ISSN 2366-7915 | WWW.DOAG.ORG

REDSTACK

MAGAZIN INKLUSIVE BUSINESS NEWS

NR. 3
MAI
2022

DOAG SOUG swiss oracle user group AOUG AUSTRIAN ORACLE USER GROUP

DATA ANALYTICS

CloudLand

www.cloudland.org

DAS CLOUD NATIVE FESTIVAL
4 TAGE – 4 THEMEN

DAS EVENT DER DEUTSCHSPRACHIGEN CLOUD NATIVE COMMUNITY

Weitere Informationen

29.JUNI BIS 2.JULI 2022

im Phantasialand in Brühl

#CloudLand2022

Eventpartner: Heise Medien

Armin Wildenberg
Vorstand Data Analytics Community, Leiter Data Analytics Community

Ausgabe Nr. 3/2022
auf Abruf!

Liebe Mitglieder, liebe Leserinnen und Leser,

die aktuelle Ausgabe des Red Stack Magazin hat Data Analytics als Schwerpunkt. Die Bereitstellung von Daten gewinnt weiterhin an Bedeutung und dies nicht nur in den Unternehmen, sondern auch vermehrt in den Behörden und Städten, wie Dirk Andres in dieser Ausgabe im Interview aufzeigt.

Data as a Service könnte der Begriff der Zukunft sein, der die schnelle und anwenderbezogene Bereitstellung des Rohstoffes Daten im aktuellen Jahrzehnt widerspiegelt. Dabei sind sowohl die Bereitstellung und Struktur der Daten als auch die automatisierte Auswertung wichtige Faktoren.

Beispielhaft dafür beschreiben Alfred Schlaucher mit Metadatenverwaltung und Andreas Buckenhofer mit Datenarchitektur sowie Jan Ott zu Data Lake in ihren Artikeln Möglichkeiten und Alternativen dazu.

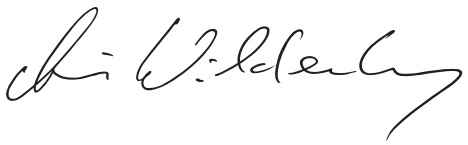
In ihren Artikeln zu Machine Learning und ML as a Service referenzieren Detlef Schröder sowie Oliver Fuhrmann und Germans Hirsch auf die aktuellen Entwicklungen bei selbstlernenden Anwendungen.

Für die kommenden Herausforderungen im Bereich Data Analytics bedarf es zusätzlich einer passenden und skalierbaren Basisarchitektur und diese vorzugsweise in der Cloud.

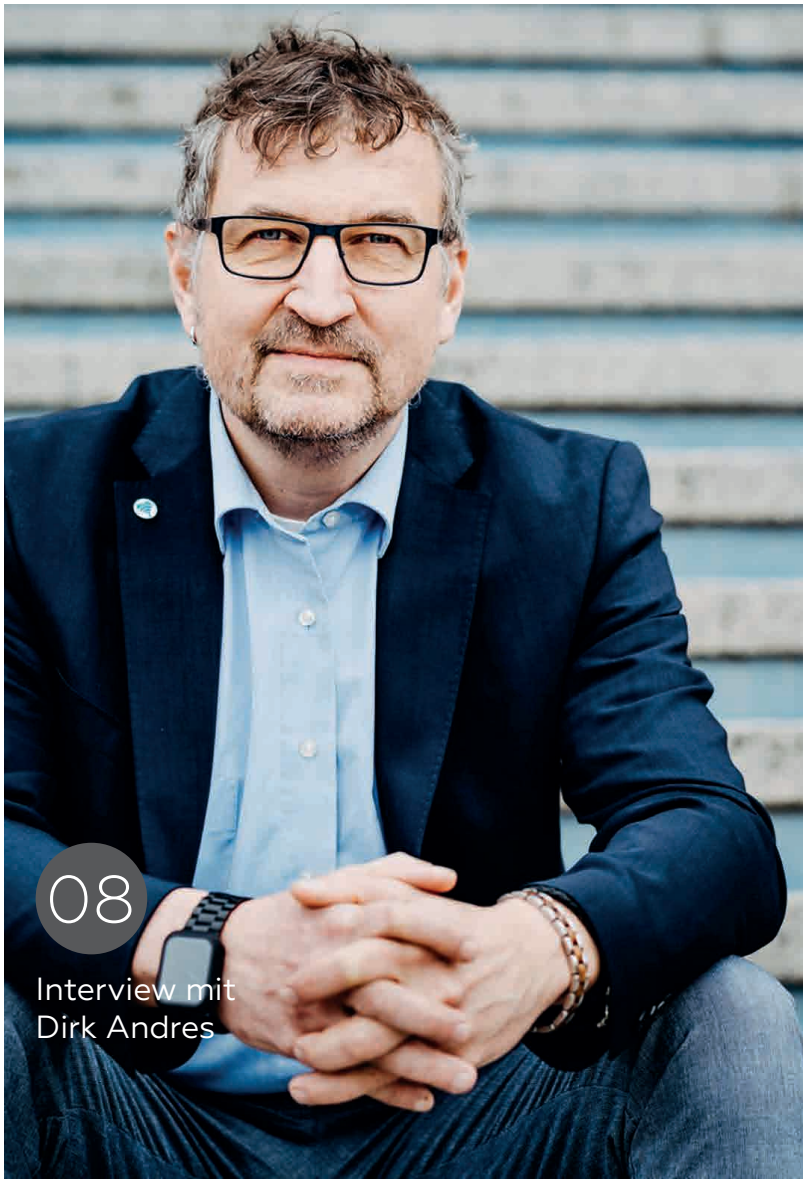
Eine fortwährend aktuelle und wachsende Bedeutung hat Ausbildung und Talentförderung in der IT. Diesen Themen mit seinen Facetten widmet sich mit vier Artikeln unsere Business News. Die Gewinnung von neuen Talenten und die Qualifikation der Mitarbeiter wird in den nächsten Jahren eine der großen Herausforderungen für die Unternehmen sein.

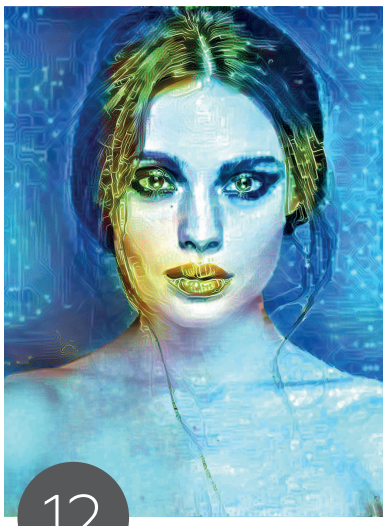
Wir freuen uns alle, in diesem Jahr wieder eine Konferenz und Ausstellung in Präsenz zu haben und auf den Thementag zu Beginn. An den beiden folgenden Konferenztagen sind unsere Autoren auch im direkten Gespräch und teilweise mit ihren Vorträgen vor Ort.
Wir sehen uns auf der DOAG 2022 Konferenz und Ausstellung im September.

Bleiben Sie gesund.



Armin Wildenberg





Chancen und Grenzen von AutoML

Metadaten helfen beim Aufbau einer Plattform für analytische Daten

Einleitung

Data Analytics & KI

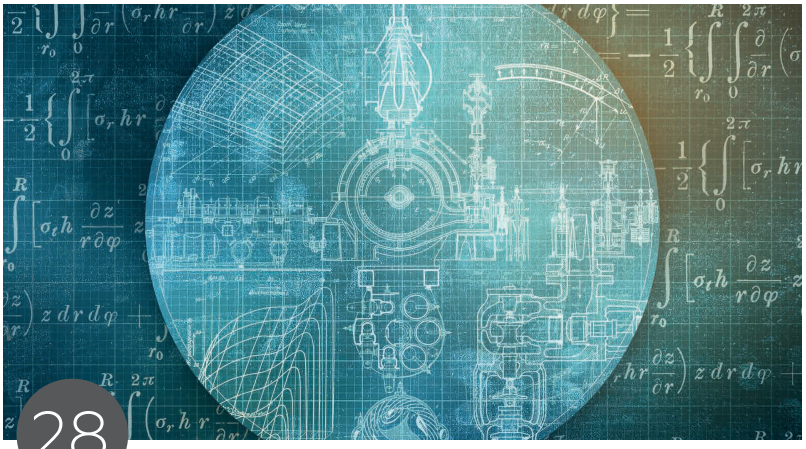
Development

Datenbank

APEX

BUSINESS NEWS

Ausbildungsberufe in der IT



28

Machine Learning as a Service zur Optimierung von Produktionsprozessen

74

Von der Motivation, nach der Schulbank einen Ausbildungsberuf in der IT zu ergreifen

66

IT-Berufe – Ausbildung und Studium im Praxisverbund

82

Von der Motivation, neben dem Beruf ein Masterstudium zu ergreifen

Intern

News

TIMELINE

7. APRIL 2022

„DB Patterns“ heißt das Thema, dem sich Moritz Klein und Niels de Bruijn im Dev Talk widmen.

8. APRIL 2022

In der WebSession mit Christian Pfundtner dreht sich alles rund um das Thema „Connect 2 worlds - Datenaustausch zwischen Oracle und PostgresSQL“.

14. APRIL 2022

Das Programm der DOAG 2022 Datenbank ist online. Im Van der Valk Airporthotel Düsseldorf findet vom 30. bis 31. Mai die DOAG 2022 Datenbank statt. Nach langem Warten können sich die Besucher in gut zwei Wochen auf ein Wiedersehen mit zahlreichen Vorträgen für Einsteiger und Experten aus den Bereichen Datenbank und Engineered Systems, einen geselligen Community-Abend sowie reichlich Gelegenheit für interessante Gespräche und Networking freuen – endlich wieder vor Ort in Düsseldorf. Fried Saacke ist erfreut über die erneut gute Resonanz bei den Anmeldungen.

21. APRIL 2022

Im DevTalk mit Sabine Heimsath und Carolin Hagemann geht es um das Thema „Frauen in der IT “.

28. APRIL 2022

Der Call for Papers für die DOAG 2022 Konferenz + Ausstellung endet. Die große Jahreskonferenz wird dieses Jahr um zwei Monate vorgezogen und findet nach zweijähriger Online-Realisierung wieder als Präsenzveranstaltung vom 20. bis 23. September im Nürnberger Convention Center statt. Save the date!

2. MAI 2022

Beim Regionaltreffen NRW heißt es am Vorabend der APEX connect 2022 im Phantasialand in Brühl „APEX Open Mic Night“, bei der jeder Teilnehmer die Gelegenheit hat, seine App oder ein Feature in fünf Minuten vorzustellen.

3. UND 4. MAI 2022

Die Besucher der APEX connect 2022 erleben zwei abwechslungsreiche Konferenztage rund um APEX und viele weitere Development-Themen! Spannende Vorträge, unterhaltsame Community-Aktivitäten, inspirierende Workshops sowie viele Gelegenheiten zu Austausch und Networking – erstmals wieder vor Ort und diesmal mit der unvergleichlichen Freizeitpark-Atmosphäre des Phantasialands in Brühl bei Köln.

5. MAI 2022

Carsten Czarski, Carolin Hagemann und Niels de Bruijn lassen in einem DevTalk die APEX connect 2022 noch einmal Revue passieren.

11. UND 12. MAI 2022

Im Berliner Expertenseminar „Oracle Database Performance: Hands-On“ lernen die Teilnehmer anhand verschiedener praktischer Beispiele unterschiedliche Techniken kennen, mit denen die Performance in der Datenbank im Detail analysiert werden kann.

CLOUDLAND 2022: CLOUD CUSTOMER STORIES, GEEK-NIGHT, ZEN UND SUMMER-PARTY

Marcos López

Das von der Deutschsprachigen Cloud Native Community organisierte Festival verspricht „die starren Grenzen klassischer Konferenzen zu sprengen".

Ein Festival als neues Konferenzerlebnis

Wie Event-Partner Heise Medien, der an der programmatische Gestaltung der CloudLand beteiligt und bei der Durchführung vor Ort aktiv eingebunden ist, auf heise online berichtete, werden auf dem viertägigen Festival neben technischen Vorträgen zahlreiche unkonventionelle Sessions und Veranstaltungen für ein gänzlich neues Konferenzerlebnis sorgen.

Unkonventionelle Formate

Workshops, Mini-Konferenzen, Cloud Customer Stories und technische Vorträge werden in der abenteuerlichen Freizeit-Kulisse des Phantasialand in Brühl begleitet von Geek-Night-Talks, Cloud-Yoga, Zen und Summer-Party. Die Deutschsprachige Cloud Native Community (DCNC) – eine Vereinigung eigenständiger Communitys, User Groups und Meetups aus Deutschland, Österreich und der Schweiz –, die DOAG (Deutsche Oracle-Anwendergruppe) und Heise Medien setzen darauf, das Programm des Events, das unter dem Motto „4 Tage – 4 Themen" stattfindet, zu einem Familienereignis werden zu lassen. Denn: Die Grenzen zwischen den Veranstaltungen des Festivals und den Attraktionen der Fahrgeschäfte sind für Festival-Besucher fließend, das heißt sie können beliebig oft hin- und her wechseln.

Ling Bao lässt grüßen

Dafür stehen im DOAG-Shop noch vergünstigte Tageskarten sowie der CloudLand Festival Pass zum Early-Bird-Preis zur Verfügung, genauso wie ein limitiertes Kontingent an Familienunterkünften im exotisch-schicken Phantasialand-Hotel Ling Bao. Hierfür im Shop einfach runter scrollen, Doppelzimmer auswählen, dann Kinder und Jugendliche dazu buchen und fertig ist das Familienzimmer für die CloudLand 2022, dem ersten deutschsprachigen Cloud Native Festival. *https://www.cloudland.org/de/home/*

„Der durch digitale Transformation einhergehende Veränderungsprozess und gesellschaftliche Wandel ist meines Erachtens höher anzusehen als die industrielle Revolution im 18./19. Jahrhundert."

Martin Meyer, Redaktionsleiter des Red Stack Magazin, sprach mit Dirk Andres, dem Leiter der Stabsstelle Digitalisierung in Kaiserslautern über Data Analytics, Digitalisierung und die datengetriebene Stadtentwicklung.

Können Sie sich kurz vorstellen? Mit was beschäftigen Sie sich beruflich?

Ich bin Dirk Andres, 53 Jahre alt, Diplom-Verwaltungswirt FH und Leiter der Stabsstelle Digitalisierung in Kaiserslautern.

Seit vielen Jahren beschäftige ich mich mit der Steuerung durch Information sowie mit Führungsinformationssystemen. Getreu dem Motto „Wissen statt Bauchgefühl" verfolge ich beruflich das Ziel, eine datengetriebene Stadtentwicklung umzusetzen.

Was sind Data Analytics? Was ist Data Science?

Die Begriffe fallen sicherlich auch in die Kategorie „Modewörter". Die Inhalte sind hingegen geläufig: Einerseits versuchen wir mithilfe von Daten Szenarien zu verstehen und gegebenenfalls auch zu prognostizieren, um daraus dann das notwendige Wissen für weitere Entwicklungsschritte zu schöpfen. Eigentlich geht es jedoch darum, wie wir die erhobenen Daten sinnvoll, effizient und realistisch in der Praxis nutzen können.

Ein Beispiel – datengetriebene Stadtentwicklung. Wir können unter anderem bestimmte Quartiersdaten erheben, die Auskunft über die Altersstruktur der Einwohner widerspiegeln und diese mit Milieudaten verschneiden. Durch Analysen können dann Aussagen darüber getroffen werden, wie das Viertel gestaltet werden sollte, damit es der Lebensrealität der Menschen entspricht und ihnen bestenfalls einen Mehrwert bietet. Stellt man zum Beispiel fest, dass viele junge Familien in ein Viertel ziehen, kann die Planung einer Kita in Betracht gezogen werden, noch bevor sich der Bedarf angekündigt hat.

Welche Rolle spielt die Digitalisierung beim Thema Data Analytics?

Der durch digitale Transformation einhergehende Veränderungsprozess und gesellschaftliche Wandel ist meines Erachtens höher anzusehen als die industrielle Revolution im 18./19.Jahrhundert. Im Rahmen dieses Prozesses haben wir nun die Möglichkeit, Prozesse selbst zu analysieren, zu gestalten und zu verbessern. Gleichzeitig ist es vor allem für die Verwaltung wichtig, die Handlungsfähigkeit kontinuierlich aufrechtzuerhalten. Dabei sind hohe Sicherheit, Datenschutz, kurze Wege, Niederschwelligkeit bei geringen finanziellen Möglichkeiten, Personalverfügbarkeit und die Qualifikation des Personals sicherzustellen.

Die Verwaltung sieht sich dabei ständig neuen Herausforderungen gegenübergestellt: die Umstellung auf Online-Dienste, die Pandemie oder Flüchtlingswellen – und das alles in kürzester Reaktionszeit. Diese Aufgaben stellen aktuell einen Balanceakt insbesondere für den öffentlichen Bereich dar.

Wie und wo werden Data Analytics heute eingesetzt?

Aus meiner Sicht noch viel zu wenig. Wir stecken hier leider noch in den Anfängen. Zunächst sind die Grundlagen zur Verfügung zu stellen, die Daten zu plausibilisieren, qualitativ zu bewerten, gegebenenfalls zu verbessern, zu kategorisieren und datenschutzrechtlich zu verorten. Viele dieser Teilaspekte wurden in der Vergangenheit nicht berücksichtigt oder nicht in dieser Gesamtheit gesehen. Eine reale Vernetzung unserer Informationen erkenne ich bislang jedoch nicht. DAS sind meiner Ansicht nach die Hausaufgaben, die wir zuerst machen müssen. Gleichzeitig müssen die Führungskräfte dahingehend sensibilisiert werden, ihre Entscheidungen auf Daten zu stützen – vieles geschieht heute noch aus dem Bauchgefühl heraus.

Worin liegen die Chancen von Data Analytics?

Hier kann ich ganz klar nochmals den Aspekt „WISSEN STATT BAUCHGEFÜHL“ wiederholen – darin liegt die Zukunft. Sowohl die Datenmenge als auch die Informationsvielfalt nehmen täglich zu. Einhergehend mit immer mehr neuen technischen Möglichkeiten werden Begehrlichkeiten geweckt. Hier wird es wichtig sein, kurzfristig auf fundierte Daten zugreifen zu können und zu dürfen. Die Voraussetzung dafür: ein offenes Mindset bei Entscheidungsträgern für solche Prozesse.

Wo sehen Sie die deutsche Wirtschaft und Verwaltung bei diesem Thema?

Zur Wirtschaft kann ich nur wenig fundierte Ansichten vertreten, die Verwaltung steht jedoch sicherlich mit der Wirtschaft noch lange nicht auf Augenhöhe – aber es tut sich was.

Als Smart City im Modellvorhaben des Bundes, aber auch im Rahmen des Interkommunalen Netzwerks Digitale Stadt (IKONE DS) stehen wir mit etlichen Kommunen im regen Austausch und merken: Die Probleme werden nicht nur gesehen, sondern auch angegangen – und das Ganze ohne Konkurrenzdenken. Was das betrifft, ist die Verwaltung im Vorteil. Die Herausforderungen der digitalen Transformation sind überall gleich, das heißt natürlich, wir müssen das Rad bei funktionierenden Herangehensweisen nicht immer neu erfinden, sondern können voneinander lernen. Voraussetzung sind natürlich eine offene Kommunikation und transparentes Handeln.

Was wünschen Sie sich für die (digitale) Zukunft? Wann ist Ihre Stadt smart?

An dieser Stelle möchte ich Eigenwerbung für unser Motto in Kaiserslautern machen, nämlich „Herzlich digital“. Damit stellen wir den Menschen in den Mittelpunkt der Digitalisierung. Wir haben die große Chance, die Weichen zu stellen und die digitale Zukunft in unserem Sinne zu gestalten. Wir haben auch die Möglichkeit, uns dabei nicht bloß, wie mehrfach erwähnt, auf ein vages Bauchgefühl zu stützen, sondern kennen Mittel und Wege, um klare Handlungsvorschläge abzuleiten, die auf Fakten beruhen. Das in Kombination mit einem offenen Mindset der Verantwortlichen öffnet uns große Türen, die volle Bandbreite der Digitalisierung auszuschöpfen. Die Definition der Smart City besteht für mich darin, dass sich die Menschen in der Stadt in erster Linie wohlfühlen. Digitalisierung ist dabei allerdings nur ein Teil-Aspekt, um dies zu erreichen.

DIRK ANDERS

Dirk Andres ist 53 Jahre alt und Diplom-Verwaltungswirt FH. Er beschäftigt sich seit Jahren mit der Steuerung durch Information sowie Führungsinformationssystemen. Seit 2020 ist er Leiter der Stabsstelle Digitalisierung in Kaiserslautern

AUS DER FERNE BETRACHTET: DEN DATENSCHATZ HEBEN

Wolfgang Taschner, Rentner und ehemaliger Chefredakteur des Red Stack Magazin

Während der Corona-Pandemie wurde ich laufend mit Informationen aus allen Kanälen überflutet. Um aus dem Daten-Stress herauszukommen, fing ich an, mich weniger auf die eingehenden Nachrichten zu konzentrieren, sondern mehr die Trends zu betrachten. Daraus ergab sich für mich ein Bild von der Lage, mit dem ich ohne Druck umgehen und das ich bei Bedarf an sich verändernde Gegebenheiten anpassen konnte.

Genauso stelle ich mir Projekte vor, die sich mit Data Analytics beschäftigen. Aus einer Fülle von Daten gilt es, die Informationen herauszufiltern und zusammenzufügen, auf deren Basis sich fundierte Entscheidungen für künftige Geschäftsprozesse treffen lassen.

Vor Kurzem habe ich das Buch „Der Alchimist" von Paulo Coelho gelesen – ein schönes Märchen über einen spanischen Hirten, der sich auf die Suche nach einem vergrabenen Schatz macht. Die spannend und abwechslungsreich geschriebene Geschichte bietet reichlich Anknüpfungspunkte, in die sich jeder sein eigenes Leben hineininterpretieren kann. Auf diese Weise habe auch ich meinen Schatz gefunden.

In diesem Sinne wünsche ich Ihnen allen Glück und Erfolg beim Suchen und Heben Ihres Schatzes.

WOLFGANG TASCHNER

DOAG DATENBANK KOLUMNE: DBA 2.0 – WIE SIEHT ES MIT SOFT SKILLS AUS?

Andreas Buckenhofer, Datenbank Community

Andreas Buckenhofer stellt Soft Skills vor und erläutert, wozu sie nützlich sind.

Technologien und Tools zu beherrschen, reicht seit langem nicht mehr aus. Wer hat es nicht erlebt? Die beste technische Lösung verkauft sich nicht von selbst, wenn man es nicht schafft, diese zu kommunizieren und andere zu überzeugen. Kommunikationsprobleme zwischen DBAs und Entwicklern sind ein altbekanntes Thema. Softskills wie Kommunikation oder agiles Mindset sind auch in technischen Rollen essentiell.

Die Arbeit eines DBAs ist durch den Rückgang von Routinetätigkeiten längst Änderungen unterworfen. Zunehmende Automatisierung wie Infrastructure as Code oder die Nutzung von Managed Services sind Beispiele. Neue Systeme werden komplexer und können nicht durch noch mehr Planung entwickelt werden. Zusammenarbeit, Agilität oder Produktorientierung sind notwendig. Nur gemeinsam können die Anwendungen in der heutigen VUCA-Welt umgesetzt werden, die durch Volatilität, Unsicherheit, Komplexität und Mehrdeutigkeit gekennzeichnet ist. Dazu notwendig ist, dass jeder im Team beiträgt und an den eigenen Fähigkeiten arbeitet. Egal ob DBA, Architekt, Entwickler, Productowner, usw.

Zu den Soft Skills gehören Sozial- und Persönlichkeitskompetenzen sowie methodische Kompetenzen. Beispiele sind:

- Sozialkompetenzen: Teamfähigkeit, Netzwerken, Kommunikationsfähigkeit, Konfliktmanagement, Empathie, uvm.
- Persönlichkeitskompetenzen: Selbstreflexion, agiles Mindset, Verantwortungsbereitschaft, Work-Life-Balance, Selbstmanagement, Glaubwürdigkeit, uvm.
- Methodische Kompetenzen: Präsentationstechniken, Analysetechniken, Rhetorik, Veränderungsmanagement, Kreativitätstechniken, uvm.

Schulungen zu einem neuen Produkt oder das jährliche Upgrade-Training zur neuesten DB-Version sind selbstverständlich. Wie sieht es mit Schulungen zu Soft Skills aus? Diese sind genauso relevant und gehören in den Entwicklungsplan eines Mitarbeiters.

Konferenzen wie die DOAG Konferenz + Ausstellung bieten seit Jahren Vorträge zu Soft Skills an. Wer das Angebot noch nicht wahrgenommen hat: einfach mal in solche Vorträge gehen und über den Tellerrand schauen.

Chancen und Grenzen von AutoML

Detlef E. Schröder, Oracle Deutschland – Technology Software Engineering

KI und ML sind zwei Schlagworte, die heutzutage nicht mehr wegzudenken sind. Dabei gibt es zwei Richtungen, in denen Lösungen angeboten werden. Zum einen sehr tief technisch und zum anderen als automatisierter Prozess. Im Folgenden wird der automatisierte Prozess am Beispiel von Oracle AutoML näher beleuchtet sowie dessen Chancen und Grenzen aufgezeigt. Dies wird mit praktischen Beispielen untermauert und somit ein Weg durch den ML-Dschungel gebahnt.

Der Machine-Learning-Prozess

Der Begriff Machine Learning ist ein Oberbegriff für die Verwendung von Algorithmen für die Gewinnung von Regeln und Modellen zur Generalisierung von Lerndaten und historisch bekannten Ergebnissen. Die gewonnenen und generalisierten Modelle und Regeln ermöglichen anschließend, aus unbekannten Daten Erkenntnisse zu erzielen oder Aktionen abzuleiten. Damit unterscheidet sich der Machine-Learning-Prozess grundlegend von der klassischen Programmierung, bei der die Regeln im Vorhinein feststehen und auf die Daten angewendet werden.

Das Ziel der Generalisierung von Mustern und Gesetzmäßigkeiten in Daten kann durch bekannte Ergebnisse aus historischen Daten erzeugt werden – überwachtes Lernen – oder ohne diese Vorgaben extrahiert werden –unüberwachtes Lernen.

Eine entscheidende Bedeutung kommt der Auswahl der Daten zu. Denn diese Daten müssen zu der definierten Frage als Ausgangspunkt des gesamten Prozesses passen. Bevor der Modellierungsprozess starten kann, ist es wesentlich, das Problem oder die Aufgabe so zu beschreiben, dass aus der fachlichen Sicht auch alle relevanten Informationen und zugrunde liegenden Daten ermittelt werden können. Erst auf dieser Basis kann mit den weiteren Schritten im Machine-Learning-Prozess, der Datenaufbereitung und -vorbereitung, weiter vorangegangen werden.

Der ML-Prozess wird wie in *Abbildung 1* dargestellt als **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining definiert.

