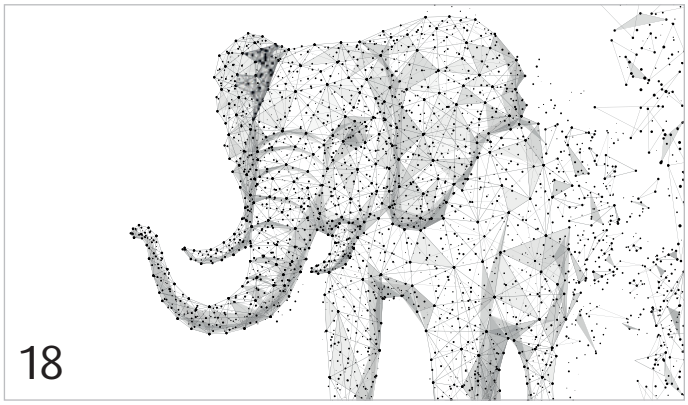
Titel, Gestaltung und Satz:
Caroline Sengpiel,
DOAG Dienstleistungen GmbH
(Anschrift s.o.)

Anzeigen:
Simone Fischer,
DOAG Dienstleistungen GmbH
(verantwortlich, Anschrift s.o.)
Kontakt: anzeigen@doag.org

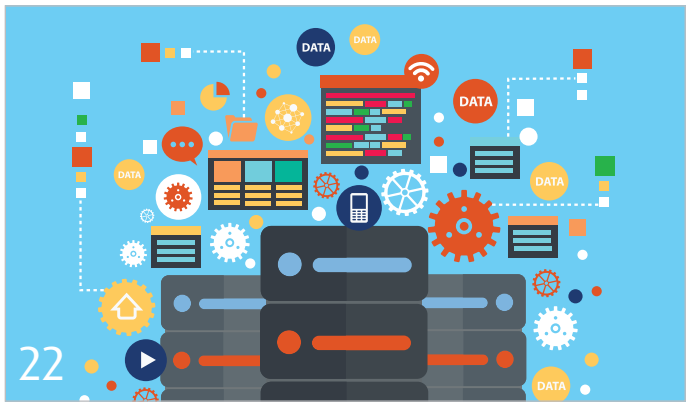
Mediadaten und Preise unter:
www.doag.org/go/mediadaten

Ein neuer, logischer Data Lake bietet Multifunktionalität.

Der Data Hub wandelt Rohdaten in nutzbare Daten um.



Wachsende Datenmengen als Chance für Automation und Prozessvereinfachung.



Ein Logistikunternehmen setzt bei System-Neuentwicklung auf agile Methoden.

Unsere Inserenten

Löst der Data Lake das Data Warehouse ab?

Sven Bosinger, Its-people, und Alfred Schlaucher, Oracle

In Zeiten von Big Data kommt immer häufiger die Frage auf, ob die klassische Data-Warehouse-Architektur noch zeitgemäß ist oder ob sie nicht besser durch neuere Ansätze abgelöst werden sollte. Hierbei wird häufig vergessen, dass es bei einem Data Warehouse (DWH) nicht allein um die Sammlung und Speicherung von Daten geht. Vielmehr sind DWHs primär dazu entwickelt worden, Wissen aus Daten zu gewinnen. Dazu müssen Daten in einen Kontext gesetzt werden, sodass aus ihnen wertvolle und verlässliche Informationen werden. Erst durch die Interpretation durch den Benutzer entsteht Wissen, das im besten Fall zu nachvollziehbaren und belastbaren Entscheidungen führt.

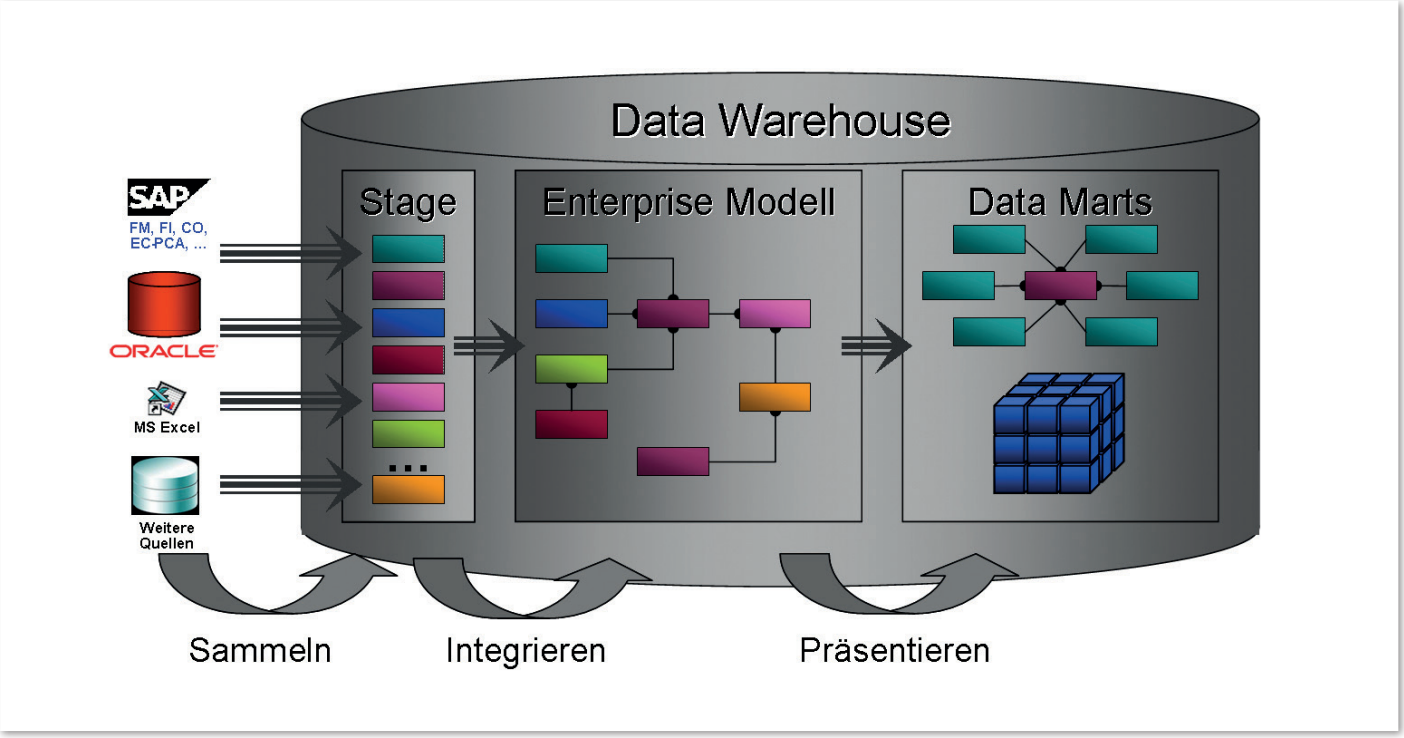


Abbildung 1: Standardmodell Data Warehouse (Quelle: Sven Bosinger)

Das klassische Standardmodell

DWH-Systeme sind bei den meisten größeren Unternehmen bereits seit Mitte der 1990er Jahre im Einsatz. Dabei hat sich eine klassische Architektur mit drei Schichten etabliert *(siehe auch Abbildung 1)*:

1. Stage-Layer: Hier erfolgen sowohl technische Integration als auch Qualitätskontrolle der eingehenden Daten. Diese werden so lange gespeichert, bis sie in den Core-Layer integriert werden.

2. Core-Layer (Enterprise-Modell): Die Daten werden mithilfe eines regelmäßigen Batch-Jobs (stündlich, täglich usw.) aus dem Stage-Layer übernommen, bei Bedarf angepasst oder transformiert und in die vorhandenen Daten integriert. Das Ziel ist ein hoch integriertes, relationales Datenmodell mit ausgeprägten Fremdschlüsselbeziehungen. Dabei werden gegebenenfalls qualitätsmäßig unzureichende Daten abgelehnt. Da ein DWH Daten über einen langen Zeitraum speichert und Änderungen nachvollzogen werden sollen, ist es notwendig, die Daten entsprechend zu historisieren.

3. Presentation-Layer (Data Marts): Die Daten des Core-Layers werden in sogenannte „Data Marts" geladen. Diese können in unterschiedlichen Technologien implementiert werden und dienen als vordefinierter Datenlieferant für die diversen Frontend-Werkzeuge. Über die Dimensionierung der Daten wird ein fachbereichsspezifischer Kontext hergestellt, der sie für eine sinnvolle Interaktion mit den DWH-Benutzern aufbereitet. Hierbei kommt es zu bewussten Aggregationen und Einschränkungen des Betrachtungszeitraums der Daten. Des Weiteren werden sogenannte „KPIs" (Key Performance Indicators) errechnet, die als Fakten in einem Dimensionsmodell bereitgestellt werden.

Warum hat sich diese Form der Architektur durchgesetzt?

Die Aufteilung in drei Schichten hat viele Vorteile. Es kommt zu einer klaren Trennung der Datenhaltungsschicht (Core-Layer), die von ihrem Modell her über eine lange Zeit hinweg konstant ist, und dem doch recht volatilen Presentation-Layer, der bei geändertem Abfrageverhalten angepasst und erweitert werden kann. Während der Core-Layer darauf perfektioniert ist, Daten schnell zu laden und Datenqualität sicherzustellen, liegt der Fokus beim Presentation-Layer auf der Performance der Abfragen und dem Kontext, der „Lesbarkeit" für den Benutzer. Im Core-Layer werden die Daten über die Integrationsmittel einer relationalen Datenbank in einem redundanzfreien sogenannten „3-Normal-Form-Modell" (3NF) gespeichert. Im Presentation-Layer werden gezielte Denormalisierungen, also Redundanzen und Aggregierungen vorgenommen, um zu einfachen und performanten Querys zu gelangen, die die Daten an die Frontend-Werkzeuge weitergeben.

So kann man die zum Teil sehr widersprüchlichen Anforderungen an Datenhaltung und -präsentation optimal unterstützen und den Konflikt elegant auflösen. Dadurch, dass alle Daten in ihrer feinsten Granularität im Core-Layer vorhanden sind, können die Data Marts jederzeit ohne Datenverlust umstrukturiert werden. Abgerundet wird die Architektur durch den Stage-Layer, der als Sammelbecken für die Datenquellen fungiert. Nachdem die Daten im Stage-Layer gespeichert wurden, kann jede Weiterverarbeitung durch die auf hoch performante Massendatenverarbeitung spezialisierte Sprache SQL durchgeführt werden. Damit werden die Daten performant und transaktionsgesichert über die nachfolgen Layer geladen. Des Weiteren entlastet das Kopieren der Daten in den Stage-Layer die Quellsysteme, da Aktionen danach nur noch im DWH stattfinden.

Wo liegen die Nachteile dieser Architektur?

Die oben dargelegte Architektur hat naturgemäß Nachteile. Davon wiegen drei besonders schwer und sind von bedeutender Relevanz im Vergleich zu Big-Data-Lösungen:

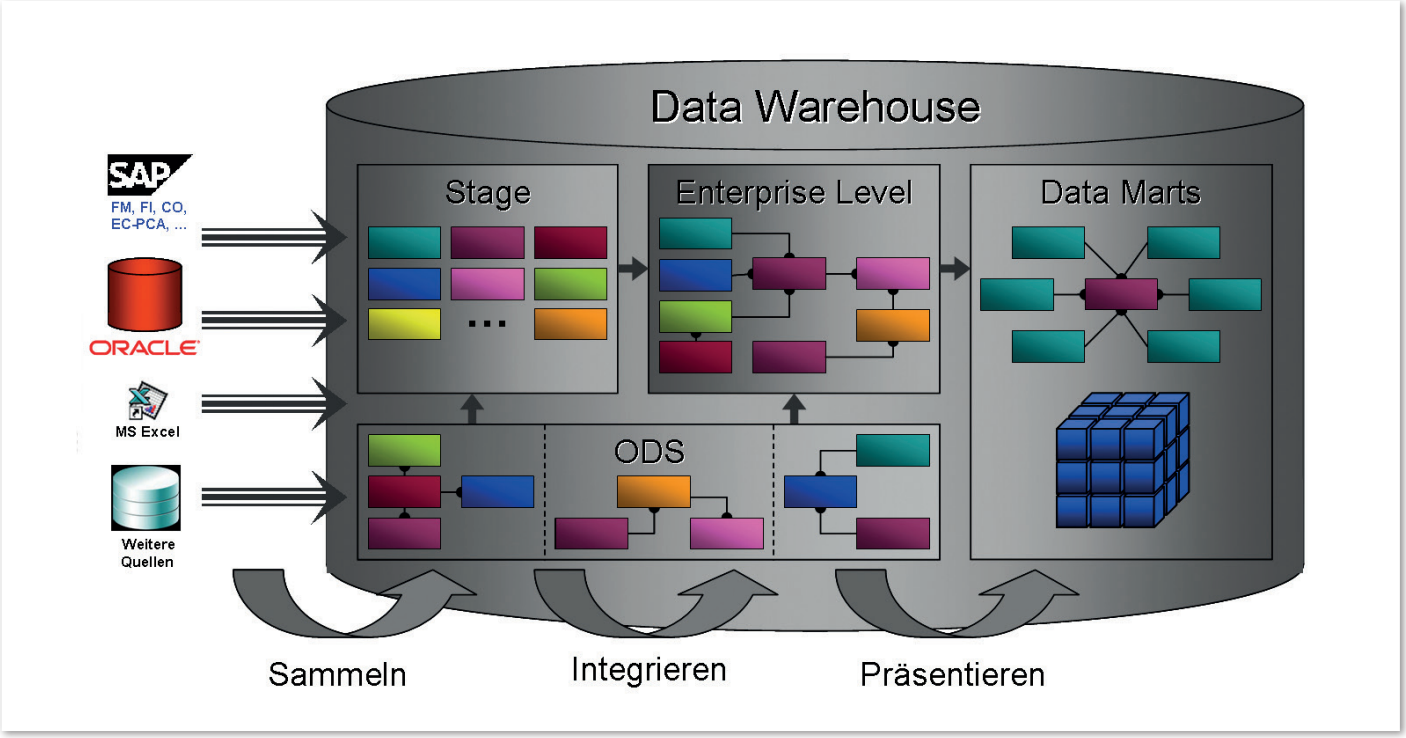


Abbildung 2: Operational Data Store (Quelle: Sven Bosinger)

1. Informationslatenz: Aufgrund des mehrschichtigen Daten-Modells und der Batch-Verarbeitung gibt es eine große Zeitspanne (Latenz) zwischen der Anlandung der Daten im Stage-Layer und dem Bereitstellen im Presentation-Layer. Da sowohl Stage- als auch Core-Layer nicht für die Abfrage durch Frontend-Werkzeuge geeignet sind, dauert es immer einen von Batchfrequenz und -laufdauer abhängigen Zeitraum, bis eine neue Information zu Auswertungen zur Verfügung steht. Um das Problem der Informationslatenz zu umgehen, gibt es eine Reihe von Erweiterungen des Standard-Modells. So können die Daten statt in einen klassischen Stage-Layer in einen sogenannten „Operational Data Store" (ODS) geladen werden (siehe Abbildung 2). Dieser ähnelt von seiner Datenmodellstruktur her dem Quellsystem. Hier kann der Benutzer anders als beim Stage-Layer sinnvolle Abfragen durchführen. Probleme entstehen, wenn der Benutzer diese Realtime-Daten mit bereits im Presentation-Layer gespeicherten Daten in Beziehung setzen will. Die Integration und Vergleichbarkeit im Rahmen der Auswertung kann sehr fehleranfällig sein und spezielles Wissen über die Daten und ihre Speicherung voraussetzen. Eine weitere Lösung ist die Schaffung eines Realtime-Bereichs *(siehe Abbildung 3)*. Ein Teil der Tabellen wird nicht über einen regelmäßigen Batch-Lauf aktualisiert, sondern durch einen ereignisgesteuerten Ladeprozess. Damit wird jeder neue Datensatz einzeln und sofort über den Core-Layer in den Presentation-Layer geladen, sobald er im Stage-Layer gespeichert wurde. Dadurch ist die Latenz nicht mehr abhängig von der Batch-Frequenz, sondern nur noch von der Verarbeitungsdauer (Near-time-Szenario). Allerdings hat auch dieses Verfahren seine Grenzen: Finden im Core-Layer aufwendige Datenqualitätsprüfungen oder im Presentation-Layer umfangreiche Aggregationen statt, steigt die Latenz wieder. Grundsätzlich lässt sich das Problem der Informationslatenz darauf zurückführen, dass der Kontext und die Struktur der Daten bereits zum Zeitpunkt des Speicherns geschaffen werden (schema on write), wohingegen sie bei Big-Data-Systemen erst zum Zeitpunkt des Lesens der Daten (schema on read) erzeugt werden.

2. Agilität: Durch das starre Gerüst einer relationalen Datenbank führt das Einbinden neuer oder das Erweitern vorhandener Quellsysteme in der Regel zu Änderungen im Datenmodell. Diese können nicht ad hoc vorgenommen werden, sondern erfordern eine sorgfältige Planung. Auch die Ladeprozesse müssen bei Datenmodelländerungen eine entsprechende Erweiterung erfahren, um neue Strukturen zu berücksichtigen. Selbst im einfach gehaltenen Stage-Layer müssen Tabellen geändert oder neu angelegt werden. Das alles bedeutet Aufwand und Zeit. Diesem Problem kann man mit einer flexiblen Datenmodellierungs-Methode, dem Data Vault, begegnen. Durch diese Technik kann eine Erweiterung effizienter durchgeführt werden. Die Schicht des Presentation-Layers hat eine größere Dynamik als die des Core-Layers. Aufgrund der Datenredundanz ist es einfach möglich, neue Data Marts zu generieren, um einem veränderten Abfrageverhalten Rechnung zu tragen. Aber auch hier ist es grundsätzlich nötig, neue Objekte in den bestehenden Layer zu implementieren.

3. Kosten: Die Kosten für Lizenzen, Hardware und Betriebsführung eines DWH sind nicht unerheblich, gerade wenn spezialisierte DWH-Appliances wie Exadata und Teradata zum Einsatz kommen. Die Kosten steigen in der Regel sowohl mit dem Datenvolumen als auch mit der durch Ladeprozesse und Abfragen erzeugten CPU-Last. Diese hohen Kosten stellen ein Hemmnis für innovative Ansätze im Bereich der Datenanalyse dar. Vor allem, wenn die benötigten Daten im DWH noch nicht vorhanden sind, führen die hohen

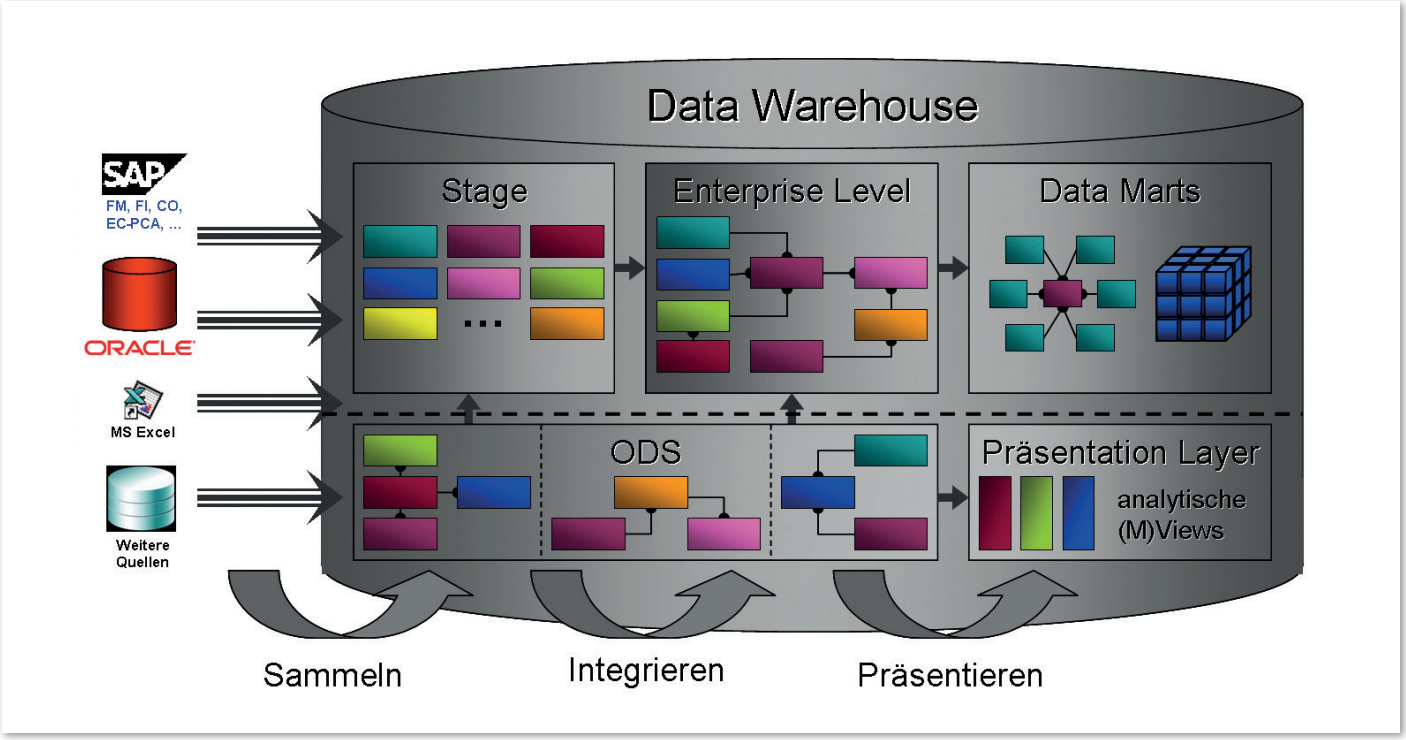


Abbildung 3: Realtime-Bereich (Quelle: Sven Bosinger)

Kosten der Integration und die nicht abschätzbaren Ergebnisse in der Return-on-invest-Betrachtung schnell zu abschlägigen Entscheidungen. Dies steht einem gelebten Data-Science-Ansatz diametral entgegen. Dadurch besteht die Gefahr, dass Datenschätze ungehoben bleiben und Unternehmen im Wettbewerb der Ideen und Innovationen zurückzufallen.

Was kann ein Data Lake besser?

Betrachtet man die drei relevanten Schwachpunkte eines DWH, stellt sich die Frage, was ein Data Lake, also eine Big-Data-Lösung, dazu beitragen kann, diese Probleme zu meistern.

Die Informationslatenz ist bei einem reinen Data Lake nicht so offensichtlich gegeben. Da in einem Distributed File System (z.B. Hadoop, HDFS) schnell und einfach Daten gespeichert werden können, sind diese auch mit dem Augenblick der Speicherung grundsätzlich auswertbar. Durch die massive Parallelisierung der Abfragen (MapReduce) ist man in der Lage, die angehäuften Datenbestände in kurzer Zeit zu durchsuchen und eine Ergebnismenge zu generieren. Damit wird die Informationslatenz auf jene Zeit reduziert, die eine Abfrage benötigt, um das gewünschte Ergebnis zu erzeugen. Es ist nicht mehr notwendig, Daten durch einen Ladeprozess über verschiedene Layer zu transportieren.

Allerdings hat auch dieser Ansatz Nachteile. Durch das Schema on read muss die innere Struktur und Konsistenz der Daten bei jedem Zugriff erneut hergestellt werden. Bei verschiedenen Zugriffen auf die Daten kann nicht sichergestellt werden, dass die Entwickler bei jeder Abfrage die gleichen Konsistenz- und Datenqualitätsregeln anwenden. Dadurch kann es zu unterschiedlichen, sich widersprechenden Ergebnissen kommen. Das Schema on write bietet hier die Sicherheit, jedes Mal zu einer gleichen Aussage zu kommen.

Auch der Zugriff auf die Daten ist komplexer als bei einer relationalen Datenbank. Während es bei Letzterer ausreicht, mit dem sehr mächtigen SQL auf die Daten zuzugreifen, muss bei einem Data Lake ein programmatischer Weg, in der Regel mittels Java, gewählt werden. Allerdings gibt es hierfür Softwareprodukte wie Spark, die es ermöglichen, wiederum mit SQL-Mitteln auf die im HDFS gespeicherten Resilient Distributed Datasets (RDDs) zuzugreifen.

Die Flexibilität eines Data Lake ist um ein Vielfaches höher. Neue Datenarten können ohne Systemänderungen gespeichert und sofort integriert werden. Lediglich bei der Auswertung durch die diversen Abfragen müssen gegebenenfalls weitere Programme implementiert werden. Dazu gibt es viele Bibliotheken und Klassen, mit denen fast jeder Dateityp gescannt und die gewünschten Informationen extrahiert werden können.

Die Kosten einer Big-Data-Lösung hängen sehr stark von den eingesetzten Werkzeugen, deren Distribution und der benötigten Hardware ab. Ist einmal eine Big-Data-Lösung implementiert, sind die Kosten für Implementierung und Speicherplatz für eine neue Analyse auf neuen Daten vernachlässigbar, solange sich dadurch der Datenbestand nicht massiv vergrößert. Da neue Dateien als RDDs einfach einem HDFS hinzugefügt werden können und dann sofort auswertbar sind, hat man lediglich die Kosten für Programmierung der Abfrage und Bewertung der Ergebnisse zu berücksichtigen. So wird ein Return on invest wesentlich günstiger als bei einem DWH.

Ein weiterer Grund für die immer größere Relevanz des Data Lake ist die durch die Big-Data-Diskussion gewachsene Wahrnehmung von Analysemethoden und nutzbarem Datenmaterial in Unternehmen. Man will heute nicht mehr nur Transaktions- und Stammdaten analysieren, sondern reichert diese mit externen Daten an. Es reicht nicht mehr aus zu wissen, welche Kundengruppe welche Produktgruppe in welchem Zeitraum und in welcher Region gekauft hat. Man will wissen, mit welcher Motivation und mit welchen Erfahrungen ein Kunde ein Produkt kauft. Entschied in der Vergangenheit das Bauchgefühl des Vertriebsmanagers über Vertriebsstrate-

gien und Produktplanungen, geht man heute methodisch vor und nutzt Analysen aus allen Erlebensbereichen eines Kunden. Unternehmensinterne Geschäftsdaten werden ergänzt mit massenhaft vor allem im Internet zur Verfügung stehenden Daten über Kunden, Trends oder Bewegungen.

Neue Fokusbereiche bei der Analyse

Vor allem Kommunikationsdaten aus den unterschiedlichsten Interaktionen eines Kunden – in der Regel Text und Sprache – rücken in den Fokus.

Wenn wir uns allein auf diesen Aspekt beschränken und nicht auch noch IoT- und Sensordaten oder Bildinformationen betrachten, so können wir bereits neue Bedarfe für Data Lakes erkennen, die über klassische DWH-Anwendungen hinausreichen. Eine Warenkorbanalyse zählt gemeinsam gekaufte Produkte beispielsweise mit einem Naive-Bayes-Algorithmus durch. Das ist eine reine Mengen-betrachtende und -vergleichende Analyse, wie sie durch einen summierenden und gruppierenden SQL-Befehl in einer relationalen Datenbank gut, wenn nicht optimal, zu erledigen ist. Die Stimmungslage eines Kunden in einem Webseiten-Kommentar zu analysieren, gelingt jedoch nicht mehr durch Zählen und Summieren. Man benötigt ein Modell mit einer möglichst großen Vielfalt von potenziellen Stimmungen.

Ein Kundenkommentar ist eine Ansammlung von Wörtern eines Wortfundus, der je nach sozialer Gruppe von wenigen hundert bis mehreren tausend Wörtern reicht. Darin liegt die Schwierigkeit: Zum einen verwendet ein Text eine große Anzahl unterschiedlicher Wortkombinationen und zum anderen muss man diese den Stimmungen bekannter sozialer Personengruppen zuordnen. Das Analysemodell wird genauer und kann besser Stimmungen vorhersagen, wenn zur Erstellung möglichst viele Texte eingeflossen sind. Zur Analyse solcher Massendaten haben sich in jüngster Zeit Deep-Learning-Methoden als besonders erfolgreich hervorgetan, die in einer hohen Anzahl von parallelen Rechenitterationen sehr gute Modelle erzeugen. Stimmungen von Menschen sind vorhersagbar. Dies ist eine neue Qualität von Analysen und auch ein Anwendungsfall für Data Lakes.

Warum ist ein klassisches DWH dennoch unverzichtbar?

Bei allen Vorzügen eines Data Lake bietet das klassische DWH einen unschlagbaren Vorteil – den der Daten-Konsistenz und damit der Datenqualität. Was ein gut designtes DWH von Hause aus liefert, ist die ständige Garantie, dass die präsentierten Daten einem definierten Anspruch an Datenqualität genügen.

Ein DWH sorgt aufgrund seiner Transaktionssicherheit dafür, dass der Anspruch an Datenqualität zu jedem Zeitpunkt gewährleistet ist. Dies erhöht die Belastbarkeit der Auswertungen und deren Interpretationen um mehrere Potenzen.

Während bei der Abfrage von Daten aus einem Data Lake aufgrund des Schema-on-read-Ansatzes ein tiefes Verständnis der Daten und ihrer Speicherform vorausgesetzt wird, ist bei einem persistierten Schema, dem Schema on write, ein derartiges Wissen nicht nötig. Das DWH-System als solches verhindert eine Fehlbenutzung und damit eine Fehlinterpretation der Daten. Der im DWH so wichtige Ansatz des sogenannten „Single point of truth“ kann mit einer Big-Data-Lösung niemals gewährleistet werden.

Synergieeffekte – hybride Architekturen

Akzeptiert man, dass ein klassisches DWH und eine Big-Data-Lösung keine konkurrierenden, sondern sich ergänzende Systeme sind, stellt sich die Frage, welches System für welche Daten am besten geeignet ist.

Kennzahlen, Transaktionen oder einheitliche Stammdaten sind in einer vorstrukturierten, relationalen Datenhaltung sicher besser aufgehoben, während Texte für eine Stimmungsanalyse besser in ein Massenspeicher-Medium mit Hadoop-Infrastruktur passen. Aber alle genannten Datenarten bilden gemeinsam eine umfassende Datengrundlage für die in der heutigen Zeit nötigen Analysen, in denen man die Ursache von Umsatzschwankungen auch in den Schwankungen der Wahrnehmung des eigenen Unternehmens durch Kunden sieht. „Harte“, unternehmensinterne Beispielkennzahlen lassen sich sinnvoll durch „weiche“ Kennzahlen aus unternehmensexternen Quellen erweitern, von denen man so viel wie möglich erschließt und als heterogene Massendaten im Data Lake sammelt.

Um diese Synergien zu erreichen, wird man Data Lakes und DWH-Systeme in einer zusammenhängenden hybriden Architektur zusammenführen *(siehe Abbildung 4)*. So nutzt man die Vorteile beider Verfahren.

Mehr noch: Wendet man beide Verfahren gemeinsam an, so entlasten sie sich gegenseitig. Die jeweiligen Lösungen müssen sich nicht mit Aufgaben beschäftigen, für die sie weniger gut geeignet sind. Nachfolgend einige Beispiele:

- Obwohl das Archivieren von operativen Quelldaten keine DWH-Aufgabe ist, wird es in einigen Unternehmen doch praktiziert, um zum einen nachweisfähig für geladene Daten zu sein, aber auch, weil es aufgrund der bereits vorhandenen ETL-Strecken praktisch ist. Die Archivierung belastet das DWH jedoch und macht es unnötig teuer. Für Archivierung bietet es sich in einer hybriden Umgebung an, Quelldaten zunächst in den Data Lake zu laden. In das DWH überführt man von dort aus nur die wirklich benötigten Daten und bereitet sie bereits beim Laden auf. Das reduziert das Volumen des DWH-Systems drastisch, spart Kosten, Ressourcen und ETL-Laufzeit.
- Umgekehrt wird man Stamm- und Referenzdaten eher in einem DWH platzieren. Die Anzahl der entsprechenden Tabellen erreicht schnell den Bereich von über 1.000. Viele dieser Tabellen sind klein mit manchmal nur wenigen hundert Sätzen. Das Vorhalten der Masse dieser Einzelobjekte in einem Data Lake kann schnell aufwendig und unübersichtlich werden. In einem DWH helfen auch ohne eine dedizierte Metadatenverwaltung schon allein der automatisch aktive Datenbankkatalog, zusätzliche Namens- und Spaltenkonventionen sowie Relationen zwischen den Tabellen oder Constraints. Stamm- sowie Referenzdaten und Data Lakes passen nicht zusammen.
- Das im DHW oft komplexe Change-Data-Capture-Verfahren, also das Identifizieren von Änderungen in operativen Anwendungen, lässt sich durch die Streaming-Möglichkeiten eines Data Lake für das DWH vereinfachen. Der Streaming-Vorgang im Data Lake (Kafka) erfasst die Änderungen der Vorsysteme und dokumentiert sie passend für das DWH, das sich die Änderungen bequem ohne weitere Prüfungen abholt. Man spart ETL-Laufzeit.
- Monoforme und voluminöse Datenarten wie Maschinendaten, Click-Traffic, Call Detail Records und Kassen-Bon-Daten bleiben im Data Lake und erfahren dort eine Use-Case-spezifische Analyse. Deren Ergebnisse und Aggregationen werden dann wieder im DWH abgelegt und stehen den Anwendern in Dashboards oder im Standardbericht gepaart mit Stamm- und Transaktionsdaten zur Verfügung.

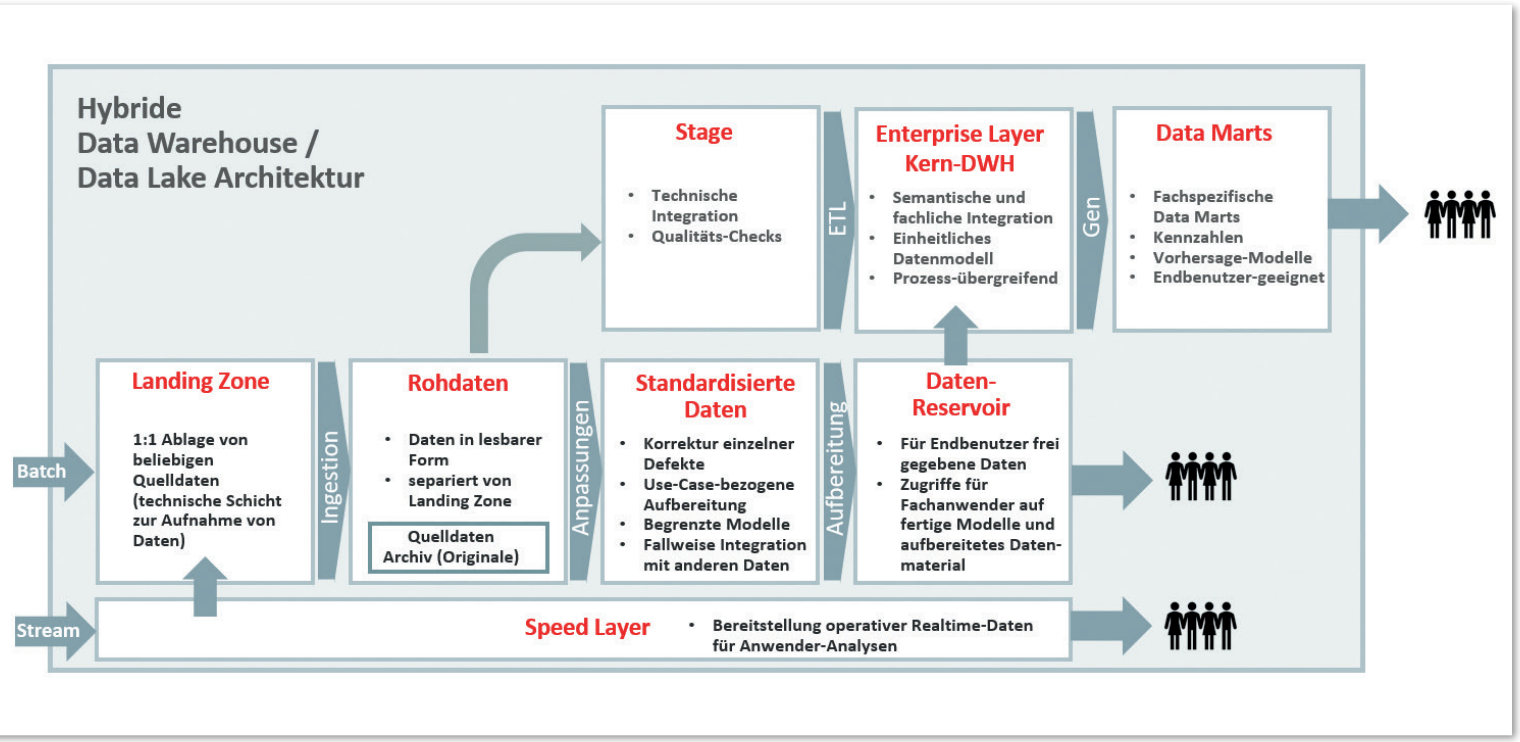


Abbildung 4: Zusammenführung Data-Warehouse-Konzept und Data Lake zu einer hybriden Architektur (Quelle: Alfred Schlaucher)

- Über ein hybrides Konzept kann man schließlich auch das Analyse- und Auskunftsmanagement leichter in Angriff nehmen, also die Steuerung der Benutzergruppen, ihre Zugriffe auf sensible Datenbereiche, das Standardberichtswesen oder die Unterstützung von Sonderwünschen in speziellen Abteilungen mit Sandboxes. Alles, was mit einem standardisierten Reporting und spontanen Auskünften zu tun hat, läuft über das DWH mit einem ausgereiften Security- und User-Management. Die eher experimentellen Analyse-Aktivitäten oder auch Sonderwünsche kann man dem weniger standardisierten Data Lake überlassen.
- Entwickelt ein Data Scientist ein neues Modell im Data Lake, kann dies auch Standardberichte im DWH anreichern. Das Modell wird in die Kern-Warehouse-Schicht überführt und optimiert die Herleitung einzelner Kennzahlen.

Ausblick: Für hybride Lösungen braucht man kein Hadoop

Unternehmen diskutieren aktuell Cloud-Strategien. Die besprochene hybride Lösung ist mit überschaubaren Mitteln heute schon als Cloud-Lösung realisierbar. Entgegen den ersten Big-Data-Projekten mit komplexen Serverfarmen belastet man mit einer Cloud-Lösung nicht gleich sein Budget. Eine Umgebung entsteht Zug um Zug über vorbereitete Cloud-Services und skaliert bei Bedarf, wenn sich Erfolge mit ersten Prototypen einstellen. Dies ist heute auch ohne eine Hadoop-Umgebung mit HDFS möglich:

- Der Data Lake wird mittels S3-Object Storage realisiert – ein ausfallsicherer, skalierbarer und kostengünstiger Massenspeicher in der Cloud.
- Machine-Learning-Algorithmen laufen in einer Spark-in-Memory-Umgebung auf einem Compute-Service.
- Zusätzliche Rechenpower entsteht durch wahlweise zugeschaltete GPUs (Graphical Processing Units).
- Das DWH wird als Warehouse Cloud Service autonom verwaltet und verfügt über einen SQL-Zugriff auf den Object Storage.

Fazit

Das DWH-Konzept ist nicht überholt, sondern wichtiger denn je. Die Big-Data-Diskussion und die dadurch geborene Idee des Data Lake hat Schwächen im DWH-Konzept erkennen lassen, denen man mit einer hybriden Lösung begegnen kann. Data Lakes ersetzen nicht das DWH, sie ergänzen es.

Die wichtige Frage der Metadatenverwaltung und Data Governance im Data Lake wurden hier nicht angesprochen. Dieses Thema ist schon im DWH eine kaum gelöste Aufgabe. Im Data Lake wird diese Aufgabe nicht einfacher, denn sie ist keine Frage von Tools und Technik, sondern eine Frage der Kultur im Umgang mit Daten und gelebtes Datenmanagement. Das sollte separat behandelt werden.

Sven Bosinger
sven.bosinger@its-people.de

Sven Bosinger ist bei der Its-people GmbH als Portfoliomanager für den Bereich Analytics verantwortlich. Des Weiteren ist er im DOAG e.V. Themenverantwortlicher für Data Warehouse. Er ist seit 1994 in der IT mit den Themenschwerpunkten DSS, DWH, BI und Analytics tätig. Seit 2006 ist er freiberuflicher Berater und Gesellschafter der Its-people GmbH. In dieser Funktion berät er Unternehmen bei der Einführung und Neustrukturierung ihrer Analytics-/BI-Landschaften. Zudem ist er regelmäßiger Referent bei diversen DOAG-Veranstaltungen.

Alfred Schlaucher
alfred.schlaucher@oracle.com

Alfred Schlaucher ist seit den 1980er Jahren in der Datenverarbeitung unterwegs. Wichtige Arbeitsfelder waren Datenmodellierung, Metadatenmanagement sowie Datenbank- und Anwendungsprogrammierung in Unternehmen der Fertigungsindustrie. In den letzten 20 Jahren arbeitete er eng mit Kunden der Firma Oracle zusammen und berät diese bei Warehouse-Architekturen. Er begleitete die Big-Data-Diskussion in Deutschland seit ihrem Entstehen 2010 und spezialisierte sich in den letzten Jahren zunehmend auf Fragen des Machine Learning vor allem mit der Sprache R.

Data Lakes neu denken

Thomas Niewel, Denodo

Obwohl ein relativ junges Konzept, erfreuen sich Data Lakes in Unternehmen enormer Beliebtheit. Sie bergen viele Vorteile, bringen aber auch neue Herausforderungen mit sich. Thomas Niewel erklärt, wo Virtualisierungstechnologien Abhilfe schaffen können, und stellt ein neues Modell vor – den logischen, multifunktionalen Data Lake.

Der traditionelle Data Lake

Die Auswahl des richtigen Datenmaterials für BI-Analysen kann viel Zeit in Anspruch nehmen – das gilt insbesondere dann, wenn Daten aus vielen internen und externen Quellen zusammengestellt werden müssen, wie etwa aus Hadoop-, relationalen und Social-Media-Quellen sowie ERP- und Cloud-Anwendungen. Denn es geht nicht nur darum, die richtigen Daten zu finden und zu kopieren. Die Aufgabe ist ungleich komplexer: Enthalten zwei oder mehr Quellen vergleichbare Daten, muss der Data Scientist entscheiden, welcher Datensatz für die Analysen besser geeignet ist. Aus manchen Quellen dürfen Daten nur nach vorher eingeholter Erlaubnis kopiert werden. Wiederum andere Daten erfordern Anonymisierung. Mitunter sind spezielle Programme nötig, um Daten aus einer Quelle zu extrahieren. Wenn es um große Datenmengen geht, stellt sich überdies die Frage nach einem geeigneten Speicherort. Mit Data Lakes möchte man diese Aufgabe meistern. Es handelt sich in der Regel um eine Hadoop-Umgebung, in der alle für den Data Scientist relevanten Daten gespeichert sind.

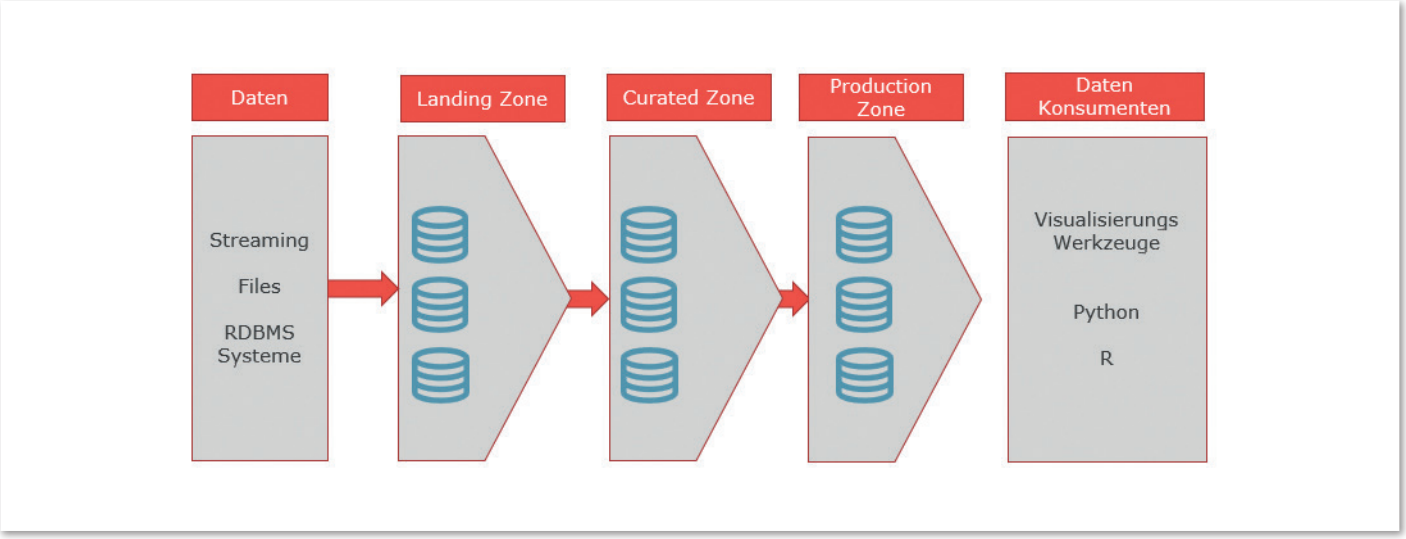


Abbildung 1: Data Lake: Beispiel einer Mehrebenen-Struktur. Verschiedene Nutzergruppen greifen auf unterschiedliche Ebenen zu (Quelle: Denodo).

Was unterscheidet das Konzept der Data Lakes von anderen Konzepten wie beispielsweise einem Data Warehouse (DWH)? In einem Data Lake werden die Daten im ursprünglichen Daten-Format abgespeichert, dem sogenannten „Raw Format". Eine Konvertierung in ein relationales Modell wird nicht durchgeführt. Im Gegensatz hierzu werden in DWH-Systemen für Abfragen optimierte Datenmodelle entwickelt (Star-Schema, Snowflake-Schema). Diese werden durch ETL-Prozesse (extract, transform, load) beziehungsweise ELT-Prozesse befüllt. Ein Beispiel für ein solches Werkzeug ist Oracle ODI.

Für den Data Scientist ist der Data Lake ein Paradies: Er kann Datenbestände analysieren, mit ihnen experimentieren und komplexe Analysen durchführen. Die Datenstruktur und andere Anforderungen definiert er von Fall zu Fall neu. So profitiert der Data Scientist einerseits von den Vorteilen eines zentralen Daten-Repository, verfügt andererseits jedoch auch über viele Freiheiten. Die Arbeit mit traditionellen Data Lakes erfordert allerdings technisches Know-how, über das viele Nutzer nicht verfügen. Wenn also alle Nutzer von der Technologie profitieren sollen, sind weitreichende Anpassungen notwendig.

Nachteile von klassischen Data Lakes

Data Lakes sind zentrale Daten-Repositorys, auf deren Inhalte mit Data-Science-Verfahren zugegriffen werden kann. Letztlich stellt sich allerdings die Frage, ob es sinnvoll ist, alle Daten an einen Ort zu kopieren. Die folgenden Punkte beleuchten problematische Bereiche eines klassischen Data Lake:

- *Große Datenmengen lassen sich schwer bewegen*
  In einigen Umgebungen verhindert bereits das Datenvolumen die Erstellung physischer Kopien. Es fehlt schlicht an Bandbreite und Speicherkapazität, um große Datenmengen in einem Data Lake zu speichern.

- *Datenschutz und Sicherheit*
  Gesetzliche Regelungen erschweren es bei vielen Daten, sie an einen zentralen Ort zu kopieren oder gemeinsam mit anderen Informationen abzuspeichern. Hier ergeben sich nicht selten Fallstricke und juristische Risiken. Aber auch grundsätzliche Sicherheitsbedenken können eine Rolle spielen. Denn der beste Schutz einer möglicherweise kritischen Datenquelle greift nicht, wenn die Daten routinemäßig in einen minder geschützten Data Lake kopiert werden.