Hierbei sind diese ersten Schritte entscheidend für den Erfolg des gesamten Prozesses. Ohne eine klare Definition können die Ergebnisse nicht entsprechend verwendet werden. Und ohne eine adäquate Datenbasis kann keine Generalisierung stattfinden. Diese vorbereitenden Aufgaben werden meist interdisziplinär von den beteiligten fachlichen Experten – den sogenannten Domain Ownern – und den statistischen Experten – den Operational Ownern – durchgeführt.

Die Anwendung von Mathematik zur Erkenntnisgewinnung und Modellerstellung ist „nur" ein Schritt in dem Prozess. Daher ist eine Gesamtbetrachtung des Prozesses zur Bewertung der Automatisierung dieses Prozesses auch entscheidend.

Was macht AutoML?

AutoML ist kein klar abgegrenzter Begriff. Aber Wikipedia definiert dies wie folgt: **„Automated machine learning (AutoML)** is the process of automating the tasks of applying machine learning to real-world problems. AutoML covers the complete pipeline from the raw dataset to the deployable machine learning model." *[1]* AutoML.org legt den Schwerpunkt auf die Verwendung durch Nicht-Experten und geht davon aus, dass ML-Methoden ohne Expertenwissen verwendet werden können.

Beide definieren auch die Prozessschritte, die aus dem ML-Prozess im AutoML behandelt und automatisiert werden. Dies beginnt bei der Aufbereitung der Daten und endet bei der Bewertung der Modelle. Bezogen auf den CRISP-DM bildet dies die Prozessschritte 3-5 ab.

Auch das AutoML von Oracle umfasst diese Prozessschritte, wie sie in der schematischen Darstellung in *Abbildung 2* abgebildet sind.

In dem OML-Python-Paket, das zur Python-Schnittstelle des Oracle Machine Learning gehört, ist die Klasse „automl" implementiert. Die „Auto Algorithmus Selection" identifiziert die erfolgversprechendsten Algorithmen anhand der ausgewählten Kennzahl Modellgüte. Das „Adaptive Sampling" bestimmt die optimale Größe der Trainingsdaten, vor allem auf der Basis von unbalancierten Datenbeständen. Der dritte Schritt „Auto Feature Selection" untersucht die Attribute auf ihre Verwendbarkeit und filtert das Rauschen heraus. Des Weiteren wird die Anzahl der verwendeten Attribute so ausgewählt, dass die Trainingszeit optimiert wird, ohne die Modellgüte zu verschlechtern. Im letzten Schritt werden die Modelle bezogen auf das Gütekriterium optimiert und dafür die Hyperparameter der Modelle angepasst. Das Ergebnis ist eine Bewertung der Modelle anhand der verschiedenen Güte-

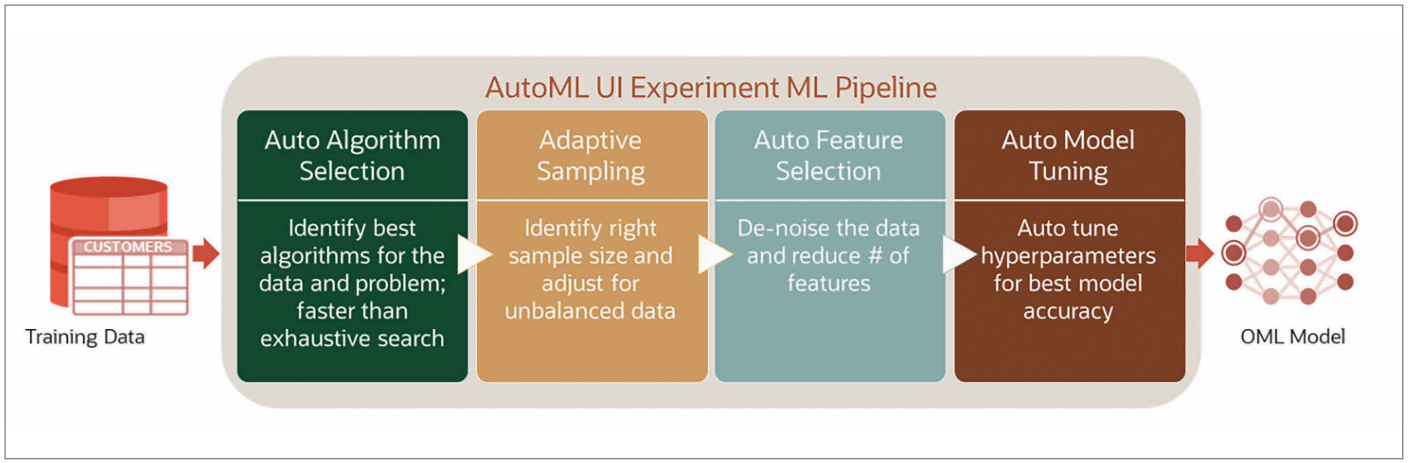


Abbildung 1: Das CRISP-DM (© Kenneth Jensen, Wikipedia.org)

kennzahlen und eine damit verbundene Auswahl des besten Modells.

Wo liegen die Grenzen von AutoML?

Bezogen auf das CRISP-DM unterstützt das AutoML vor allem die Prozessschritte, die direkt mit den Daten zusammenhängen. Die Schritte zu Beginn – die Definitionsphase und die fachliche Datenphase – werden nicht unterstützt. Ebenso ist ein AutoML nicht in der Lage, die Interpretation der Generalisierung zu bewerkstelligen. Diese Schritte, an denen im Wesentlichen die Domainexperten beteiligt sind, verbleiben außerhalb der Automatisierung. AutoML deckt im Wesentlichen die Schritte ab, die datenintensiv sind.

Die Grenzen stellen jedoch keine Bewertung dar. Ob ein AutoML hilfreich oder wertvoll eingesetzt werden kann, liegt an der Gewichtung der Prozessschritte. Die Unterstützung in der Anwendung der Mathematik und bei Algorithmen stellt in vielen Fällen eine enorme Erleichterung dar, vor allem, wenn das Wissen darüber bei den Domainexperten nicht vorhanden ist.

Ebenso vereinfacht AutoML viele Schritte eines Data-Scientisten, die sonst per Hand durchgeführt werden müssten. Somit steigert es die Effizienz und Produktivität im Prozess.

Ob und inwieweit AutoML unterstützend einzusetzen ist, liegt also an der Gewichtung der Schritte, die unterstützt werden können. Dies wird in der folgenden Grafik in *Abbildung 3* verdeutlicht.

Risiken von AutoML

Die Verkürzung des Prozesses auf wenige Schritte stellt das größte Risiko dar. Es geht oft bei der Lösung einer Fragestellung nicht um die Erstellung eines singulären Modells zur Generalisierung der Datenwelt, sondern um eine variable Kombination von mehreren voneinander abhängigen Modellen, die jeweils einen Teil der Fragestellung bearbeiten. Dies macht auch die Aufgabe der Interpretation so wichtig und entscheidend.

Eine weiteres Risiko, laut AIINSIGHTS.COM.AU, ist „Garbage in – Garbage out". *[2]* Machine Learning lebt von den Daten. Wenn die in den Prozess gegebenen Daten nicht in der Lage sind, die Datenwelt zu generalisieren, können die gefundenen Zusammenhänge und Modelle dies ebenfalls nicht. An dieser Stelle zeigt sich die hohe Relevanz zum einen der richti-

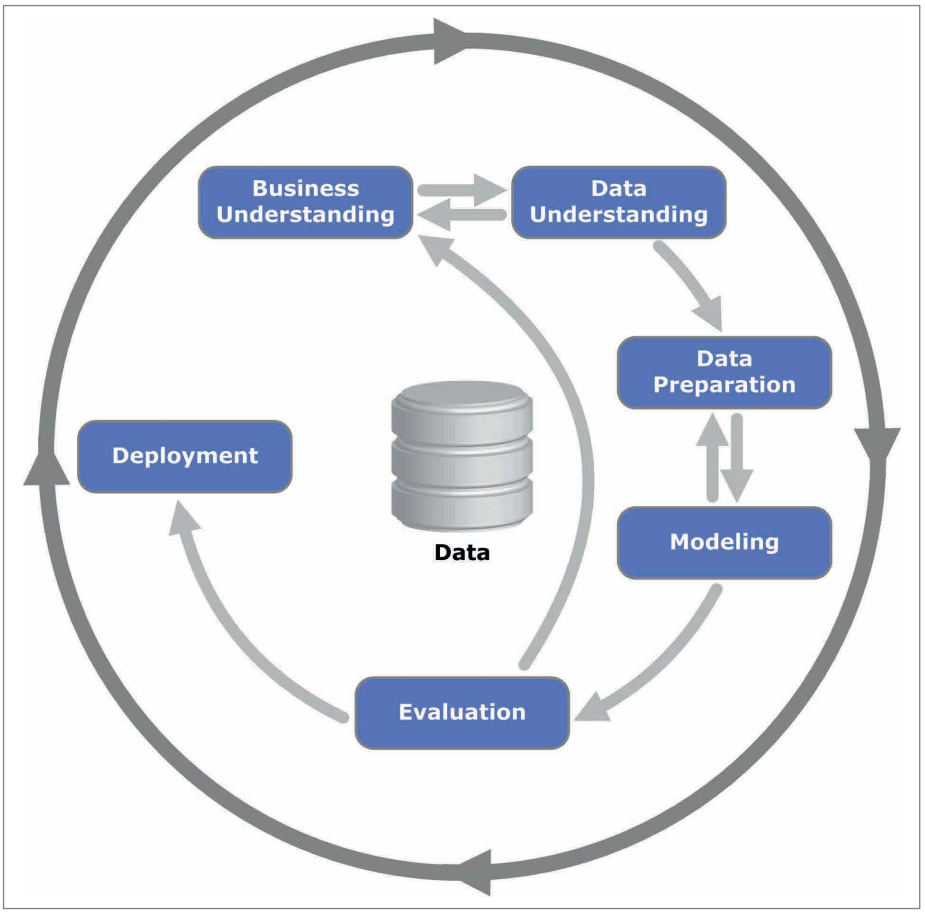


Abbildung 2: Der Oracle-AutoML-Prozess (© oml-automl-ui-tech-brief.pdf, oracle.com)

gen Fragestellung und zum anderen der dafür notwendigen, richtigen Datenselektion und -verwendung.

Denis Vorotyntsev macht in seinem Blog „AutoML is overhyped“ *[3]* deutlich, dass nur etwa 20% der für einen ML-Prozess verwendeten Zeit durch AutoML unterstützt werden und „dieser Motor noch kein Auto ausmacht“. Dazu gehört, wie schon beschrieben, einiges mehr.

Immer wieder taucht in der Diskussion um ML auch der „Bias“ auf. „Bias“ meint dabei die Verzerrung, die in den Daten steckt. Wenn in der Vergangenheit nur die sprichwörtlich „alten, weißen Männer“ ihre Datenspuren hinterlassen haben, wird kein Algorithmus und AutoML-Prozess der Welt in der Lage sein, etwas anderes vorherzusagen oder zu generalisieren. Das Bias-Problem bleibt vom Einsatz von AutoML unberührt. Daher stellt AutoML hier zwar kein größeres Risiko dar, aber der automatische und unabhängige Prozess könnte dies verschleiern.

Chancen von AutoML

Und doch hat AutoML seine Berechtigung gerade dort, wo den Domain-Experten das Wissen und die Unterstützung bei der Umsetzung fehlen. Diese Nicht-Data-Scientisten können natürlich davon profitieren. Allerdings ist auch hier das Verständnis für die Prozesse notwendig. Dieses jedoch vorausgesetzt, kann AutoML dann auch helfen, die Schnittstelle zwischen den fachlichen Experten und Data-Science-Experten zu bilden. Gerade dieser Dialog stellt oft einen wesentlichen Erfolgsfaktor da.

Auch wenn nur ein Teil der Prozessschritte durch AutoML abgedeckt werden, liegt in jedem Fall eine Arbeitserleichterung vor. Denn ohne AutoML müssten diese Prozesse manuell erledigt werden. Des Weiteren kann der AutoML-Einsatz die Qualität und Dokumentation verbessern.

Fazit

Auch wenn hier mehr Risiken als Chancen aufgezeichnet sind, hängt die Bewertung von AutoML an der Gewichtung der Unterstützungsleistung und muss daher individuell von den Projektteams vorgenommen werden. Wie immer ist auch AutoML kein Allheilmittel, aber eine sinnvolle Unterstützung.

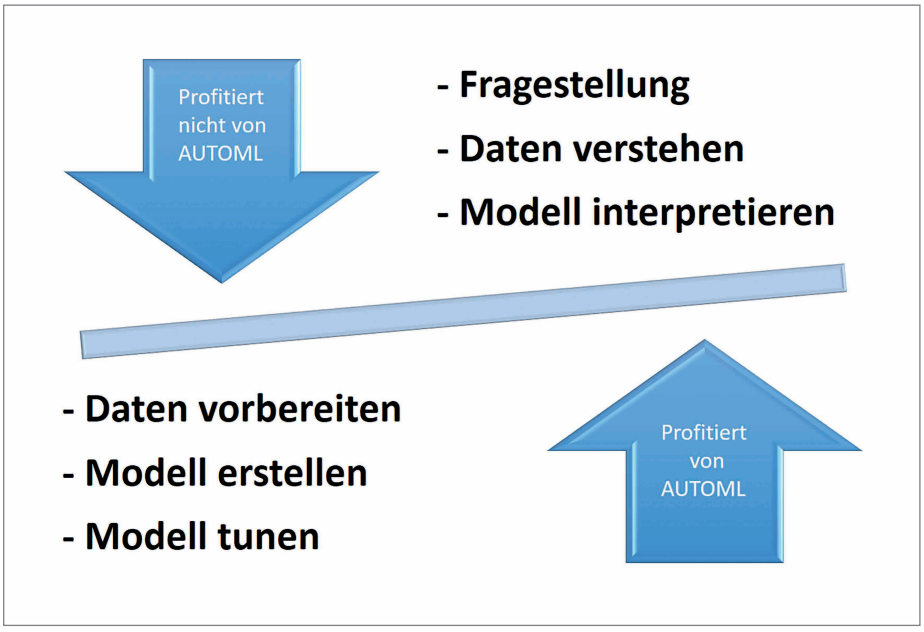


Abbildung 3: Gewichtung der AutoML-Unterstützung (© Detlef E. Schröder)

Quellen

[1] Wikipedia – Definition AutoML
[2] 5 Dangers of AutoML – AIINSIGHTS.com.au
[3] AutoML is Overhyped – Denis Vorotyntsev, towordsdatascience.com

Über den Autor

Detlef E. Schröder arbeitet als Master Principal Solution Engineer bei Oracle Deutschland B.V. & Co KG. In dieser Rolle unterstützt er Kunden bei analytischen Fragestellungen und KI-Aufgabestellungen sowie DWH-Architekturen. Des Weiteren entwickelt und hält er Seminare zu den Themen und ist in der Redaktion der „Oracle DB Monthly News“.

Detlef E. Schröder
Detlef.E.Schroeder@oracle.com

Abbildung 1: Data Lake (Quelle: https://www.flickr.com/photos/ian-arlett/34233379390)

Data Lake – Wer braucht das schon?

Jan Ott, Trivadis

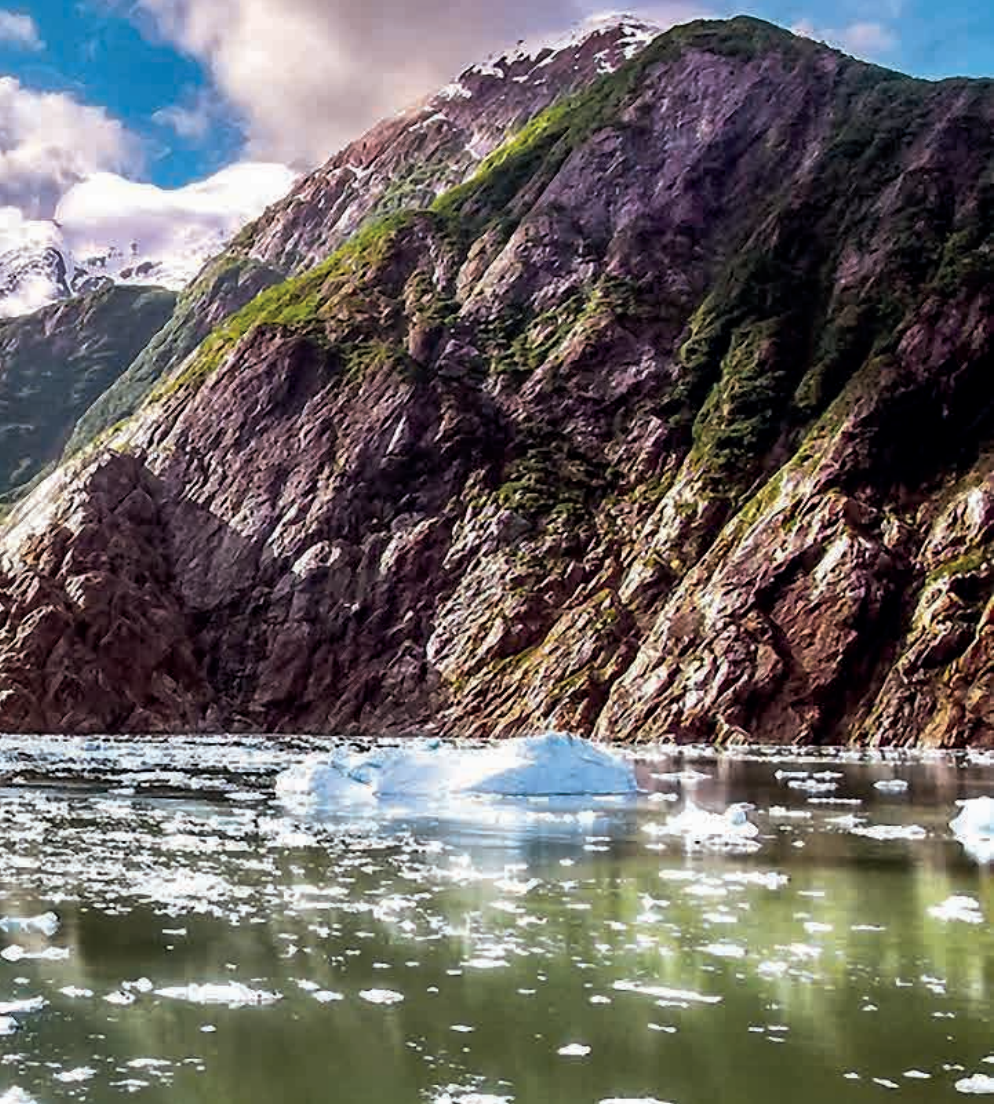
Data Warehouses sind heutzutage etabliert. Seit einiger Zeit schwirrt der Begriff Data Lake durch die IT-Welt. Ist das nur ein Modetrend oder muss man sich damit befassen? In diesem Artikel möchte ich die Idee eines Data Lakes vorstellen und mit dem Data-Warehouse-Konzept vergleichen.

Definition

Um ein gemeinsames Verständnis zu erreichen, benötigen wir Definitionen für die zentralen Begriffe Data Lake und Data Warehouse.

Data Lake

Ein Data Lake (Datensee) ist ein System, in dem alle Daten eines Unternehmens gesammelt werden. Diese Daten sind in ihrer rohen Form abgelegt, meist als Dateien. Die Daten stammen von den operativen Systemen der Unternehmung. Es können jedoch auch Daten aus anderen Quellen sein, zum Beispiel aus sozialen Netzwerken. Einfach alles, was für die Berichterstattung, Analysen und maschinelles Lernen verwendet werden kann. Die Daten



können klar strukturiert sein, also beispielsweise aus einer relationalen Datenbank stammen, semistrukturiert sein wie E-Mails, Twitter oder unstrukturiert wie binäre Daten, Bild und Ton. Dies bedeutet, dass sie die verschiedensten Formate haben können: CSV, XML, JSON und andere.

Data Lake – James Dixon

James Dixon hat das bildlich sehr gut zusammengefasst. Er schreibt sinngemäß:

„Wenn Sie sich einen Data Mart als einen Vorrat an abgefülltem Wasser vorstellen – gereinigt, verpackt und strukturiert für den einfachen Verbrauch –, dann ist der Data Lake ein großes Gewässer in einem natürlicheren Zustand. Der Inhalt des Data Lake strömt aus einer Quelle ein, um den See zu füllen, und verschiedene Benutzer des Sees können kommen, um ihn zu untersuchen, einzutauchen oder Proben zu nehmen."

Data Warehouse

Demgegenüber haben wir das Data Warehouse.

Ein Data Warehouse ist ein System, in das gezielt Daten geladen werden.

Diese Daten werden aus den Quellsystemen extrahiert und in mehreren Schritten (ETL) transformiert, damit sie am Ende geprüft und bereit für unterschiedliche Verwendungszwecke sind. Es wird gezielt nur das geladen, was auch verwendet wird.

Gerne würde ich noch einen weiteren Begriff erläutern:

Data Lakehouse

Data Bricks hat diesen Begriff mitgeprägt. Sie schreiben auf ihrer Seite sinngemäß:

„Ein Data Lakehouse ist eine neue, offene Datenmanagement-Architektur, die die Flexibilität, Kosteneffizienz und Größenordnung von Data Lakes mit dem Datenmanagement und den ACID-Transaktionen von Data Warehouses kombiniert und Business Intelligence (BI) und maschinelles Lernen (ML) auf allen Daten ermöglicht. Das Ziel ist Einfachheit, Flexibilität und tiefe Kosten."

In den folgenden Absätzen stelle ich diese drei Architektur-Ansätze einander gegenüber.

Data Lake

Die Idee

Der Definition des Data Lake sind folgende Kernaussagen zu entnehmen:

- ein Datenspeicher für alle Daten des Unternehmens
- Basis für „alles" im analytischen Bereich
  - Reporting
  - Visualisierung
  - Analyse
  - Machine Learning
  - ...

Ein Data Lake ist eine Infrastruktur. Diese soll alle Daten des Unternehmens enthalten. Die Daten sollen möglichst performant aufgenommen und weitergeben werden können. Damit soll Silos in den Unternehmen, die sich in der Vergangenheit gebildet haben, entgegengewirkt werden. Die Daten dürfen in den verschiedensten Formaten angeliefert werden. Dies auch, um möglichst viele Daten des Unternehmens in diesem einen Topf zu sammeln.

Damit dieses Ziel erreicht werden kann, wird ein weiteres Paradigma eingeführt – Schema on Read.

Unter Schema on Read versteht man die Anwendung von Strukturen und Datenprüfungen zum Zeitpunkt des Lesens der Daten. Im Gegensatz dazu werden beim Schema on Write Strukturen und Prüfungen bereits beim Schreiben der Daten in den Datenspeicher angewendet.

Der Vorteil des Schema on Read besteht darin, dass wir alle Daten laden. Es kann allerdings sein, dass sie Fehler enthalten. Doch dies wird erst beim Lesen – Schema on Read – erkannt und dann entsprechend behandelt. Dies bedingt eine neue Fehlerkultur.

Um die Freiheit so groß wie möglich zu halten, werden alle Daten aus allen Source-Systemen geladen. Dies führt zum Teil zu redundanten Daten und zu großen Datenmengen.

Somit kommen wir nun zu der Verwendung der Daten. Dadurch, dass wir alles einfach geladen haben, können wir nun entscheiden, was wann und wie verwendet wird. Damit haben wir die Freiheit, aus allen Daten zu wählen. Das heißt, wir können nun entscheiden, ob

Abbildung 2: Data Swamp (Quelle:https://www.flickr.com/photos/82134796@N03/10603438015)

wir aggregieren, filtern oder andere Manipulationen vornehmen wollen.

Da die Daten beim Laden nicht geprüft wurden, müssen wir das nun beim Lesen der Daten – Schema on **Read** – nachholen. Die möglichen Fehler müssen abgehandelt werden.

Diese rohen Daten sind nicht unbedingt für den normalen Benutzer geeignet. Doch für Ad-hoc- und One-Shot-Implementationen, Self Service Data Labs, Queries und Advanced Analytics schon.

Das Paradigma, alles zu laden, hat natürlich seine Schattenseite: Data Swamp.

Data Swamp

Die ersten Data Lakes haben gezeigt, dass mit all den Daten, die unstrukturiert und ohne Kontrolle in diesen See geworfen werden, Probleme entstehen. Das Größte davon ist, dass die Daten wie in einem Sumpf verschwinden und dann nicht mehr gefunden beziehungsweise genutzt werden können. Daher kommt der Name Data Swamp beziehungsweise Datensumpf, wie die *Abbildung 2* sehr schön zeigt.

Die Ursache davon ist, dass jeder macht, was er will.
Hier die Ursachen, die zu einem Datensumpf führen:

- keine Metadaten
- keine Dokumentation
- keine Qualitäts-Sicherung
- keine Datenflussanalyse
- keine vereinheitlichte Veredelung der Daten
- keine Standard-Formate
- keine Versionierung
- keine einheitliche KPI-Berechnung

Zusammengefasst: Es gibt keine Datenarchitektur.

Data Warehouse

Ein Data Warehouse ist das totale Gegenteil eines Data Lake. Hier wird nur geladen, was auch verwendet wird. Die Daten werden zwar in verschiedenen Formaten geliefert, doch anschließend in einem ETL-Prozess (Extract, Transform und Load) in eine einheitliche Struktur gebracht. So sind die Daten hochstrukturiert und bereit für die Verwendung in Reports, für Visualisierungen und Weiteres. Zudem wird alles geprüft und somit sichergestellt, dass die Qualität der Daten über jeden Zweifel erhaben ist. Daten aus Data Warehouses werden unter anderem an interne wie externe Kontrollorgane geliefert.

Abgrenzung Data Lake – Data Warehouse

Datenstruktur: Roh – Verarbeitet

In einem Data Lake lädt man die Daten möglichst roh. Dies hat den Vorteil, dass die Verarbeitung minimale Ressourcen benötigt und die Daten sehr schnell geladen werden können. Es wird auch keine Zeit verschwendet, um die Datenstrukturen zu prüfen (Schema on Read). Dies hat zur Folge, dass die Daten bei Lieferung verfügbar sind. Andererseits müssen die Strukturen zu einem späteren Zeitpunkt geprüft werden.