- *Mangelnde Kooperationsbereitschaft*
  Auch abseits der Sicherheitsbedenken können verschiedene Abteilungen innerhalb eines Unternehmens zu dem Schluss kommen, dass sie ihre Daten nicht teilen möchten – sei es aus Unsicherheit oder aus mangelndem Gemeinsinn.

- *Komplexe Transformation*
  Weil die Daten im Data Lake im Ausgangsformat gespeichert werden, erfordert ihre Nutzung immer eine vorherige Transformation. Dieser Prozess kann sich sehr kompliziert gestalten und viel Zeit in Anspruch nehmen.

- *Fehlende Metadaten*
  Bei längst nicht allen Quelldaten werden erklärende Metadaten mitgeliefert. Für Data Scientists ist es deshalb oft nicht einfach, die Bedeutung bestimmter Datenelemente zu verstehen. Fehlinterpretationen verfälschen jedoch sämtliche Analysen.

- *Aufwendiges Management des Data Lake*
  Der Data Lake ist eine Technologie, deren Pflege mit einem gewissen Aufwand verbunden ist. Dabei geht es zum Beispiel um den Betrieb einer Hadoop-Cluster-Umgebung und um die fortlaufende Aktualisierung der gespeicherten Daten.

Zusammengefasst erschweren die genannten Punkte die Nutzung eines Data Lake.

Die Use Cases eines Data Lake

Vielfach werden Data Lakes ausschließlich für Anwendungsfälle im Advanced-Analytics-Bereich genutzt – Data Scientists sind hier die wesentlichen Nutzer der Data Lakes. Die einseitige Nutzung wirkt sich jedoch auf die Wirtschaftlichkeit aus, die Erschließung zusätzlicher Nutzungsszenarien für Data Lakes ist daher durchaus erstrebenswert.

Aktuell existieren in vielen Unternehmen Datensilos, da verschiedene Nutzergruppen innerhalb eines Unternehmens ganz eigene dedizierte Quellen für ihre Analysen benutzen. Während Data Scientists auf Data Lakes vertrauen, greifen die Verantwortlichen für traditionelles Reporting auf ein DWH zurück, wohingegen Business-User Data Marts verwenden. Dadurch entstehen Silos und redundante Systeme, die auf unterschied-

lichen Technologie-Stacks wie Datenbanken, Hadoop-Umgebungen, ETL-Lösungen und Governance-Werkzeugen beruhen und zwangsläufig zu Kostenexplosionen sowie Inkonsistenzen zwischen einzelnen Datenquellen führen.

Würde ein zentraler und gemeinschaftlich genutzter Data Lake Abhilfe schaffen? Ja, unter den richtigen Voraussetzungen kann ein Data Lake durchaus als Quelle für verschiedenste Analysen dienen. Dazu zählen etwa Echtzeitanalysen, Advanced Analytics, operationelles Reporting und Self-Service BI. Visualisierungswerkzeuge wie der Oracle Data Visualization Desktop können von Konsumenten genutzt werden. Das erfordert jedoch eine flexiblere Architektur, die es ermöglicht, einen Data Lake für breitere Benutzergruppen zu erschließen. Diese Architektur sollte sowohl den Zugriff auf alle Daten, die im Data Lake gespeichert sind, als auch den Zugriff auf weitere Unternehmensdaten (z.B. aus operativen Systemen) ermöglichen. Des Weiteren sind Governance und Security-Mechanismen notwendig, um einen Data Lake für zusätzliche Benutzergruppen zu öffnen. Ein weiterer wichtiger Punkt ist eine gute Performance für alle Analysen.

In einer klassischen Data-Lake-Architektur werden in der Regel verschiedene Zonen genutzt *(siehe Abbildung 1)*.

Die Landing Zone enthält die Rohdaten. In der zweiten Zone, der Curated Zone, liegen kopierte Daten aus der Landing Zone. Im Rahmen des Kopiervorgangs werden diese Daten verarbeitet, beispielsweise bereinigt oder integriert. Anschließend wandern sie weiter in die Production Zone, wo Nutzer auf sie zugreifen können. Auf jeder Ebene steigt der Grad der Verarbeitung, wodurch es für Nutzer einfacher wird, die Daten weiterzuverwenden. Data Scientists können Daten somit direkt aus der Landing Zone ziehen, während weniger technisch versierte Nutzer Daten aus der Production Zone nutzen. Das Problem dieses Ansatzes besteht jedoch darin, dass jede einzelne Ebene Kopien der Daten enthält. Die geschilderten Probleme – kostenintensive Redundanz, Inkonsistenzen und problematische Governance – nehmen somit eher zu als ab. Die Umsetzung dieser Struktur ist deshalb in vielen Fällen wenig vorteilhaft.

Der logische Data Lake

Eine Alternative, die die beschriebenen Nachteile ausräumt, ist der logische Data Lake. Er präsentiert sich Nutzern in einer Weise, die sie glauben lässt, alle Daten seien zentral in einem globalen Data Lake vorhanden. Erst bei einem Blick unter die Haube offenbart sich die Wirklichkeit: Lediglich ein Teil der Daten ist in dem Data Lake gespeichert, ein anderer Teil verbleibt in den Quellsystemen und wird in den logischen Data Lake integriert. Für Optimierungen kann ein weiterer Teil der Daten in einem Cache gespeichert werden.

Ermöglicht wird diese neue Art des Datenzugriffs durch Technologien wie der Datenvirtualisierung. Diese erlaubt eine Einbindung sämtlicher Datenquellen, von einfachen Dateien bis hin zu SQL-Datenbanken, Hadoop-Clustern und Software-as-a-Service-Anwendungen. Beliebige Datenquellen können – unabhängig von dem Format und des API – in einen logischen Data Lake integriert werden.

Der Zugriff der Anwender beziehungsweise von Anwendungen auf den logischen Data Lake kann über klassische SQL-Schnittstellen (JDBC, ODBC) oder über den Aufruf von Web Services (REST, SOAP) erfolgen. Durch diese Schnittstellen ist es möglich, die Daten über beliebige Werkzeuge und Programme zu konsumieren. Die Datenvirtualisierung rückt so alle technischen Komplexitäten der Datenquellen in den Hintergrund – Nutzer können sich auf das für sie Wesentliche konzentrieren.

Die Kernbausteine der Datenvirtualisierung sind Views. Unterschieden wird zwischen Base Views und Derived Views. Base Views stellen ein Mapping zu externen Daten her – zum Beispiel zu einer Tabelle in einer Oracle-Datenbank, zu Daten in einer XML-Datei oder aus einem Web Service. Mithilfe von Derived Views kann ein semantisches Datenmodell bereitgestellt werden, das von unterschiedlichsten Werkzeugen und Programmen genutzt wird. In Derived Views können Transformationen und Kombinationen von verschiedensten Datenquellen durch SQL-Operationen abgebildet werden.

Beispiele für diese Transformationen sind Typ-Konvertierungen, Aggregationen, Filter, Joins, Unions und Berechnungen. Durch Views können Quelldaten in jede erdenkliche Form umgewandelt werden. Darüber hinaus bietet Datenvirtualisierung ausgefeilte Caching-Mechanismen. Durch das Caching wird die Ergebnismenge einer View optional in einer relationalen Datenbank oder einer Hadoop-Plattform gespeichert. Dieses Vorgehen ist sinnvoll, wenn Datenquellen keine effizienten Zugriffe er-

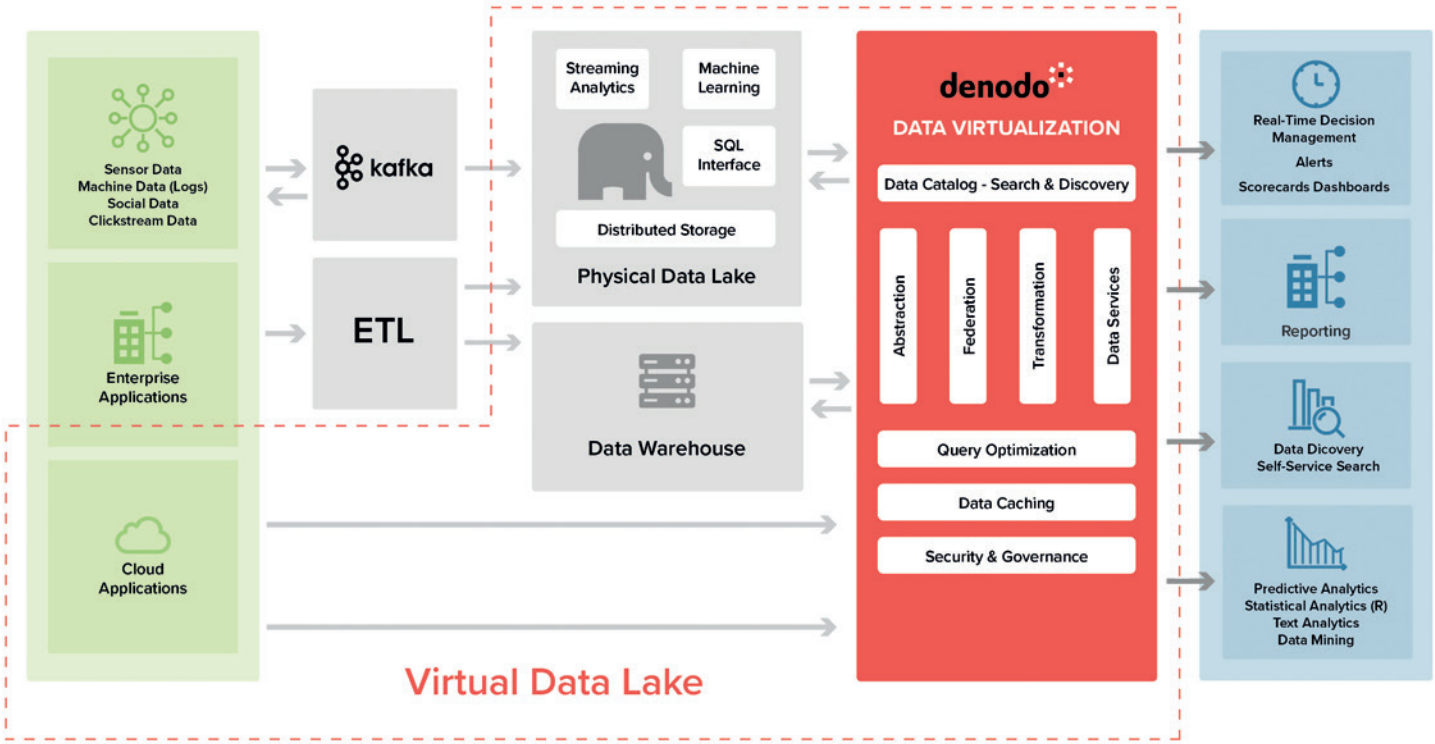


Abbildung 2: Multi Purpose Data Lake (Quelle: Denodo)

lauben oder die Quellplattform keine zusätzliche Last mehr verarbeiten kann. Das Caching ist transparent für den Anwender – bei Cached Views werden Abfragen automatisch auf den Cache umgeleitet.

Zusätzlich wird die Verarbeitung von Querys beschleunigt, da die Transformationsoperationen nicht mehr ausgeführt werden müssen.

Die Datenvirtualisierung ermöglicht das schnelle Einbinden neuer Datenquellen, da keine komplexen ETL-Mechanismen benötigt werden. Durch den in der Datenvirtualisierungsplattform integrierten Security-Layer ist es möglich, Security- und Governance-Regeln auf alle integrierten Datenquellen anzuwenden. Um effiziente Zugriffe auf Datenquellen durchzuführen, verfügt die Datenvirtualisierungsplattform über einen Cost Based Optimizer, der heterogene Abfragen auf die unterschiedlichsten Datenquellen optimiert; durch diese Optimierungen wird die Verarbeitungszeit von Querys minimiert.

Mithilfe der Datenvirtualisierung lässt sich die beschriebene Mehrebenen-Struktur eines Data Lake durch eine Kombination von Views abbilden.

Die Schwachstellen eines physischen Data Lake können durch die Datenvirtualisierung behoben werden. Die Datenvirtualisierung erlaubt die Implementierung der in ETL-Prozessen angewandten Transformationen in Views. Auf diese Weise erschließt sich eine Vielzahl von Nutzungsmöglichkeiten für unterschiedliche Nutzergruppen.

Fazit

Die vorgeschlagene Architektur lässt sich als logischer, multifunktionaler Data Lake bezeichnen *(siehe Abbildung 2)*. Im Vergleich zu herkömmlichen Data Lakes sind zahlreiche Vorteile mit multifunktionalen Data Lakes verbunden. Zum einen werden die Probleme einer zentralen Datenspeicherung umgangen, weil sich Daten auch dann analysieren lassen, wenn sie in ihren Quellen verbleiben müssen – sei es aus technischen Gründen oder in Hinblick auf die Leistung, Bandbreite, Sicherheit oder gesetzliche Vorgaben. Zum anderen kann ein logischer, multifunktionaler Data Lake sämtliche anderen Data-Delivery-Systeme innerhalb eines Unternehmens ersetzen, da er Nutzern mit unterschiedlichem Know-how Zugriff auf verschiedene Interfaces mit jeweils variierenden Analysetiefen ermöglicht. Dadurch werden Kosten gesenkt, Inkonsistenzen und Redundanzen abgebaut und die Data Governance erleichtert.

Quellen

Rick F. van der Lans: Architecting the Multi-Purpose Data Lake with Data Virtualization

Thomas Niewel
tniewel@denodo.com

Thomas Niewel ist Technical Sales Director DACH bei Denodo, wo er Kunden bei der Architektur und Implementierung von Datenvirtualisierungsprojekten berät. Er verfügt über mehr als 20 Jahre Erfahrung als Berater in den Bereichen RDBMS-Systeme, Systemarchitektur, -integration, -migration und -tuning sowie Benchmarking. Nach Abschluss seines Studiums der allgemeinen Informatik begann er seine Karriere als Anwendungsentwickler bei den Continentale Gummiwerken in Hannover. Denodo ist führend in der Datenvirtualisierung und bietet agile, leistungsstarke Datenintegration, Datenabstraktion sowie Echtzeit-Datendienste für verschiedenste Datenquellen.

Organisatorische und infrastrukturelle Anforderungen im Unternehmensumfeld für Big Data Services

Dr. Sebastian Appelhans, Thyssenkrupp

Damit aus der wachsenden Rohdatenmenge Informationen gewonnen werden können, müssen diese Daten in ihrer Qualität gesichert und dokumentiert sowie zentral zur Verfügung gestellt werden. Für diesen Umwandlungsprozess muss die technische Infrastruktur für Datenzugriff und Datenverarbeitung mit Data Governance zusammenwirken. Dies ermöglicht die Einführung einer zentralen Zugriffs- und Verwaltungsschicht für Datenquellen – den Data Hub. Mit geführten Freigabeprozessen interagiert der Data Hub mit der Data Governance, um einen qualitätsgesicherten Datenstand bereitzustellen.

Wachsende Datenmenge und Vielfalt

Die Datenmenge wächst von Jahr zu Jahr schneller und schneller – weltweit und auch bei Thyssenkrupp. Damit wachsen auch die Verwendungsmöglichkeiten. Eine große Chance öffnet sich, wenn man die taktische mit der operativen Prozessebene zusammenführt: Dabei bucht die operative Ebene die Belege eines Prozesses in die Systeme, während die taktische Ebene diese Daten zur Optimierung und Richtungsentscheidung nutzt. Damit die taktische Ebene nicht

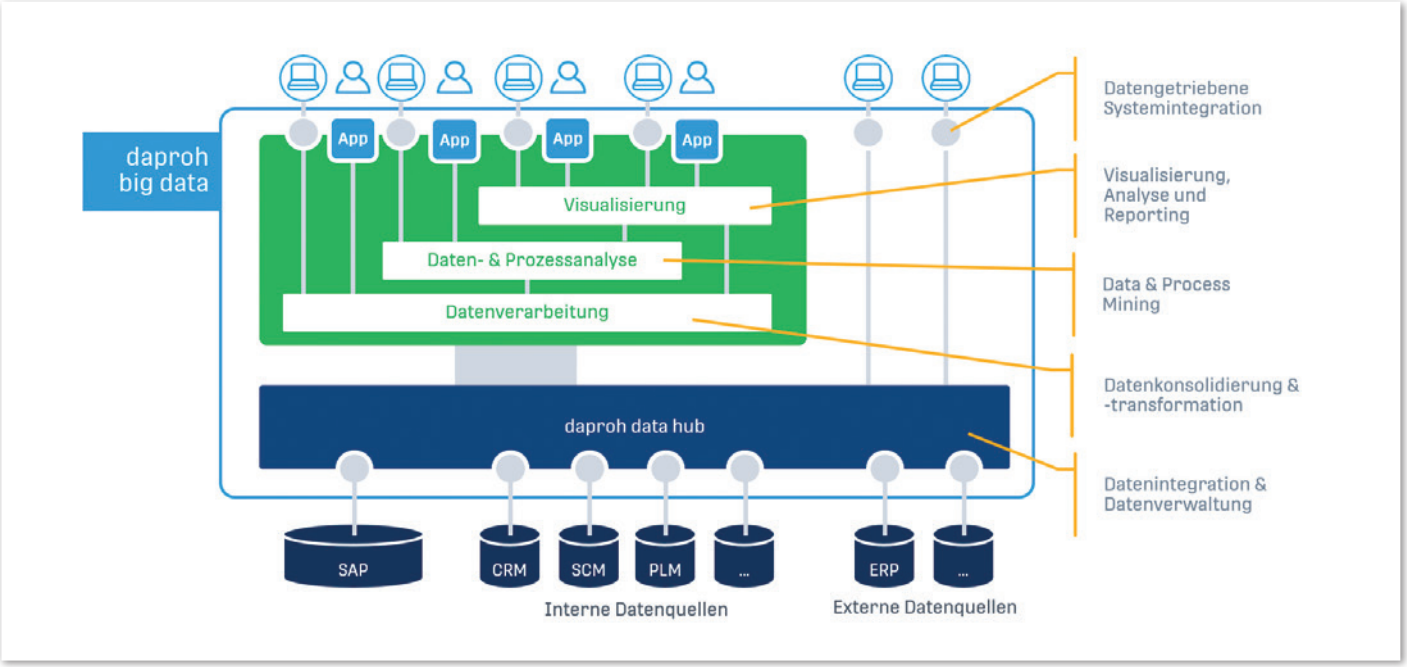


Abbildung 1: Die Daproh-Big-Data-Plattform bietet fortgeschrittene Prozessanalysemethoden, die von nicht-linearer Datenbereinigung über Process Mining bis zu Predictive Analytics in Kombination mit stufenlos skalierbarer, intuitiver Visualisierung reichen (Quelle: Thyssenkrupp).

nur Einzelbetrachtungen, sondern ein kontinuierliches Monitoring durchführen kann, ist eine effiziente Datenversorgung nötig. Aus dem Projekt Daproh (data and process harmonization) wird hierfür die Grundlage geschaffen.

Die Plattform „Daproh Big Data" hilft Thyssenkrupp dabei, diese Herausforderungen umzusetzen. Die Big-Data-Plattform basiert auf One Logic's Analytics Software One Data und dem Technologiestack der Oracle BDA. Diese Lösung bietet sowohl Reports für Nutzer als auch Microservices für Anwendungen zur effizienten Prozessoptimierung *(siehe Abbildung 1)*.

Das rasante Wachstum der Datenmengen geht mit einem ebenso schnellen Anstieg der Datenvielfalt einher. Das Prozess- und Datenmanagement eines Unternehmens wird so vor organisatorische und infrastrukturelle Anforderungen gestellt. Aufseiten der Organisation ist dafür starke Prozesskompetenz nötig, um zu identifizieren, welches Optimierungspotenzial in den Prozessen vorliegt. Die technische Infrastruktur für Extraktion und Analyse muss dabei Daten aus den vielfältigen operativen Systemen zusammenbringen, bereitstellen und aufbereiten, sodass sie analysiert werden können.

Organisation und Struktur innerhalb von Thyssenkrupp

Für das Prozessmanagement gibt es bei Thyssenkrupp eine etablierte Organisation mit klar abgegrenzten Rollen und Gremien. Die Analyse von Prozessen und Entscheidungen findet in den lokalen Gesellschaften statt, da dort das Verständnis der operativen Prozesse am größten ist.

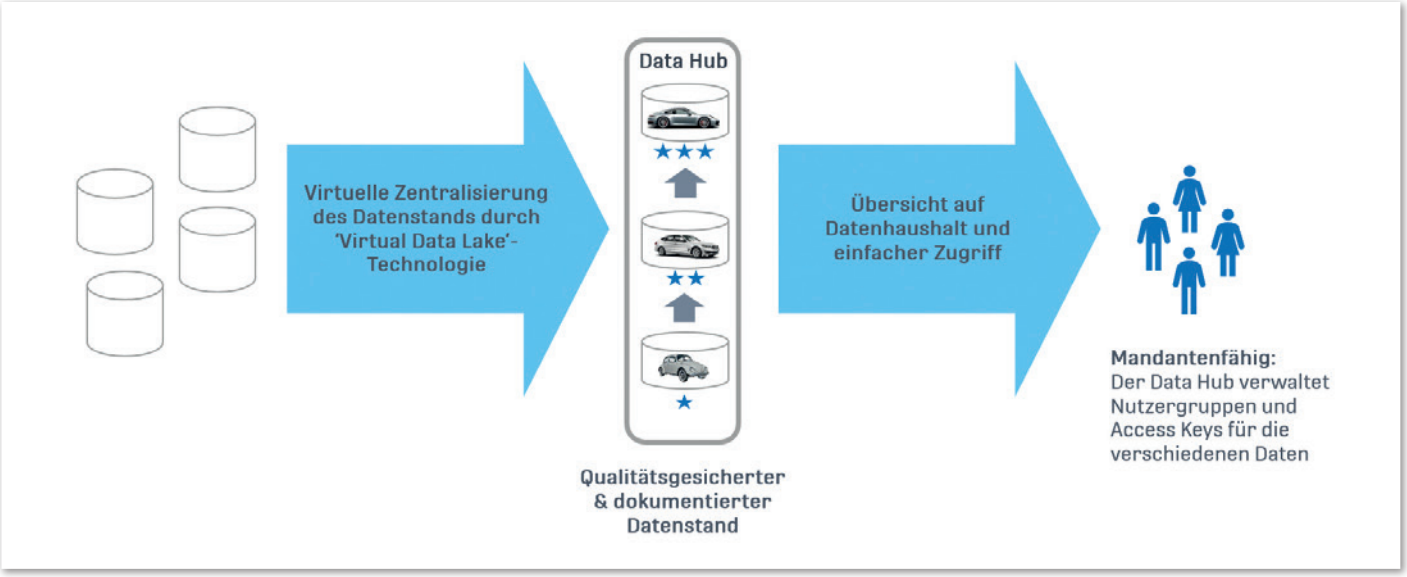


Abbildung 2: Der Data Hub stellt verschiedenen Nutzerkreisen einen qualitätsgesicherten Datenstand für produktive Applikationen, Analysen und Reports bereit (Quelle: Thyssenkrupp).

Diese Initiativen werden durch die zentrale Prozessorganisation begleitet.

Infrastruktur für Extraktion und Analyse der Daten

Aus Sicht der Infrastruktur stellt sich die Herausforderung, die Diversität der Daten in den Griff zu bekommen. Denn um die Daten für produktive Anwendungen nutzbar zu machen, müssen immer erst die richtigen Daten aus verteilt liegenden Datenquellen zusammengesucht und aufbereitet werden. Hierbei erschwert die Diversität von Datenquellen, -arten und -qualität sowohl den Zugang zu den Daten als auch deren Transparenz.

Der zentral gemanagte Datenzugriff und die qualitätsgesicherte Bereitstellung der Daten bei großer Datenmenge und Komplexität sind daher eine maßgebliche Herausforderung. Dabei sind folgende Themen zu beachten: zentraler Zugriff über einen Access Layer, Data Wrangling, Datenqualität und Dokumentation sowie Verteilung der Datenlast.

Um einen zentralen Datenzugriff zu ermöglichen, müssen diverse Datenquellen zusammengebracht werden. Abhängig von ihrer Natur liegen die Daten auf spezialisierten Quellsystemen: Streaming-Daten auf Kafka, unstrukturierte Daten in Big-Data-Clustern oder NoSQL-Datenbanken und strukturierte Daten auf klassischen relationalen Datenbanken. Die Spezialisierung der Systeme auf bestimmte Datenarten erlaubt dabei eine sehr hohe Effizienz in der Speicherung und Verarbeitung. Eine Zentralisierung des Datenzugriffs sollte daher nicht mit einer physischen Zentralisierung der Daten einhergehen.

Die Virtual-Data-Lake-Technologie bietet eine Möglichkeit, den Datenzugriff virtuell zu zentralisieren und dennoch die Daten weiterhin auf den Ursprungssystemen zu belassen. Dabei werden in einem Access Layer die Zugangsdaten für die Quellsysteme hinterlegt und Zugangsberechtigungen für verschiedene Nutzergruppen zentral verwaltet.

Um die Qualität dieser bereitgestellten Daten sicherzustellen, müssen sie einen Data-Wrangling-Prozess durchlaufen. Dieser besteht im einfachsten Fall nur aus einem Check, ob die Daten die Qualitätsstandards erfüllen.

Für den Anwender der Daten muss dabei die Qualität der Daten transparent sein. Die Qualitätsmerkmale der Daten müssen den Gütestatus des Preprocessings und den Grad der Datendokumentation reflektieren. Um dies zu erreichen, muss es einen Publishing-Prozess geben, der beispielsweise durch ein Vier-Augen Prinzip die Qualität von Datenständen und automatisierten Data-Wrangling-Prozessen sicherstellt. Diese Qualitätsmerkmale sowie die Dokumentation und weitere Metainformationen über die Daten müssen in einem Datenkatalog auffindbar sein und zentral zur Verfügung gestellt werden.

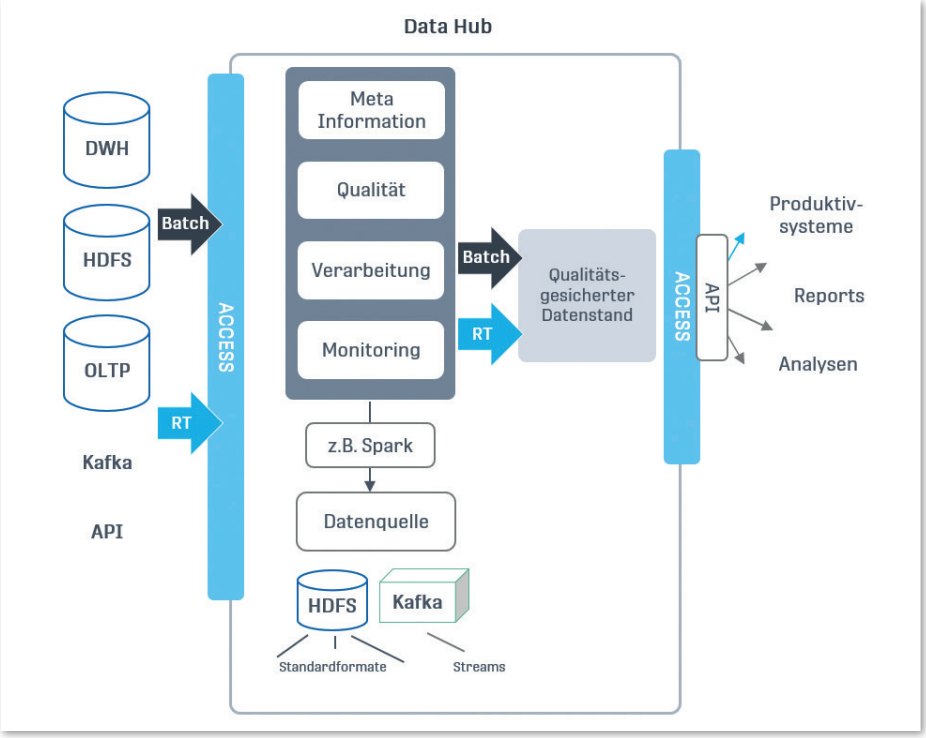


Abbildung 3: Über einen Access Layer bindet der Data Hub Rohdaten an (sowohl Batch- als auch Realtime-RT-Daten). Im Data Hub werden diese Rohdaten automatisch überprüft (Monitoring) und verarbeitet (Verarbeitung). Zudem werden hier zentral Metainformationen zu den Daten sowie Informationen zur Datenqualität hinterlegt. Das Ergebnis ist ein sauberer Datenstand, der wiederum über einen Access Layer zur Verwendung in Analysen, Reports und Produktivsystemen freigegeben wird (Quelle: Thyssenkrupp).

Der Data Hub führt Governance und Datenverarbeitung in einer Software zusammen

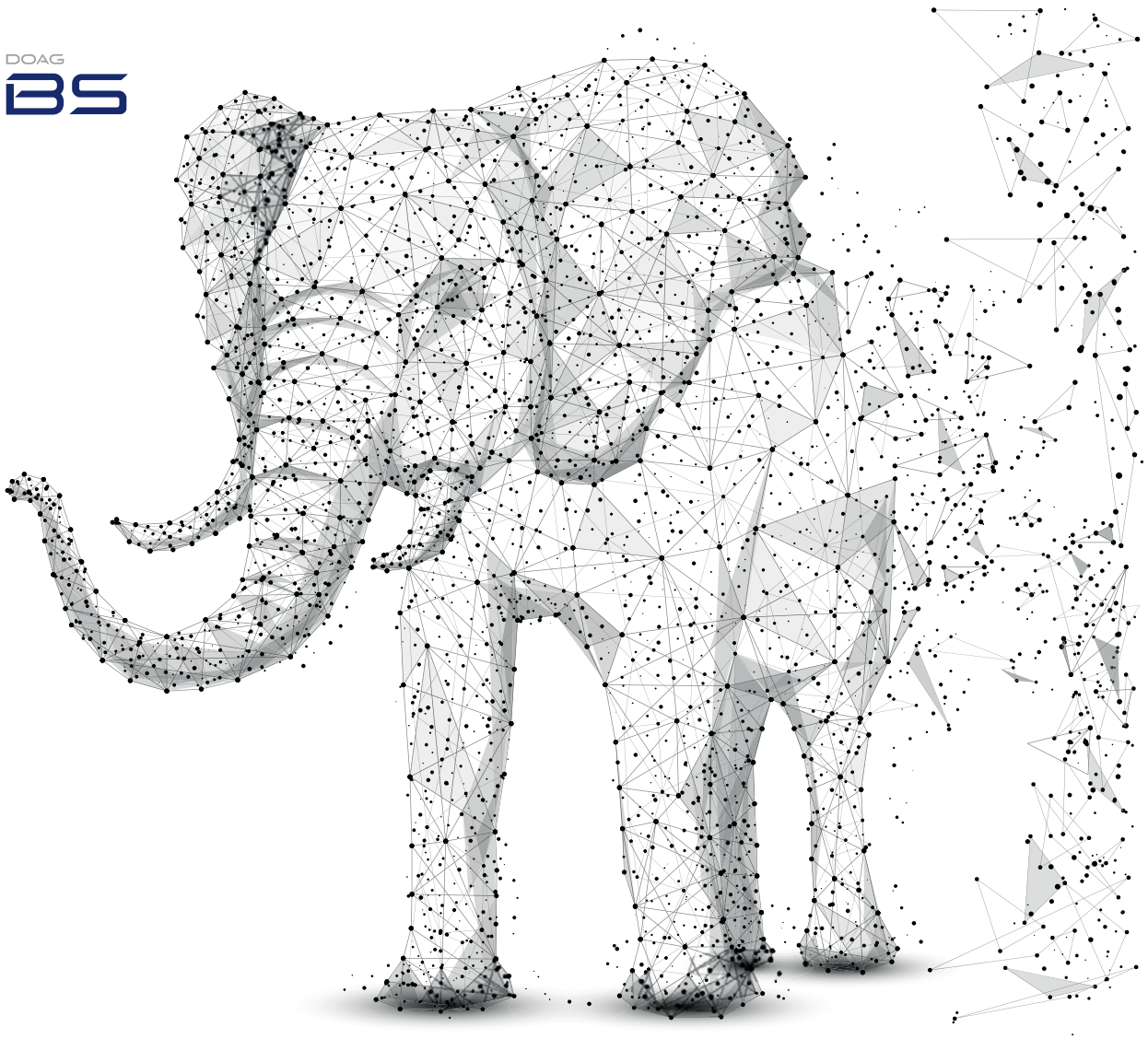
Die genannten Anforderungen werden durch einen Data Hub erfüllt. Er dient als zentrale virtuelle Zugriffs- und Verwaltungsschicht von qualitätsgesicherten Datenquellen und bringt dabei die technische und organisatorische Seite des zentralen Datenmanagements zusammen *(siehe Abbildung 2)*. Für die organisatorische Seite, die Data Governance, ist dabei eine hohe Prozesskompetenz notwendig, was bei Thyssenkrupp von der Daproh-Organisation übernommen wird. Im Data Hub wird dabei die Data Governance durch geführte Freigabeprozesse abgebildet. Um die Skalierbarkeit des Data Hub zu gewährleisten, muss die Last der Daten verteilt werden. Eine Plattform wie Oracle BDA bietet sich daher als technische Basis für ein solches System an.

Fazit

Der Data Hub ermöglicht zusammen mit Data Governance eine effiziente Umwandlung von Rohdaten in nutzbare Daten. Die zentrale Bereitstellung dieser Daten im Data Hub erlaubt es, Ergebnisse aus der Proof-of-Concept-Ebene hochzuskalieren und unternehmensweit produktiv zu setzen. Damit wird die Vision einer auf Echtzeitdaten basierenden Entscheidungsfindung Realität. Thyssenkrupp vertraut dabei auf den One-Logic-Ansatz „from garage to production", um das Potenzial im Datenwachstum produktiv nutzbar zu machen.

Dr. Sebastian Appelhans
sebastian.appelhans@thyssenkrupp.com

Dr. Sebastian Appelhans ist Head of tk Business Process Management bei der Thyssenkrupp AG. Nach Abschluss seiner Promotion an der WWU Münster im Jahr 2000 bekleidete er verschiedene leitende Positionen innerhalb der Siemens AG. 2015 wechselte er zur Thyssenkrupp AG. Neben der Optimierung der Geschäftsprozesse und der Implementierung der konzernweiten „big data"-Plattform ist er insbesondere für die konzeptionelle Weiterentwicklung sowie inhaltliche Umsetzung des Konzernoptimierungsprogramms Daproh verantwortlich.



Eine Safari durch den Datendschungel am Beispiel eines Recommender-Systems

Matthias Hofmaier und Dr. Arthur Varkentin, Novatec Consulting

Täglich wachsende Datenmengen stellen Unternehmen oftmals vor große Herausforderungen, sind jedoch auch Basis für Automation und Vereinfachung von Prozessen. Eine Methode dafür wollen wir anhand eines Fallbeispiels aufzeigen und Sie ein wenig wie auf einer Safari durch die Thematik begleiten. Ein Empfehlungssystem, das die Rechnungsstellung bei Versicherungen erleichtern soll, dient dabei als unser gedanklicher Anker.

Aktuell produzieren wir weltweit jeden Tag ca. 2,5 Exabyte (2,5 * 1018 Byte!) an Daten. Damit machen die Daten, die in den letzten zwei Jahren generiert wurden, über 90 Prozent des gesamten Datenbestands aus [1]. Gerade durch die vermehrte Nutzung von Technologien wie dem Internet of Things (IoT) ist die Tendenz steigend. Neben den direkten Daten aus Geräten entstehen Statistiken und Metadaten, Analysen und Kennzahlen, also weitere Daten. Dass dies kein lineares Wachstum ist, liegt auf der Hand. Wo wir heute sind und erst recht wo wir morgen stehen werden, lässt sich wohl nur schwerlich abschätzen.

Rechenleistung und insbesondere Speicher werden immer günstiger, sodass es für Unternehmen oftmals kein nennenswerter Kostenfaktor mehr ist, Daten zu speichern und vorzuhalten. Schon längst positionieren sich Anbieter von Cloud und Infrastructure as a Service (IaaS), um den steigenden Bedarf am lukrativen, wachsenden Markt zu decken. Die Frage, die sich für Unternehmen mehr und mehr stellt, ist: Wie können wir diese Daten verwenden, um für uns daraus einen Mehrwert zu generieren? Sei es, um

die internen Prozesse zu verbessern oder die Kunden und Zielgruppen besser zu verstehen. Es ist weitestgehend unstrittig, dass in vielen Daten erhebliches Potenzial schlummert, doch das ist nur die halbe Wahrheit. Der Weg, dieses Potenzial zunächst zu erkennen und dann auch zu bergen, kann mitunter sehr steinig sein. Und es ist bei Weitem nicht jede Datenhalde eine Goldgrube.

Ist das Potenzial erst einmal identifiziert, so gibt es inzwischen gute Möglichkeiten, Ordnung in den Datendschungel zu bringen. Glücklicherweise oder vielleicht auch notwendigerweise entwickelt sich parallel zur Datenexplosion auch die Rechenleistung weiter und ermöglicht es, (schon lange existierende) Algorithmen aus dem Bereich des maschinellen Lernens (ML) effizient zu implementieren. Diese können eindrucksvoll schnell riesige Datenmengen prozessieren und Zusammenhänge offenlegen, die ansonsten verborgen geblieben wären.

Auf die nunmehr verfügbaren Technologien und Plattformen einzugehen, würde den Rahmen bei Weitem sprengen und könnte ganze Bücher füllen – was es im Übrigen auch tut.

Wir möchten Sie auf eine kleine Safari durch den Datendschungel mitnehmen und Ihnen anhand eines Beispiels eine Möglichkeit aufzeigen, mit der ein Mehrwehrt aus vorhandenen Daten gewonnen werden kann.

Dschungel in Sicht: Rechnungsstellung in der Versicherungsbranche

Eine typische Domäne für eine große Anzahl von Daten ist die Versicherungsbranche. Erfolgt ein Schaden, etwa durch einen Brand oder eine Überschwemmung, wird dieser zunächst bei der Versicherung gemeldet. Anschließend wird der Schaden durch einen Dienstleister instandgesetzt und abgerechnet. So gehen bei Versicherern täglich Tausende von Rechnungen ein, die reguliert werden müssen. Je nach Versicherer tauchen verschiedene Rechnungspositionen auf, die bestimmten Leistungsposten entsprechen und in Katalogen vorgegeben werden. Dabei können zwei Rechnungsdokumente an verschiedene Versicherer, die eine vergleichbare Leistung abrechnen, durchaus unterschiedlich aussehen. Der Katalog von Versicherer A beinhaltet für das Streichen einer Fassade beispielsweise die Rechnungspositionen *Reinigungsarbeiten, Abklebearbeiten, Grundieren* und *Streichen*, wohingegen bei Versicherer B nur die Positionen *Vorarbeiten* und *Streichen* abgerechnet werden können. Gerade aus Sicht eines Dienstleisters wird die Rechnungsstellung so zu einem sehr komplexen Prozess, der nicht nur zeitintensiv, sondern auch sehr fehleranfällig ist *(siehe Abbildung 1)*.

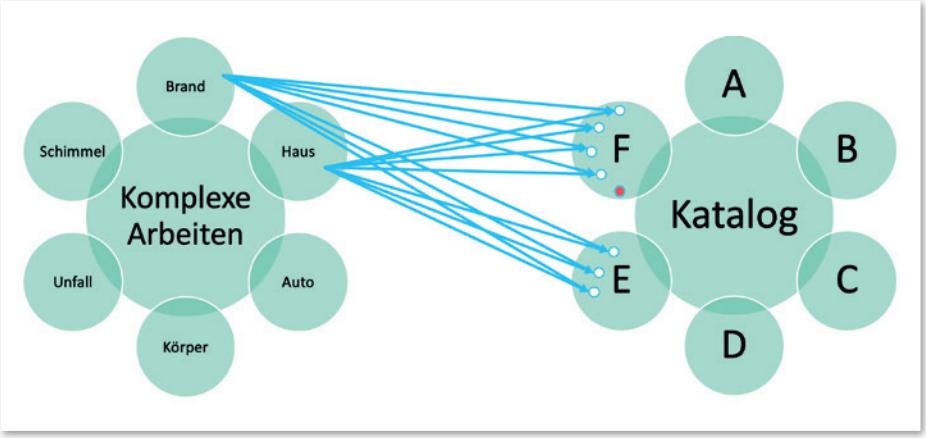


Abbildung 1: Abhängigkeiten zwischen verrichteten Dienstleistungen und Abrechnungskatalogen. Für einen Posten aus dem Katalogbereich F (rot markiert) ist keine Rechnung für eine entsprechende Arbeit eingegangen oder wurde anders bzw. falsch aufgeschlüsselt (Quelle: Novatec Consulting).

Rettung naht: das Recommender-System

Recommender-Systeme helfen, den oben genannten Prozess zu vereinfachen. Sie nutzen Daten aus der Vergangenheit und machen so Vorschläge für aktuelle Problemstellungen. Ein typisches Beispiel für ihre Anwendung sind Warenkörbe von Onlinehändlern. Hier werden oft Produkte empfohlen, die in ähnlichen Kombinationen auch schon in Warenkörben anderer Kunden gelandet sind. Ein anderes Beispiel sind Freundschaftsvorschläge auf Social-Media-Portalen. Dabei dienen die Vernetzungspfade als Anhaltspunkte für die Empfehlung. In unserem Anwendungsfall sollen bei der Erstellung der Rechnung automatisch Rechnungspositionen vorgeschlagen werden, die einer zu berechnenden Dienstleistung zugeordnet werden können oder eventuell vergessen wurden.

Recommender-Systeme selbst können in drei verschiedene Bereiche gegliedert werden: inhaltsbasiert, kollaborativ und hybrid. Inhaltsbasierte Systeme nutzen Nutzerprofile und Objektbeschreibungen, um Vorschläge zu machen. Die Profile entstehen dabei aus bereits vergangenem Verhalten von Nutzern und bilden deren Präferenzen ab. Typische Mechanismen zur Erstellung solcher Nutzerprofile sind beispielsweise Ratings oder das aus Social-Media-Portalen bekannte *Gefällt mir*. Sobald ein solches Profil erstellt ist, kann es mit verschiedenen Objektbeschreibungen verglichen werden. Wenn nun eine hohe Ähnlichkeit zwischen Profil und Beschreibung besteht, wird dies als Vorschlag angeführt.

Kollaborative Systeme hingegen berufen sich nicht auf Nutzerprofile und Beschreibungen, sondern analysieren das Vorgehen von Nutzern auf ähnliche Verhaltensmuster und nutzen die so erkannten Regelmäßigkeiten als Vorschlagskriterien. Herausfordernd hierbei ist das sogenannte Kaltstart-Problem. Wenn keine Daten für Entitäten mit ähnlichen oder gleichen Verhaltensmustern vorhanden sind, kann auch kein Vorschlag für aktuelle Situationen gebildet werden. Ein Weg, um dieses Problem zu überwinden, kann ein hybrider Ansatz sein. Wie der Name verspricht, werden hier inhaltsbasierte und kollaborative Verfahren genutzt und später über Techniken wie Voting miteinander kombiniert.

Im Falle unserer Problemstellung können wir auf eine große Menge vorhandener Daten zurückgreifen. Diese bestehen aus ca. 50.000 Rechnungen, die sich insgesamt aus ungefähr einer Million Rechnungspositionen (Datenpunkte) zusammensetzen. Im folgenden Abschnitt wird nun näher beleuchtet, wie wir mithilfe verschiedener Techniken und den genannten Daten ein kollaboratives Vorschlagssystem realisieren können.

Assoziationsanalyse, Autoencoder oder neuronales Netzwerk – Buschmachete, Taschenmesser oder doch lieber eine Motorsäge?

Da die Datenbasis und die verschiedenen Arten von Recommender-Systemen nun bekannt sind, gilt es, sich die Frage nach der konkreten Umsetzung zu stellen. In unserem

Fall haben wir uns dazu entschieden, drei verschiedene Techniken zu verwenden und deren Ergebnisse zu vergleichen.

Die erste Version unseres Vorschlagssystems wurde durch eine Assoziationsanalyse umgesetzt. Hierbei werden die Vorkommen von Rechnungspositionen und deren Kombinationen im Datensatz gezählt. Die Häufigkeit eines Elements oder einer Kombination von Elementen über den Datensatz wird dabei als *Support* bezeichnet. Anhand des Supports kann die bedingte Wahrscheinlichkeit für das Vorkommen von Elementen, meist Konfidenz genannt, berechnet werden. Mit der berechneten Konfidenz können nun klare Aussagen getroffen werden *(siehe Abbildung 2)*: „Wenn Rechnungsposition A aufgeführt wurde, wurde in X Prozent der Fälle auch Rechnungsposition B aufgeführt." Für unser System wurden also alle bisherigen Positionen auf einer Rechnung getrennt nachgeschlagen und ab einem gewissen Konfidenz-Schwellwert als Vorschläge akzeptiert. Mithilfe dieser Technik kann ein Vorschlagssystem sehr schnell und wenig komplex realisiert werden. Warum dies jedoch nicht alle Probleme löst, wird später im Vergleich der Ergebnisse klar.

Variante zwei unseres Systems wurde durch einen sogenannten *Autoencoder* umgesetzt *(siehe Abbildung 3)*. Hierbei handelt es sich um ein Vorgehen aus dem maschinellen Lernen, bei dem eine komprimierte Repräsentation der Eingabe erlernt wird. Der Lernprozess sieht dabei so aus:

- Komprimiere Eingabe auf Kern fixer Größe.
- Versuche, die Eingabe aus diesem Kern zu rekonstruieren.
- Vergleiche Eingabe mit Rekonstruktion.
- Passe Gewichte an, sodass Distanz zwischen Eingabe und Rekonstruktion minimal wird.

Durch die Kompression versucht der Autoencoder, die Semantik einer Rechnung zu erfassen. Dabei werden Kennzahlen, die für die Rekonstruktion relevant sind, erhalten und die redundanten Informationen verworfen. In der Rekonstruktion lernt der Autoencoder, wie die typische Rechnung für die gegebenen Kennzahlen erstellt werden kann, für die der Fehler im Schnitt am geringsten ist. Was hier zunächst komplex klingen mag, ergibt für unser Vorschlagssystem durchaus Sinn. Hier trainieren wir den Autoencoder anhand unserer verschiedenen vollständigen Rechnungen. Wenn wir nun eine unvollständige Rechnung eingeben, kann diese mit der vom Autoencoder erstellten Rekonstruktion verglichen und deren Abweichung als Vorschlag verwendet werden. Dieses Vorschlagssystem ist allerdings nur funktionsfähig, wenn eine solche Rekonstruktionslogik auch ausreichend erlernt werden konnte.