	Data Lake	Data Warehouse
Datenstruktur	Roh bis wenig verarbeitet	Hochverarbeitet
Welche Daten	Alles, was gebraucht werden könnte	Nur was gebraucht wird
Benutzer	Data Scientis	Business Anwender
Zugänglichkeit	Begrenzt zugänglich Günstig und schnell erweiterbar	Einfach, breit verfügbar Teuer und komplex zu ändern
Mögliche Fragestellung	Offen	Vordefiniert
Integrationsgrad	Gering	Hoch
Design	Undefiniert ggf. Data Vault	3NF, Dimensional, Data Vault

Abbildung 3: Gegenüberstellung von Data Lake und Data Warehouse (Quelle: Jan Ott)

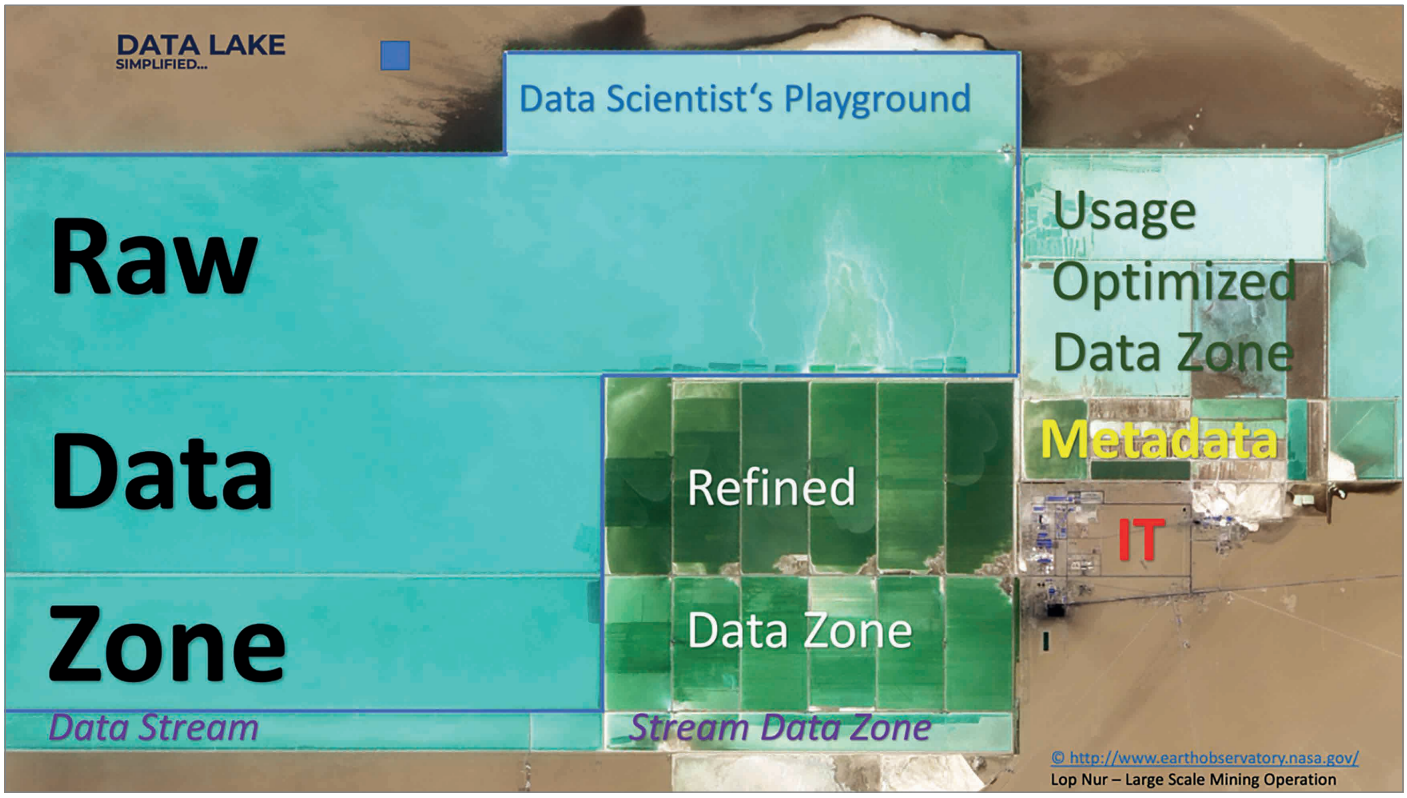


Abbildung 4: Data Lake – Simplified (Quelle: http://www.earthobservatory.nasa.gov/)

Bei einem Data Warehouse ist es genau umgekehrt. Hier wird gleich zu Beginn alles geprüft und entsprechend verarbeitet (Schema on Write). Zudem werden schon Berechnungen gemacht und die Daten für den Zugriff optimiert. Das alles braucht Zeit und daher ist ein Data Warehouse meist auch erst mit einem Tag Verzögerung verfügbar *(siehe Abbildung 3)*.

Datenmengen: Alles – Teil

In einen Data Lake wird möglichst alles geladen. Alle Daten aus allen verfügbaren Quellen werden berücksichtigt. Einfach alles, was vielleicht einmal von Nutzen sein kann. Es werden daher auch keine Filter appliziert. All dies führt dazu, dass wir in Data Lakes sehr große bis riesige Datenmengen anhäufen. Dies ist beides – ein Segen, wir haben alles – und ein Fluch, wir haben alles – riesige Datenmengen.

In ein Data Warehouse laden wir nur, was wir wirklich brauchen. Das heißt, dass wir Filter applizieren, nicht verwendete Felder weglassen und zum Teil die Daten aggregieren. Das schlägt sich in kleineren Datenmengen nieder und kann zu einem Nachteil werden. Wenn wir zu einem späteren Zeitpunkt feststellen, dass ein Feld fehlt oder wir Daten auf einem tieferen Aggregationslevel benötigen, dann geht das nicht mehr. Die Daten sind nicht vorhanden.

Benutzergruppe: Data Scientist – Business Anwender

Der Data Lake ist möglichst roh. Dies bedeutet, dass es einiges an Wissen braucht, um die Daten zu verwenden. Daher werden sie meist nur vom Data Scientist benutzt, also von einer kleinen Gruppe von Usern. Das hat auch damit zu tun, dass die Benutzer hochprivilegiert sein müssen. Zudem werden meist spezielle Tools eingesetzt, die auch entsprechendes Wissen voraussetzen.

Ein Data Warehouse nimmt dem Benutzer all das ab. Die Daten sind spezifisch auf die entsprechenden Benutzergruppen zugeschnitten. Es ist somit nur inhaltliches beziehungsweise Fachwissen erforderlich. Zudem sind die Reports nur mit entsprechenden Privilegien verfügbar.

Zugänglichkeit

Die Daten in einem Data Lake sind nur schwer zugänglich. Nicht, weil sie nicht verfügbar sind, sondern weil sie roh vorliegen. Sie sind allerdings schnell und günstig erweiterbar. Die Daten sind außerdem noch nicht gefiltert. Zudem sind die Beziehungen der Daten untereinander noch nicht abgebildet. Das erschwert den Zugang.

Bei einem Data Warehouse sind die Daten vorbereitet für die spezifische Verwendung, zum Beispiel für Reports oder Dashboards. Es müssen aber Releases, Prüfungen und so weiter abgewartet werden, was die Verfügbarkeit verzögert.

Rückschlüsse aus dem Vergleich

Der Vergleich macht klar, dass beides seine Berechtigung hat. In einer ersten Euphorie-Welle wurden die ganzen Data Warehouses abgeschafft. Man ging davon aus, dass man nur einen Data Lake braucht, weil dieser ja alle Daten enthält. Doch so einfach ist es nicht, wie obige Ausführungen zeigen.

Aufbau eines Data Lake (vereinfacht)

Ein Data Lake im ursprünglichen Sinn besteht aus den beiden Teilen „Raw Data Zone" und „Data Scientist's Playground", wie in *Abbildung 4* dargestellt. Stellen wir dies nun dem Data Warehouse gegenüber.

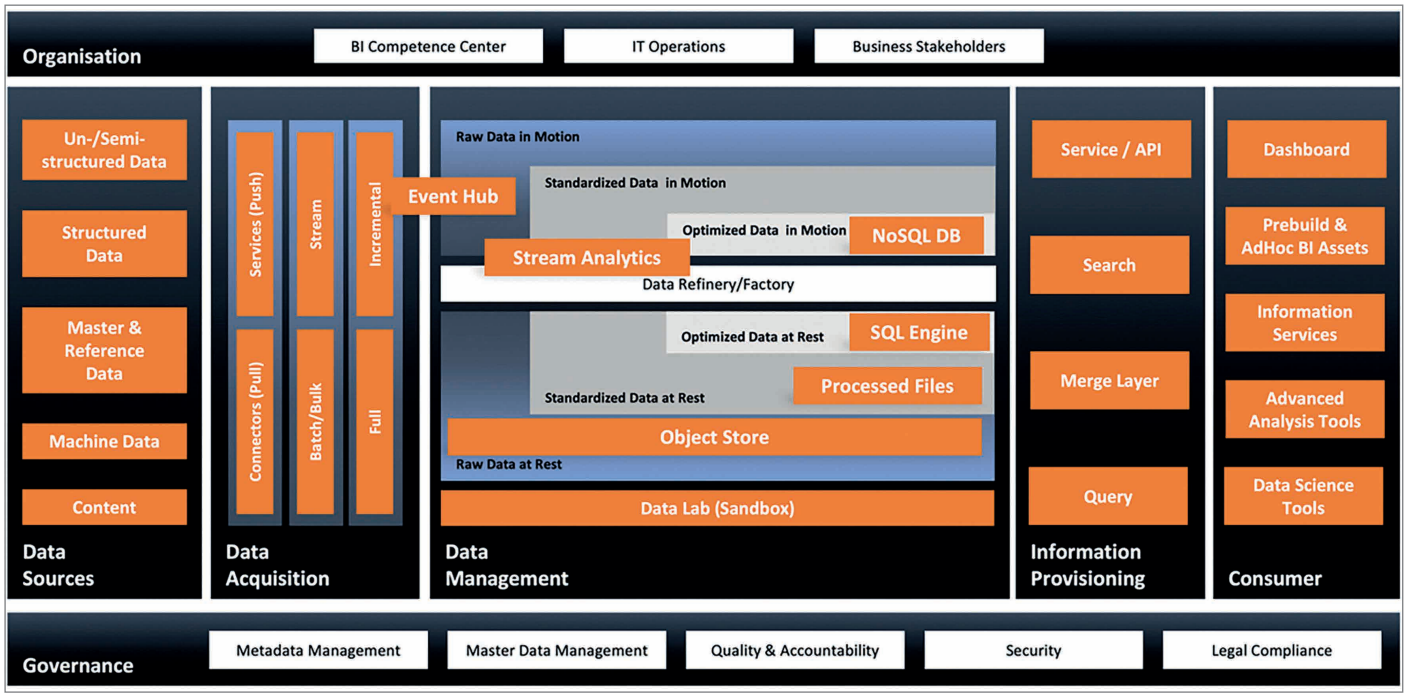


Abbildung 5: Data Lake – Design (Quelle: Jan Ott)

Der Data Lake wandelt sich

Das Konzept des Data Lake hat sich in den letzten Jahren verändert. Es sind dazugekommen:

- Refined Data Zone
- Usage Optimized Data Zone
- Metadata
- Data Stream / Stream Data Zone

Die neuen Bereiche decken nun mehr ab. In der „Refined Data Zone" sind die Daten beispielsweise immer noch nahe beim Original, doch sind sie geprüft, ein Schema wurde appliziert und sie wurden in ein Standardformat, etwa AVRO, konvertiert. Dies macht man, um den Zugriff zu erleichtern und sicherzustellen, dass diese Schritte nicht von jedem Benutzer der Daten nochmals implementiert werden müssen.

Design

Data Lake

(siehe Abbildung 5).

Data Warehouse

An den verschiedenen Designs sieht man die Unterschiede der Komponenten, die angesprochen/verwendet werden sowie der Herangehensweise.

In den Grafiken bedeutet die Farbe Weiß, dass diese Komponente nicht zur Anwendung kommt, die Farbe Orange, dass die Komponente verwendet wird, und das helle Orange, dass die Komponente optional ist *(siehe Abbildung 5 und 6)*.

Wieso?

Ein Data Lake deckt einen anderen Bedarf ab als ein Data Warehouse. In einem Data Lake werden zwar auch die traditionellen Quellen wie zum Bespiel die Daten aus den operativen Systemen geladen. Doch dazu kommen noch Log Files, Click-Streams, Social Media und vieles mehr. Diese Datenmengen würden ein Data Warehouse überfordern. Zudem will man möglichst alle Daten des Unternehmens erfassen. Auch dies führt zu großen Datenmengen. Wenn dann noch die unterschiedlichen Datentypen wie JSON, CSV, XML, Bit-Stream, Bilder, Video und Weitere angeliefert werden, dann wird das Data Warehouse inperformant. Genau das hat in der Vergangenheit dazu geführt, dass dann Silos entstanden sind.

Die Skalierung ist ein weiteres Thema. Ein Data Lake ist vom Design her so ausgelegt, dass bei Bedarf einfach skaliert werden kann. Dies ist bei einem Data Warehouse meist nicht möglich. Zudem ist ein Data Lake in der Lage, riesige Datenmengen zu verarbeiten, da hier die Datenverarbeitung zurückgestellt wird (Schema on Read).

Der Aufbau der Infrastruktur eines Data Lake ist prädestiniert, um darauf Tools für Machine Learning und Analytics aufzubauen. Da unkompliziert und schnell eine neue Quelle in den Data Lake aufgenommen werden kann, ist man viel flexibler.

Weil es nur noch eine Plattform für die rohen Daten gibt und diese von allen Systemen, die sie benötigen, verwendet wird, werden dadurch die operativen Systeme entlastet.

Dadurch ist ein Data Lake der Grundstein für die Zukunft – für Prognosen von zukünftigen Erkenntnissen. Diese kann man meist nicht aus einem Data Warehouse heraus erarbeiten.

Für ein Data Warehouse ist ein Data Lake allerdings eine hervorragende Grundlage. Die Daten wurden schon aus dem operativen System extrahiert und können gleich verwendet werden. Wenn der Data Lake schon eine gewisse Zeit existiert, können auch die historischen Daten entfernt werden. Es muss also keine neue Schnittstelle zum produktiven System gebaut werden. Mit der Sicherheit, dass alle Daten im Data Lake liegen, kann im Data Warehouse nur das eingelesen und verarbeitet werden, was wirklich gebraucht wird. Damit kann das Data

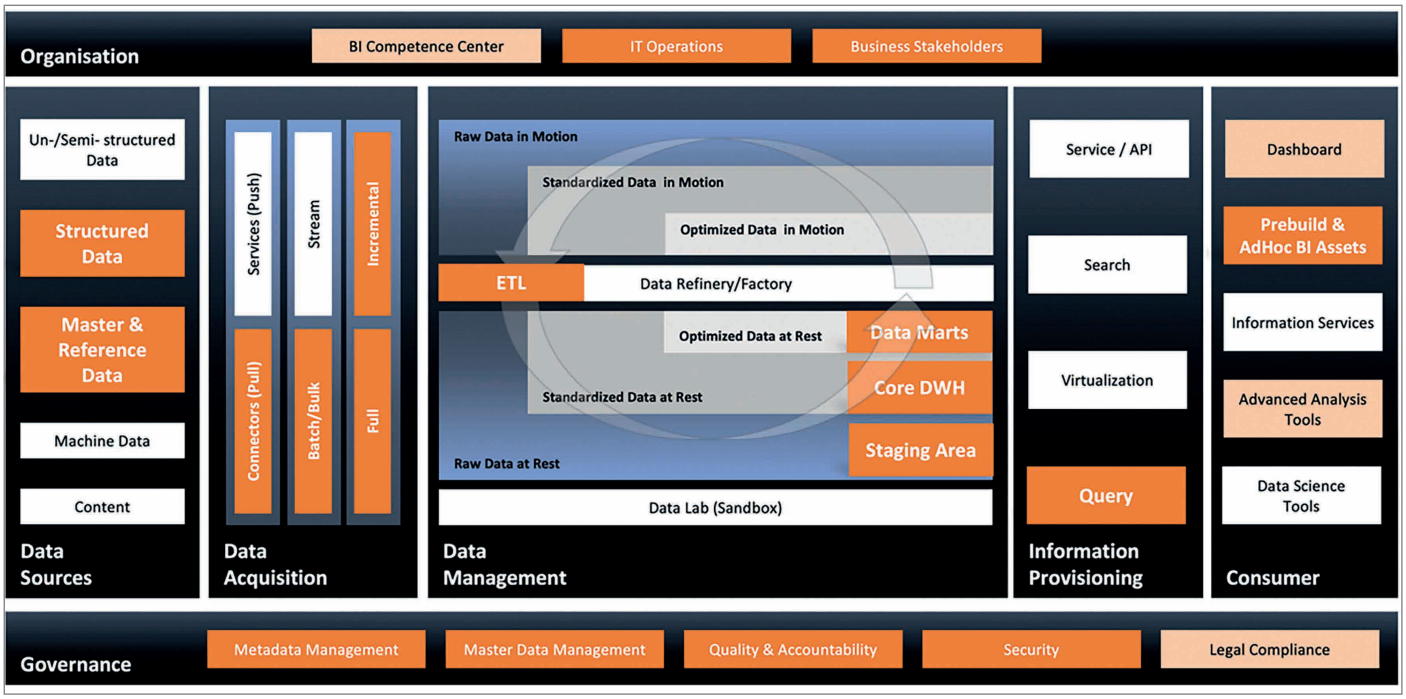


Abbildung 6: Data Warehouse – Design (Quelle: Jan Ott)

Warehouse klein, performant und spezifischer gebaut werden.

Ein Data Lake kann auch für Echtzeit-Daten verwendet werden.

Wieso nicht?

Der wohl wichtigste Grund ist, dass alle mit dem existierenden Data Warehouse zufrieden sind. Die Latenz, Performance und der Inhalt sind gut. Die Akzeptanz in der Firma ist groß.

Zudem ist klar, dass ein Data Lake mehr Aufwand und Kosten nach sich zieht. Auch fehlt im Unternehmen häufig das Wissen zur Data-Lake-Technologie und dies müsste aufgebaut werden. Dabei hat sich gezeigt, dass es nicht einfach damit getan ist, dass das Data-Warehouse-Team einen Kurs besucht. Das Team hat eine gute Grundlage, doch es braucht neues, zum Teil komplett anderes Wissen.

Fazit

Ein Data Lake ist geeignet für

- Data Science
- Daten in Bewegung
- viele Quellen mit großer Datenvielfalt
- die Ergänzung zum Data Warehouse
- Sandboxing

Ein Data Lake ist zusammen mit einem Data Warehouse ein tolles Team. Sie ergänzen sich in vielerlei Hinsicht:

- Der Data Lake ist für das Data Warehouse eine gute Basis für sämtliche Unternehmensdaten.
- Mit einem Data Lake im Hintergrund kann sich das Data Warehouse auf seine Stärken konzentrieren, was ihm ermöglicht, kompakter und performanter zu arbeiten.
- Weitere Data Warehouses können bei ihrer Entwicklung von den bestehenden Schnittstellen zum Data Lake profitieren.
- Prozesse in Data Science, Machine Learning oder Data Analytics belasten die Data Warehouses nicht mehr, weil die Infrastruktur eines Data Lake genau für solche Prozesse optimiert ist.

Einige Firmen bieten da schon Lösungen, die in diese Richtung gehen. Hier ein paar Beispiele:

- Data Lakehouse (Databricks)
- Snowflake/Redshift – S3
- Azure Synapse – Azure Data Lake Store

Jan Ott
Jan.Ott@trivadis.com

Metadaten helfen beim Aufbau einer Plattform für analytische Daten

Alfred Schlaucher, Oracle Deutschland

Was haben gelbe Fahrräder mit Metadaten zu tun? Zunächst nichts. Aber wenn es darum geht herauszufinden, warum eine Fahrradhändlerkette plötzlich insgesamt mehr verkauft, nur weil neonfarbene gelbe Fahrräder zusätzlich im Sortiment sind, benötigt derjenige, der die Ursache herausfinden will, möglichst viele Daten aus dem Einkauf, der Logistik, dem Vertrieb und zusätzlich auch noch Markt- beziehungsweise Trenddaten. Und das möglichst schnell. Denn vielleicht gibt es ja eine Regel, nach der ein hippes, schwarzes Rennrad als „Hingucker" im Schaufensterregal den Verkauf noch mehr ankurbelt? Da zählt jede Woche, die Saison ist begrenzt. Hier können Metadaten helfen. Der Analyst findet über Metadaten geleitet die passenden Daten, ohne viele Personen direkt kontaktieren zu müssen. Die Metadaten verraten ihm auch, in welchem Zustand die Daten sind, und er muss lediglich öffentlich zugängliche Statistikdaten über Trends hinzufügen ***(siehe Abbildung 1)***. Nach wenigen Tagen Analysetätigkeit mit einer Handvoll Geschäftsobjekten und 100 Einzelmerkmalen empfiehlt er, neben den gelben auch noch rote Fahrräder in die Regale zu stellen.

Eine etwas mühsame Geschichte

So ähnlich könnte ein Metadaten-Einsatz aussehen, wenn alles optimal läuft. Die Geschichte der Metadaten ist allerdings eine Abfolge von immer wieder gescheiterten Versuchen, durch zusätzliche (Meta-) Beschreibungen der Komplexität von wachsenden Unternehmensdaten Herr zu werden. Dabei klangen die Versprechungen immer sehr gut: Man legt ein sogenanntes Metadaten-Repository über Unternehmensdaten an, und schon finden alle Mitarbeiter die Daten, die sie benötigen, fast automatisch. Das hat in den wenigsten Fällen funktioniert und wenn, dann nur in überschaubaren technischen Bereichen, aber nicht flächendeckend für alle Unternehmensdaten und Benutzergruppen, die mit Daten arbeiten wollen. Die Ursachen hierfür liegen nicht in der Technologie, ausgereifte Metadaten-Werkzeuge gab es immer. Die Ursachen lagen und liegen in der Art, wie Unternehmen beziehungsweise Mitarbeiter mit Metadaten umgegangen sind beziehungsweise umgehen. „Beschreibungen" und „Dokumentation" waren immer zusätzliche Arbeiten in den Projekten, die Kosten jedoch nicht eingeplant, oder die Metadaten-Dokumentation wurde als Erstes gestrichen, wenn ein Projekt die Budgetgrenzen sprengte. Bereits erfolgte Dokumentation veraltete im Verlauf der Zeit durch viele Änderungen in den Anwendungen.

Aus Management-Sicht wurde schließlich oft nicht verstanden, dass Metadaten-Management ein zusätzlicher Kostenfaktor in den IT-Budgetplanungen ist. Denn man bezahlt nur etwas, wenn man den Nutzen davon sieht. Das ist bei Metadaten allerdings schwer. Bereits in den 1980er Jahren gab es Abhandlungen, die versuchten, den Wert von Metadaten durch eingesparte Suchzeiten der Anwender zu berechnen: „Wenn ein Mitarbeiter pro Tag nur 5 Minuten Zeit bei der Suche nach Informationen spart, macht das bei 1000 Mitarbeitern und 200 Arbeitstagen im Jahr…". Solche Rechnungen waren für die meisten dann doch zu abstrakt und haben nicht überzeugt. Metadaten-Initiativen scheiterten oft an der fehlenden Unterstützung im Unternehmen.

Dabei hatten es die ersten Dictionaries (so hießen Metadaten-Repositories in den 1980er und 1990er Jahren noch) in der ursprünglich zentralisierten Großrechner-IT recht einfach, denn die zu beschreibenden Daten lagen oft nur in einem einzigen System und an einem physischen Ort. Aus dieser Zeit sind erfolgreiche Metadaten-Einsätze bekannt. Die Datenfelder aller Datenbanken, Cobol- und Assemblerprogramme waren in großen Unternehmen in einem Dictionary zentral zu finden. Per Knopfdruck konnte man spontan Datenstrukturen und Zusammenhänge abrufen, was sich etwa bei der PLZ-Umstellung von 4 auf 5 Stellen und beim Jahrtausendwechsel positiv bemerkbar machte. Das änderte sich, als in den 1990ern die Client-Server-Architekturen mehr und mehr die Großrechner ablösten. Es gab nicht mehr nur ein Metadaten-Repository, sondern jede Anwendung hatte ihre eigene Metadaten-Verwaltung. Die Suche über Metadaten endete dann an den Grenzen der Anwendung. Das war das Ende einer zentral organisierten Metadaten-Verwaltung zu Unternehmensdaten.

Wiedergeburt der Metadaten?

Aktuell erleben wir in der IT zwei tiefgreifenden Veränderungen, die den Einsatz von Metadaten befördern können:

- Unternehmen ersetzen zunehmend die eigene IT durch Cloud-Lösungen.
- Ein neues Bewusstsein für den Umgang mit Daten ist entstanden.

Für einen potenziellen Metadaten-Einsatz bedeutet dies:

1. In Cloud-Umgebungen lassen sich Metadaten leichter verwalten, denn Datenbestände liegen nicht in verteilter Hardware, sondern in einer Software-gesteuerten, virtuellen Umgebung. Die Aktualisierung von Metadaten lässt sich technisch leichter automatisieren. Das Architekturmodell ähnelt wieder der alten Großrechnerwelt; auch wenn eine Cloud-Infrastruktur sich über mehrere Rechenzentren rund um den Globus verteilt, sind die virtuellen Wege kurz.
2. Viel diskutierte Konzepte wie Data Lake und Data Mesh zeigen ein neues Datenbewusstsein in den Unternehmen. Die Idee des „Data as a Product" funktioniert nur, wenn Daten umfangreich beschrieben werden. Metadaten sind schon fast ein Muss (!) bei der Beschreibung von Datenobjekten. Metadaten helfen beim Auffinden und Verstehen von Datenobjekten in einem Data Warehouse, einem Data Lake, einem Data-Fabric-Konstrukt oder Data-Mesh-Datenraum. Ein Data Catalog Tool wird bei den aktuellen Architektu-

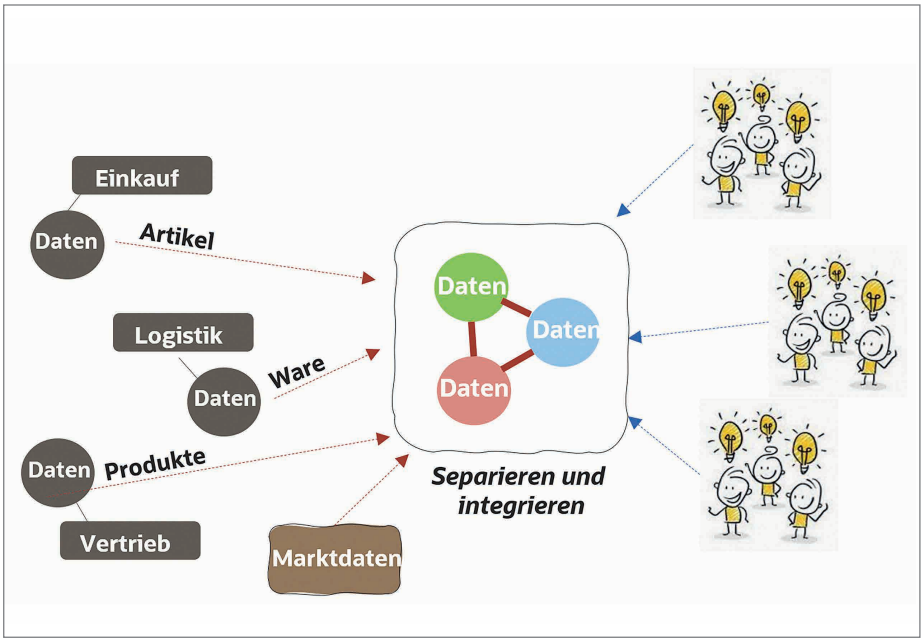


Abbildung 1: Verteilte Daten müssen gefunden und integriert werden, um sie nutzbar zu machen (Quelle: Alfred Schlaucher)

ren von Datenplattformen in der Regel miteingeplant.