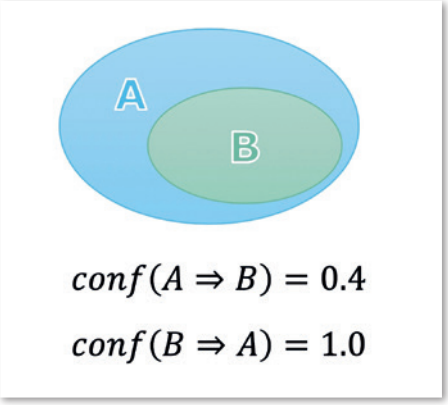


Abbildung 2: Assoziationsanalyse – Konfidenzberechnung anhand eines Euler-Diagramms (Quelle: Novatec Consulting)

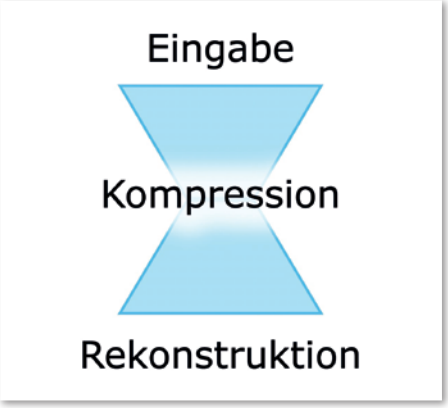


Abbildung 3: Beispielhafte Darstellung eines Autoencoder (Quelle: Novatec Consulting)

Zu guter Letzt: das neuronale Netzwerk *(siehe Abbildung 4)*. Auch dieses generische Lernmodell eignet sich, um Vorschlagssysteme abzubilden. Das Netz muss entsprechend unserem Problem modelliert und anschließend trainiert werden. In unserem Fall modellieren wir unsere Vorschläge als Klassifikationsproblem, das heißt, jede Rechnungsposition wird in der Ausgabe als Klasse repräsentiert. Diese jeweiligen Klassen zeigen nun an, welche Position in der eingegebenen Rechnung fehlt oder fehlen könnte. Die Lernaufgabe des Netzwerks besteht also darin, alle Rechnungen zu klassifizieren. Als Trainingsdaten für das neuronale Netzwerk nutzen wir die bestehenden Rechnungen, aus denen wir für alle Rechnungen mit mehr als zwei Positionen jeweils eine Position entfernen und sie als Label hinzufügen. Die verbleibenden Posten werden dann als Eingabe verwendet. Wir wissen also genau, welche Position bei einer Eingabe fehlt, weil wir sie selbst herausgenommen haben.

Das Netzwerk hat den Vorteil, dass auch nichtlineare Abhängigkeiten erlernt und je nach Trainingsdaten höherwertige Kombinationen von Positionen in den Vorschlagprozess einfließen können.

Qualität der Recommender

Abschließend stellt sich nun die Frage: Wie haben unsere verschiedenen Recommender-Systeme abgeschnitten? Um dies zu beantworten, müssen wir zunächst nach geeigneten Evaluationsmetriken Ausschau halten oder aber eigene definieren. Ein mögliches Vorgehen ist dabei die Ermittlung des bereits entstandenen Schadens. Also des Betrags, der dem Dienstleister bisher durch vergessene Rechnungspositionen entgangen sein könnte. Dafür werden für alle Rechnungen Vorschläge erstellt. Diese werden dann ab einem gewissen Konfidenzschwellwert als korrekt akzeptiert und aufsummiert. Der Schwellwert muss später je nach Einsatzdomäne und in Absprache mit dem jeweiligen Fachbereich festgelegt werden, da zunächst nicht klar ist, welche Vorschläge wirklich als korrekt zu betrachten sind. Die Ergebnisse *(siehe Abbildung 5)* zeigen, dass der Recommender basierend auf der Assoziationsanalyse den größten Schaden schätzt. Er hätte im realen Fall also mehr Vorschläge als die anderen beiden Vorgehensweisen geliefert. Der Autoencoder liegt bei der Vorschlagsakzeptanz in etwa zwischen der Assoziationsanalyse und dem neuronalen Netzwerk.

Abbildung 4: Neuronales Netzwerk zur Vorschlagsfindung (Quelle: Novatec Consulting)

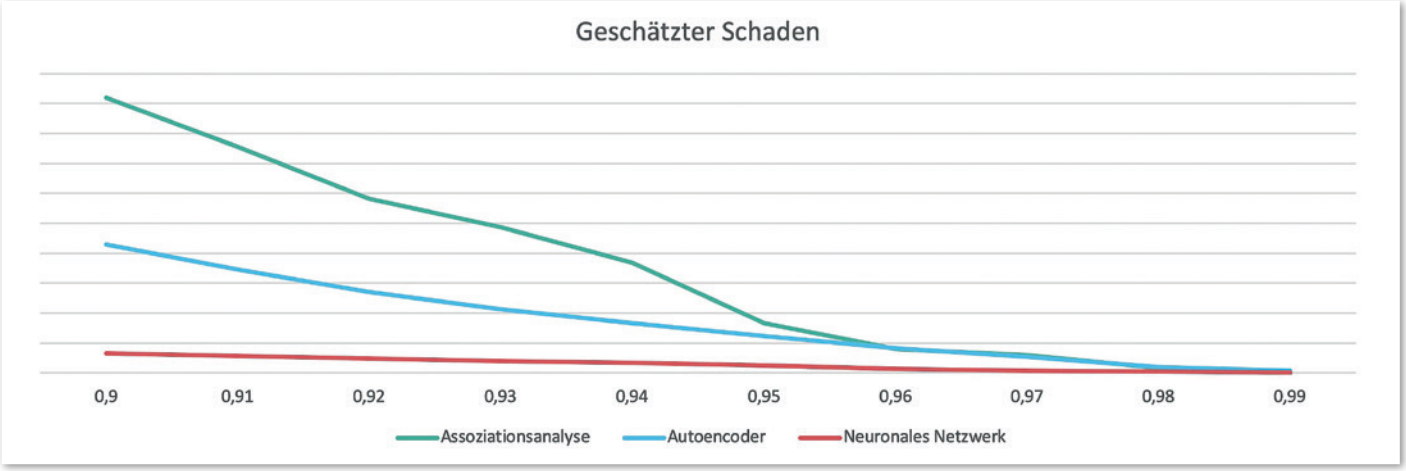


Abbildung 5: Geschätzter Schaden aufgrund von entgangenen Rechnungspositionen über verschiedene Konfidenzschwellwerte (Quelle: Novatec Consulting)

Im Kontext von Vorschlagssystemen heißt es allerdings nicht immer, dass viele Vorschläge auch gleich viel Mehrwert liefern müssen. In manchen Domänen ist sogar eher das Gegenteil der Fall: lieber keine Vorschläge als falsche. Daher wird eine weitere Metrik benötigt, um die Güte der einzelnen Systeme zu evaluieren. Eine typische Methode hierfür ist die sogenannte *Top-N-Metrik*. Bei der Erstellung der Testdaten wird analog zu der Trainingsdatenerstellung für das neuronale Netzwerk vorgegangen und für Rechnungen mit mehr als zwei Positionen jeweils eine Position entfernt (Leave-one-out cross validation). Anschließend wird geprüft, ob sich die entfernte Position unter den N besten Vorschlägen befunden hat. Die Verteilung zwischen Trainings- und Testdaten liegt dabei in einem Verhältnis von 80 zu 20.

Wenn Sie nun diese Metrik betrachten *(siehe Abbildung 6)*, können Sie klare Unterschiede sehen und feststellen, dass nicht zwingend das System mit der größten Vorschlagsakzeptanz am besten abschneidet.

Gerade der Recommender, der durch ein neuronales Netz umgesetzt wurde, überzeugt trotz seines geringen Vorschlagsakzeptanzwertes mit einem hohen Top-N-Score. Die Assoziationsanalyse mit ihrem hohen Akzeptanzwert liegt im Mittelfeld und der Autoencoder kann am wenigsten mit seiner Genauigkeit überzeugen. Auch bis hin zum Top-5-Score (Vorschlag an fünfter Stelle gefunden) verändert sich diese Rangfolge nicht. Hier zeigt das neuronale Netzwerk eine absolute Genauigkeit von circa 50 Prozent, das System lag demnach in der Hälfte der Fälle mit einem seiner Vorschläge richtig.

Mit diesen Ergebnissen können wir also durchaus funktionierende Umsetzungen eines Recommender-Systems präsentieren. Für eine konkrete Auswahl des Verfahrens muss nun zwischen Vorschlagsmenge und -güte abgewogen werden. Dies entscheidet sich je nach Rahmenbedingungen der Anwendung.

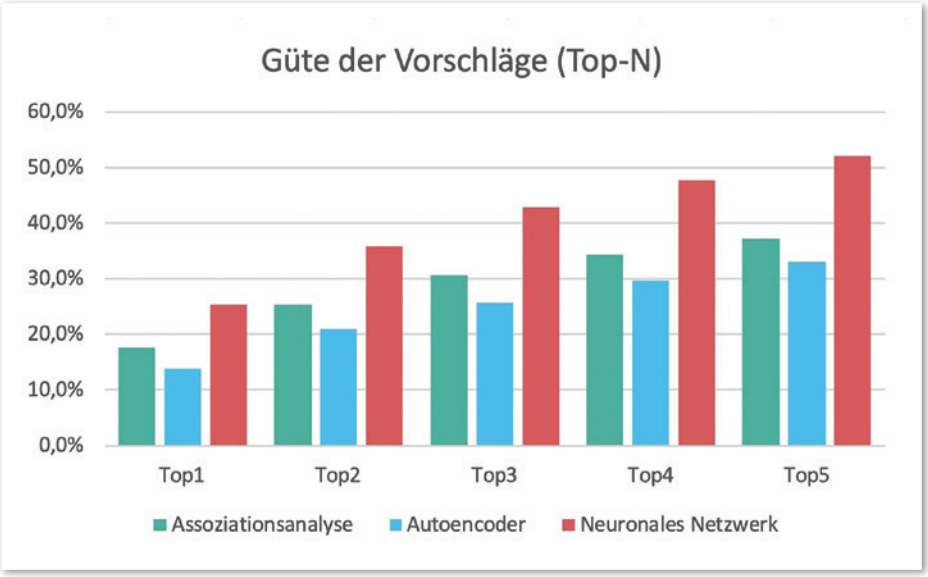


Abbildung 6: Vorschlagsgüte der einzelnen Systeme über eine Top-N-Metrik (Quelle: Novatec Consulting)

Die Safari neigt sich dem Ende

Sie konnten nun beobachten, wie Daten genutzt wurden, um ein reales Business-Problem anzugehen und einen Mehrwert aus vorhandenen Daten zu generieren. In diesem Fall speist sich dieser Wert hauptsächlich aus der Zeitersparnis und der Entlastung bei der Rechnungsstellung sowie gegebenenfalls einer höheren Erkennungsrate von vergessenen Positionen. Für Ihr jeweiliges Business-Problem kann der Wert anders geartet sein. Nicht jedes Problem lässt sich mit Daten lösen, doch es empfiehlt sich, ein Problem darauf mindestens zu prüfen.

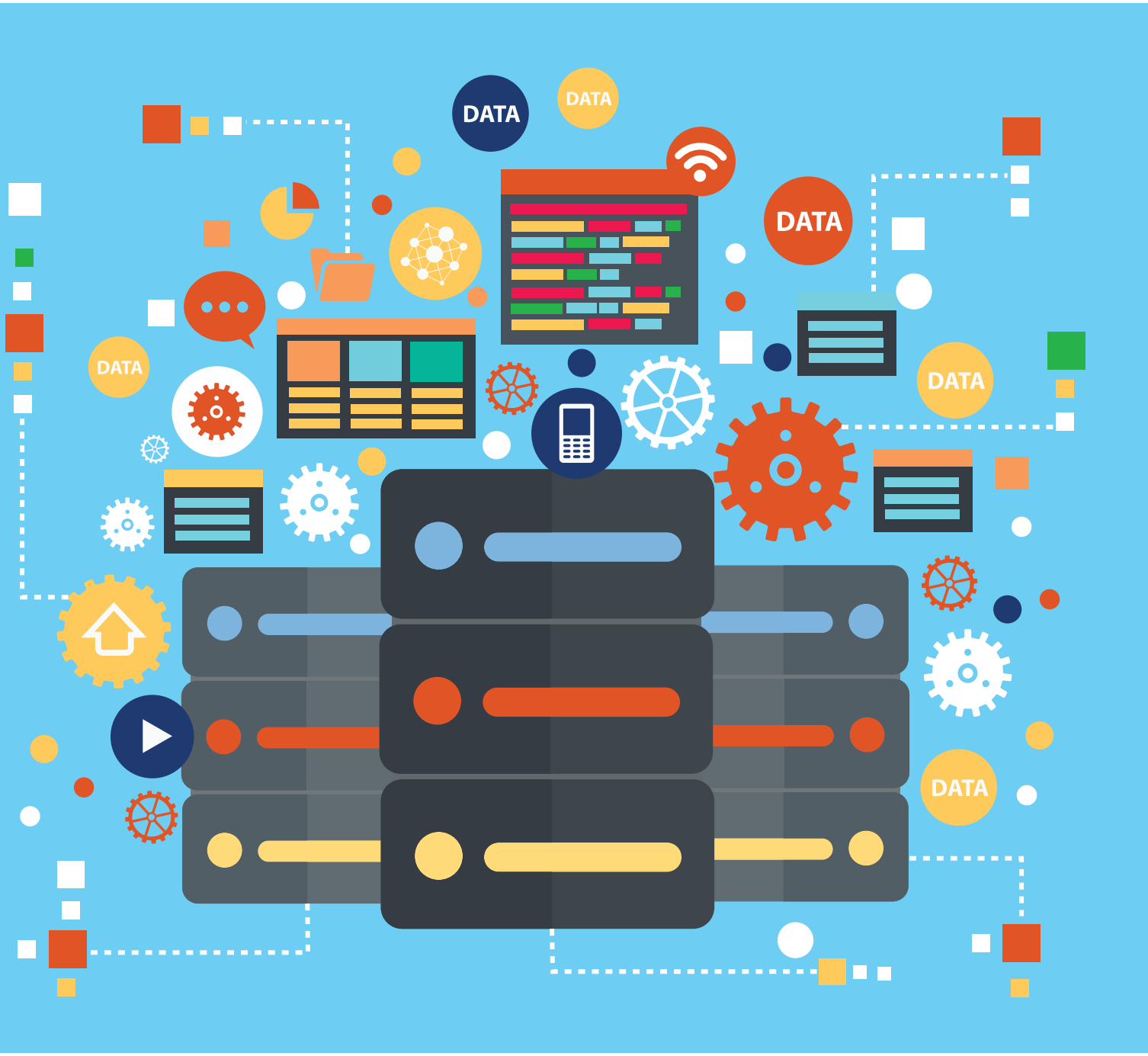
Quelle

[1] *https://www.domo.com/learn/data-never-sleeps-5*

Matthias Hofmaier
matthias.hofmaier@novatec-gmbh.de

Dr. Arthur Varkentin
arthur.varkentin@novatec-gmbh.de

Dr. Arthur Varkentin promovierte 2018 in Physik und setzte sich im Rahmen seiner Doktorarbeit mit Segmentierungs- und Klassifizierungsproblemen auseinander. Matthias Hofmaier studierte Medien und Kommunikationsinformatik und beschäftigte sich im Rahmen seiner Abschlussarbeit mit maschinellem Lernen auf mobilen Architekturen. Beide wirken seit 2019 bei der Novatec GmbH in den Bereichen „Maschinelles Lernen" und „Künstliche Intelligenz" mit.



Agile Methoden und Data Warehouse – ein Praxisbericht

Dr. Susanne Bosinger, DP DHL IT Services

Während sie in der klassischen Softwareentwicklung längst etabliert sind, trifft man agile Methoden im Umfeld von Data Warehouse (DWH) noch eher selten an. Die Methode selbst soll in diesem Artikel nicht im Vordergrund stehen. Vielmehr soll es ein Erfahrungsbericht aus der Perspektive eines technischen Teams sein, ein Bericht über Freud und Leid mit dem Einsatz agiler Methoden. Dabei liegt der Fokus auf den täglichen Herausforderungen auf technischer und menschlicher Ebene sowie darauf, wie es dennoch funktionieren kann.

Das Projekt war angesiedelt in einem Großunternehmen in der Logistikbranche mit verbreitetem Einsatz des Wasserfallmodells und teils monolithischen Projekten. Wir sollten als Teilprojekt einer größeren Plattform unser System quasi auf der grünen Wiese neu entwickeln. Die Vorgehensweise im Gesamtprojekt war Scrum.

Besonderheiten in DWH- und BI-Projekten

Die klassischen Projektmanagementmethoden und Vorgehensmodelle sind oft nicht optimal, wenn die Anforderungen sich während des Projektes ändern oder im Projektverlauf neue Erkenntnisse gewonnen werden, aus denen wiederum neue Anforderungen resultieren. Beides ist jedoch typisch für DWH- und BI-Projekte.

Gerade in DWH-Projekten besteht die Herausforderung, dass es viele Layer und vernetzte technische Abhängigkeiten gibt. Während Anpassungen einer gut gekapselten Funktionalität leicht zu schätzen und mit überschaubarem Aufwand umzusetzen sind, kann die Anforderung, neue Attribute aus einer Datenquelle zu übernehmen, leicht den Rahmen eines Sprints sprengen, da mehrere Layer betroffen sind und es weitreichende technische Abhängigkeiten gibt.

Das Team

Ein neues Projekt mit einem neuen Team quasi auf der grünen Wiese zu starten, kann Fluch und Segen zugleich sein. Bis auf den Scrum Master und den „Teilzeitarchitekten" hatte das Team (inklusive Product Owner und vier Entwicklern) bis dato keine Erfahrungen mit Scrum gemacht.

Wir waren verteilt über vier Standorte, hatten eine Altersspanne von Ende 20 bis Mitte 60, vier verschiedene Nationalitäten und die verschiedensten technologischen und persönlichen Hintergründe. Während die einen jahrzehntelang mehr oder weniger glücklich mit dem Wasserfallmodell gelebt hatten, brannten andere darauf, sich mit dem neuen Vorgehen auseinanderzusetzen.

Eine der größten Herausforderungen bestand für uns darin, dass Scrum-Teams selbst organisiert sind und kein Projektleiter einem sagt, was man tun soll. Auch gab es weder eine klar definierte Vorgehensweise noch ein detailliertes technisches Design. Einen Lösungsansatz zu erarbeiten, lag in der Verantwortung des Teams.

Anfänglich gab es oft abweichende Einschätzungen, welche Lösung die beste wäre und welcher Weg dorthin führen sollte. In regelmäßigen Abstimmungsmeetings haben wir schließlich ein gemeinsames Verständnis darüber entwickelt, wie wir miteinander arbeiten wollen und wann etwas als erledigt anzusehen ist. Das war zeitintensiv und herausfordernd, aber letztendlich der Schlüssel, da wir uns aus eigener Kraft und eigenem Antrieb auf unser Vorgehen und auf den Anspruch an uns selbst geeinigt hatten. Dieser Invest hat sich später mehr als ausgezahlt.

Es gab auch unlösbare Konflikte innerhalb des Teams. So hat ein Kollege auf eigenen Wunsch das Team verlassen, da er sich mit der Methode nicht anfreunden konnte. Fluktuation und Unruhe in der Besetzung im Team sind ein Störfaktor. Für Scrum-Teams gilt das vielleicht noch ein wenig mehr als für normale Entwicklerteams, da die Zusammenarbeit und das Zusammenspiel innerhalb des Teams enger und verzahnter sind und personelle Wechsel somit gravierende Auswirkungen haben.

Die Meetings

Unsere Sprints hatten eine Länge von 14 Tagen mit regelmäßigen Sprintwechseln (Review). Im Anschluss fanden die Retrospektive und das Planning für den nächsten Sprint statt. Damit war dieser Tag weitgehend mit Meetings verplant, was andererseits jedoch für die restlichen Tage des Sprints mehr unterbrechungsfreie Zeit für das Umsetzen mit sich brachte. Nach jeweils der Hälfte eines Sprints führten wir unser Story Refinement durch.

Als Tool für das Backlog und unsere Tasks war Jira gesetzt. Da wir an verschiedenen Standorten arbeiteten, war der Einsatz eines virtuellen Boards unerlässlich. Die Dailys fanden vormittags via Skype statt. Wir waren angehalten, das Board vor den Meetings zu aktualisieren.

In maximal 15 Minuten berichtete jeder, was er gemacht hatte, was geplant war und wo gegebenenfalls Fragen oder Probleme aufgetaucht waren. Darüber hinaus legte der Product Owner dar, ob es vom Kunden neue Informationen oder Rückmeldungen gab. Konsequentes Zeitmanagement ist dabei notwendig, um das vereinbarte Timeboxing einzuhalten. Wenn es technische Details zu besprechen gab oder ein Peer Review anstand, wurde dies mit den betroffenen Personen direkt im Anschluss an das Daily besprochen.

Das Sprint Planning war angesetzt auf zwei Stunden, wurde aber nach Bedarf länger oder kürzer gestaltet. Hier wurden Storys aus dem Backlog besprochen, hinterfragt und bewertet. Bei der Schätzung gab es anfänglich Irritationen, da es beim Planning Poker ja um die Komplexität der Storys beziehungsweise der Tasks geht und nicht um die tatsächliche Dauer. Das war ungewohnt. Nach ein paar Sprints hatte sich das jedoch eingeschwungen. Es war hilfreich, Storys aus vergangenen Sprints als Referenz zu nehmen.

Alle zwei Wochen kamen wir zum Sprint Review zusammen, um die Ergebnisse des Sprints vorzustellen. Neben dem Team war auch die Kundenseite beteiligt, die in Polen, Belgien und den Niederlanden saß. Zu Beginn des Projekts gab es einige On-site-Meetings, zu denen alle Beteiligten von allen Standorten zusammenkamen. Das war zwar initial aufwendig, letztendlich aber unerlässlich, um eine gute Basis für die Zusammenarbeit zu schaffen. Gerade vor dem Hintergrund eines verteilten Teams war es essenziell, alle Beteiligten mehrfach persönlich zu treffen und zu sprechen.

Für die späteren Reviews wurden die Kunden und Kollegen von anderen Standorten per Skype dazu geschaltet. Dabei wurden die Ergebnisse des Sprints von den jeweiligen Teammitgliedern vorgestellt und erläutert. Während der Präsentation konnten Fragen gestellt und Punkte diskutiert werden, das Klima war hier offen. Am Ende der Vorstellung konnte der Kunde dann sagen, ob er den Sprint abnimmt.

Dem Review folgte direkt die Sprint-Retrospektive. Generell wurde nach der Stimmung im Team, Verbesserungspotenzial und Dingen, die besonders gut gelaufen sind, gefragt. Wir listeten auf, was wir beibehalten oder loswerden sollten und was wir neu hinzunehmen beziehungsweise ausprobieren wollten.

Dieses Meeting sollte dazu da sein, die Arbeitsweise, die Zusammenarbeit und die Gesamtsituation im Projekt zu beleuchten. Die Stimmung sollte unbedingt offen sein, um ehrliche und damit spannende, wertvolle, reinigende sowie kreative Beiträge zu bekommen. Ein solches Meeting steht und fällt mit dem Scrum Master. Wir haben dabei sehr unterschiedliche Vorgehensweisen erlebt.

Meiner Erfahrung nach ist es eines der wichtigsten Meetings und alle Beteiligten sollten bemüht sein, das Meeting zu nutzen, um Probleme anzusprechen, Konflikte zu klären sowie Ideen und Vorschläge anzubringen. Wann sonst hat man die Gelegenheit, offen zu diskutieren und das Vorgehen auf den Prüfstand zu stellen?

Herausforderungen

Das Schätzen der User Storys war zu Anfang etwas holprig, da die Teilnehmer nicht nur Aufwand, sondern auch Komplexität und Risiko schätzten. Für die kleineren und übersichtlicheren Storys hatte sich das nach wenigen Sprints eingeschwungen. Schwierig wurde es hingegen bei fachlichen Anforderungen, die einen größeren technischen Umfang hatten.

Der initiale Aufwand für die ersten KPIs beziehungsweise User Storys war sehr viel höher als für die folgenden, da zu Beginn noch keine vorhandenen Strukturen und Prozesse wiederverwendet werden konnten. Später waren es dann geänderte Anforderungen, die sich durch alle Layer zogen und verzweigte technische Abhängigkeiten hatten, die größeren Aufwand nach sich zogen.

Hier half es, die Punkte, die nicht in einer gewissen Zeit zu klären oder halbwegs belastbar zu schätzen waren, in ein separates Meeting auszulagern. So hatten wir anfangs sehr viele Meetings. Rückwirkend betrachtet hätte es hier allerdings auch keine Abkürzung gegeben.

Eine Schätzung gerade größerer Storys ohne vorherige, detaillierte Auseinandersetzung mit den möglichen technischen Lösungsmöglichkeiten war oft sehr schwierig. Auch die grundlegenden Architekturentscheidungen zu Beginn des Projekts erforderten diverse Abstimmungsmeetings, da im Team unterschiedliche Ideen zur Realisierung bestanden. Die Selbstorganisation und Eigenverantwortung von Scrum-Teams machten es jedoch notwendig, sich auf ein gemeinsames Verständnis zu einigen.

Es hat sich für uns bewährt, zu Anfang eines neuen Sprints ein längeres Design-Meeting abzuhalten, in dem architekturelle Entscheidungen getroffen und bis zu einer gewissen Detailtiefe die Vorgehensweise festgelegt wurden. Gerade im DWH-Umfeld sind zu Beginn oft größere Design-Entscheidungen zu treffen, die der generellen Zielarchitektur zugrunde liegen. Es hat sich entgegen der anfänglichen Erwartung auch als effizienter herausgestellt, mit dem ganzen Team zu eruieren, ob größere Änderungen mit Abhängigkeiten und Betroffenheit aller Layer in einem Sprint unterzubringen sind, als hinterher alle Betroffenen einzeln abzuholen und eventuell neue Aspekte wiederum zu kommunizieren.

Die User Storys umfassten die Business-Anforderungen der Fachseiten. Hier war ein Umdenken erforderlich, um die fachlichen Formulierungen aus den Storys in technische Tasks herunterzubrechen, die vom Team umgesetzt werden konnten. Gelegentlich war es notwendig, eine fachliche Story aufzuteilen, da der Umfang zu groß gewesen wäre, um die technische Umsetzung in wenigen Tagen zu gewährleisten.

Wir hatten auch rein technische Storys. Bei einer gravierenden Änderung in der Anforderung haben wir sogar einen ausschließlich technischen Sprint eingezogen, weil alle Strukturen wie Tabellen und View sowie die Stored Procedures für die Datenbewirtschaftung angepasst werden mussten.

In der Auswahl der Storys für einen Sprint haben wir eher konservativ geplant. Wenn es zeitlich möglich war, haben wir in Abstimmung mit dem Project Officer noch weitere Storys in den aktuellen Sprint aufgenommen.

Ganz wichtig ist es, quer durch das gesamte Team einvernehmlich zu klären, wann ein Punkt als erledigt gilt. Diese Definition of Done ist vom Team zu erarbeiten und abzusegnen. Wir haben es generell formuliert und für die User Storys gegebenenfalls ergänzt. Das schließt Test und Dokumentation ein, sodass die Software lauffähig ist und zurückgerollt werden kann. Gerade wenn es plötzlich keinen Projektleiter mehr gibt, ist es unerlässlich, eine abgestimmte Arbeitsweise zu haben – sowohl in Bezug auf das Programmieren als auch auf das Testen, das Einchecken und das Deployen sowie nicht zuletzt das Kommunizieren.

Freud und Leid – die Erfahrungen

In der Regel war der Kunde sehr zufrieden und durch die kurzen Sprintzyklen immer über den aktuellen Entwicklungsstand informiert. Dennoch gab es bei den Reviews, in denen das Sprintergebnis vorgestellt wurde, auch frustrierende Momente. So hatten wir einen technisch sehr anspruchsvollen Sprint, dessen Kernpunkte an der Oberfläche nicht zu visualisieren waren und damit vom Kunden nicht so wahrgenommen und gewürdigt wurden wie von uns erwartet. An anderer Stelle mussten unter größten Anstrengungen eingebaute Features im nächsten Sprint wieder ausgebaut werden, weil sich die Spezifikation geändert hatte. Einmal wurde ein Sprint nicht vorgeführt, obwohl er bis auf wenige, kleine Details fertig war. Ein Teammitglied bestand darauf, dass das Sprintziel nicht erreicht war, und konnte auch den Project Officer überzeugen. Letztlich formal völlig richtig, aber für das (noch unerfahrene) Entwicklerteam sehr unbefriedigend. Bei näherer Betrachtung sind das allerdings alles Punkte, die – unabhängig von der Methode – mit guter Kommunikation innerhalb des Teams und durch Lernen aus den gemachten Erfahrungen gut in den Griff zu bekommen sind.

Eine gravierende und positive Neuerung war, dass die Verantwortung für die ausgelieferte Software beim Entwicklungsteam selbst, also bei uns lag. Aufgrund des überschaubaren Lieferumfangs pro Sprint und des eigenverantwortlichen Deployments entfielen sehr zeitintensive Elemente wie separate Tests, Build, Paketierung, betriebliche Abnahme und Wartungsfenster. Das versetzte uns in die Lage, in kleinen Zeitintervallen Software auszuliefern und sehr zeitnah auf Rückmeldungen des Kunden zu reagieren.

Peer Reviews oder Peer Programming haben wir zu Anfang wenig gemacht, was sich im Laufe des Projekts jedoch deutlich änderte. Nach und nach haben wir gemerkt, wie wichtig es ist, keine Wissensinseln im Team zu haben. Wir haben sehr gute Erfahrungen damit gemacht, sowohl bezüglich Know-how-Verteilung im Team als auch bei der Fehlerfindung beziehungsweise dem Hinterfragen von Vorgehensweisen. Da wir keine reinen Tester im Team hatten, gewann das Peer Review immer mehr an Bedeutung.

Wenn das Projekt den Einsatz unterschiedlicher Technologien erfordert, ist es unabdingbar, dass alle Beteiligten ein gesundes Grundverständnis davon haben, an welchen Themen die anderen konkret arbeiten und wie die einzelnen Funktionen und Technologien zusammenspielen. Kopfmonopole mit zentralem Know-how bei nur einer Person sind in jedem Fall zu vermeiden. Dennoch hat sich mit der Zeit herauskristallisiert, dass einzelne Teammitglieder aufgrund ihrer Kenntnisse und Erfahrungen in bestimmten Bereichen den Lead für Architekturthemen oder Datenbankanalysen an sich zogen.

Als wichtigstes Meeting haben wir die Retrospektive erlebt. Mit der Zeit war es deutlich mehr, als nur die Stimmung darüber abzugeben und zu sammeln, was gut gelaufen ist und was nicht. Es war die Plattform, offen und frei zu diskutieren, Dinge auf den Prüfstand zu stellen und Ideen zur Optimierung gleich im nächsten Sprint auszuprobieren. Wie ein Mini-Lessons-Learned, dessen Ergebnisse nicht in der Ablage landen, sondern gleich im nächsten Sprint umgesetzt und ausprobiert werden.

Da auch die Kunden der einzelnen Teilprojekte nicht vor Ort saßen, war es wichtig, dass sich das ganze Entwicklungsteam zu Beginn des Projekts mehrfach mit den Kunden traf, um sich persönlich kennenzulernen und die Anforderungen zu besprechen. Entgegen ursprünglichen Befürchtungen war es irrelevant, ob das Team räumlich eng beieinandersaß. Vielmehr war es entscheidend, dass sich mit der Zeit eine gute Kommunikationskultur und abgestimmte Arbeitsweise etablierten. Ohne das hätte es nicht funktioniert.

Angestrebt wurde ein möglichst hoher Grad an Automatisierung bei Build und Deployment, aber auch beim Testen (Unit Test, Integrationstest gegen Expected Values). Das haben wir bisher nicht in dem gewünschten Umfang realisieren können. Auch wenn Entwickler Testaufgaben übernehmen können und sollen, so wäre es hilfreich gewesen, wenigstens einen dedizierten Tester im Team zu haben.

Zu Beginn gab es Klärungsbedarf beim richtigen Umgang mit Bugs. Wichtig war es, die Kriterien zur Einschätzung abzustimmen, ob eine sofortige Reaktion erforderlich ist oder nicht. Für kritische, produktive Themen wurde ein Bug-Eintrag erstellt und das Problem sofort behoben. Je nach Komplexität wurde der Fix im Vier-Augenprinzip getestet und direkt auf der Produktion eingespielt.

Scrum löst nicht alle Probleme des Wasserfallmodells. Themen wie Anforderungen, Budget, Ressourcen und Zeit sind bei Scrum-Projekten genauso ein Thema wie bei allen anderen Vorgehensmodellen auch. Es wird nur sehr viel schneller transparent, wo das Projekt steht und ob etwas aus dem Ruder läuft.

Agiles Vorgehen ist kein Wunderheilmittel. Es ist auch keine wohlklingende Ausrede für Projektanarchie in Form fehlender Planung, mangelnder Steuerung, unklarer Absprachen, halbgarer Anforderungen und nicht abgestimmten Vorgehens. Im Gegenteil: Scrum fordert Engagement und Eigenmotivation, Selbstorganisation des Teams und klare Absprachen, gute Kommunikation sowie die Bereitschaft, sich auf Veränderungen einzulassen.

Agilität passt nicht zu jedem Projekt. Agile Methoden sollen helfen, schnellstmöglich auf sich verändernde Bedingungen zu reagieren. Je genauer im Vorfeld klar ist, wie es hinterher aussehen soll, und je länger die Releasezyklen sind, desto eher kann man in Erwägung ziehen, auch beim Wasserfallmodell zu bleiben. Aber sehr komplexe, volatile Probleme bekommt man mit analytischen Wasserfallmethoden nicht mehr in den Griff.

Umdenken! Möchte man agile Methoden einführen, so ist die Bereitschaft wesentlich, sich von bestehenden Prozessen, Hierarchien, Strukturen und Checklisten zu lösen und sich auf Neues einzulassen. Will man schnell auf Veränderungen reagieren können, muss man intensiv in Kommunikation investieren. Wenn man Lernen ermöglichen will, darf man Fehler nicht bestrafen.

Das gilt für die ganze Organisation. Selbstorganisierte Teams, die Verantwortung übernehmen, empathisch miteinander umgehen und sich engagieren, fallen nicht vom Himmel. Solche Teams müssen sich entwickeln dürfen. Da gibt es keine Abkürzung.

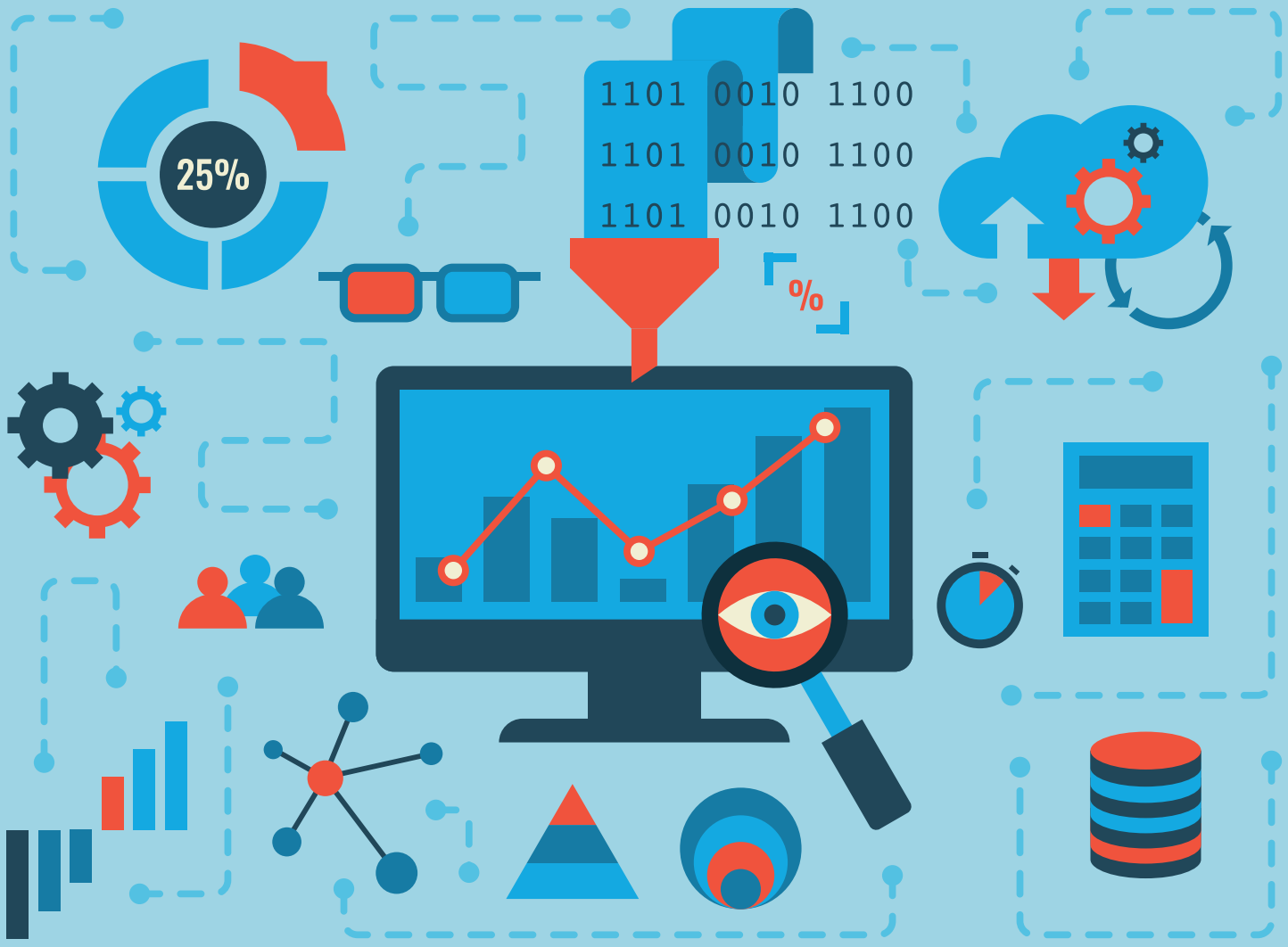
„Kein Plan überlebt die erste Feindberührung."
Helmuth von Moltke (1800 – 1891).

Fazit

Zu Beginn des Projekts standen zugegebenermaßen Zweifel im Raum, ob in so kurzen Sprints tatsächlich etwas „zum Anfassen" geliefert wird und man nach 14 Tagen schon erste KPIs liefern kann. Aber ja, es geht! Es braucht allerdings den Mut und die Offenheit, eingetretene Pfade zu verlassen und um die Ecke zu denken.

Dr. Susanne Bosinger
susanne.bosinger@dpdhl.com

Dr. Susanne Bosinger arbeitet als Softwareentwicklerin und Beraterin bei der Deutschen Post DHL IT Services. Sie ist seit fast 20 Jahren in der IT tätig und hat ihren Schwerpunkt in der datenbanknahen Entwicklung. Vor 8 Jahren hat sie die Oracle-Welt gegen die des SQL-Servers eingetauscht. Die Begeisterung für Datawarehouse-Themen ist geblieben. Vor ein paar Jahren hat sie erste Erfahrungen mit Scrum gemacht und engagiert sich seitdem für die Methode. Von Zeit zu Zeit schreibt sie Artikel oder tritt bei Konferenzen auf.



Wissen statt Bauchgefühl

Dirk Andres, Stadt Kaiserslautern

Den Veränderungen weltweiter Vernetzung und einer immer komplexer werdenden Gesellschaft müssen auch Kommunalverwaltungen Rechnung tragen. Die Datenmengen, die heutzutage in den Rathäusern zusammenlaufen, sind gewaltig und hochgradig divers. Daraus die wesentlichen Informationen herauszufiltern und diese für die Entscheidungsträger anschaulich aufzubereiten, ist eine Mammutaufgabe, der sich die Stadtverwaltung Kaiserslautern seit einigen Jahren erfolgreich stellt. Mithilfe der Oracle Business Intelligence Standard Edition One 11g hat die Stabsstelle Zentralcontrolling die inzwischen mehrfach ausgezeichnete Steuerungssoftware KLAR (KaisersLautern Analyse Recherche) entwickelt.

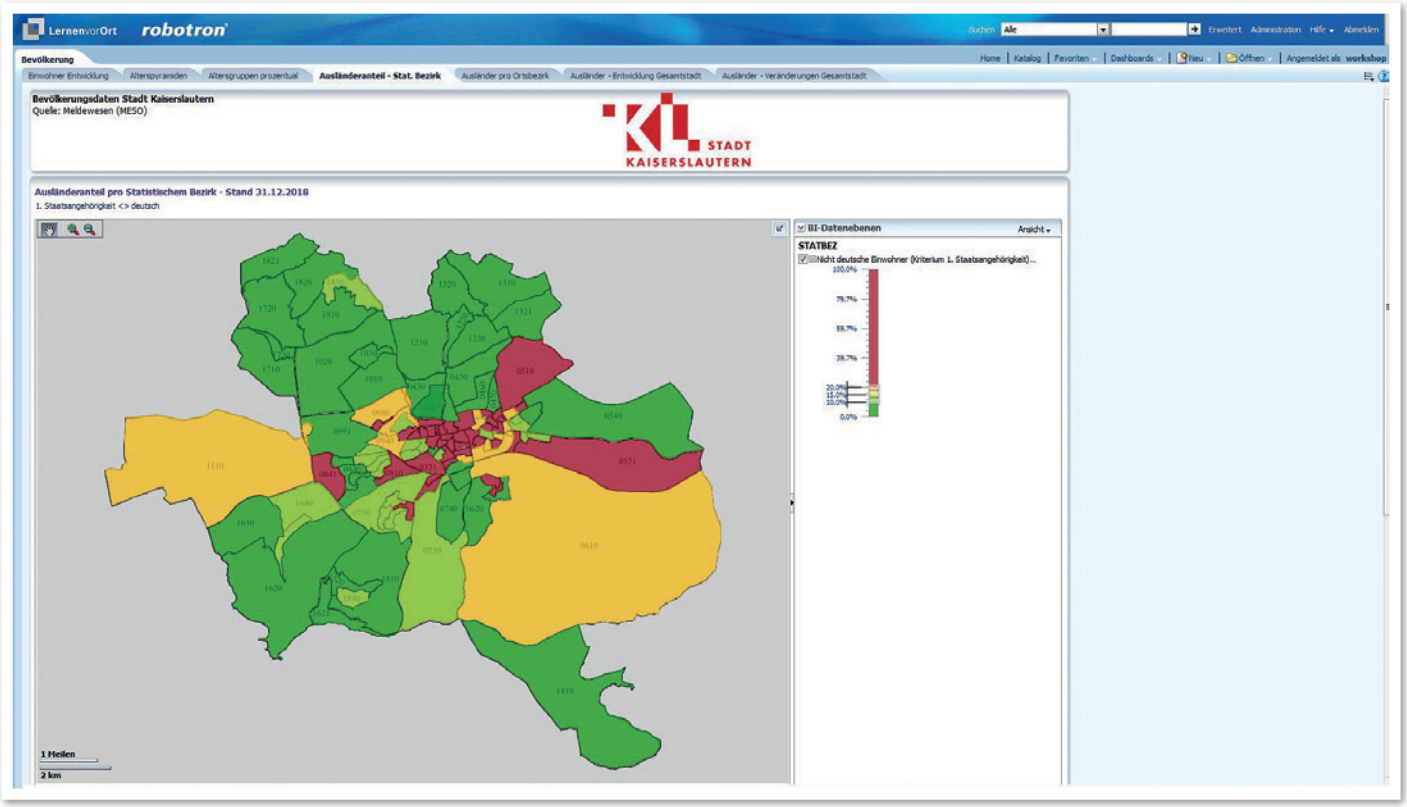


Abbildung 1: Der Ausländeranteil der statistischen Bezirke in der Kartendarstellung (Quelle: Stadt Kaiserslautern)

Egal ob Friedhöfe, Kitas, Schulen oder Steuern: Die Themenfelder, die von einer Verwaltung abgedeckt werden, sind enorm vielfältig. Wie in vielen anderen Kommunen gab es auch in der Stadt Kaiserslautern bis vor wenigen Jahren keine breite Informationsfläche, die Daten jederzeit abrufbar macht und Vergleiche ermöglicht. Daten wurden in der Vergangenheit meist nur problembezogen vom jeweiligen Fachreferat erhoben, ausgewertet und verarbeitet – in ganz unterschiedlichen Formaten und Qualitäten. Steuerung und Vorbereitung politischer Entscheidungen durch die Stadtspitze oder den Rat erforderten dadurch die individuelle Aufbereitung durch die jeweiligen Sachbearbeiter. Müssen unterschiedliche Datenformate aus unterschiedlichen Fachreferaten zusammengeführt werden, stößt dieses System schnell an seine Grenzen.

„Wenn Kaiserslautern wüsste, was es weiß"

Initiiert vom Oberbürgermeister und Stadtvorstand, arbeitet die Stabsstelle Zentralcontrolling seit dem Jahr 2014 daran, diesen Missstand zu beheben. Konzentrierte man sich anfangs noch auf Teilbereiche der Verwaltung – die Stabsstelle entwickelte etwa Visualisierungstools für den städtischen Haushalt oder die Zuweisung und Unterbringung von Asylsuchenden –, entschied man sich vor nunmehr drei Jahren für die große Lösung. Ausgehend von der grundsätzlichen Fragestellung „Wenn Kaiserslautern wüsste, was es weiß…" wurde das Ziel fokussiert: Verwaltung, Entscheidungsgremien, Politik und die Bürger in die Lage zu versetzen, alle für sie relevanten Informationen aus einem einheitlichen, maximal einfach aufgebauten System zu gewinnen.

Mit „Wissen statt Bauchgefühl" gilt es nun, die unterschiedlichen Prozesse und Informationen der vielfältigen Verwaltungsaufgaben qualitativ, effizient und niederschwellig zugänglich zu machen und sinnvoll zu verknüpfen.

Auf die ideal geeignete Software, um diese Idee in die Tat umzusetzen, stieß man letzten Endes per Zufall. Bei einem gemeinsamen Bildungsprojekt mit dem städtischen Referat für Jugend und Sport und der städtischen Bildungskoordination hatte die Transferagentur Rheinland-Pfalz/Saarland ihre hauseigene Software „Kommunal-Kombi" im Gepäck. Die war – wie sich schnell herausstellen sollte – perfekt geeignet, um nicht nur Bildungsdaten, sondern Daten beliebiger Art zusammenzuführen und zu analysieren. Als die Mitarbeiter der Stadt Kaiserslautern bei einer Tagung der Transferagentur auf das Tool aufmerksam wurden, erkannten sie schnell dessen Potenzial für eine ganzheitliche Steuerung.

Für die Analyse und übersichtliche Darstellung von Daten auf Dashboards greift die Anwendung auf die Standard-Software Oracle Business Intelligence Standard Edition One 11g zurück. Die „Kommunal-Kombi" wurde ursprünglich im Rahmen des Förderprogramms „Lernen vor Ort" (2009 bis 2014) des Bundesministeriums für Bildung und Forschung (BMBF) für das kommunale Bildungsmonitoring entwickelt. Das IT-Instrumentarium wurde von der Firma Robotron im Auftrag des BMBF programmiert und wird für das Einpflegen, Halten und Aktualisieren von Daten genutzt. Als integrierte Software-Lösung ermöglicht das aus zwei Anwendungsteilen bestehende IT-Instrumentarium, Bildungsdaten in einer zentralen Datenbank systematisch zusammenzuführen und die Auswertung zentral zu organisieren. Die Mitarbeiter können über ihren Webbrowser jederzeit auf einen systematisierten Datenkatalog zugreifen und bei Bedarf selbstständig individuelle Analysen erstellen. Dabei können die Daten nicht nur in Form verschiedener Diagramme, sondern auch in Karten dargestellt werden *(siehe Abbildung 1)*.