Plattformen für analytische Daten und Metadaten

Die Architektur der Plattformen zur Verwaltung von analytischen Daten hat sich in den vergangenen Jahrzehnten in Etappen weiterentwickelt. Zu dem bereits seit den 1990er Jahren etablierten zentralen und unternehmensweit aufgestellten Data Warehouse hat sich in den 2010er Jahren der Data Lake als eine universelle, auch für „spontane" und Originaldaten geeignete Datenablage gesellt. Während beide Konzepte noch durch eine zentrale IT verwaltet und entwickelt werden, entstanden in jüngster Zeit Data-Fabric- und Data-Mesh-Architekturen mit dem Ziel, Aufbau und Verwaltung der Datenobjekte näher zu den Anwendern zu bringen. Die Fachabteilungen selbst kennen ihren Datenbedarf und ihre Daten besser. Wie meist in der IT, so entstehen auch hier neue Konzepte und Ideen in Wellen und erfahren einen gewissen Hype. Alle vier heute anerkannten Konzepte haben jedoch ihre Vorteile, die man für den Einsatz im eigenen Unternehmen bewerten sollte:

Data Warehouse:

- Leicht zugängliche Daten (i. d. R. SQL)
- Daten für bestimmte Berichte bereits aufbereitet
- HistorischeSichten(ermöglichenTrenderkennung und Planung)
- Bereichsübergreifendes Sichten durch semantischen Integrationsschritt
- Standardisierte Herleitungsverfahren (Nachvollziehbarkeit)

Data Lake:

- Kurzfristig bereitstehende Daten (Fast Data)
- Daten im Originalzustand (Daten offen für noch unbekannte Analysen)
- Breiteres Datenangebot wegen geringeren Aufbereitungsaufwands
- Kostengünstige universelle Ablage

Data Fabric

- Schnellere Datenbereitstellung (Daten müssen nicht bewegt werden)
- Einfache Zugriffe auf Originaldaten (ohne komplexe technische Realisierung)
- Verteilte Lagerung
- Wegfall von Datentransporten

Data Mesh

- Verantwortlichkeit für Daten besser geregelt
- Qualität der Daten
- Breiteres Datenangebot durch Daten-als-Produkt-Denken

Bei allen vier Konzepten stellt man Daten einer breiteren Benutzergruppe zur Verfügung und bei allen Konzepten stellen sich automatisch zum Beispiel die folgenden Fragen:

- Welche Daten stehen zur Verfügung?
- Passen die Daten zu meinen Aufgaben?
- Wo liegen diese Daten?
- Bin ich berechtigt, die Daten zu lesen?
- Wie kann man auf die Daten zugreifen?
- Sind die Daten aktuell?
- Sind die Daten vollständig?
- ...

Solche Fragen beantworten Metadaten. Metadaten beschreiben Unternehmensdaten so umfänglich, dass die jeweilige Zielgruppe als Datenkonsument die Daten in einem Data Warehouse, Daten in einem Data Lake, Daten als remote liegendes Datenobjekt (Data Fabric) oder Daten als Datenprodukt („Data As A Product" – Data Mesh) nicht nur auffindet, sondern auch erkennen kann, ob die gefundenen Daten den Bedürfnissen entsprechen.

Der Vorgang klingt einfacher, als er sich in der Praxis darstellt. Wir kennen aus der Linguistik das Phänomen des „semiotischen Dreiecks", wonach die reine Benennung eines realen Objektes (Sachverhalt, Geschäftsobjekt) bei verschiedenen Personen unterschiedliche Assoziationen (Bedeutungen und damit Inhalt) hervorrufen *(siehe Abbildung 2)*.

Das verhält sich auch mit Datenobjekten in einem Unternehmen so. Die Mitarbeiter in unterschiedlichen Abteilungen sind in unterschiedliche Prozesse eingebunden und mit teilweise komplett unterschiedlichen Aufgaben betraut. Geschäftsobjekte sind für verschiedene Mitarbeiter aus unterschiedlichen Blickwinkeln heraus interessant. Die Einkaufsabteilung des Fahrradhändlers interessiert günstigste Einkaufspreise, die Logistikabteilung interessiert die Größenmasse eines Fahrrades, um Lager und Transportkosten zu sparen, die Vertriebsabteilung interessiert eher, ob sich rote Fahrräder besser verkaufen lassen als gelbe. Und die Controlling-Abteilung interessieren Kosten und Erlöse über den gesamten Vertriebsprozess. Das Beispiel klingt einfach, aber es ist auf viele Situationen in Unternehmen im Umgang mit Daten übertragbar.

Abbildung 2: Jeder versteht unter einem Objektnamen etwas anderes (Quelle: Alfred Schlaucher)

Metadaten erfüllen folgende Aufgaben:

- Sie müssen Daten (Geschäftsobjekte) aus verschiedenen Perspektiven heraus beschreiben, sodass möglichst alle Nutzergruppen den für sie relevanten Inhalt wiederfinden.
- Verschiedenste Suchstrategien müssen möglich sein, die Abfrage eines einzelnen Namens reicht nicht aus: Zugang über synonyme Begriffe, Assoziationen, verwandte Begriffe, Kategorie-Begriffe, Suchen über Hierarchien, Strukturauflösungen usw.
- Eignung der Daten aus Datenqualitätsgesichtspunkten heraus. Das sind unter anderem Korrektheit, Stimmigkeit, Vollständigkeit, Aktualität, Verfügbarkeit.
- Fundstellen und technische Zugriffsmöglichkeiten (im einfachsten Fall ein Link)
- Formatbeschreibung, technische Merkmale wie Größe, Dateiart etc.
- Security-Richtlinien

Semantische Integration durch Metadaten

Data Warehouse, Data Lake, Data Fabric und Data Mesh haben die gemeinsame Herausforderung, Daten aus unterschiedlichen Sachzusammenhängen oder Unternehmensprozessen zu integrieren. Aber wie entsteht ein integriertes Geschäftsobjekt, wenn seine Daten in unterschiedlichen Prozessen unterschiedlich gehandhabt werden? Ein Fahrrad wird in einer Vertriebsabteilung modern klingend als „Produkt mit vielen Vorteilen" gesehen und als solches im Vertriebs-IT-System gelistet. Das gleiche Fahrrad ist aber auch im Lagerverwaltungssystem der Logistikabteilung als schlichter „Artikel" mit Maßen wie „Gewicht", „Anzahl pro Palette" usw. enthalten. Die Einkaufsabteilung kooperiert täglich mit einer Vielzahl von Lieferanten, die teilweise komplette Fahrräder oder auch nur Fahrradteile liefern. Im Einkaufsverwaltungssystem nennt man das, was eingekauft wird, daher einfach nur „Ware". Will die Controlling-Abteilung jetzt eine übergreifende Ertragsrechnung pro verkauftes Fahrrad durchführen, bedarf es Daten aus den drei Systemen, und das am besten an einer einzigen Stelle und in einer einzigen Tabelle oder Datei mit Verkaufserlös, Vertriebskosten, Lager- und Transportkosten sowie Einkaufswert.

Dieses simple Fahrradbeispiel ist nur ein einfaches Erklärungsmuster, um die Problematik zu verdeutlichen, die bei nahezu allen Geschäftsobjekten in einem Unternehmen auftritt. Es lässt sich übertragen. Das Beispiel wird schnell komplexer, wenn wir den Data Scientist hinzunehmen, der ein Markttrend-Projekt zu der Frage durchführt: „Warum verdient ein Konkurrenzunternehmen mit schwarzen Rennrädern und Neonfarben-E-Bikes im Sortiment mehr pro Einheit?" Der Data Scientist ist möglicherweise nicht mit Fragen des Vertriebs, des Einkaufs oder der Logistik vertraut. Er soll aber dennoch alle relevanten Daten aus diesen Geschäftsprozessen finden und zusammentragen, um neue Erkenntnisse zu erlangen.

Hier helfen Metadaten. In Metadaten führt man alle Aspekte rund um ein Geschäftsobjekt („das Verkaufsprodukt Fahrrad") zusammen. Wenn Fahrräder als „Ware", „Artikel" und „Produkte" in verschiedenen Systemen vorkommen, dann kreiert man ein Metadaten-Objekt mit beliebigem Namen „XY4711" mit Synonymen „Ware", Artikel" und „Produkt". Will man zusätzliche marktbezogene Aspekte ausdrücken, dann definiert man ein übergreifendes „Superobjekt", zum Beispiel „Freizeitsportgerät", das alle Fahrrad-Varianten von Rennrad bis zu E-Bikes auflistet, und schließlich noch ein weiteres „Superobjekt" mit Trends des Jahres, beispielsweise mit Farben. Ein Metadaten-Repository verbindet alle Objekte zum Beispiel über sogenannte „Links" oder über schlichte Attribute eines Metadaten-Eintrags oder eines Business-Terms. Dies geschieht auf einer abstrakten Ebene, eben in dem Meta-Layer innerhalb des Metadaten-Repository. Man beschreibt Beziehungen, die so in der Realität nicht zu sehen sind. Metadaten integrieren die Dinge und Sachverhalte der realen Welt.

Umsetzung einer Metadaten-Verwaltung der Cloud

In modernen Cloud-Umgebungen wie Oracle Cloud Infrastructure (OCI) gibt es für diese Aufgabenstellung seit 2020 den neu entwickelten Data Catalog als kostenfreien Service. Man kann an diesem Service eine ideale Umsetzung der oben beschriebenen Metadaten-Anforderungen exemplarisch ablesen. Zunächst ist dieser Service ein vollständig nativer Cloud-Service, also völlig eingebettet in die Cloud-Infrastruktur *(siehe Abbildung 3a und 3b)*, das heißt:

- Er ist vollständig mit Cloud-Mitteln realisiert und benötigt nur im Hintergrund virtuell bereitgestellte und automatisch skalierende Ressourcen.
- Der Service ist in das Security- beziehungsweise Policy-System (IAM) eingebettet. Er hat also kein eigenes Security-System sowie andere Repository-Tools, sondern nutzt die Cloud-User-Berechtigungen.
- Der Service ist mit den gleichen (Python- / CLI-) API-Mitteln programmier- und automatisierbar wie alle Services der OCI (z. B. Data Science Service), wenn es denn sein muss.
- Oberfläche und Online-Bedienung sind die des OCI GUI.
- Aus anderen Services (z. B. Data Science Service) kann man Metadaten mit Cloud-Mitteln abrufen.
- Zugriffe auf Datenobjekte zum Sammeln von technischen Metadaten, ob Data Lake (Object- / S3-Storage) oder Data-Warehouse-Datenbank, werden teils automatisch bereitgestellt.

Fachbegriffe abbilden

Physikalische Datenobjekte (Dateien, Tabellen) sollten in einem gepflegten Datenhaushalt nur einmalig vorhanden sein. Wie oben beschrieben, haben verschiedene Benutzergruppen allerdings verschiedene Sichten auf dasselbe Objekt. Diese Sichten spiegeln sich in der Sprache wider. Jede Abteilung, jedes Sachgebiet nutzt eigene Fachbegriffe. Diese Fachbegriffe greift man etwa mit einem Glossar ab. Ein Glossar sammelt und beschreibt zum einen die wichtigsten Fachbegriffe, gleichzeitig gruppiert das Glossar die Begriffe über Kategorien. Man erhält also eine sortierte und kategorisierte (geordnete) Übersicht über alle relevanten Fachbegriffe und damit eine Beschreibung der Fachlichkeit in einem Geschäftsprozess. Der OCI Data Catalog Service nimmt ein solches Glossar komplett auf. Es ist die Grundlage für die weitere Beschreibung der technischen Objekte.

Weil es mehrere Fachabteilungen beziehungsweise Geschäftsprozesse gibt, kann man auch mehrere Glossare anle-

gen. Und weil einzelne Begriffe (Terms) in unterschiedlichen Glossaren vorkommen und dort jeweils anders definiert werden (weil die Fachabteilungen Begriffe unterschiedlich deuten), gibt es auch Links zwischen Begriffen in verschiedenen Glossaren. Das heißt, jede Fachabteilung (Daten-Owner) definiert einen Begriff zunächst für sich. Dass es diesen Begriff als Variante aber noch woanders gibt, wird durch einen Link im Glossar des Data Catalog ausgedrückt.

Fachbegriffe mit physischen Daten verknüpfen

Reale Datenobjekte (Dateien, Tabellen, Streams) kann man jetzt mit den Begriffen (Terms) der Glossare verknüpfen. Mehrere Terms zeigen auf ein oder mehrere Datenobjekte (n:n- Beziehung). Die Einträge der Datenobjekte in den Data Catalog entstehen automatisch durch einen Harwesting-Prozess. Dieser scannt regelmäßig Datenbanken (Data Warehouse), Data Lake (Object Storage) oder andere Datenablagen (auch On-Premises), um die reale Datenwelt mit dem Metadaten-Repository synchron zu halten. Die Verbindungen zwischen den Begriffen der Glossare und Datenobjekten erfolgt entweder manuell durch den Datenadministrator (oder Data Owner), durch automatisierte Link-Läufe oder mithilfe von Machine Learning – basierten Vorschlägen des Data-Catalog-Systems aufgrund von Ähnlichkeiten zwischen Objektnamen und Glossarbegriffen. Dieser Vorgang kann neben dem Erfassen von Fachbegriffen der Glossare mit der aufwendigste Teil der Metadaten-Pflege sein. Die investierte Arbeit lohnt sich allerdings später, weil die Leistungsfähigkeit der Metadaten gesteigert wird. In einem Data-Mesh-Szenario wird diese Aufgabe durch das Data-Owner-Team übernommen.

Spezifische Beschreibungsmöglichkeiten

Über die Glossare ist bereits ein Suchzugang für Fachanwender auf Datenobjekte ermöglicht. Einfacher zu realisierende Suchzugänge stellen die benutzerdefinierten Attribute (Custom Properties) dar. Man kann damit ein Datenobjekt über seine physikalischen Eigenschaften hinaus beliebig auch fachlich beschreiben. Hier eine Vorschlagsliste für mögliche Beschreibungen (Custom Properties) in einem Einsatzszenario:

Custom Property	Erklärung, mögliche Werte
Domain	Fachbereich, Sachgebiet [abc...]
Sub Domain	Teilgebiet [abc...]
Ownership	Verantwortlicher Fachbereich, Daten-Team
Beschreibung	Allgemeine sprachliche Beschreibung
Technische Struktur	CSV, Parquet, Table, Text, Key Value, Graph
Zugriffs-Variante	SQL, Data-Fabric-Zugang, Access Tool, Python-Code
Beispieldaten	Exemplarische Daten [abc...234…&%$$]
Qualitätslevel	Vereinbarung zwischen Datenerzeuger und Datenkonsumenten, z. B. „Originalzustand", „enthält Doubletten", „doppelte Einträge entfernt", „fehlende Werte ergänzt", „Glaubwürdigkeits-Check gemacht", „realistische Werte", „Umrechnungsfaktor",…
Quelle / Quell-Domain	Hier wird der Ursprung, die Herkunft, die Quelle des Objektes angegeben. Dieser Ursprung muss nicht unbedingt mit der technischen Quelle übereinstimmen.
Cloud-Event	Protokoll, Version, Python-Zugriffsbeispiel
Identifier-Feld	Angabe, über welche Spalte zugegriffen wird
Federated-Entity	Referenz-Objekt für übergreifende Beziehungen
Bearbeitet am	Bearbeitungsdatum
Freigabestatus	Freigabestempel, in Arbeit, geprüft, abgenommen, freigegeben
Security Level	Öffentlich, restricted

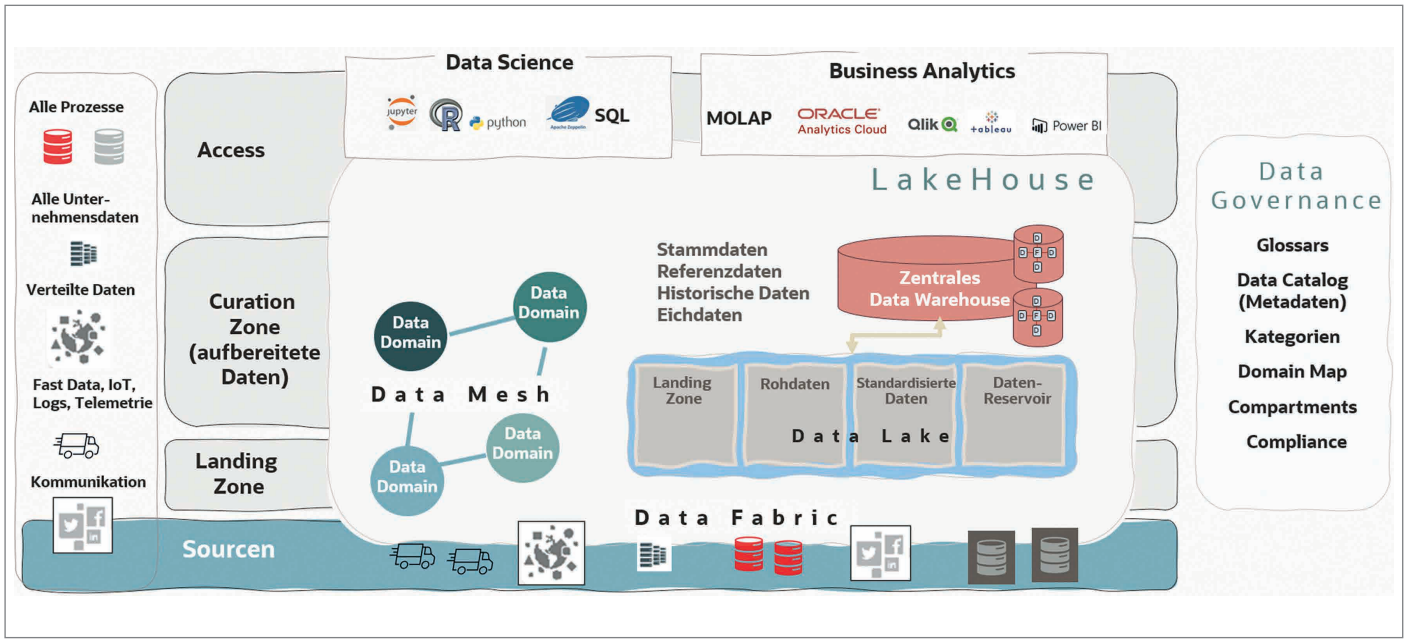


Abbildung 3a: Funktionsbereiche einer Plattform für analytische Daten am Beispiel Oracle Cloud (OCI) (Quelle: Alfred Schlaucher)

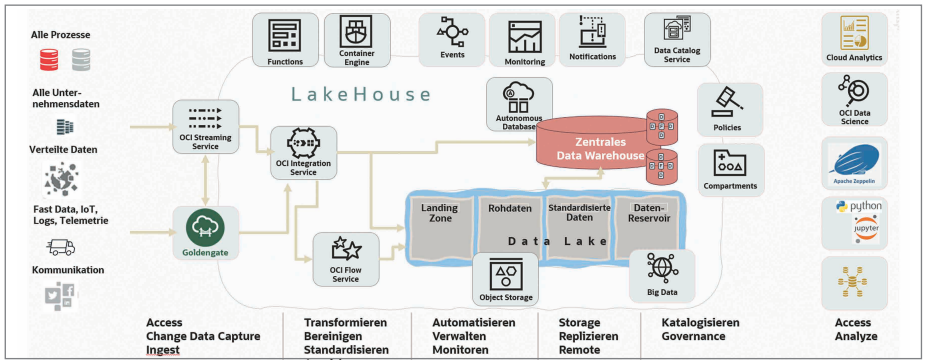


Abbildung 3b: Komponenten (Services) einer Plattform für analytische Daten am Beispiel Oracle Cloud (OCI) (Quelle: Alfred Schlaucher)

Die potenziellen Inhalte der Properties sind formatierbar. Auch Selektionslisten können vorgegeben werden, damit die Inhalte möglichst standardisiert und nicht wahlfrei erfasst werden. Möglich ist auch eine automatische Pflege der Properties durch Programmierung, etwa über das Python Cloud API oder automatisiert im Verlauf des Harwesting-Laufs.

Eine zusätzliche und völlig freie Beschreibungsvariante sind die Tags. Hier hinterlegt man einen oder mehrere Begriffe, die später die Suche der Anwender unterstützen.

Deskriptoren-Methode

Die etwas aufwendigere, aber sehr effiziente Deskriptoren-Methode hilft, wenn ein noch feingliedrigeres Beschreibungsverfahren benötigt wird. Will man ein Datenobjekt bezüglich eines Aspektes beschreiben, dann definiert man etwa als Werte für eine Custom Property eine Reihe von fixen Kategorien, mit denen man den zu beschreibenden Aspekt gliedert. Zu jeder Kategorie kann man schließlich noch eine Reihe von fixen Unterkategorien festlegen. Theoretisch sind Datenobjekte damit über Hunderte von speziellen Fachbegriffen beschreibbar.

Anwender suchen und finden passende Daten

Mit diesem Beschreibungs-Set sind Datenobjekte, ob im Data Warehouse, im Data Lake, in einem Data-Fabric-System oder in einem nach Data-Mesh-Vorstellungen organisierten Pool von Datenprodukten, universell und detailgenau beschreibbar. Anwendern, die Datenobjekte nutzen wollen, stehen jetzt unterschiedliche Suchstrategien offen. Ein Anwender, der Daten sucht, hat meist nur einzelne Begriffe in seinem Fokus, über die er suchen will. Er gibt diese Begriffe vollständig oder auch mit Asterisk-Stern gekürzt, in ein Suchfeld ein und der Data Catalog sucht über

- Namen der Objekte
- Business Names der Objekte
- Links zu den Glossar Terms
- Einträge in den Properties
- Tags

Das Suchergebnis ist eine Liste möglicher Datentreffer, in der Regel verbunden mit einem Link für den Zugriff auf das Datenobjekt, zum Beispiel in einem Data Lake.

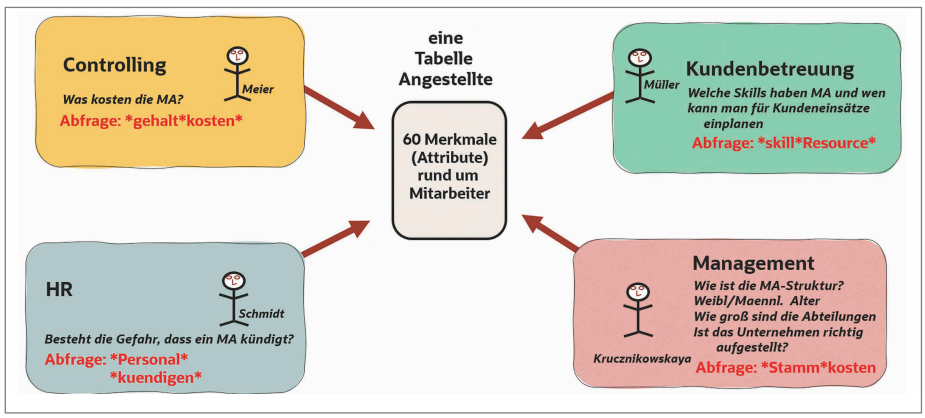


Abbildung 4: Jeder sucht mit seinen Begriffen. Gefunden wird immer dasselbe Objekt (Quelle: Alfred Schlaucher)

Damit ist das Ziel erreicht: Mitarbeiter können aus ihrer jeweiligen individuellen Perspektive heraus nach Datenobjekten suchen, ohne dass sie über Speicherort, IT-System oder den Namen des Suchobjektes Bescheid wissen müssen. Ein letztes Beispiel verdeutlicht diese Praxis:

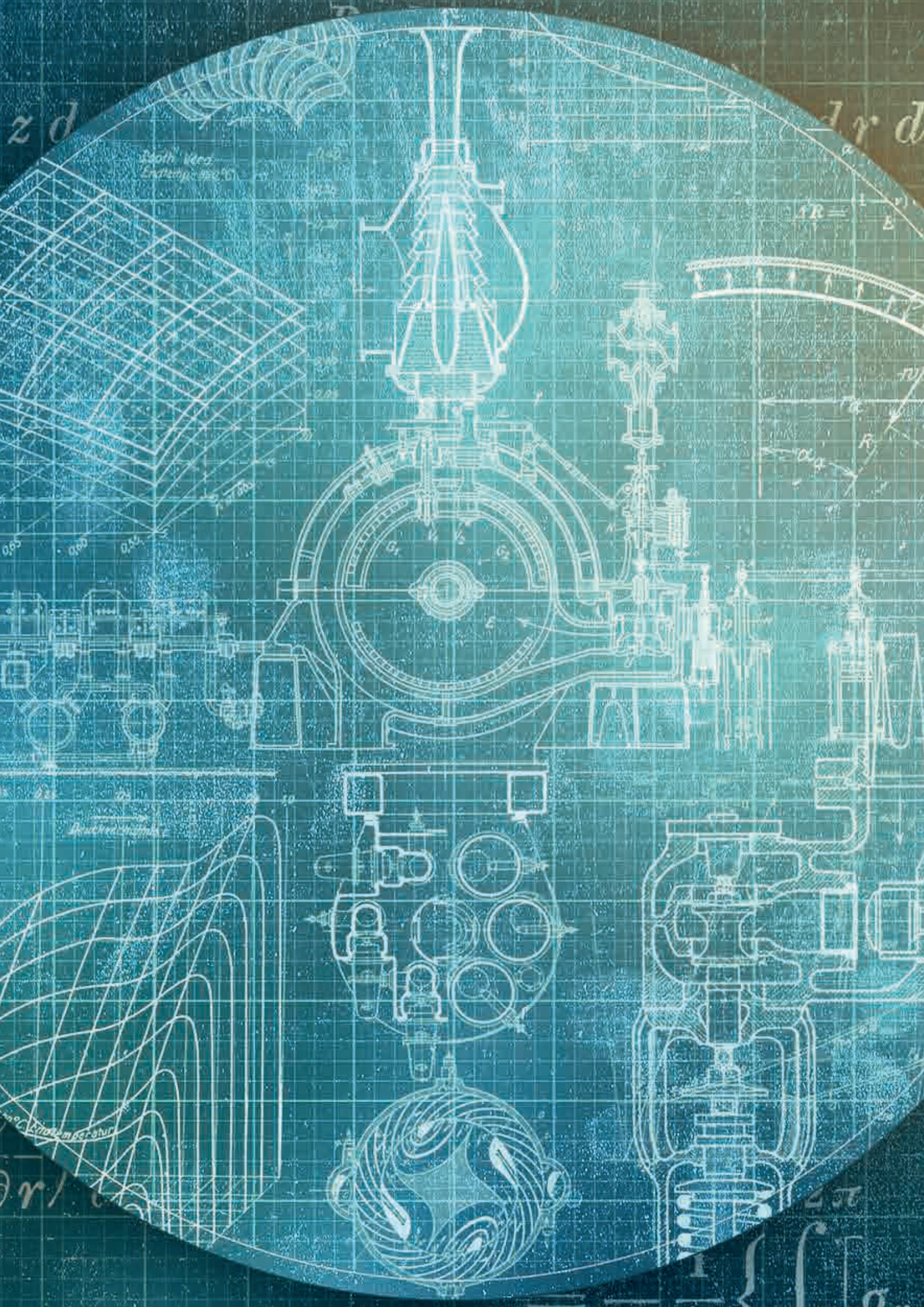
Ein Unternehmen ist ständig bemüht, die Mitarbeiter-Ressourcen so optimal wie möglich aufzustellen. Mitarbeiter sind zum einen ein Kostenfaktor, zum anderen aber auch Leistungsträger:

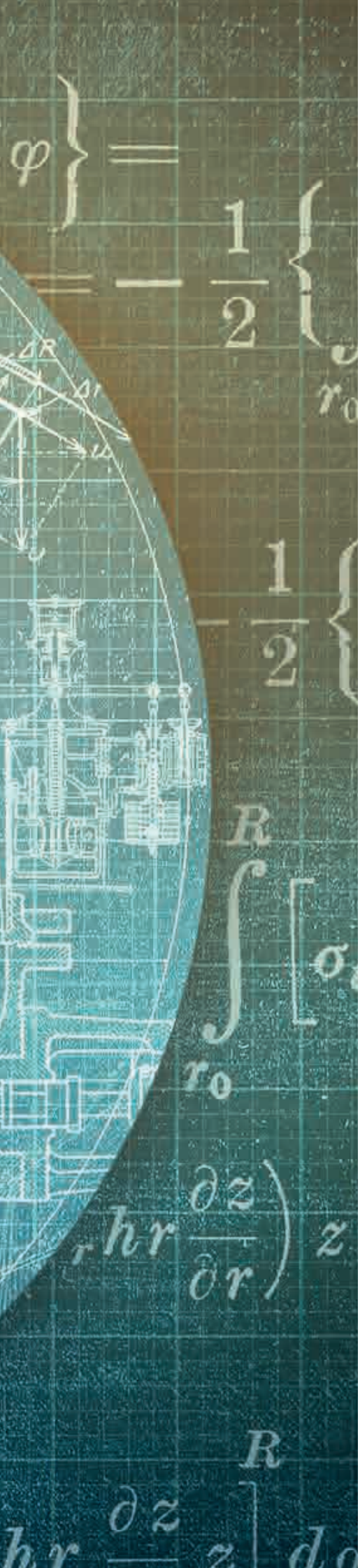
Der Vertrieb sucht für Kundenprojekte Mitarbeiter mit den passenden Skills und ist an einer Liste mit detaillierten Erfahrungen interessiert (Suchwort: *skill*). Die Controlling-Abteilung ist neben einer Gehaltstaufstellung auch noch an zusätzlichen Mitarbeiterkosten interessiert (Suchwort: *gehalt* *kosten*). Das Management sucht Gesamtaufstellungen zu den Mitarbeiterkosten, um sie mit ihrer Budgetplanung in Einklang zu bringen (Suchwort: *Mitarbeiter*kosten*). Die Personalabteilung sucht Faktoren, die Mitarbeiter dazu bewegen, entweder zu kündigen oder bei dem Unternehmen zu bleiben (Suchwort: *kuendigen*). Alle Suchvorhaben treffen auf eine einzige Mitarbeiterliste in einem Data Lake, in der neben dem Gehalt auch 30 weitere Mitarbeitermerkmale zur Art der Beschäftigung enthalten sind *(siehe Abbildung 4)*.

Metadaten unbedingt mit einplanen

Wer aktuell ein neues Projekt zu Data Mesh, Data Lake oder Data Warehouse startet, sollte unbedingt die Chance nutzen, Metadaten mit einzuplanen. Einfacher kann der Start nicht sein, als es gleich zu Beginn zu versuchen. Zum OCI Data Catalog Service können vom Autor detaillierte Konzepte, Spezifikationen und Python-Routinen kostenfrei bezogen werden.

Alfred Schlaucher
alfred.schlaucher@oracle.com





Machine Learning as a Service zur Optimierung von Produktionsprozessen

Oliver Fuhrmann und Germans Hirsch, Trevisto

Komplexe Produktionsprozesse unterliegen einer Vielzahl von Einflussfaktoren. Die in einer Produktion anfallenden Daten sind mannigfaltig und von Menschen alleine nicht zu durchdringen. Enorme Ressourcen und Zeit werden darauf verwendet, um in industriellen Prozessen die Qualität zu verbessern oder Kosten zu minimieren. Die Produktionsprozesse besser zu verstehen, war daher schon immer ein Anliegen von Produktionsleitern und Qualitätsmanagern. Die Herausforderung besteht darin, dass viel Fachwissen, Erfahrung sowie Experten aus unterschiedlichen Bereichen erforderlich sind, um wichtige Erkenntnisse zusammenzutragen. Heute, mit viel leistungsfähigerer Hardware, größeren zur Verfügung stehenden Datenmengen und besseren Algorithmen, schreitet Machine Learning weiter vor und vereinfacht nun die Suche nach Antworten auch in Bereichen, die zuvor als „undurchschaubar" galten. In diesem Beitrag wird anhand eines konkreten Beispiels aufgezeigt, wie künstliche Intelligenz kombiniert mit aussagekräftigen Daten Mehrwert in der Optimierung von Produktionsprozessen erzeugen kann.

KI-Potenziale in der Produktion erkennen und heben

Im Siemens-Gerätewerk in Amberg wollen die Projektpartner Fraunhofer, Siemens und Trevisto durch den Einsatz von künstlicher Intelligenz (KI) mehr Transparenz in den Produktionsprozessen erzeugen und Steuerungsbedarfe sowie Verbesserungspotenziale identifizieren. Dabei soll eine automatische Aufnahme und Analyse von Prozessmodellen durch den Einsatz von künstlicher Intelligenz (KI) erfolgen. Prozessschritte werden in Echtzeit identifiziert und es werden produktionsrelevante Ereignisse abgeleitet – durch die Daten der in der Smart Factory existierenden Technologien. Am Standort Cham fertigt Siemens die Produktreihen Leistungsschalter und Schütze unter den bekannten Marken Sirius und Sentron. Die Durchführung des Projekts wird gefördert im Rahmen des Programms Informations- und Kommunikationstechnologie durch das bayerische Staatsministerium für Wirtschaft, Landesentwicklung und Energie.

Um das Potenzial zu heben, wurden die Produktionsprozesse der Leistungsschalter zunächst fachlich und später unter Zuhilfenahme der auf künstlicher Intelligenz basierenden Methoden

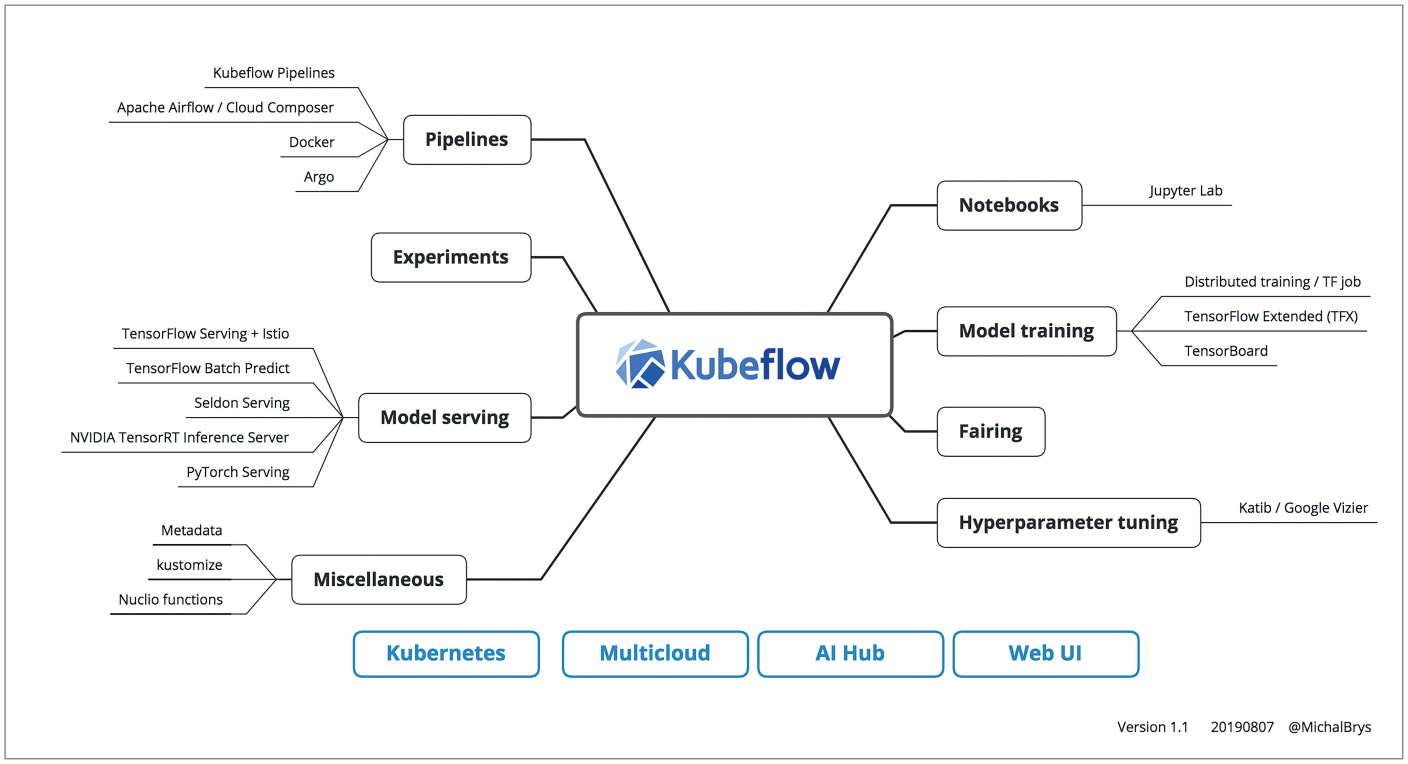


Abbildung 1: Kubeflow Toolkit ermöglicht eine Ende-zu-Ende-Modellentwicklung durch die bereitgestellten Module. (Quelle: https://medium.com/@michal.brys/kubeflow-a-machine-learning-toolkit-for-kubernetes-d8686f6c91b6)

analysiert. Ein Leistungsschalter ist ein mechanischer Schutzschalter, der Überlast- und Kurzschlussschäden an elektronischen Komponenten verhindert. Eine Herausforderung bei der Produktion der Leistungsschalter ist, dass diese aus einer Großzahl von Bauteilen bestehen, die unterschiedliche Toleranzen haben. Diese Toleranzen wirken sich auf eine komplexe Art und Weise aus, sodass es kaum möglich ist, Aussagen über das Auslöseverhalten durch einfache Gleichungen zu modellieren. Um dem entgegenzuwirken, werden im Laufe des Produktionsprozesses Messungen durchgeführt und Daten gesammelt, um den Leistungsschalter am Ende zu justieren. Die abschließende Prüfung ist dabei die zeitaufwendigste und sagt am Ende aus, ob die Justierungen korrigiert werden sollten oder nicht.

Experten haben Hypothesen aufgestellt, wie die Justierung genau eingestellt werden kann. Es wurden Daten zu den einzelnen Leistungsschaltern erfasst, sodass am Ende ein qualitativ gutes Dataset entstanden ist, das sich für Machine Learning eignet. Künstliche Intelligenz soll an genau dieser Stelle anknüpfen und nach möglichen Mustern, Interaktionen und Besonderheiten in den Daten suchen mit dem Ziel, die Ausschüsse, Nacharbeiten und Prüfkosten zu reduzieren.

Das eingesetzte Machine Learning Toolset

Zur Lösung der gestellten Aufgabe und zum Erreichen der gesetzten Ziele kamen folgende Tools zum Einsatz:

Kubernetes: Ist ein Open-Source-System zur Automatisierung der Bereitstellung, Skalierung und Verwaltung von containerisierten Anwendungen. Es ist die Basis für alle weiteren Bausteine und somit ein Kernelement des Trevisto-KI-Stacks.

Kubeflow: Ein Open Source Machine Learning Toolkit für Kubernetes, mit dem der komplette Workflow von Datenanbindung bis Modellbereitstellung umgesetzt werden kann. Durch die Transformation des Codes in eine, von Kubeflow voraus-

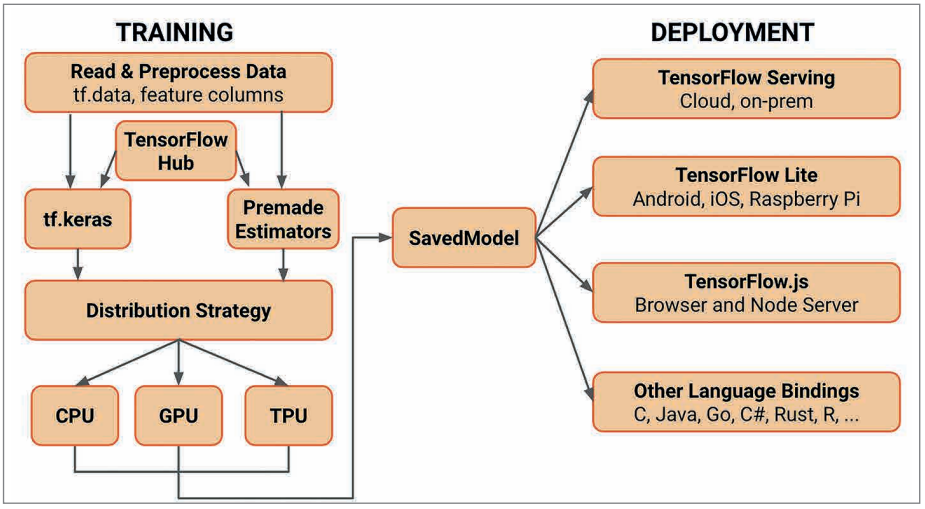


Abbildung 2: TensorFlow Modelltraining und -bereitstellung (Quelle: https://blog.tensorflow.org/2019/01/whats-coming-in-tensorflow-2-0.html)

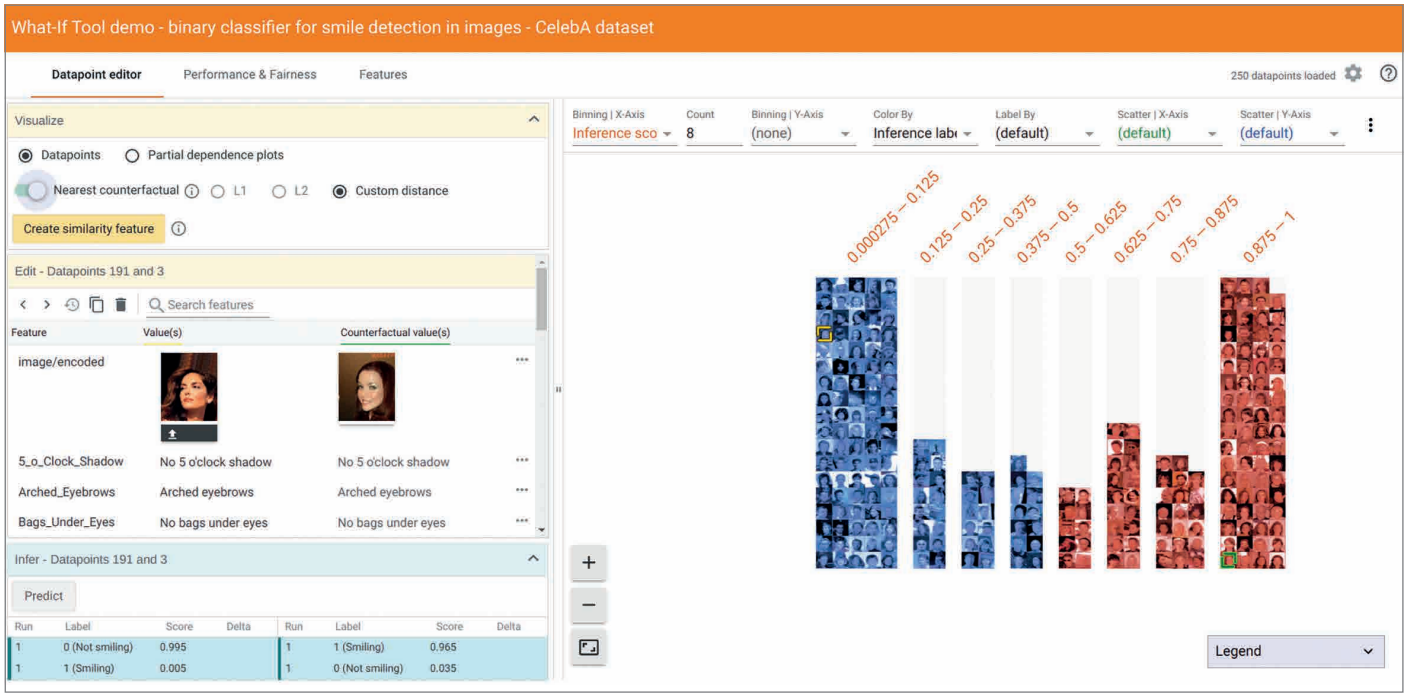


Abbildung 3: What-If Tool Dashboard zur Visualisierung der Daten und zum Testen der Modelle (Quelle: https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html)

gesetzte, Pipelinestruktur kann auf die von Kubeflow bereitgestellte Funktionalität zugegriffen werden. Diese vereinfacht die Wartbarkeit, Entwicklung, Validierung und Bereitstellung der KI-Modelle deutlich *(siehe Abbildung 1)*.

TensorFlow: Eine Ende-zu-Ende-Machine-Learning-Plattform, die zum Erstellen, Trainieren und Validieren der Modelle im Zusammenspiel mit Kubeflow genutzt wird. TensorFlow erlaubt, auf vorhandene Topologien aus dem Hub zuzugreifen, aber auch eigene zu erstellen, dabei können auch beliebige Machine-Learning-Algorithmen wie zum Beispiel neuronale Netze oder Decision Trees verwendet und kombiniert werden. Am Ende des Trainings oder der Validierung können durch TensorFlow die Modelle für spätere Bereitstellung und Nutzung gespeichert werden *(siehe Abbildung 2)*.

What-If-Tool: Zur Analyse des Modells und der Daten setzt Trevisto ein What-If-Tool *(siehe Abbildung 3)* ein. Das What-If-Tool ermöglicht, Hypothesen aufzustellen und diese schnell zu prüfen, die Wichtigkeit der einzelnen Datenspalten zu analysieren und zu bewerten, Modelle einfach gegenüberzustellen und Modellverhalten zu visualisieren.

Python: Die Skriptsprache Python zählt aktuell mit zu den wichtigsten Programmiersprachen und ist auch im Bereich Machine Learning sehr stark vertreten. Die Sprache funktioniert sehr gut mit anderen Tools, sodass sehr häufig die gesamte Entwicklung zum großen Teil in Python durchgeführt wird. Ein weiterer Punkt für Python ist auch die große Auswahl an wissenschaftlichen Bibliotheken und Modulen, auf die immer wieder zurückgegriffen werden kann. Diese vereinfachen das Visualisieren und Analysieren der Daten und der Ergebnisse und vereinfachen auch das Prüfen der Modelle durch eine Großzahl an bereitgestellten Metriken und Verfahren.

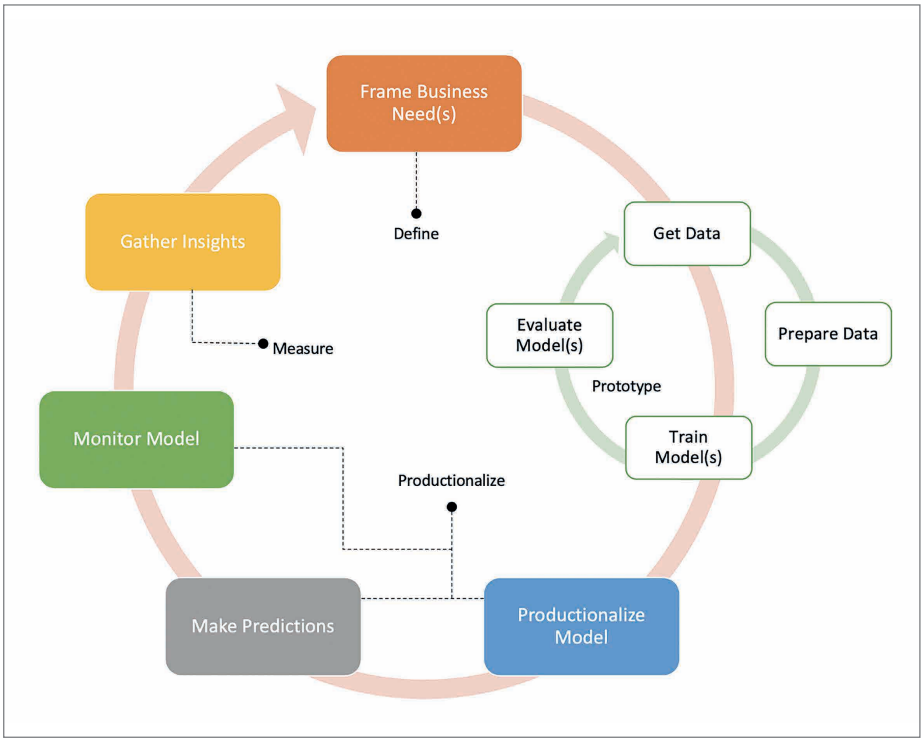


Abbildung 4: Vorgehensweise bei Machine Learning as a Service (Quelle: Oliver Fuhrmann und Germans Hirsch)

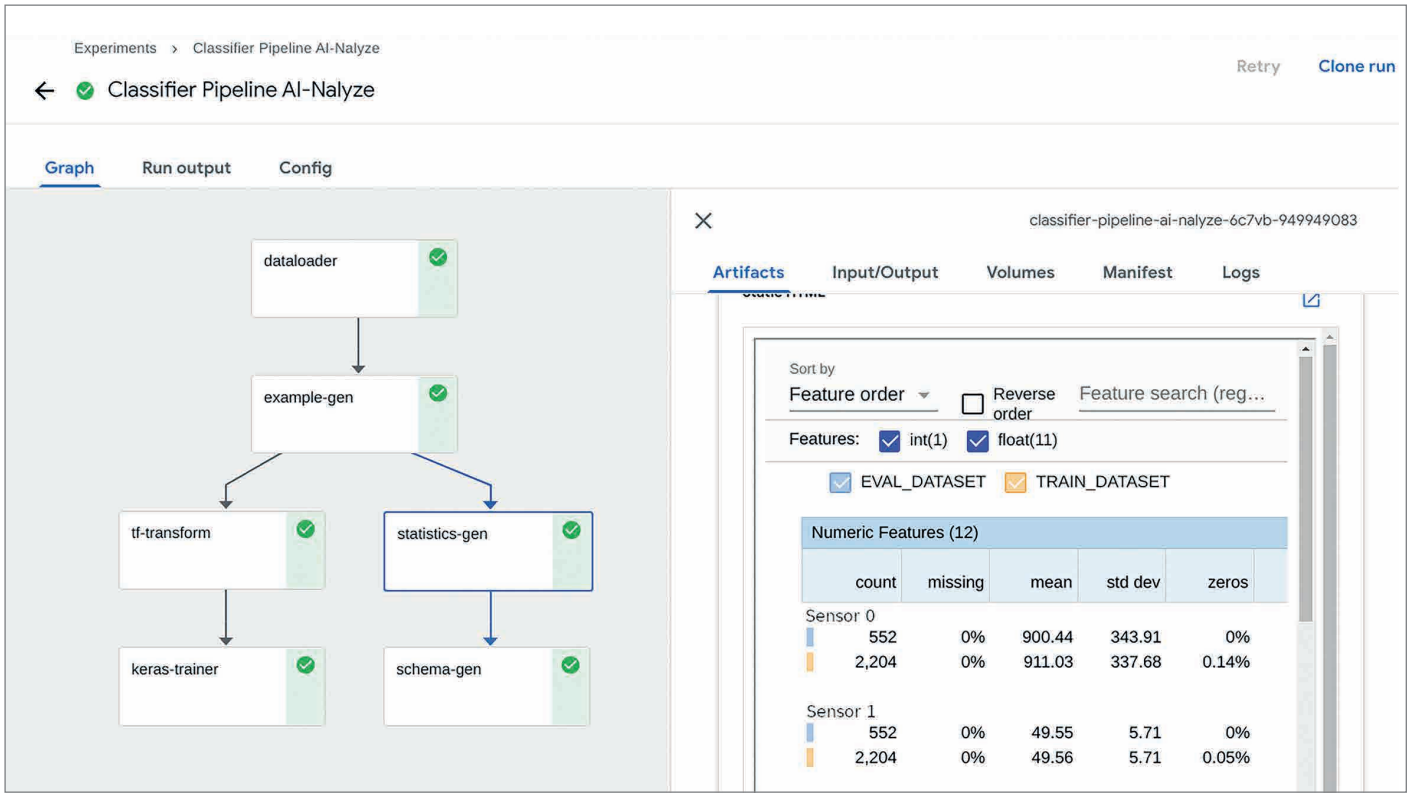


Abbildung 5: Kubeflow Machine Learning Pipeline mit bereitgestelltem Artefakt (Quelle: Oliver Fuhrmann und Germans Hirsch)

Ein bewährtes Vorgehen beim Machine Learning erleichtert den Einstieg

Machine Learning as a Service setzt ein bewährtes Vorgehen voraus. Ein solches Machine-Learning- Vorgehen besteht aus mehreren, sich wiederholenden Arbeitsschritten *(siehe Abbildung 4)*. Jede Iteration führt dazu, dass die Wünsche des Kunden immer im Fokus bleiben und dass die Entwicklung des Modells oder des Produkts weiterhin optimal läuft.

Der äußere Kreislauf bildet das Projekt ab und lässt sich in vier Bereiche unterteilen. Der erste Bereich, die Definition der Business-Ziele, ermöglicht es, die gewünschten Ergebnisse, notwendige Maßnahmen und sinnvolle Metriken zu erarbeiten, diese in den Workflow einzubauen und stets bei der Durchführung des Projekts gegenzuprüfen. Sollte sich im Laufe des Projekts einer der oben genannten Punkte ändern, kann das problemlos angepasst werden. Der nächste Bereich ist die Entwicklungsphase, in dieser wird zunächst eine Sichtung nach relevanten Daten durchgeführt. Danach werden die Daten analysiert und passend zu den vorher definierten Zielen transformiert. Die ersten Modelle werden anhand der aufbereiteten Daten erstellt und mit einer Baseline verglichen. Die besten Modelle werden anschließend im What-If-Tool analysiert und anhand zurückgehaltener Daten nochmals gegengeprüft. Das finale Modell wird je nach Vorgabe in die Produktion eingebunden und mit einem AB-Test anhand neuester Daten im „realen Einsatz" gegen das bestehende System geprüft. Im letzten Abschnitt kann das Modell zu neuen Erkenntnissen in verschiedenen Bereichen führen und weitere Veränderungen sowie Verbesserungen anstoßen.