Bundesweit wurde das IT-Instrumentarium bisher vor allem genutzt, um im

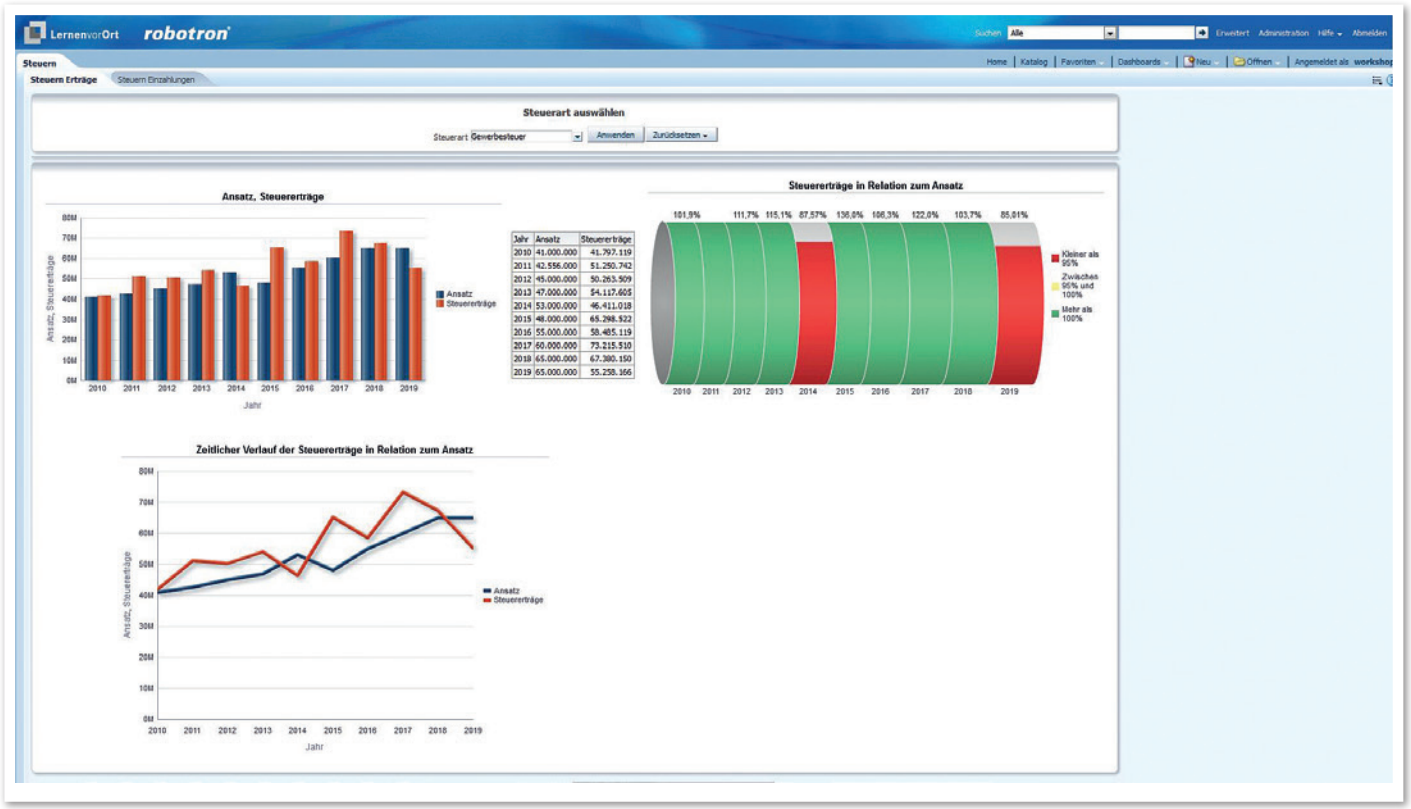


Abbildung 2: Darstellung der Gewerbesteuerzahlen in KLAR (Quelle: Stadt Kaiserslautern)

Rahmen eines kommunalen Bildungsmonitorings regionale Entwicklungen in der Bildung zu beobachten und so auf Basis belastbarer Daten das Bildungswesen vor Ort zu steuern. Die Anwendungsmöglichkeiten für das IT-Instrumentarium gehen jedoch weit über das Monitoring im Bildungsbereich hinaus. Unterstützt durch die Transferagentur Rheinland-Pfalz/Saarland und später auch durch Oracle, machte sich die Stabsstelle Zentralcontrolling daran, die „Kommunal-Kombi" für die Zwecke der Stadtverwaltung Kaiserslautern anzupassen. KLAR war geboren.

Unter dem Namen KLAR werden in der Stadtverwaltung Kaiserslautern nun seit 2017 vorhandene Informationsstrukturen, aber auch Strukturdaten aus den Bereichen Bevölkerung, Bildung und Soziales unter einem Dach zusammengeführt. In einem ersten Schritt soll es den verwaltungsinter-

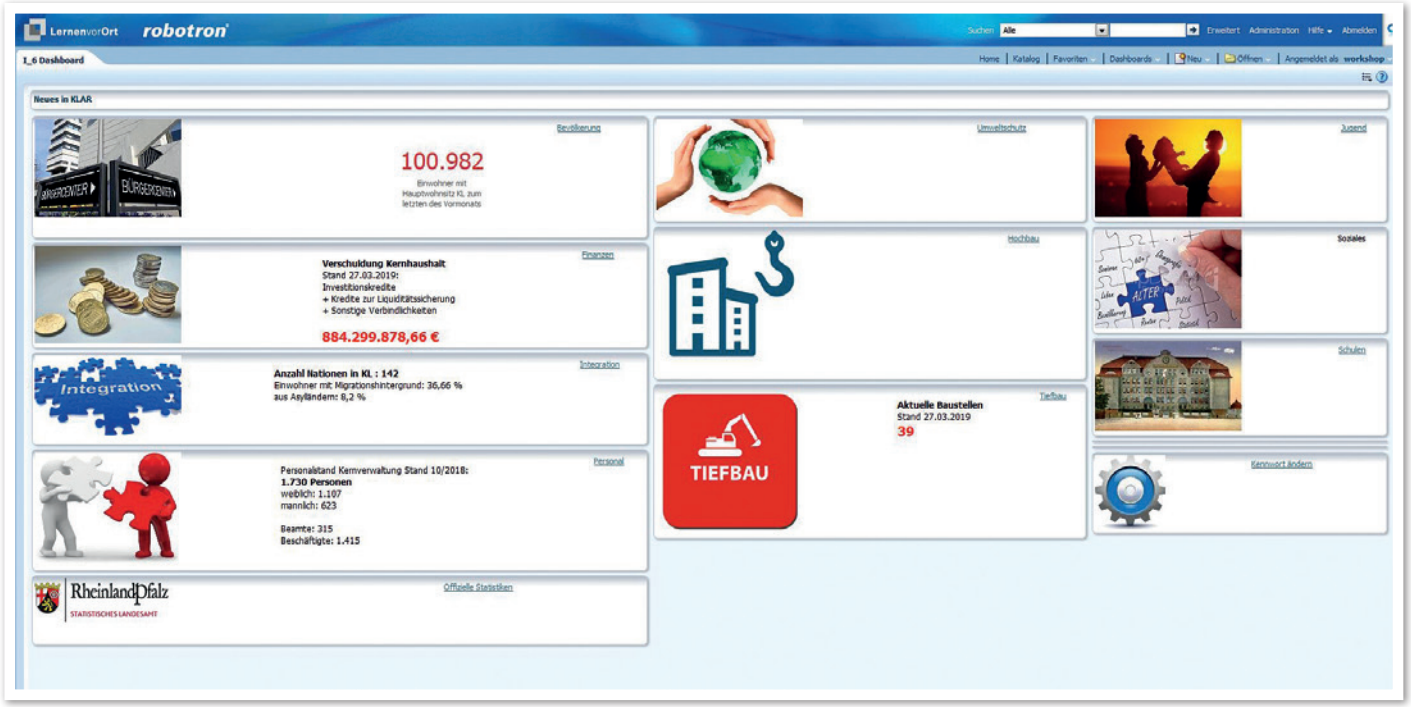


Abbildung 3: Darstellung der Themenfelder im KLAR-Dashboard (Quelle: Stadt Kaiserslautern)

nen Entscheidungsträgern ermöglicht werden, kontinuierlich und kurzfristig qualitativ hochwertige Informationen abzurufen und in ihre Entscheidungen einfließen zu lassen.

Die Benutzung soll dabei neben der Datensicherheit auch einem simplen Bedienungskonzept gerecht werden, um Hemmschwellen klein zu halten. Durch eine benutzerorientierte Berechtigungsstruktur sind die Informationen bedarfsgerecht abrufbar. Hilfreich ist hierbei, dass Kommunal-Kombi webbasiert ist. Da es über den Browser bedient wird, ist der mobile Einsatz von KLAR auch über Smartphone und Tablet problemlos möglich.

Nach und nach sollen nun Informationen aus allen Bereichen der Stadtverwaltung in KLAR einfließen. Das Referat für Jugend und Sport ist eine der ersten Einheiten, die Informationen in das System einpflegen. Damit ist es beispielsweise möglich, in einer Sitzung des Jugendhilfeausschusses kurzfristig Zahlen zu den Hilfen zur Erziehung mit stadtteilgenauen sonstigen Strukturdaten abzugleichen. Ebenso kann durch KLAR die Transparenz im Finanzbereich erhöht werden. So können bereits zu diesem Zeitpunkt Finanz- und Ergebnishaushalt in gleicher Zeitreihe gegenübergestellt und rückblickend mit dem Vorjahreszeitraum verglichen werden *(siehe Abbildung 2)*.

Ein gesunder Personalmix

Der Aufbau von KLAR ist eine komplexe Aufgabenstellung mit einem hohen internen Abstimmungsbedarf. Um diesen zu bewältigen, setzt man in der Stabsstelle Zentralcontrolling bewusst auf eine gesunde Personaldurchmischung. Die derzeit drei Mitarbeiter kommen aus drei unterschiedlichen Bereichen und bringen neben dem klassischen Verwaltungshintergrund auch die notwendigen IT- und BWL-Kenntnisse in das System ein. Darüber hinaus sind Mitarbeiter aus dem ganzen Haus beteiligt. Das sind zum einen der städtische Datenschutzbeauftragte, die Informations- und Kommunikationsabteilung, das Referat Organisationsmanagement sowie der Personalrat. Zum anderen natürlich die Sachbearbeiter der jeweiligen Fachabteilungen, die die entsprechenden Daten zur Verfügung stellen.

Unter ihrer Federführung werden alle relevanten Daten und Informationen zunächst für die jeweilige Abteilungsleitung zusammengestellt. Während diese sich somit ein Bild über die eigene Abteilung verschaffen kann, ist der Sachbearbeiter gleichzeitig in der Lage, seinen operativen Verantwortungsbereich für sich transparent zu gestalten. Die Abteilungsleitungen wiederum ziehen aus den zur Verfügung gestellten Daten Entscheidungshilfen für die ihnen übergeordneten Referatsleitungen und ergänzen sie gegebenenfalls um weiterführende Informationen. Dieser Prozess setzt sich von unten nach oben fort bis hin zur Verwaltungsspitze.

Als große Schwierigkeit hat sich anfangs die Qualität der verfügbaren Daten erwiesen. Bei deren Plausibilisierung stieß man schnell auf fehlerhafte Inhalte und Datenlücken. Eine Nachbearbeitung durch die Fachreferate war erforderlich. Inzwischen hat sich die Datenqualität in den jeweiligen Referaten deutlich verbessert – auch das ist ein Erfolg der KLAR-Einführung und der Arbeit des Zentralcontrollings. Alle im System verfügbaren Daten können nun als absolut zuverlässig und hinreichend verifiziert eingestuft werden.

Nachdem die Verwaltungsspitze, mehrere Referate und Abteilungen den Zugang zu KLAR hatten, wurde dies auch der Politik ermöglicht. In einem nächsten Schritt werden nun die Informationstiefe und die Zugangsmöglichkeit für den Bürger in den Fokus rücken.

Kooperation mit externen Instituten

Eines der Hauptziele von KLAR ist eine ansprechende und leicht zugängliche optische Aufbereitung komplexer Datenbestände. Um dieses Ziel zu erreichen, arbeitete die Stabsstelle unter anderem Ende 2017 mit Till Nagel, Professor für Informationsvisualisierung an der Hochschule Mannheim, und dem Forschungsbereich „Smarte Daten & Wissensdienste“ am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Kaiserslautern zusammen. In einem Semesterprojekt entwickelten Studierende anhand „echter“ Daten aus Kaiserlautern bereits Ideen für eine mögliche Visualisierungssoftware. Themenschwerpunkt war dabei das Zuzugsverhalten der Einwohner im Stadtgebiet.

Weitere Kooperationen mit ansässigen Institutionen wie Universität, Fraunhofer Institut und Wirtschaftsförderung sind bereits angelaufen oder stehen in den Startlöchern.

Aktuell werden in KLAR in der jeweiligen Erstansicht („Dashboard“) die Themenfelder abgebildet, mit Kerninformationen versehen und anhand von Kennzahlen dargestellt *(siehe Abbildung 3)*. Auf die üblichen Visualisierungspraktiken wie Sunburst, Drehzahlmesser und Live-Ticker kann selbstverständlich zurückgegriffen werden. Weniger ist dabei jedoch oftmals mehr. Diese Erstinformationen reichen, ähnlich wie in einem Kraftfahrzeug-Cockpit, in den meisten Fällen schon aus, um Rückschlüsse zu ziehen oder Sachstände zu erkennen.

Tiefergehende Informationen sind dann durch Drill-down zu erhalten, um Analysen, Recherchen und Prognosen durchzuführen. Die Tiefe der Informationsebene kann dabei ebenfalls vom Nutzer individuell ausgewählt werden.

Mehrfach ausgezeichnet

Das Projekt KLAR hat 2018 bundesweit Anerkennung gefunden. So erlangte man beim eGovernment-Wettbewerb der Firmen BearingPoint und Cisco Platz 1 in der Kategorie „Bestes Modernisierungsprojekt 2018“ sowie den ersten Platz beim Publikumspreis. Diesem zugrunde lag ein für alle Internetnutzer offenes Onlinevoting. Auf KLAR entfielen mehr als 1.400 der über 4.000 abgegebenen Stimmen, insgesamt also mehr als ein Drittel. An dem Wettbewerb nahmen 59 Institutionen aus Deutschland, Österreich und der Schweiz teil, 14 davon schafften es ins Rennen um den Publikumspreis. Schirmherr des Wettbewerbs war Prof. Helge Braun, Chef des Bundeskanzleramtes.

Bei den renommierten Best Practice Awards des Business Application Research Center (BARC) hat die städtische Controlling-Software Ende 2018 dann erneut für Aufsehen gesorgt und in der Kategorie Mittelstand gemeinsam mit der Hannover Rück den zweiten Platz erzielt. KLAR unterlag im Finale lediglich dem Internetportal Scout24.

Dirk Andres
dirk.andres@kaiserslautern.de

Dirk Andres ist Dipl. Verwaltungswirt und leitet seit 2014 die Stabsstelle Zentralcontrolling der Stadt Kaiserslautern. Zuvor war er 20 Jahre im Sozialreferat tätig, 4 Jahre in der wirtschaftlichen Hilfe und im Anschluss als Leiter Datenmanagement. Bis Ende 2018 war er zudem Geschäftsführer der Public Development Consulting – PDC. Er initiierte für die Stadtverwaltung Kaiserslautern das Informationstool KLAR, das im eGovernment-Wettbewerb als „Bestes Modernisierungsprojekt 2018“ ausgezeichnet wurde. Privat interessiert sich der 3-fache Familienvater für Handwerk, Sport und Jugendarbeit."

Geschäftsvorfälle flexibel und dynamisch steuern

Evgenia Rosa, Oracle

Geschäftsprozesse können sich je nach Einsatzgebiet und Anwendungsfall in ihrer Art wesentlich unterscheiden: von vorhersehbaren, strukturierten Workflows (zum Beispiel Dokumenten-basierte oder Genehmigungsprozesse) bis zu dynamischen, unstrukturierten Prozessen, deren Ablauf erst zur Ausführungszeit bestimmt werden kann. Für die IT-technische Unterstützung und Automatisierung dieser Prozesse sind unterschiedliche Ansätze und Werkzeuge notwendig. Oracle Integration Cloud (OIC) bietet Modellierungs- und Ausführungsumgebung sowohl für strukturierte als auch für unstrukturierte Prozesse. Außerdem bietet OIC die Möglichkeit, den Automatisierungsgrad der Prozesse durch Robotic Process Automation (RPA) zu erhöhen.

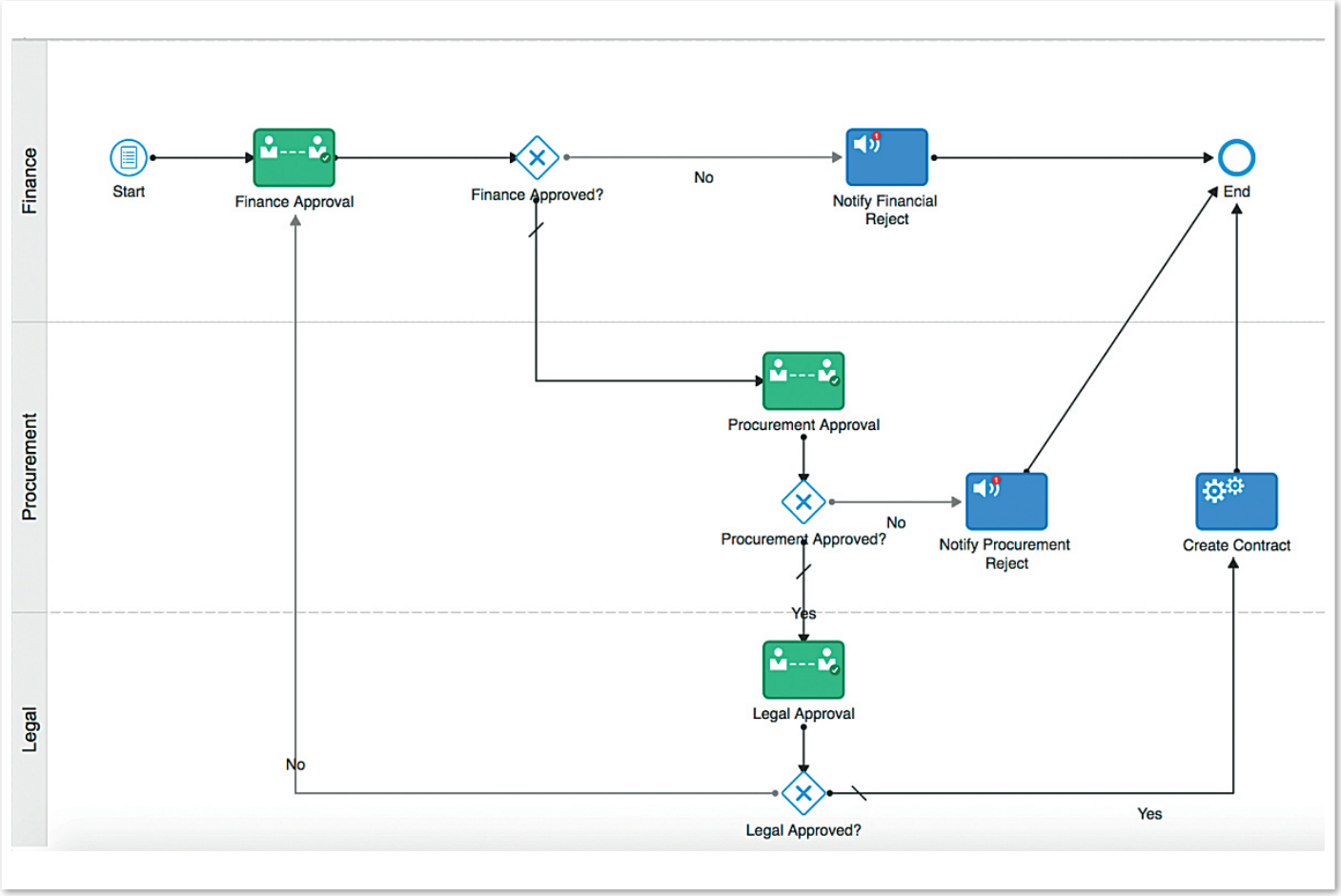


Abbildung 1: Strukturierter BPMN-Prozess (Quelle: Oracle)

Für die Modellierung und Automatisierung der strukturierten Prozesse hat sich der Industrie-Standard „Business Process Model and Notation" (BPMN) bewährt und durchgesetzt. Allerdings kann man mit BPMN nur vorhersehbare, also im Voraus bekannte, strukturierte Prozesse umsetzen. Ein Genehmigungsprozess wäre ein typisches Beispiel für einen strukturierten Prozess. Der Ablauf kann fest definiert und dann immer wieder ausgeführt werden. Ändert sich der Ablauf, wird das Prozessmodell angepasst und neu installiert. Oracle unterstützt BPMN seit mehr als zehn Jahren in seiner BPM-Plattform. So ist BPMN ein integraler Bestandteil der Prozess-Komponente der Oracle Integration Cloud. *Abbildung 1* zeigt einen strukturierten BPMN-Prozess.

Es kann vorkommen, dass der Prozess in bestimmten Fällen unter verschiedenen Bedingungen (Daten) anders ablaufen soll. Man kann zwar in diesem Fall mit Verzweigungen (BPMN-Gateways) arbeiten, aber das Prozessmodell wird schnell unübersichtlich, wenn zu viele Bedingungen abgeprüft werden müssen. Für mehr Flexibilität und Übersichtlichkeit in strukturierten Prozessen kann der Einsatz von Regeln sorgen. Die Bedingungen werden anstatt in Gateways zum Beispiel in Entscheidungstabellen untergebracht. Dafür gibt es einen weiteren Industrie-Standard, den „Decision Model and Notation" (DMN). Oracle unterstützt DMN mit dem Decision Service, der in der Prozesskomponente der Oracle Integration Cloud zur Verfügung steht.

Zur Modellierungs-Best-Practice gehört, dass die komplexe Entscheidungslogik nicht mit dem BPMN-Prozessfluss, sondern mit Regeln beziehungsweise Entscheidungstabellen abgebildet werden soll. *Abbildung 2* zeigt einen strukturierten Prozess mit Benutzung einer DMN Decision in BPMN.

Dynamische Prozesse

Was geschieht allerdings, wenn der Prozess nicht vorhersehbar ist? Wenn je nach Situation ganz anders vorgegangen werden muss, um ein Ziel zu erreichen? Dafür steht der Begriff „Case Management", der sich auf eine solche Problematik fokussiert. Im Case Management steht nicht der Prozess selbst, sondern die Erreichung eines Ziels im Vordergrund. So wird zum Beispiel ein Fall („Case") angelegt und bearbeitet, wenn ein Kunde eine Störung des Service seinem Service-Provider meldet. Das zu erreichende Ziel wäre die Behebung der Störung und dieses Ziel wird erreicht, indem der Service-Provider bestimmte Aktionen vornimmt, die für die gegebene Kundensituation am besten geeignet sind. Es sind also für die Zielerreichung Aktivitäten vorgesehen; deren Reihenfolge ist allerdings nicht festgelegt und die Notwendigkeit der Ausführung nicht immer gegeben. Menschen entscheiden, welche Aktionen unter welchen Bedingungen ausgeführt werden. Zu den typischen Beispielen für Fallbearbeitung zählen polizeiliche Ermittlungen, Schadensabwicklung bei einer Versicherung, medizinische Rehabilitation, Kontoeröffnung bei der Bank und Vertragsabschluss bei einem TK-Provider.