Der innere Kreislauf bezieht sich auf die Entwicklungsphase; wie schon oben genannt, geht es hier hauptsächlich um die Datenbereitstellung und das Modelltraining. Jedoch können auch hier mehrere Iterationen stattfinden, wenn beispielsweise neue Datenquellen dazukommen oder verschiedene Algorithmen, Topologien oder Konfigurationen ausprobiert werden sollen. Ein weiterer Punkt ist die Erstellung einer Baseline, falls keine bereitgestellt werden kann. Eine Baseline erlaubt es, grobe Schätzungen zur Genauigkeit des finalen Modells zu liefern, aber auch schon sehr früh erste, rudimentäre Analysen durchzuführen. Am Ende jedes Entwicklungszyklus wird eine Impactanalyse durchgeführt, anhand der die wichtigsten Datenspalten entdeckt werden können. Dies ist hilfreich, um die Qualität der Daten gegenzuprüfen, aber auch um mögliche Fehler bei der Entwicklung zu entdecken.

Machine Learning as a Service implementieren

Ziel war es, die Produktion von Leistungsschaltern zu optimieren. Dazu wurde, wie im Vorgehen beschrieben, zunächst das Business-Ziel definiert. Als Ziel wurde zunächst die Reduktion der Ausschüsse, der Nacharbeiten und der Prüfungskosten der Leistungsschalter gewählt. Als passende Metrik wurde die Anzahl der Leistungsschalter innerhalb und außerhalb des Toleranzbereichs ausgesucht. Dies ermöglicht es, ein Modell zu trainieren, das in der Lage ist vorherzusagen, wie gut ein Leistungsschalter wahrscheinlich sein wird.

Die Suche nach passenden Datenquellen zeigte zwei wichtige Quellen auf. Zum einen die Grenzstromdaten kombiniert mit den Prüfdaten der Anlagen und zum anderen die Kraft-Weg-Diagramme der

verbauten Schaltschlösser. Die erste Datenquelle ist in tabellarischer Form vorhanden und kann ohne größere Transformationen für das Training des Modells eingesetzt werden. Die zweite Datenquelle besteht aus vielen Messpunkten, zusammen mit den Experten aus der Domäne wurden wichtige Merkmale der Diagramme erarbeitet und extrahiert. Das neu gewonnene, reduzierte Dataset konnte dem Modell in tabellarischer Form für das Training bereitgestellt werden.

Aufgrund der fehlenden Baseline wurde zunächst ein Random-Forest-Modell entwickelt und initial zur Bestimmung der Impacts und erster Interaktionen genutzt. Die komplette Durchführung des Trainings fand im Trevisto-KI-Stack statt, dafür wurde eine Pipeline *(siehe Abbildung 5)* erstellt, mit der Daten eingelesen, transformiert und der Trainingskomponente bereitgestellt werden können. Die Trainingskomponente erstellt das trainierte Modell und gibt es an die Evaluierungskomponente weiter, die eine statistische Auswertung der Vorhersagen durchführt. Zum Schluss speichert die Pipeline das Modell samt allen Auswertungen im MinIO Object Storage, sodass weiterer Zugriff außerhalb der Pipeline möglich ist.

Anhand der Ergebnisse der Baseline und der ersten Auswertungen konnten weitere notwendige Datenanpassungen identifiziert und durchgeführt werden. Darüber hinaus dienten die ersten Ergebnisse als Basis für komplexere Modelle wie zum Beispiel Feed Forward Neural Networks, Residual Neural Networks und TabNets *(siehe Abbildung 6)*, die in weiteren Iterationen umgesetzt und geprüft wurden. Die Evaluierung der trainierten Modelle wurde wie im What-If-Tool durchgeführt *(siehe Abbildung 7)*.

Zukünftig soll ein neuronales Netz selbstständig nach nützlichen Informationen suchen und diese extrahieren mit dem Ziel, eine höhere Genauigkeit des Modells zu erreichen und weitere Erkenntnisse aus den Diagrammen für die Fertigung zu gewinnen.

Schritt für Schritt von Daten zu Erkenntnissen

Seit dem Start des Vorhabens konnten wichtige Erkenntnisse aus den Rohdaten gesammelt werden, dazu gehörten Histogramme *(siehe Abbildung 8)*, Gruppierungen, Analysen der Cluster, Bestimmung des Informationsgehaltes der Daten und der Dimensionsreduktion, Suche nach Ausreißern und möglichen Ursachen dafür und die Prüfung der linearen Korrelationen der einzelnen Datenspalten.

Die trainierten Modelle haben viele Insights in Bezug auf Wichtigkeit der einzelnen Datenspalten *(siehe Abbildung 10)* und deren Interaktionen *(siehe Abbildung 9)* geliefert sowie eine Möglichkeit gegeben, realistische Aussagen zu der erwartbaren Genauigkeit zu machen *(siehe Abbildung 11)*.

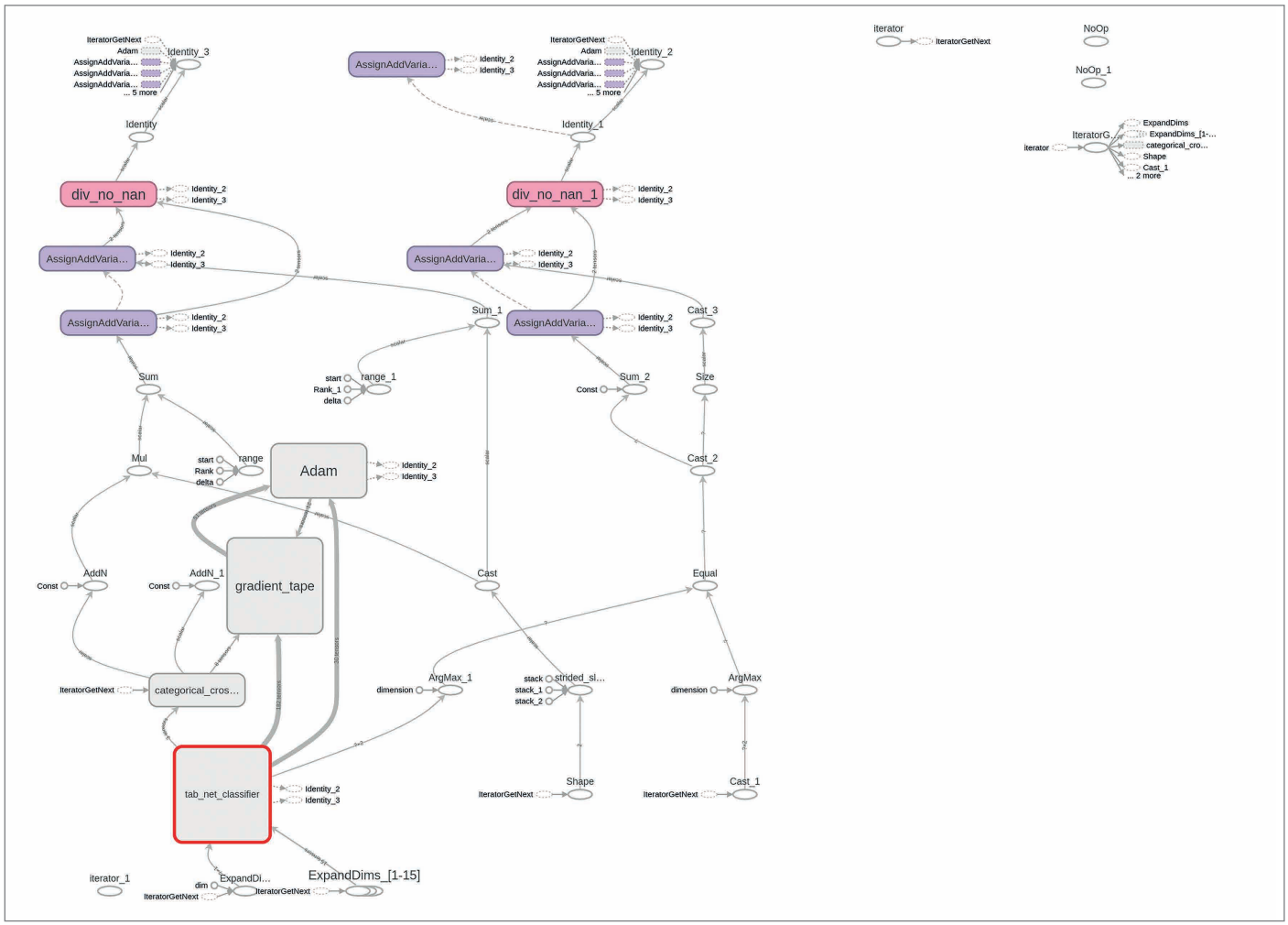


Abbildung 6: TabNet-Modellarchitektur (Quelle: Oliver Fuhrmann und Germans Hirsch)

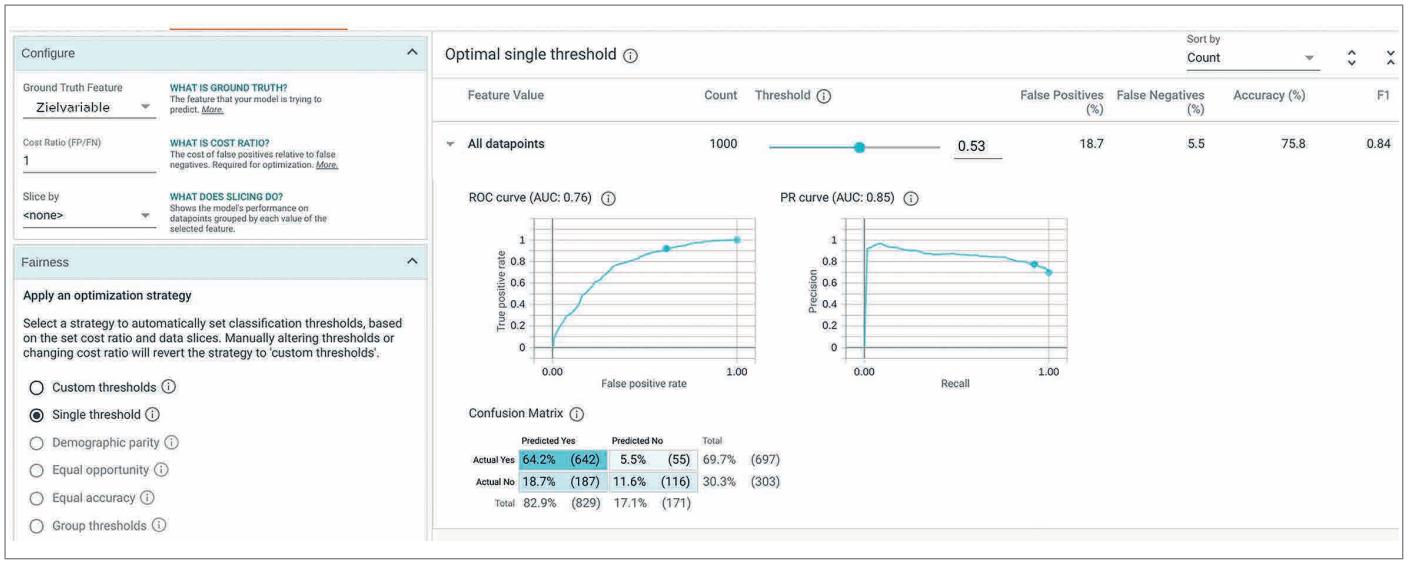


Abbildung 7: Modellanalyse im What-If Tool (Quelle: Oliver Fuhrmann und Germans Hirsch)

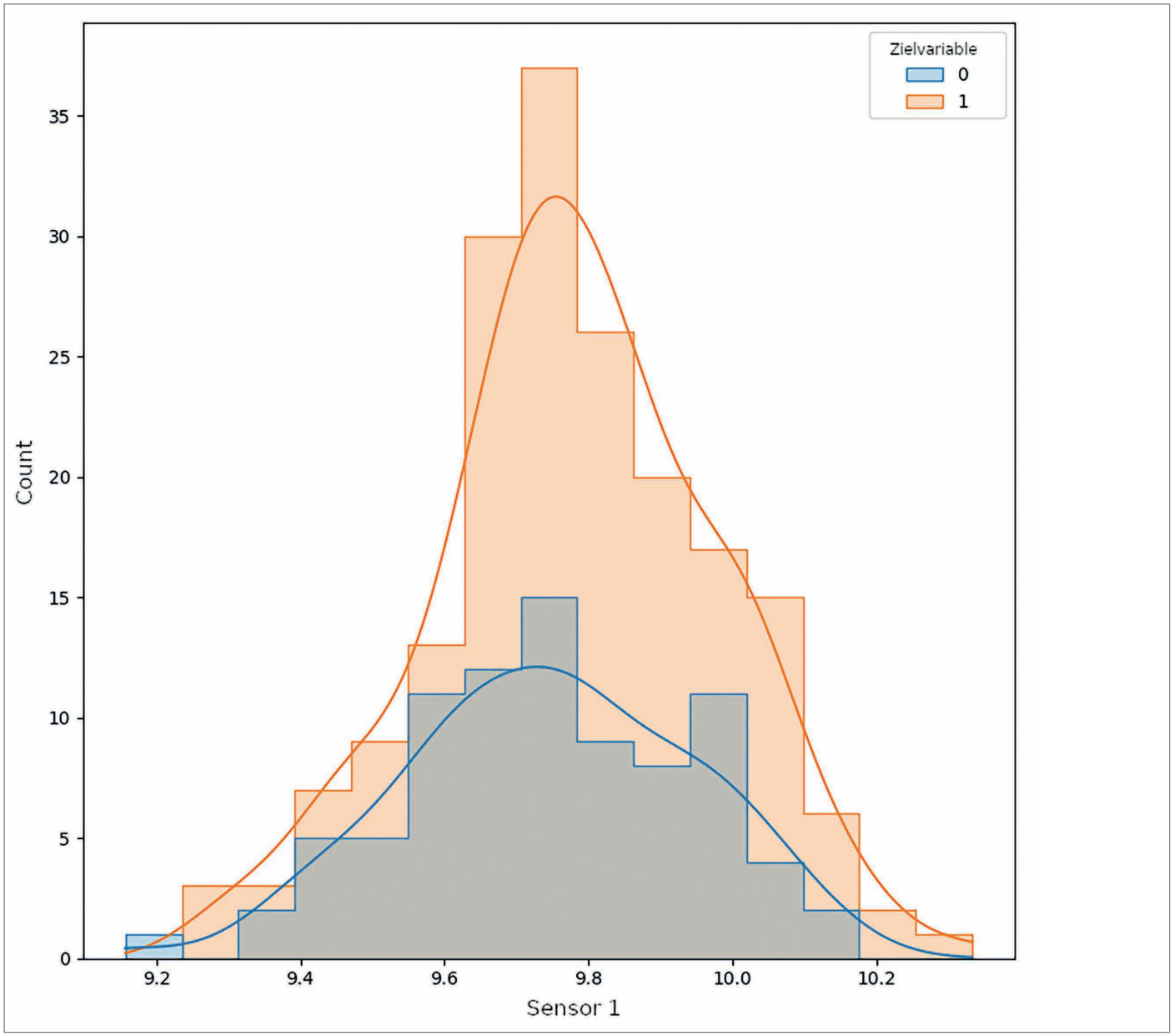


Abbildung 8: Verteilung der Sensorwerte gruppiert nach der Zielvariable (Quelle: Oliver Fuhrmann und Germans Hirsch)

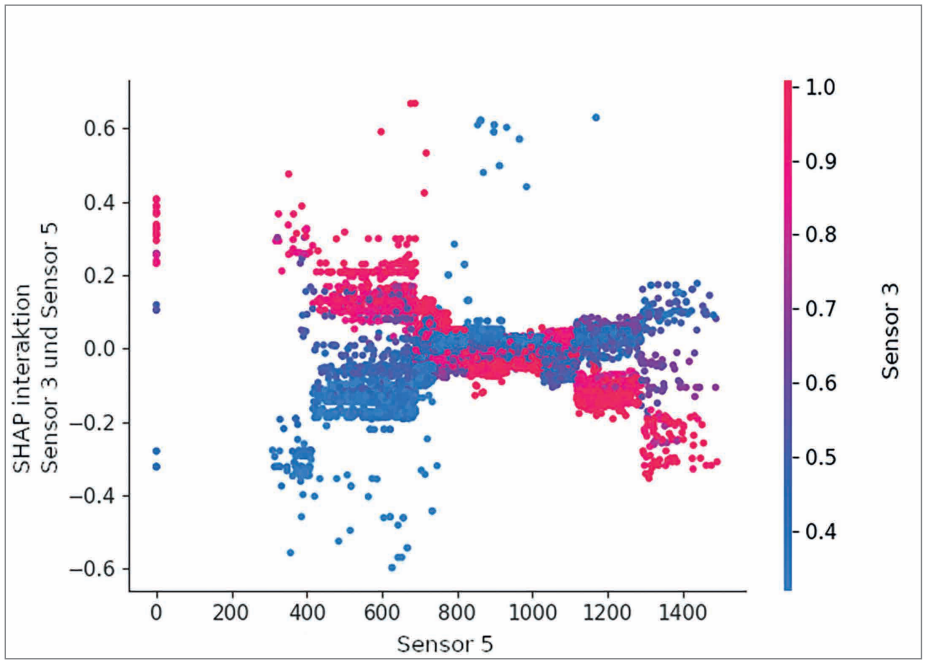


Abbildung 9: Visualisierung der Interkationen (Sensor 3 und Sensor 5). Höhere Werte im Sensor 3 und niedrigere Sensor-5-Werte in Kombination erhöhen die Wahrscheinlichkeit für optimale Leistungsschalter (Quelle: Oliver Fuhrmann und Germans Hirsch)

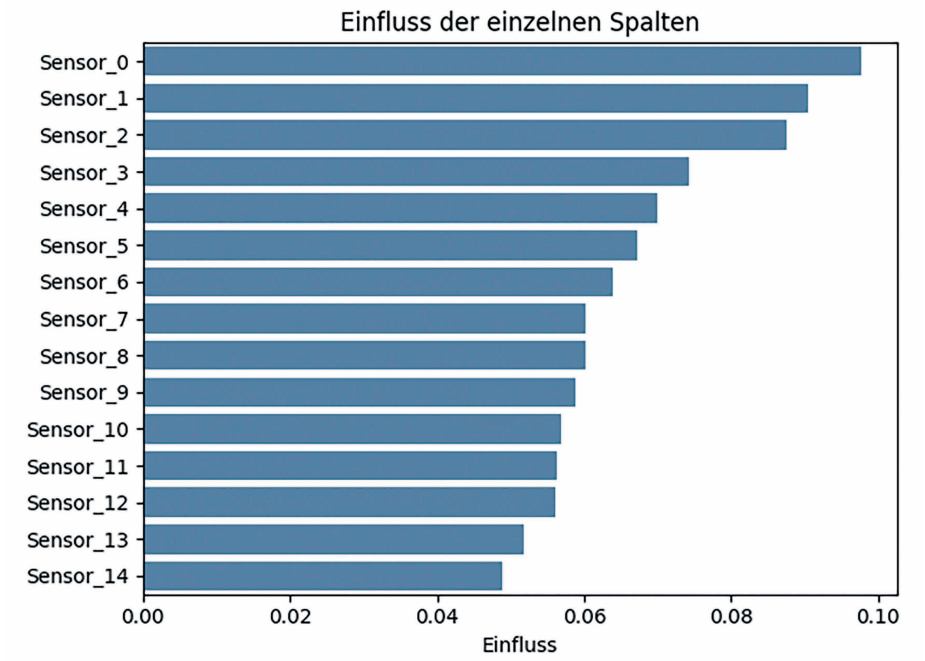


Abbildung 10: Einfluss der einzelnen Sensoren auf das Gesamtergebnis. Sensor 0 hat den größten Einfluss und Sensor 14 den geringsten. (Quelle: Oliver Fuhrmann und Germans Hirsch)

Vorhersage

Realität		Schlecht	Gut
	Schlecht	287	110
	Gut	144	495

Abbildung 11: Confusion-Matrix der Vorhersageergebnisse (Quelle: Oliver Fuhrmann und Germans Hirsch)

Die Ergebnisse der Modelle und die parallel dazu durchgeführte Analyse der Daten haben einige Hypothesen bestätigt und belastbar gemacht, einige neue Effekte und weitere potenzielle Einsatzgebiete aufgezeigt. KI kann erfolgreich genutzt werden, um Produktionsprozesse zu analysieren, daraus neue Erkenntnisse zu ziehen und diese weiter zu optimieren mit dem Potenzial, auch weitere Bereiche der Produktion oder des Betriebs zu verbessern. Durch trainierte Modelle können Insights sehr schnell generiert werden, jedoch brauchen größere Veränderungen und die Validierung der neuen Hypothesen Experten sowohl aus der Produktion als auch aus dem Bereich der künstlichen Intelligenz.

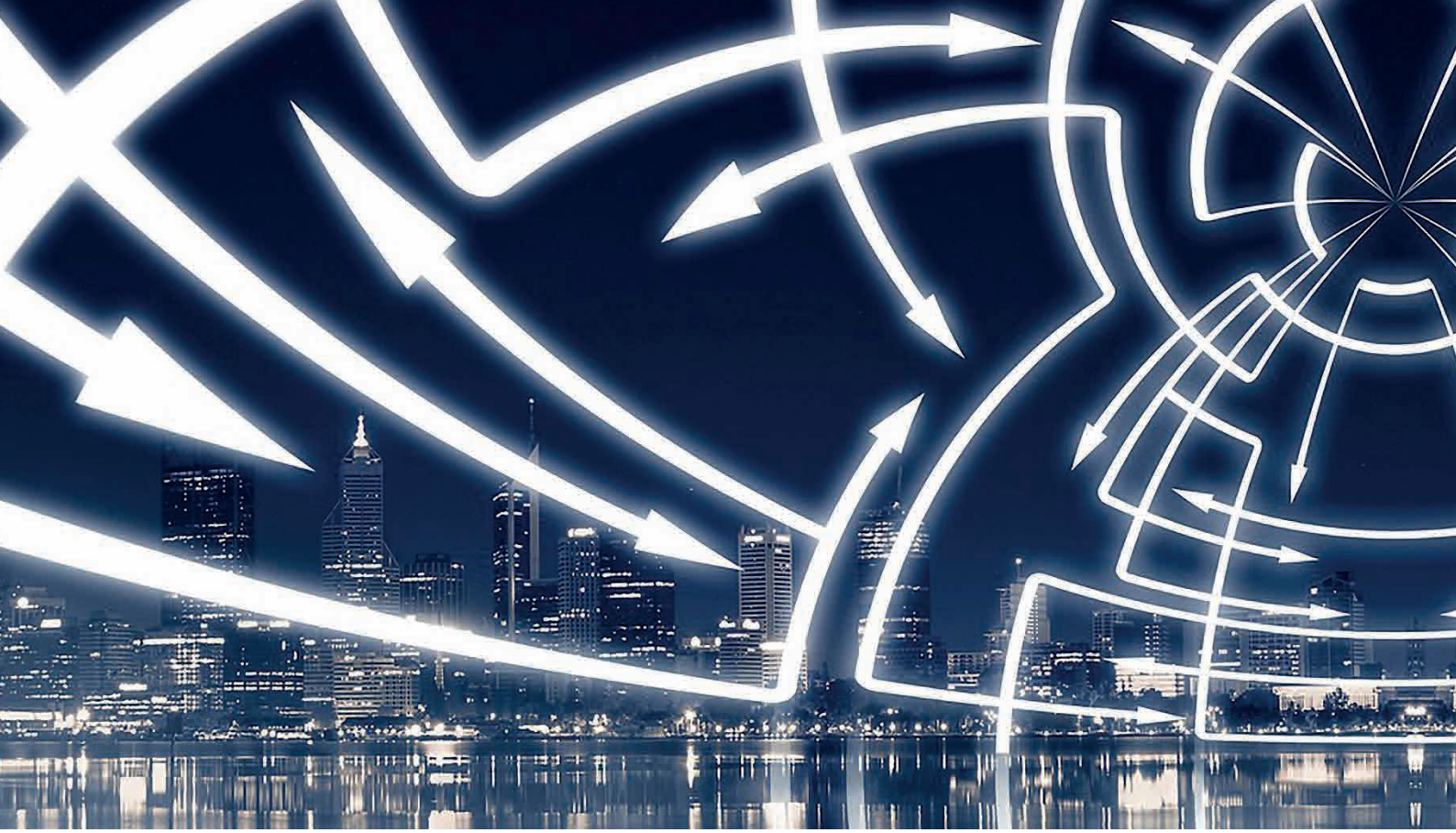
Über die Autoren

Oliver Fuhrmann verantwortet bei der Trevisto die Ressorts Business Development und Marketing. Er ist über 20 Jahre in der ITK-Branche tätig und hat zahlreiche Projekte in der Industrie, im Handel und im Dienstleistungssektor geleitet.

Germans Hirsch ist KI-Experte und Spezialist für neuronale Netze. Er ist als Consultant bei der Trevisto AG beschäftigt.

Oliver Fuhrmann
oliver.fuhrmann@trevisto.de

Germans Hirsch



Datenarchitekturen von Lakehouse bis Data Mesh: Evolution, Revolution, Chaos?

Andreas Buckenhofer, Daimler TSS

Ein Data Warehouse ist eine bekannte und bewährte Datenarchitektur. Auch Data Lakes änderten nichts an der Notwendigkeit eines Data Warehouse in einem modernen Datenmanagement. In den letzten Jahren kamen die Begriffe Lakehouse und Data Mesh auf: Was verbirgt sich dahinter?

Das klassische Data Warehouse (DWH) wurde von Bill Inmon und Ralph Kimball geprägt. In der *Abbildung 1* ist eine Beispielarchitektur dargestellt. Interne und externe Quellsysteme liefern Daten, die in den folgenden Schichten verarbeitet werden:

- Staging Layer zur temporären Aufnahme der Daten
- Core Warehouse Layer zur langfristigen Speicherung der Daten inklusive Integration und Historisierung der Daten
- Mart Layer für die Bereitstellung der Daten in einem für Endanwender verständlichen dimensionalen Modell (meist Starschema) für performante Auswertungen

Integration Layer und Aggregation Layer sind optional und kommen nur dann zum Einsatz, wenn der Transfer der Daten zwischen zwei Schichten zu komplex und eine Zwischenspeicherung notwendig ist.

Im Core Warehouse Layer werden die Daten aus den verschiedenen Quellsystemen zusammengeführt. In der klassischen Definition von Bill Inmon ist diese Schicht gekennzeichnet durch vier Kriterien:

- subject-oriented: Daten sind anhand von fachlichen Themen organisiert, etwa Kunde, Produkt, Lieferant etc.
- integrated: Daten sind standardisiert, das heißt, dass beispielsweise ver-

schiedene Kodierungen in den Quellsystemen vereinheitlicht werden.
- non-volatile: In einem DWH werden Daten nicht gelöscht oder verändert (mit Ausnahmen wie etwa rechtliche Anforderungen aus dem Datenschutz).
- timevariant: Daten können als Snapshots zu bestimmten Zeitpunkten aufgefasst werden.

Mit der Zunahme der Datenmenge (Volume), der Geschwindigkeit, mit der Daten entstehen (Velocity), und der Vielfalt der Datenformate (Variety) rückte der Fokus immer mehr in Richtung Data Lakes. Merkmale von Data Lakes, die auf Dateisystemen wie HDFS, S3, ADLS usw. basieren, sind:

- Trennung der Verarbeitung und Speicherung
- Flexibilität bei Schemaänderungen
- Anwendung des Schemas beim Lesen (schema-on-read)
- Speichern von Daten ohne Modellierung im Voraus
- Speichern beliebiger Datenformate
- Hohe Skalierbarkeit

Auch ein Data Lake erfordert eine Organisation der Daten – sonst landet man schnell in einem Data Swamp. Ein auf HDFS, S3, ADLS und Ähnlichem basierender Data Lake wird häufig mit einem auf einer relationalen Datenbank (RDBMS) basierenden DWH kombiniert, da RDBMS nach wie vor eine geringere Latenz bei Abfragen aufweisen und umfangreiche Funktionen (Security, Constraints etc.) haben.

Lakehouse

Lakehouse setzt sich zusammen aus Data Lake und Data Warehouse und erweckt dadurch den Anschein, dass eine neue Architektur definiert wird. Dies ist jedoch nicht der Fall.

Einige dieser Merkmale für ein Lakehouse sind aus verschiedenen Artikeln zusammengetragen:

- Einfaches Schemamanagement
- Metadatenmanagement
- Nutzung von Constraints in Dateisystemen
- Schema-Versionierung und Time-Travel-Abfragen
- ACID(-ähnliche) Transaktionen
- Offene Datenformate wie Parquet

Den Begriff Lakehouse findet man in der Literatur insbesondere bei Herstellern wie in *Abbildung 2* dargestellt. Es gibt keine einheitliche Definition, sondern einzelne Hersteller verwenden den Begriff für Features in ihren Produkten oder gar nur für die Vermarktung eines speziellen, aktuellen Produkts.

Diese Features sind zweifelsohne eine gute Weiterentwicklung einzelner Produkte – viele dieser Features nutzt man in relationalen Datenbanken schon seit Langem. Das Ziel ist, mit einem Produkt die Anforderungen eines klassischen DWH und eines Data Lake zu bedienen.

Data Mesh

Data Mesh wurde erstmals von Zhamak Dehghani im Artikel „How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh" vorgestellt. Dabei kritisiert sie DWHs und Data Lakes, die zu einem Flaschenhals beim zentralen Data Engineering führen. Laut Dehghani wird zu viel versprochen und zu wenig realisiert. Als Lösung schlägt sie den Data Mesh vor, der einen Wandel in der Denkweise voraussetzt. Sie gibt an, dass die nächste Datenplattform in der Konvergenz von verteilter domänengesteuerter Architektur, Self-Service-Plattformdesign und Produktdenken mit Daten liegt.

In *Abbildung 3* sind die vier Prinzipien des Data Mesh aufgeführt. Bisher hatten Datenteams keinen Domänenkontext und waren von den Geschäftsbereichen getrennt. Dies ändert sich mit dem ersten Prinzip, das sich darauf konzentriert, das

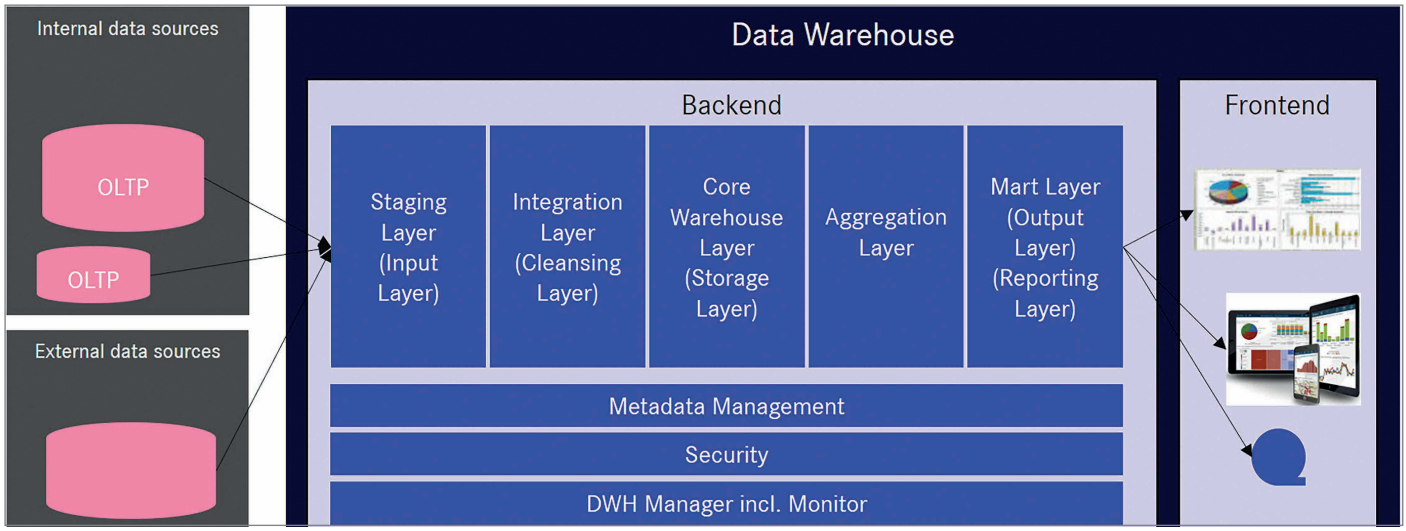


Abbildung 1: Data Warehouse (Quelle: Andreas Buckenhofer)

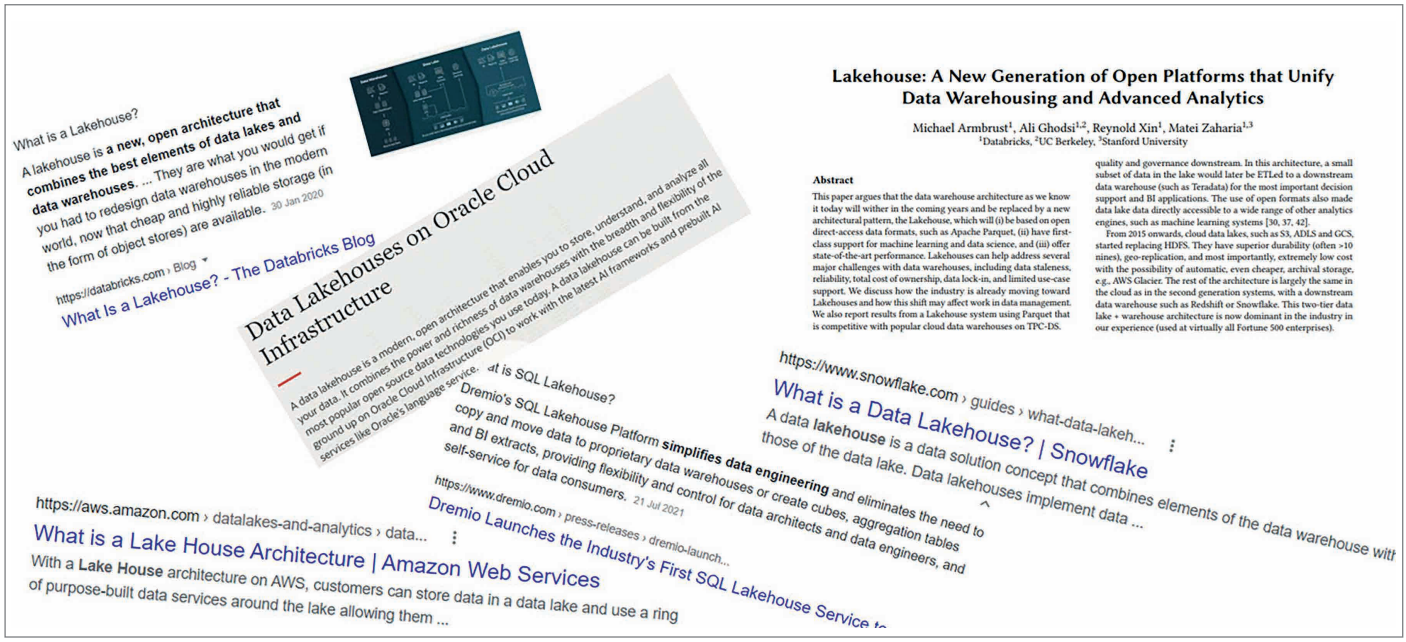


Abbildung 2: Artikel zu Lakehouse (Quelle: Andreas Buckenhofer)

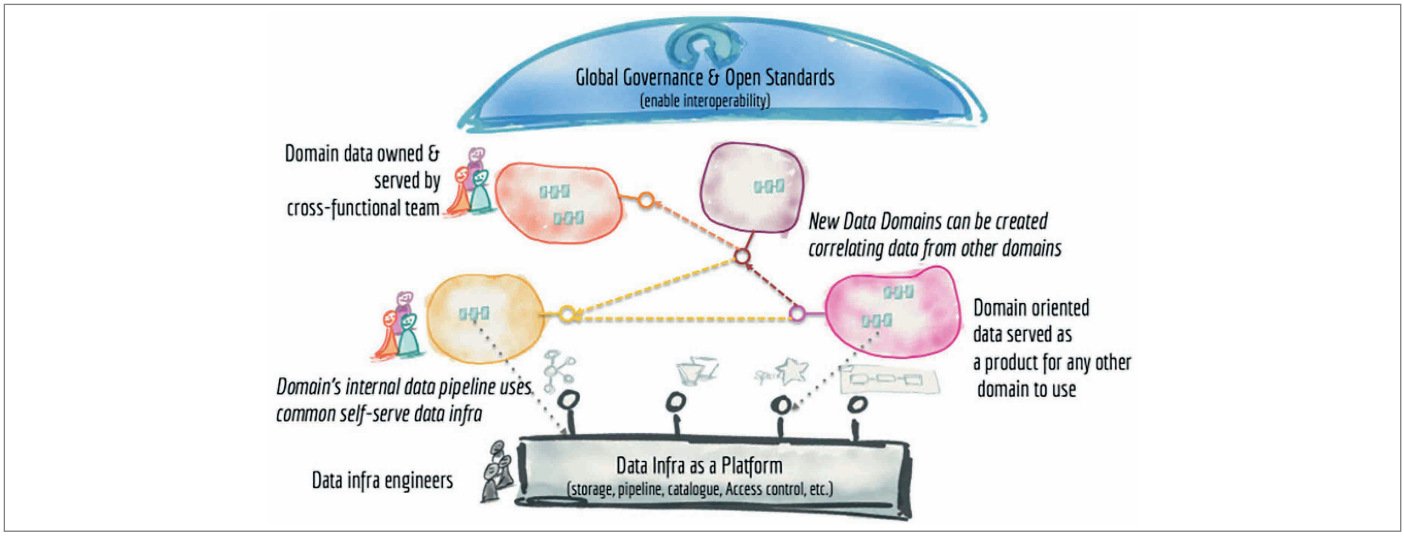


Abbildung 3: Data Mesh (entnommen aus [1]) (Quelle: Andreas Buckenhofer)

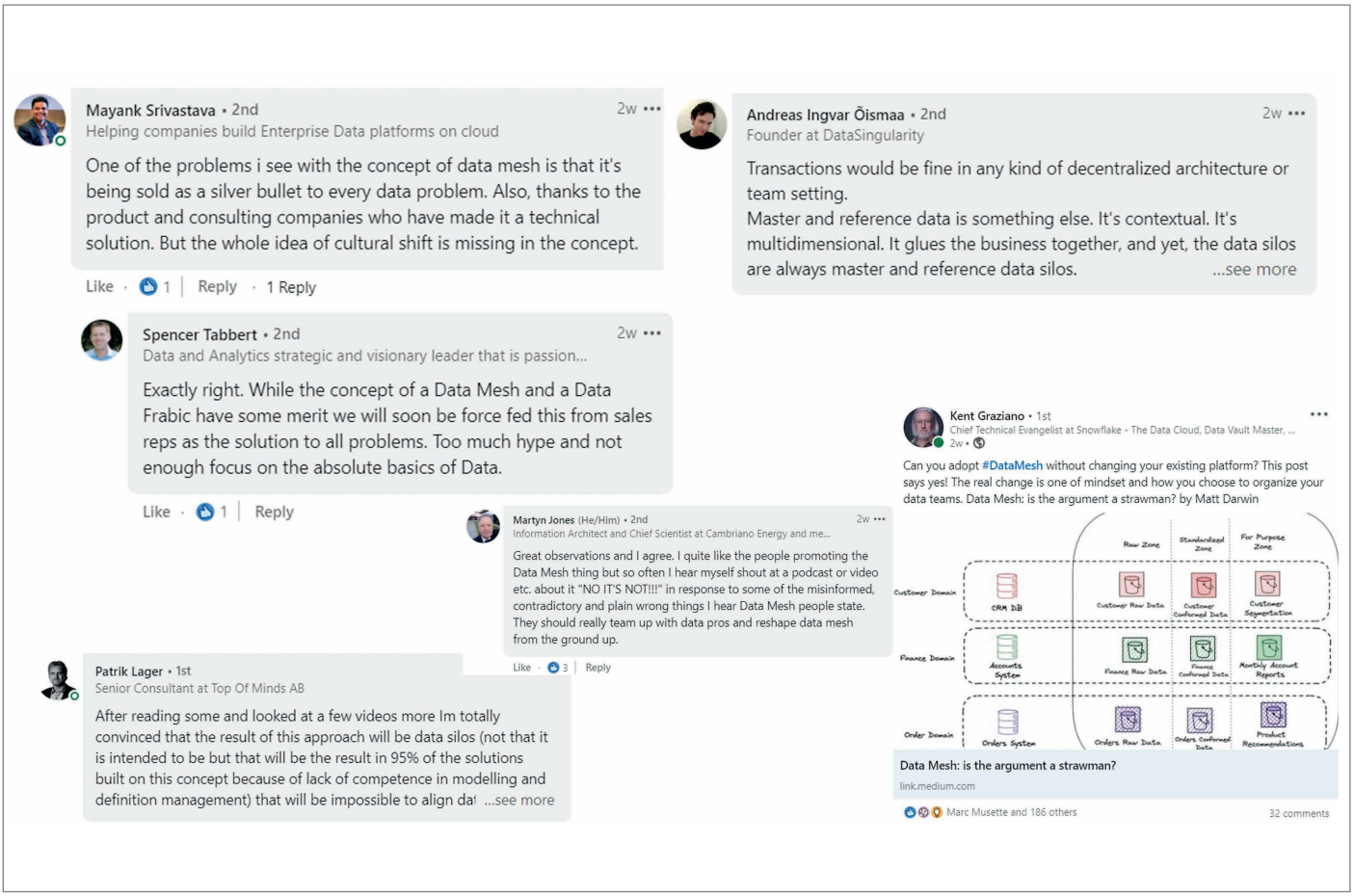


Abbildung 4: Data Mesh – kontroverse Diskussionen (Quelle: Andreas Buckenhofer)

Dateneigentum der Domäne oder dem Geschäftsbereich zu geben. Das zweite Prinzip besagt, dass Daten als Produkt zu sehen sind. Das heißt, durch Daten lassen sich Produkte oder Dienste schaffen, die nützlich sein oder monetarisiert werden können. Die Self-Service-Dateninfrastruktur als drittes Prinzip reduziert organisatorische Engpässe. Diese Plattform soll für eine breite Masse an Generalisten mit Geschäfts- oder Domänenkontext sein. Das vierte Prinzip befasst sich mit der Umsetzung von Governance in einem föderierten Modell.

Ein Data Mesh führt dazu, dass eine datengetriebene Denkweise mehr in Richtung Domäne rückt: Bereitstellung von Datenprodukten oder Klärung der Verantwortung (Domain-Ownership). Daten(qualität) entsteht in den Quellsystemen – dort muss der Umgang mit Daten verbessert werden.

Über Data Mesh wurde in den letzten Monaten in den sozialen Medien sehr viel geschrieben und es entstanden teilweise kontroverse Diskussionen. *Abbildung 4* enthält eine beliebige Auswahl von Screenshots. Kritiker sehen im Data Mesh einen Rückschritt in Richtung Silo-Denken und fehlende Datenintegration.

Revolution, Evolution, Chaos?

Im Falle eines Lakehouse werden durch Hersteller Produkte um Funktionen erweitert, die längst in relationalen Datenbanken etabliert sind wie ACID-ähnliche Transaktionen, Metadaten-Management u. a. Die Artikel zu Lakehouses stammen vorwiegend von Herstellern und führen Produktfeatures auf. Durch das Wort Lakehouse könnte man meinen, dass eine neue Architektur definiert werden könnte. Dies ist nicht der Fall – Lakehouse ist vielmehr ein Marketingbegriff (der auf sinnvollen Features von weiterentwickelten Tools basiert).

Mit Data Mesh wird ebenfalls keine Architektur definiert. Data Mesh ist aus meiner Sicht ein Mindset, das für ein Datenmanagement relevante Themen hervorhebt:

- Produktorientierung in den Domänen: Quell-Systeme haben sich bisher zu wenig um Datenbereitstellung oder Datenqualität gekümmert. Data Mesh nimmt diese Systeme in die Pflicht, in Datenprodukten zu denken – jedoch fehlt nach wie vor eine Incentivierung für die Datenbereitstellung.
- Datenbesitzer (Data Owner) und deren Verantwortung für die Daten in der eigenen Domäne müssen klar bestimmt werden.
- Datengetriebene Softwareentwicklung in der Domäne muss verstärkt werden. Für Auswertungen innerhalb der Domäne beziehungsweise für Machine Learning auf Daten in der Domäne wird kein zentrales System benötigt; dies muss lokal erfolgen.

Ein ausschließliches Denken in Domänen führt zu Silos und zu einem Rückschritt (Chaos?) im Datenmanagement: Eine Verarbeitung aller Daten in den je-

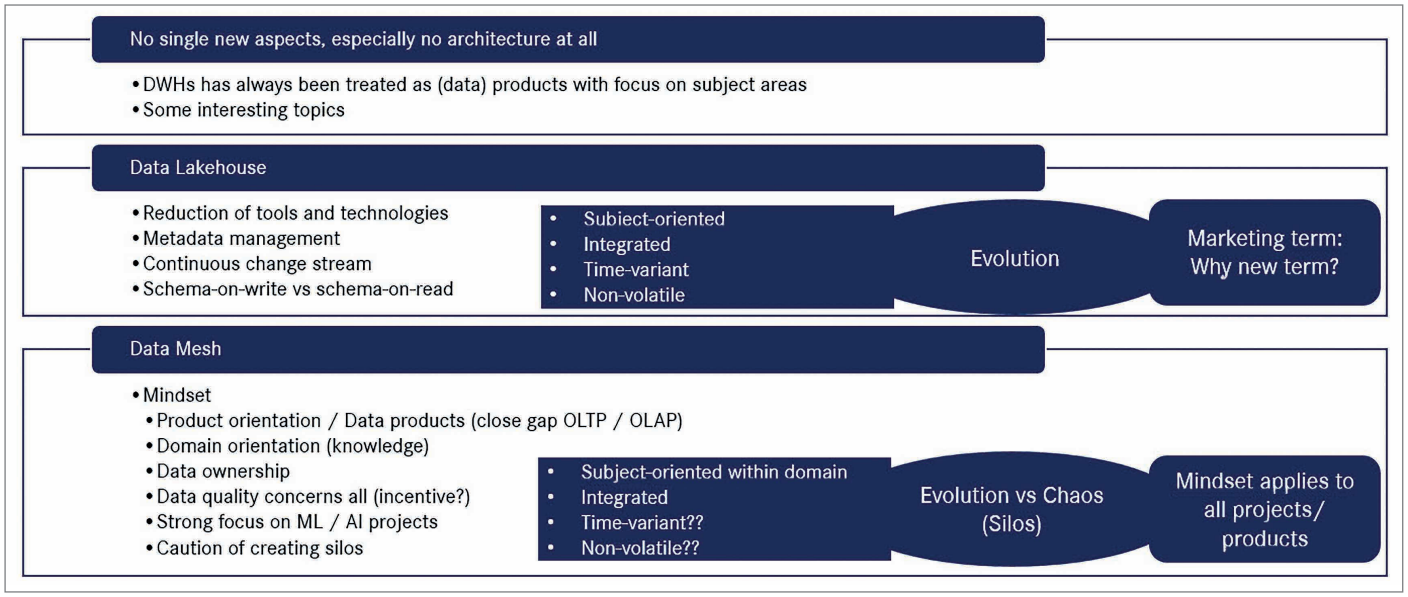


Abbildung 5: Zusammenfassung (Quelle: Andreas Buckenhofer)

weiligen Domänen ist nicht ausreichend. 360°-Kunde oder Supply Chain Management als Beispiele erfordern weiterhin ein zentrales, Domänen-übergreifendes DWH. Ein Data Mesh und eine darauf bauende Datenvirtualisierung sind bei diesen Beispielen nicht zielführend.

Weder Lakehouse noch Data Mesh bringen neue Themen, die nicht schon bekannt sind. Bestenfalls kann von einer Evolution gesprochen werden, da jeweils wichtige aktuelle Entwicklungen betont werden. Während ein Lakehouse aktuell ein Hersteller-Thema ist, so begegnet man dem Begriff Data Mesh insbesondere in Gesprächen mit IT-Architekten. Die *Abbildung 5* fasst meine Einschätzung zusammen.

Literatur

[1] Zhamak Dehghani: How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh, 20.05.2019, *https://martinfowler.com/articles/data-monolith-to-mesh.html*, abgerufen am 12.02.2022

[2] Zhamak Dehghani: Data Mesh: Delivering data-driven value at scale, O'Reilly 2022

[3] Ben Lorica, Michael Armbrust, Ali Ghodsi, Reynold Xin and Matei Zaharia: What is a Lakehouse?, 30.01.2020, *https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html*, abgerufen am 12.02.2022

[4] Jacek Majchrzak, Sven Balnojan, and Marian Siwiak: Data Mesh in action, Manning 2022

[5] Oracle: Data Lakehouse, *https://www.oracle.com/data-lakehouse/*, abgerufen am 12.02.2022

Über den Autor

Andreas Buckenhofer arbeitet bei Daimler TSS in der Business Unit „Vehicle Platforms" und verfügt über mehr als 20 Jahre Erfahrung in datenintensiven Anwendungen. Seine praktischen Erfahrungen gibt er gerne in internen Vorträgen und als Sprecher auf internationalen Konferenzen weiter. An der Dualen Hochschule Baden-Württemberg hält er regelmäßig eine Vorlesung über Datenmanagement. Er ist aktives Mitglied in der Datenbank-Community der DOAG und wurde von Oracle zum ACE Pro ernannt.

Andreas_Buckenhofer
andreas.buckenhofer@daimler.com

Robotron Datenbank-Software GmbH

Maßgeschneiderte Analytics-Lösungen

Ihre Daten sind ein Schatz, den Sie jetzt effektiv heben wollen? Robotron bietet Ihnen ein breites Spektrum an Werkzeugen und Erfahrungen im Umgang mit großen Datenmengen.

Durch die Spezialisierung unserer Experten im Bereich Business Intelligence und Data Warehouse (BI/DWH) profitieren Sie von kostengünstigen und intelligenten Anwendungen für Datenmanagement und Informationsbereitstellung.

Mit dem ***robotron*BIArchitect*** stellt Robotron ein effizientes und intelligentes Werkzeug zur Erstellung von prozess- und fachspezifischen Business-Intelligence-Auswertungen zur Verfügung.

Jetzt mehr erfahren:
www.robotron.de/biarchitect

Ihre Vorteile:

- **kein** explizites BI-Technologiewissen erforderlich
- **vollständige Business Intelligence-Anwendung**, die von einer **Webanwendung** zur intuitiven Pflege ergänzt wird
- Einbeziehung lokaler Datenbestände und Zusammenführung aller Daten in **einer zentralen Datenbank**

Passende Kurse im
Robotron Schulungszentrum

Ein bewährtes Tool für die Analyse und Auswertung großer Datenmengen bietet Oracles strategische Lösung für Unternehmensberichte und Dokumentenausgaben – der ***Oracle BI Publisher***.