Im Case Management drehen sich die Aktivitäten also um einen Fall und nicht um einen Prozess. Der dynamische Prozess ist nicht vordefiniert, er entwickelt sich „on the fly". Über den nächsten Schritt entscheidet der Mensch – der sogenannte „Knowledge Worker" (Wissensarbeiter). Er zeichnet sich durch seine Expertise im jeweiligen Fachge-

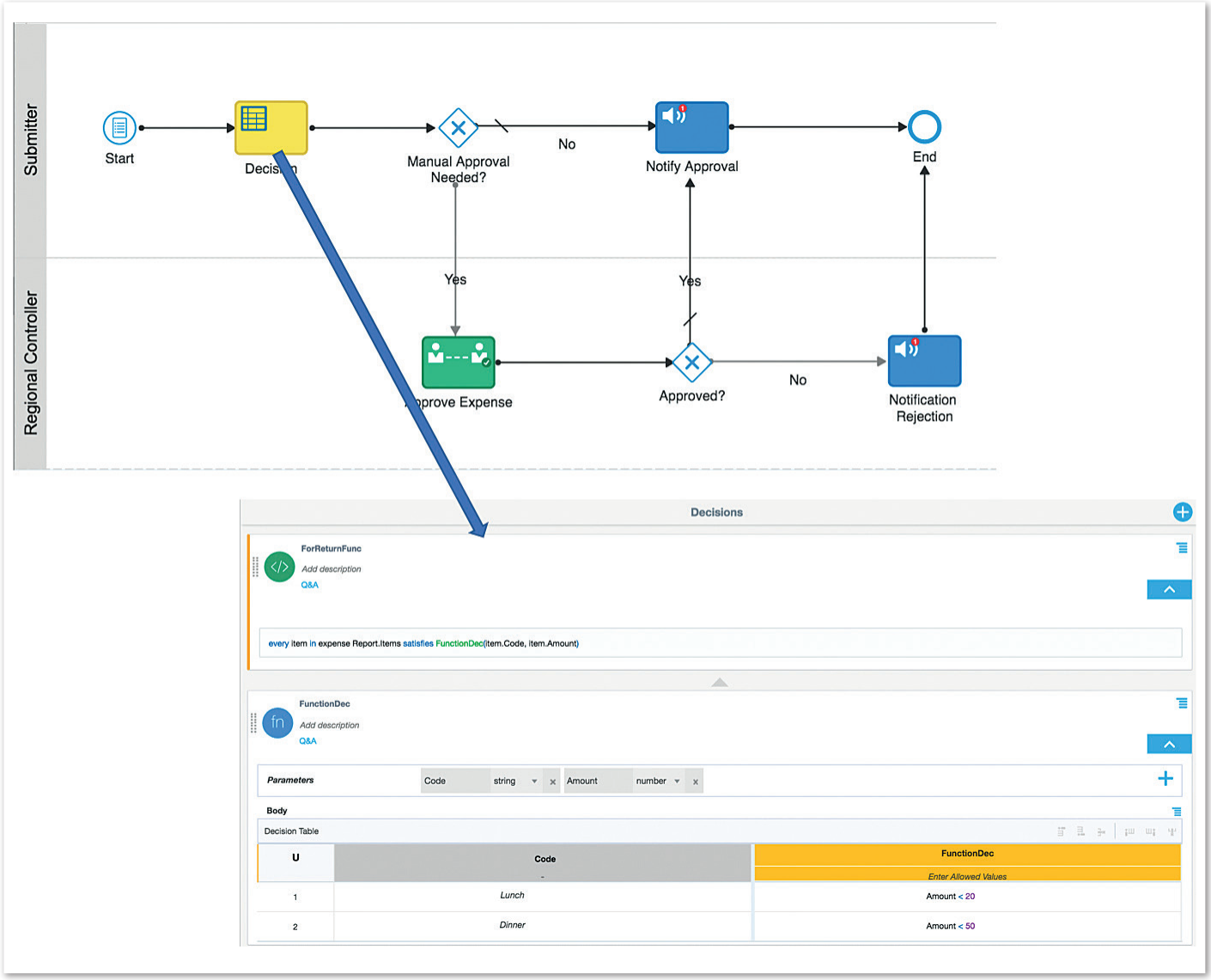


Abbildung 2: Strukturierter BPMN-Prozess mit DMN-Entscheidungsservice (Quelle: Oracle)

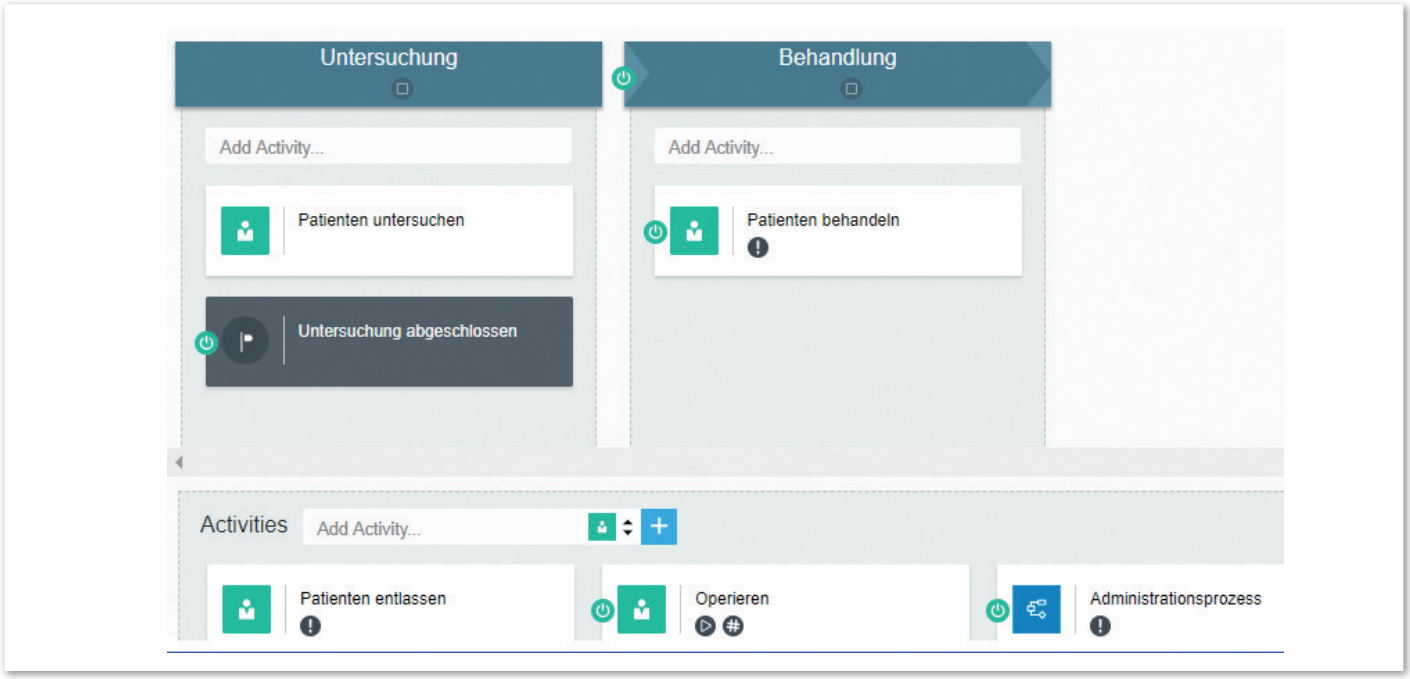


Abbildung 3: Modellierung der Fallbearbeitung (Quelle: Oracle)

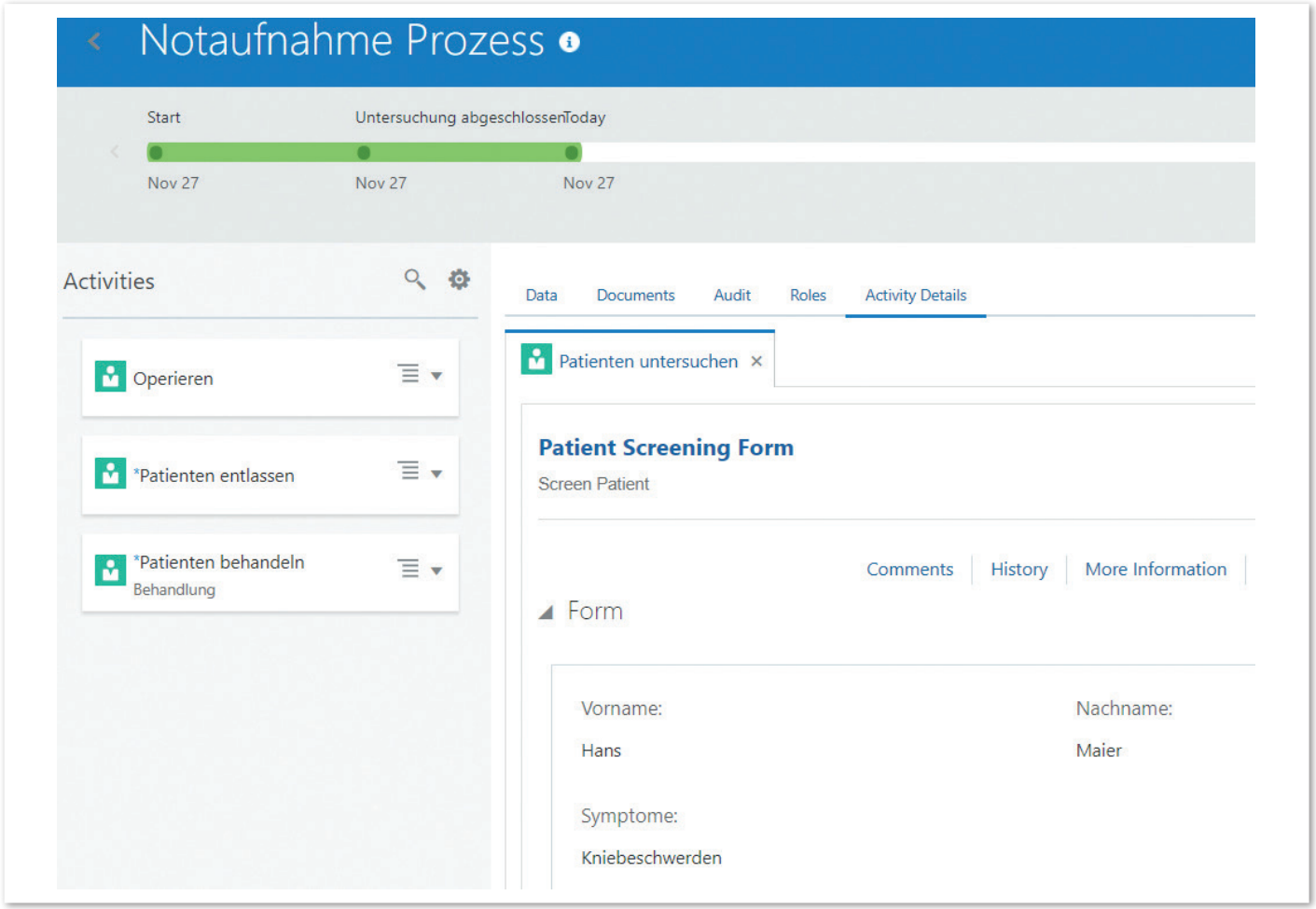


Abbildung 4: Fallbearbeitung zur Laufzeit (Quelle: Oracle)

biet aus, die ihn befähigt, individuelle Entscheidungen für jeden Case zu treffen.

Ein Case setzt sich in der Regel aus einer ungeordneten Menge von Aktivitäten und Dokumenten zusammen. Aktivitäten können obligatorisch oder optional sein. Zum Grundgedanken des Case Management gehört, dass es einen Kern von Aktivitäten gibt, die sich immer ähneln. Dieser Kern wird durch die obligatorischen Aktivitäten abgedeckt. Die individuelle Ausprägung entsteht durch die optionalen Aktivitäten des Case.

Modellierung und Umsetzung von dynamischen Prozessen

Für die Modellierung und Ausführung dynamischer Prozessabläufe sind ein anderes Vorgehen und andere Werkzeuge als für strukturierte Prozesse notwendig. In der Prozessumgebung der Oracle Integration Cloud steht neben BPMN auch eine Case-Management-Modellierungs- und -Ausführungs-Umgebung zur Verfügung.

Man beginnt mit der Definition der einzelnen Aktivitäten und der Phasen („Stages"), die der Fall bis zur abschließenden Bearbeitung durchläuft. Innerhalb der einzelnen Phasen werden die Arbeitsschritte beziehungsweise Aktivitäten als Tasks, Service-Aufrufe oder strukturierte Prozess-Bausteine zusammengefasst, die notwendig sind, um einen Case in die nächste Phase seines Lebenszyklus zu bringen.

Phasen können sequenziell oder parallel ablaufen. Einzelne Aktivitäten und Phasen lassen sich unter bestimmten Bedingungen aktivieren, deaktivieren oder terminieren. Die Bedingungen können Daten- oder Ereignis-basiert sein. Bei der Definition der Bedingungen kann der Decision Service eingesetzt werden. Zum Nachvollziehen des Fallverlaufs werden Meilensteine definiert, die die Erreichung eines bestimmten Zustands manifestieren. Die Abbildungen 3 und 4 zeigen jeweils die Modellierungsumgebung beziehungsweise die Runtime für die Fallbearbeitung.

Adaptive Prozesse

Während dynamische Prozesse (DCM) dem Benutzer viel Freiheit beim Erreichen des Zieles (Lösung des Case) lassen, heißt dies noch lange nicht, dass der Vorgang optimal bearbeitet wird. Wenn der Wissensarbeiter die Erfahrungen aus bereits abgeschlossenen Fällen nutzen kann und weitere Aktivitäten für die optimale Fallbearbeitung empfohlen bekommt, spricht man von „Adaptive Case Management" (ACM). Der Begriff „adaptive" bedeutet so viel wie „anpassungsfähig", also die Lern- beziehungsweise Anpassungsfähigkeit des Systems.

Für die technische Umsetzung eines solchen Adaptive-Case-Management-Konzepts könnten intelligente (AI/ML-)Lernalgorithmen verwendet werden, die ein System dazu befähigen, Informationen aus vergangenen Cases automatisiert zur Verfügung zu stellen oder zu nutzen. Demnach können beispielsweise Algorithmen zur Anwendung kommen, die anhand von Informationen vergangener Prozessinstanzen Vorschläge für künftige Entscheidungen generieren oder sogar eigenständige Entscheidungen treffen.

Robotic Process Automation

Der Begriff „Robotic Process Automation" oder Roboter-gesteuerte Prozess-Automa-

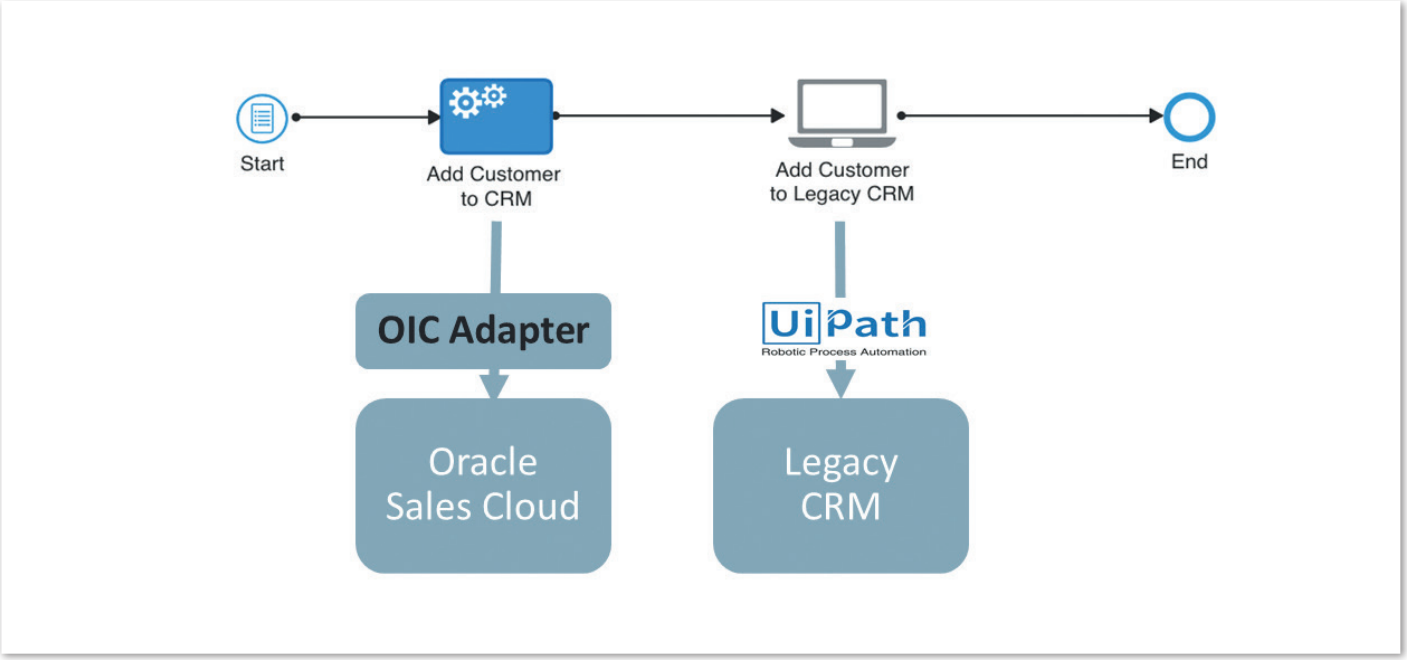


Abbildung 5: UiPath-RPA-Technologie zur Einbindung eines Legacy-CRM in einen Prozess (Quelle: Oracle)

tisierung (RPA) beschreibt eine innovative Technologie, die Automatisierung von strukturierten Geschäftsprozessen durch Software-Roboter ermöglicht. Dabei handelt es sich um die automatisierte Bearbeitung von sich wiederholenden Aufgaben, die von Menschen ausgeführt werden. Diese Roboter sind Software-Anwendungen, die eine menschliche Interaktion mit Benutzerschnittstellen von Software-Systemen nachahmen. Der Software-Roboter kann sich in eine Anwendung einloggen, Daten eingeben, Berechnungen durchführen, Dateien hoch- und herunterladen oder sich abmelden.

Die Vorteile von RPA sind vor allem Kostenersparnis durch Effizienz und Fehlervermeidung, lückenlose Dokumentation der Prozesse und steigende Mitarbeiter-Zufriedenheit (Befreiung von Routinearbeiten). Außerdem lässt sich diese Technologie schnell und effizient implementieren, ohne bestehende Infrastruktur und Systeme zu verändern.

Um seinen Kunden die Möglichkeit zu bieten, die RPA-Technologie mit der Oracle Integration Cloud zu nutzen, ist Oracle eine Partnerschaft mit dem Hersteller UiPath, einem führenden RPA-Technologie-Anbieter, eingegangen. Zusätzlich zu der beschriebenen Effizienzsteigerung ermöglicht die UiPath-RPA-Plattform den Zugang zu Systemen, die kein API haben (zum Beispiel Mainframes), oder wenn es für ein bestimmtes System noch keinen Adapter gibt. Aber auch für den Fall, dass es Integrationsmöglichkeiten gäbe, ein komplexes Integrationsprojekt allerdings nicht wirtschaftlich wäre, weil etwa die System-Landschaft in der nahen Zukunft infolge einer Akquisition konsolidiert werden soll, bietet der Einsatz von RPA eine kostengünstige Alternative.

Um den Einsatz von UiPath RPA zu vereinfachen, bietet Oracle in der Integration Cloud einen UiPath-Adapter an. So lässt sich zum Beispiel mit dessen Hilfe ein Legacy-System in einen BPMN-Prozess einbinden *(siehe Abbildung 5)*, daneben kann ein UiPath-Roboter einen BPMN-Prozess aufrufen, um etwa eine automatische Ausnahmebehandlung zu starten.

Fazit

Oracle Integration Cloud enthält neben der Integrationskomponente auch die Möglichkeit, sowohl strukturierte Prozesse mit BPMN und DMN als auch dynamische/unstrukturierte Prozesse (Dynamic Case Management) zu modellieren und auszuführen. Die Integration mit UiPath RPA bietet zusätzlich den effizienten und einfachen Einsatz von Robotic Process Automation.

Weitere Informationen zu UiPath und der Partnerschaft mit Oracle unter *„https://www.uipath.com/partners/technology-partners/oracle“* und *„http://www.oracle.com/us/products/middleware/bpm/process-automation-rda-ds-4101031.pdf“*.

Evgenia Rosa
evgenia.rosa@oracle.com

Evgenia Rosa arbeitet als Principal Solution Engineer bei Oracle Deutschland B.V. & Co. KG. In dieser Rolle unterstützt sie Oracle-Kunden und -Partner zu Themen rund um Cloud (PaaS), Integration, Prozessmanagement und Blockchain. Frau Rosa hat Informatik an der TU Berlin studiert und vor ihrer Tätigkeit bei Oracle am Forschungsinstitut für offene Kommunikationssysteme (Fraunhofer Fokus) an der Konzeption und Entwicklung von verteilten IT- und Telekommunikationsanwendungen gearbeitet.

Das E-3 Magazin

Information und Bildungsarbeit von und für die SAP-Community

Überfordert?

Wir bieten Information und Bildungsarbeit von und für die SAP-Community

e-3.de | e3zine.com

DOAG 2019 Logistik + IT

17. September 2019 in Frankfurt

DOAG

logistik.doag.org

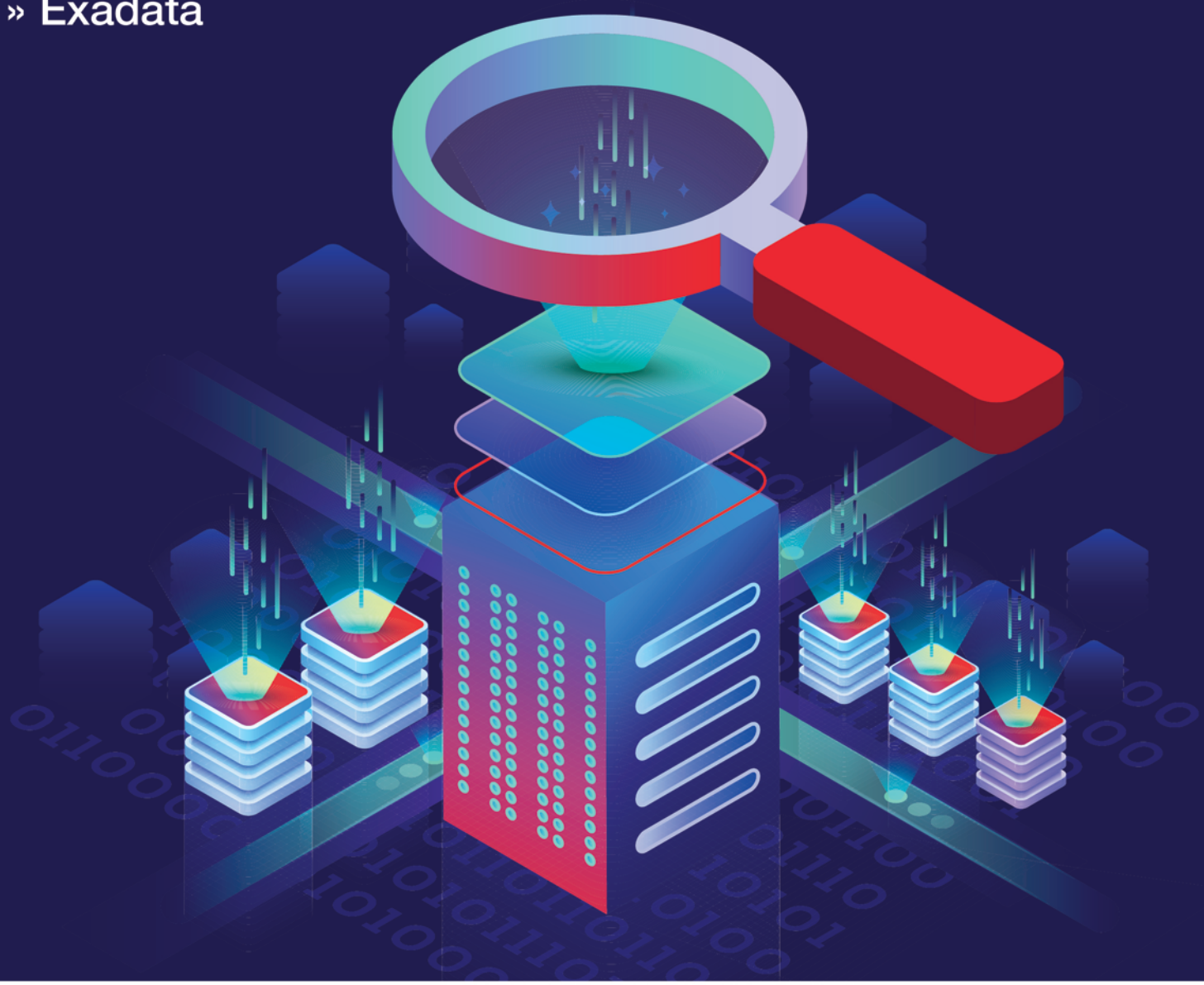
ORACLE Platinum Partner

LOGICALIS
Business and technology working as one

Machen Sie jetzt mehr aus Ihren Daten!

Mit Oracle Data Analytics

- » Autonomous Database
- » Big Data
- » Exadata



de.oracle@logicalis.de

www.logicalis.de