Im Robotron Schulungszentrum machen wir Sie fit für den Umgang mit diesem Tool. Sichern Sie sich Ihren Platz in einem der folgenden Kurse:

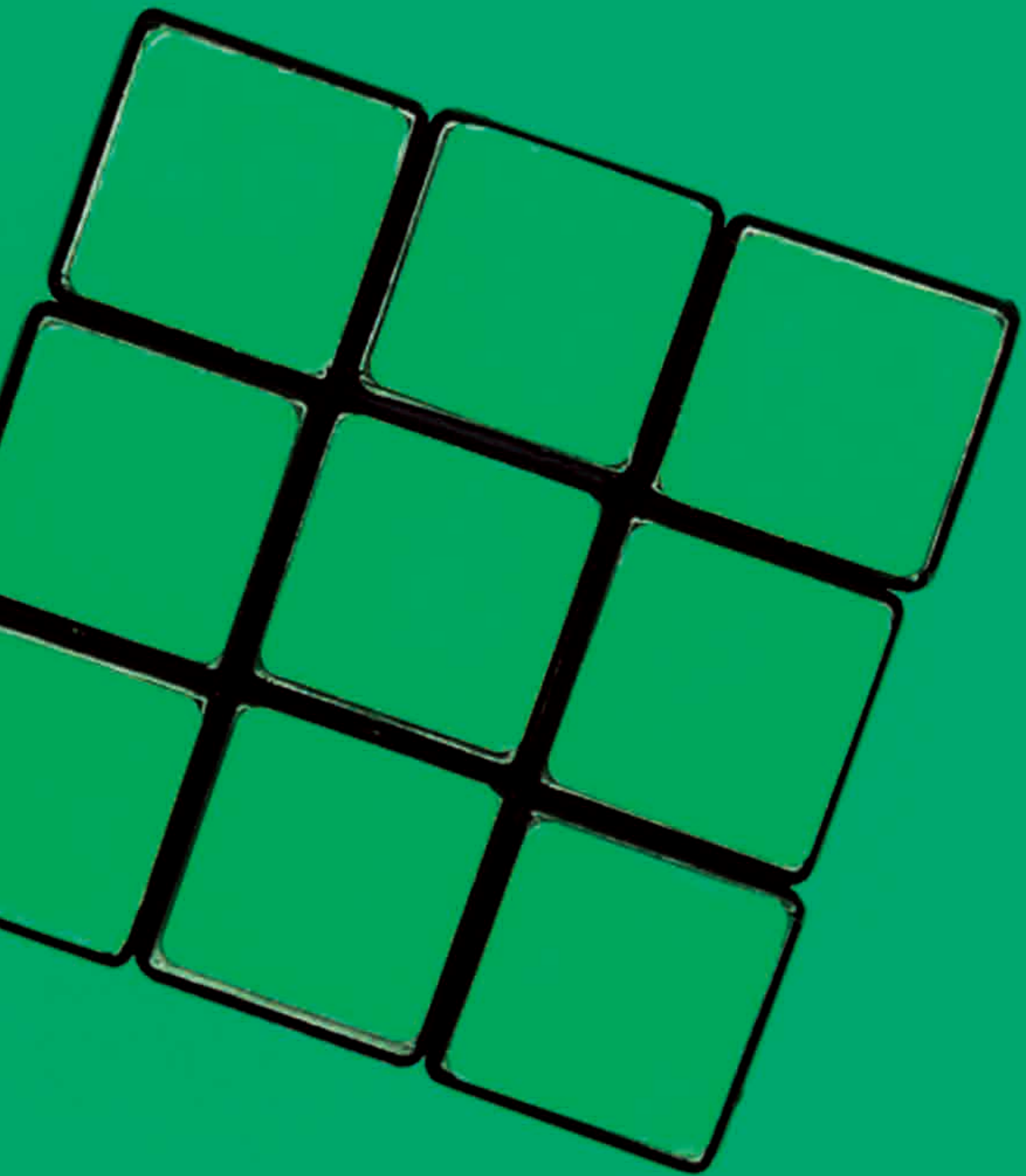
Praxisworkshop Grundlagen der Administration von Oracle BI 12c

Oracle BI 12c: Create Analyses and Dashboards Ed 1

Oracle BI 12c: Build Repositories

Eine Übersicht aller Oracle-Kurse im Schulungszentrum finden Sie auch unter: www.oracle-schulung.de

robotron schulung



Tipps und Tricks für Entwickler – Teil 2

Lothar Flatz, DBConcepts

Seit über zwanzig Jahren verbessere ich Software, die mit der Oracle-Datenbank kommuniziert. Bei jedem meiner Einsätze erstelle ich ein Arbeitsprotokoll, eine rudimentäre Dokumentation, die mir persönlich erlaubt, die Analysevorgänge auch im Nachhinein nachzuvollziehen.
Diese Dokumentation hilft mir einerseits, meine Gedanken zu klären, und erlaubt es mir auch anderseits, nach einer Unterbrechung die Arbeit lückenlos fortzusetzen. Nun habe ich außerdem einen Fundus von Hunderten praktischen Beispielen, aus denen ich Anregungen für einen Vortrag schöpfen kann.
Ich möchte mich bei Christian Schwitalla für die Idee bedanken, einen solchen Vortrag zu halten. Anfang dieses Jahres hat die DOAG mich gebeten, einen Artikel zu dem Vortrag zu schreiben.
Aufgrund der Länge des Textes habe ich ihn für das Red Stack Magazin in zwei Teile aufgespalten. Im ersten Teil ging es vor allem um Analytic Functions. Im zweiten Teil, der im Anschluss folgt, geht es um das Thema PL/SQL versus SQL.

Spaltenwertige Selects

Grundsätzliches

Sie sind sehr angenehm zu codieren, die Selects in der Select-Klausel, die genau einen Wert zurückgeben. Auf der einen Seite sind sie schön übersichtlich. Der ganze Text steht meist in einigen wenigen Zeilen. Man braucht sich über Outer Joins und Non Outer Joins keine Gedanken zu machen; wenn ein spaltenwertiges Select kein Resultat liefert, wird dadurch die Anzahl der Datensätze nicht kleiner. Zudem kann es keine Duplizierung von Datensätzen im Resultat über falsch geschriebene Joins geben.

Der große Nachteil der spaltenwertigen Selects ist, dass sie den Optimizer stark einschränken. Anders als beim Subselect in der Where-Klausel kann kein „unnest" stattfinden. Ein „unnest" ist eine Transformation, die der Optimizer vornimmt, bei der ein Subselect in die From-Klausel geschrieben wird.

Dadurch wird jeder beliebige Join möglich, zum Beispiel auch ein Hash Join.

Ein spaltenwertiges Select muss immer wie mit einem Einzelsatzzugriff ähnlich wie beim Nested Loop Join durchführt werden. (Jedenfalls ist mir keine Ausnahme bekannt.) Das spaltenwertige Select wird für jeden Datensatz wieder erneut durchgeführt. Ich nenne es deshalb auch ein Huckepack-Select, weil es gleichsam auf dem Rücken des HauptSelect mitgetragen wird.

In *Listing 1* sehen Sie ein Beispiel, bei dem ich ein spaltenwertiges Select in einem Count geschrieben habe. (Das Beispiel ist einem realen Fall aus dem medizinischen Bereich nachempfunden, deshalb die Tabellennamen.)

Hier sieht man die Buffergets und die Laufzeit mit Operation 2 schlagartig ansteigen (Operation 1 und 2 werden überlappend gemacht, die Zahlen sind nicht klar getrennt). Ich nenne dieses Muster nach der optischen Erscheinungsform scherzhaft den „Atompilz". Der Atompilz wird zu 99% durch spaltenwertige Selects oder Funktionen in der Select-Klausel verursacht.

Nun, so dumm muss man das ja nicht machen. Wir sollten einen Index auf Medi_C anlegen *(siehe Listing 2)*.

Das ist schon viel besser. Kann ein Join aller drei Tabellen das toppen? *(siehe Listing 3)*

In der Tat ist der dreifache Join schon bei nur einem spaltenwertigen Select schneller. Werden mehrere spaltenwertige Selects auf dieselbe Tabelle ausgeführt, genügt nicht ein Zugriff aus dieser Tabelle, sondern es werden mehrere gemacht. Dann wäre der Unterschied noch deutlicher.

Summa summarum: Spaltenwertige Selects sind in der Abarbeitung fast immer ineffizient. Das gilt auch dann, wenn sie auf dem Umweg über PL/SQL-Funktionen gerufen werden (siehe nächstes Kapitel).

Es bieten sich zwei Lösungsmöglichkeiten an:

1. Umschreiben in einen klassischen Join. Dies ist die beste Lösung und gibt dem Optimizer alle Freiheiten zurück. Jedoch muss man darauf achten, dass das Resultat wirklich gleichbleibt. Das ist nicht immer einfach.
2. Konsequentes Indexieren des Zugriffs. Dies ist mehr ein Kaschieren des Problems als eine eigentliche Lösung. Nach wie vor ist nur ein Nested Loop möglich. Der Aufwand ist aber oft deutlich geringer als der beim Umschreiben, da das Resultat unverändert bleibt.

```
select count(( select substr(mc.data ,1,1) from MEDI_C mc where mb.c_id = mc.id) and mc.status='N')
    from MEDI_b mb,
      MEDI_A m
 where  mb.id = m.MEDI_A_NR
    and (creation_date>'03.01.14' OR last_update_date>'03.01.14')
/

-----------------------------------------------------------------------------------------------------------
| Id  | Operation                | Name       |  Starts | E-Rows | A-Rows |   A-Time     | Buffers |
-----------------------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT         |            |       1 |        |      1 |00:00:47.62 |     13M |
|*  1 |  TABLE ACCESS FULL       | MEDI_C     |   25917 |      1 |   5183 |00:00:44.59 |     13M |
|   2 |  SORT AGGREGATE          |            |       1 |      1 |      1 |00:00:47.62 |     13M |
|*  3 |   HASH JOIN              |            |       1 |   6390K|   4924K|00:00:02.57 |  23124  |
|   4 |    INDEX FAST FULL SCAN | MEDI_B_I2_X|       1 |   386K |    386K|00:00:00.05 |   1128  |
|*  5 |    INDEX FAST FULL SCAN | MEDI_A_I01 |       1 |   6472K|   4924K|00:00:00.86 |  21996  |
-----------------------------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------

    1 - filter(("MC"."ID"=:B1 AND "MC"."STATUS"='N'))
    3 - access("MB"."ID"="M"."MEDI_A_NR")
    5 - filter(("LAST_UPDATE_DATE">'03.01.14' OR "CREATION_DATE">'03.01.14'))
```

Listing 1: Beispiel mit einem spaltenwertigen Select

Scalar Subquery Caching

Es gibt allerdings eine Optimierung, die sogar dafür sorgen kann, dass spaltenwertige Selects schneller sind als Joins.

Dabei wird das Ergebnis einer Unterabfrage gecacht und die Abfrage wird mit den gleichen Suchkriterien nur einmal aufgerufen.

Am besten sehen wir das an einem einfachen Beispiel in unserem altbekannten Scott-Schema. Zunächst die klassische Lösung *(siehe Listing 4)*.

Dann das Gleiche mit einem spaltenwertigen Select *(siehe Listing 5)*.

Wie man sieht, ist die Lösung mit dem spaltenwertigen Select sogar schneller. Die Funktionsweise und die Grenzen des Scalar Subquery Caching bespricht Tom Kyte in seinem Artikel *[4]*. Wie immer weist Tom darauf hin, dass reines SQL immer noch die beste Lösung ist.

Funktionsaufrufe in SQL

Context Switch

PL/SQL und SQL laufen auf verschiedenen Umgebungen. Wenn aus SQL eine PL/SQL-Funktion aufgerufen wird, dann müssen alle relevanten Daten, wie Bindevariablen, Cursor und Resultate, zwischen den beiden Umgebungen übermittelt werden. Den Aufwand, den dies verursacht, nennt man einen Context Switch.

Wie lange ein Context Switch dauert, ist schwer zu sagen. Zunächst einmal ist es für unterschiedliche Versionen von Oracle verschieden. Er wurde immer wieder an einer Verbesserung gearbeitet. Dann hängt es auch davon ab, wieviel und welche Information ausgetauscht werden muss.

Ein Context Switch ist sehr kurz, jedoch wird die Anzahl der Context Switches gerne unterschätzt. Nehmen wir ein Beispiel: In der Select-Klausel eines SQL-Statements, das 10000 Zeilen zurückliefert, wird eine PL/SQL-Funktion aufgerufen. Somit haben wir 10000 Context Switches, richtig? Nein, es sind zumindest doppelt so viele, denn das Ergebnis der Funktion muss ja auch an SQL zurückgeliefert werden. So werden also 10000-mal Suchparameter übergeben und 10000 Ergebnisse zurückgeliefert. Im besten Fall. Wenn in der PL/SQL-Funktion aber wieder SQL-Statements vorkommen, geht das muntere Spielchen von vorne los. Sollte in der PL/SQL-Funktion ein Cursor Loop aufgerufen werden, dann erzeugt jedes Fetch eines Resultates einen Context Switch.

Sehr gut erklärt das Steven Feuerstein in seinem Blog Post *[8]*. Ehe man es sich versieht, hat man Millionen von Context Switches und diese kosten dann auch wirklich substanziell Zeit.

Der Context Switch kann vermindert werden. Die Präsentation von Bryn Llewellyn *[9]* fasst diesbezüglich alles zusammen.

Im Folgenden das wichtigste Slide daraus. Bryn hat „SQL ruft PL/SQL“ getestet. Das heißt, es wurde in der PL/SQL kein weiteres SQL mehr gerufen. Es gibt in diesem Beispiel also nur zwei Context Switches pro PL/SQL-Funktionsaufruf.

Ganz links steht die Laufzeit in reinem SQL. Diese Zeitangabe ist in Hundertstelsekunden, die restlichen Zeiten sind Sekunden. Bitte beachten Sie, um wie viel schneller reines SQL ist.

Es stehen folgende Strategien zur Reduktion des Context Switch zur Verfügung:

1. Pragma UDF (Optimizer erfährt, dass SQL -> PL/SQL beabsichtigt ist)
2. Funktion in der With-Klausel (schnellerer Aufruf, aber nicht wiederverwendbar)
3. With-Klausel ruft eine gespeicherte PL/SQL-Funktion (etwas langsamer als 2., aber wiederverwendbar)

Wie gut die Optimierung gelingt, hängt jeweils vom Datentyp der Parameter und vom Datentyp des Funktionsresultates ab.

```
-------------------------------------------------------------------------------------------------------------------
| Id  | Operation                           |   Name     | Starts |  E-Rows | A-Rows |   A-Time   | Buffers|
-------------------------------------------------------------------------------------------------------------------
| 0   | SELECT STATEMENT                    |            |      1 |         |      1 |00:00:02.81 |  46261 |
|* 1  |  TABLE ACCESS BY INDEX ROWID BATCHED|MEDI_C      |  25917 |     1   |   5183 |00:00:00.05 |  23137 |
|* 2  |   INDEX RANGE SCAN                  |MEDI_C_I1   |  25917 |     1   |  25917 |00:00:00.02 |    414 |
|  3  |  SORT AGGREGATE                     |            |      1 |     1   |      1 |00:00:02.81 |  46261 |
|* 4  |   HASH JOIN                         |            |      1 |  6390K  |   4924K|00:00:02.35 |  23124 |
|  5  |    INDEX FAST FULL SCAN             |MEDI_B_I2_X |      1 |   386K  |    386K|00:00:00.03 |   1128 |
|* 6  |    INDEX FAST FULL SCAN             |MEDI_A_I01  |      1 |  6472K  |   4924K|00:00:00.83 |  21996 |
-------------------------------------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------

   1 - filter("MC"."STATUS"='N')
   2 - access("MC"."ID"=:B1)
   4 - access("MB"."ID"="M"."MEDI_A_NR")
   6 - filter(("LAST_UPDATE_DATE">'03.01.14' OR "CREATION_DATE">'03.01.14'))
```

Listing 2: Anlegen eines Index auf Medi_C

```
select count ( substr(mc.data ,1,1))
   from MEDI_A m,
        MEDI_b mb,
        MEDI_C mc
 where  mb.id = m.MEDI_A_NR
   and mc.status='N'
   and mb.c_id = mc.id
   and (creation_date>'03.01.14' OR last_update_date>'03.01.14')
/
-------------------------------------------------------------------------------------------------------------
| Id | Operation                | Name         |  Starts |  E-Rows  | A-Rows |   A-Time   | Buffers |
-------------------------------------------------------------------------------------------------------------
| 0  | SELECT STATEMENT         |              |     1   |          |      1 |00:00:02.08 |   23642 |
| 1  |  SORT AGGREGATE          |              |     1   |      1   |      1 |00:00:02.08 |   23642 |
|* 2 |   HASH JOIN              |              |     1   |   1267K  |    984K|00:00:02.04 |   23642 |
|* 3 |    HASH JOIN             |              |     1   |  76667   |  77300 |00:00:00.11 |    1646 |
|* 4 |     TABLE ACCESS FULL    | MEDI_C       |     1   |   7730   |   7730 |00:00:00.01 |     518 |
| 5  |     INDEX FAST FULL SCAN | MEDI_B_I2_X  |     1   |    386K  |    386K|00:00:00.03 |    1128 |
|* 6 |     INDEX FAST FULL SCAN | MEDI_A_I01   |     1   |   6472K  |   4924K|00:00:00.80 |   21996 |
-------------------------------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------

   2 - access("MB"."ID"="M"."MEDI_A_NR")
   3 - access("MB"."C_ID"="MC"."ID")
   4 - filter("MC"."STATUS"='N')
   6 - filter(("LAST_UPDATE_DATE">'03.01.14' OR "CREATION_DATE">'03.01.14'))
```

Listing 3: Join aller drei Tabellen

So steht beispielsweise Num_Date für numerischen Parameter und Resultattyp Date *(siehe Abbildung 1)*.

PL/SQL Versus SQL

Hier gibt es einige Gemeinsamkeiten zu spaltenwertigen Selects, weshalb es sinnvoll ist, das Thema im Anschluss daran zu behandeln. Aus Sicht der Effizienz gilt auch hier, alles was möglich ist, mit SQL zu machen. Dies ist um ein Vielfaches effizienter als PL/SQL. Wenn man SQL-Code durch Verwendung von PL/SQL-Funktionen in Teile aufsplittet, kann der Optimizer nur Teile, nicht aber die Gesamtaufgabe optimal lösen. Dass das in der Regel nicht optimal sein wird, versteht sich von selbst. Im schlimmsten Fall ist das Resultat sogar falsch *[2]*.

Wie man auch mit PL/SQL-Funktionen das Scalar Subquery Caching nutzen kann, zeigt Mohamed Houri sehr schön *[5]*. Den Vergleich zwischen PL/SQL über den Scalar Subquery Cache und reinem SQL zieht Mohamed Houri in einem anderen Blog *[6]*. Auch hier scheidet reines SQL klar besser ab.

Trotz der genannten erheblichen Nachteile gibt es ein wesentliches Argument,

```
select ename,
       dname
from emp e,  dept d
where d.deptno=e.deptno
;
-----------------------------------------------------------------------------------------------------
| Id  | Operation                     |  Name     | Starts  |  E-Rows  |   A-Rows |  A-Time    |Buffers|
-----------------------------------------------------------------------------------------------------
| 0   | SELECT STATEMENT              |           |    1    |          |    14    |00:00:00.01|   9   |
| 1   |  MERGE JOIN                   |           |    1    |    14    |    14    |00:00:00.01|   9   |
| 2   |   TABLE ACCESS BY INDEX ROWID |  DEPT     |    1    |     4    |     4    |00:00:00.01|   2   |
| 3   |    INDEX FULL SCAN            |  PK_DEPT  |    1    |     4    |     4    |00:00:00.01|   1   |
|* 4  |   SORT JOIN                   |           |    4    |    14    |    14    |00:00:00.01|   7   |
| 5   |    TABLE ACCESS FULL          |  EMP      |    1    |    14    |    14    |00:00:00.01|   7   |
-----------------------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------

   4 - access("D"."DEPTNO"="E"."DEPTNO")
       filter("D"."DEPTNO"="E"."DEPTNO")
```

Listing 4: Klassische Lösung in altbekannten Scott-Schema

```
select ename,
       (select dname from dept d where d.deptno=e.deptno)
from emp e
;

-----------------------------------------------------------------------------------------------------
| Id  |  Operation                    |  Name     | Starts | E-Rows |A-Rows |   A-Time   |Buffers |
-----------------------------------------------------------------------------------------------------
| 0   |  SELECT STATEMENT             |           |   1    |        |   14  |00:00:00.01 |   7    |
| 1   |   TABLE ACCESS BY INDEX ROWID |  DEPT     |   3    |    1   |    3  |00:00:00.01 |   5    |
|* 2  |    INDEX UNIQUE SCAN          |  PK_DEPT  |   3    |    1   |    3  |00:00:00.01 |   2    |
| 3   |   TABLE ACCESS FULL           |  EMP      |   1    |   14   |   14  |00:00:00.01 |   7    |
-----------------------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------

   2 - access("D"."DEPTNO"=:B1)
```

Listing 5: Lösung mit spaltenwertigem Select

dass für den Einsatz von PL/SQL-Funktionen spricht: Modularisierung. Während überall der Grundsatz gilt, Programmieraufgaben und leichter überschaubare Teilstücke aufzubrechen und Basis-Module zu schaffen, soll dieses Prinzip ausgerechnet hier nicht gelten?

Ich habe seit Jahren immer wieder nachgedacht, wie dieser Widerspruch aufzulösen ist. (Mein Kollege Björn Finke meint, dass er im Zweifel der Performance den Vorrang einräumt, weil die der Kunde wahrnimmt.) Im Endeffekt kam ich zu dem Ergebnis, dass nur Oracle eine Lösung bieten kann.

Erst seit Version 19 existiert eine Lösung, die SQL Macros. Diese bieten sogar einen brauchbaren Migrationspfad für bestehenden Code *[3]*.

Schlussbemerkungen

Gerade die SQL Macros haben mir gezeigt, dass es selbst nach vielen Jahren Oracle-Erfahrung immer noch positive Überraschungen geben kann.

Mit dem Thema sind wir noch lange nicht am Ende. Ich hoffe, dass es im Herbst 2022 einen zweiten Teil geben wird.

Quellen

[1] Hall, Tim, Analytic Functions, *https://oracle-base.com/articles/misc/analytic-functions*

[2] Saxon, Chris, The Problem with SQL Calling PL/SQL Calling SQL, *https://blogs.oracle.com/sql/the-problem-with-sql-calling-plsql-calling-sql*

[3] Schwinn, Ulrike, Parametrisierte Views mit SQL Macros, *https://blogs.oracle.com/coretec/parametrisierte_views_sql_macros*

[4] Klyte, Tom, On Caching and Evangelizing SQL, *https://blogs.oracle.com/oraclemagazine/on-caching-and-evangelizing-sql*

[5] Houri, Mohamed, Scalar subquery caching: the select from

Full presentation

	Centisec	Ratio to Pure_SQL time			
	Pure_SQL	Plain_Plsql	UDF_Plsql	With_Clause	With_Wrapper
Num_Num	71	13.6	2.6	3.2	4.9
Num_VC	354	4.2	1.8	1.9	2.5
Num_Date	428	11.3	11.2	7.7	8.0
VC_Num	125	10.2	10.2	2.2	3.2
VC_VC	81	15.4	15.5	3.6	5.2
VC_Date	488	9.9	9.8	6.7	7.1
Date_Num	215	7.5	7.4	2.4	3.1
Date_VC	498	4.3	4.3	2.0	2.3
Date_Date	65	18.8	19.3	3.4	5.2
BF_BF	74	16.5	2.4	3.0	4.5
BD_BD	77	15.7	2.5	3.0	4.6

Abbildung 1: Bryn Llewellyn, Ten Rules for Doing a PL/SQL Performance Experiment (Quelle: https://community.oracle.com/docs/DOC-1018914)

dual trick, *https://hourim.wordpress.com/2019/12/18/scalar-subquery-caching-the-select-from-dual-trick/*

[6] Houri, Mohamed, PUSH SUBQUERY, *https://hourim.wordpress.com/2019/12/23/push-subquery/*

[7] Feuerstein, Steven, Minimize context switches and unnecessary PL/SQL code: an example from the PL/SQL Challenge, *http://stevenfeuersteinonplsql.blogspot.com/2016/04/minimize-context-switches-and.html*

[8] Feuerstein, Steven, On Cursors, Context Switches, and Mistakes, *https://blogs.oracle.com/oraclemagazine/on-cursors-context-switches-and-mistakes*

[9] Bryn Llewellyn, Ten Rules for Doing a PL/SQL Performance Experiment, *https://www.youtube.com/watch?v=9HCMgUHy5O8*

Lothar Flatz
l.flatz@bluewin.ch

DOAG Online

Auf dem deutschsprachigen Oracle-Blog ist die März-Ausgabe der News-Serie erschienen.

Das sechsköpfige Redaktionsteam von Oracle Deutschland hat wieder Neuigkeiten rund um die Datenbank zusammengetragen und in einem rund 15-minütigen Video sowie einem dazugehörigen PDF aufbereitet.

Was hat sich getan im Cloud- und On-Premises-Umfeld für Datenbank-Administratoren und Entwickler? Welches sind die aktuellen Patches und Release-Updates? Was gibt es an aktuellen Postings und Videos zur Oracle-Datenbank?

Welche Termine gibt es in den nächsten Monaten? Diesmal wieder mit zusätzlichem Quick Link Posting (in Englisch). *https://www.doag.org/de/home/news/oracle-datenbanken-monthly-news-8/*

Migrationen mit Oracle: Szenarien und Lösungen

Dierk Lenz, Herrmann & Lenz Services

Migration als Thema im Oracle-Umfeld ist – wie in praktisch allen anderen Software-Bereichen ebenfalls – ein immer aktuelles Thema. Nicht zuletzt, weil ältere Versionen ihren Support-Status verlieren und man daher gezwungen ist, eine aktuelle Version einzusetzen.

Die aktuelle Situation

Das Hauptziel für aktuelle Migrationen ist die Oracle Database 19c. Ältere Versionen sind entweder bereits vollständig ohne Support oder kurz davor. Oracle Database 21c hat als „Innovation Release" zwar die gleiche Laufzeit wie 19c, nämlich April 2024. Allerdings gibt es für Innovation Releases keinen Extended Support. 19c hingegen bietet als „Long Term Release" die Möglichkeit, bis April 2027 Extended Support zu erwerben.

Oracle Database 19c läuft in vielen Installationen stabil und performant. Außerdem wird das Release auch bei vielen Software-Anbietern unterstützt.

Mit 19c sind zudem sowohl die traditionelle Non-CDB-Architektur als auch die seit 12c alternativ mögliche CDB-Architektur mit Pluggable Databases möglich (Stichwort „Multitenant Database"). Falls möglich, sollte eine Migration auch dazu genutzt werden, in die CDB-Architektur zu wechseln. Dieser Schritt ist mit allen Versionen nach 19c alternativlos. Daher: lieber jetzt mit Fallback-Möglichkeit!

Gründe für eine Migration

Die Gründe für eine Migration sind vielfältig. Neben dem Support-Status der aktuell verwendeten Version sind oft Hardware-Neuanschaffungen, Betriebssystem-Upgrades oder Plattformwechsel, Anwendungs-Upgrades oder Zeichensatzänderungen der Grund. Häufig ist auch eine Kombination dieser Gründe ausschlaggebend.

Zu beachten ist bei einer Migration immer auch die gesamte Infrastruktur. Passen die geplanten Versionen von Datenbank-Software und Betriebssystem zusammen? Wurde daran gedacht, die Client-Versionen anzupassen?

Migrationswege

Durch diese Rahmenbedingungen werden auch Migrationswege bestimmt. Bei einem Plattformwechsel oder einer Zeichensatzmigration ist eine In-Place-Migration oft nicht möglich. Stattdessen wird

eine Migration über Export/Import erforderlich.

Im Allgemeinen ist eine In-Place-Migration die schnellste Möglichkeit zur Migration. Hierbei wird letztendlich das Data Dictionary inklusive der installierten Komponenten wie zum Beispiel Java auf die neue Version migriert. Die Inhalte der Datenbank werden dabei nicht betrachtet, sodass die Dauer einer In-Place-Migration keine Abhängigkeit von der Datenbankgröße hat.

Neben der manuellen Möglichkeit zur In-Place-Migration gibt es mit dem AutoUpgrade Tool eine komfortable Möglichkeit zur Automation. Diese ist insbesondere zu empfehlen, wenn diverse Datenbanken zur Migration anstehen. Es besteht u.a. die Möglichkeit, von der Non-CDB- in die CDB-Architektur zu migrieren. Weitere Informationen hierzu sind in folgendem MOS-Dokument zu finden: AutoUpgrade Tool (Doc ID 2485457.1).

Die Migration mittels Export/Import bietet sich immer dann an, wenn strukturelle Änderungen in der Datenbank erfolgen sollen. Das kann beispielsweise die Änderung der Datenbankblockgröße, der Wechsel von Smallfile zu Bigfile Tablespaces oder der von Bascifile zu Securefile LOBs sein. Allerdings ist beim Export/Import immer die erforderliche Auszeit zu beachten, die bei großen Datenbanken überproportional wächst. Dies ist im Neuaufbau sämtlicher Indizes begründet. Andererseits bekommt man durch den Import eine frisch reorganisierte Datenbank.

An dieser Stelle eine Anmerkung: Die hier angegebenen Möglichkeiten sind nicht vollständig. Es gibt eine recht große Menge an zusätzlichen Möglichkeiten, etwa Rücksicherungen mit dem Recovery Manager inklusive Plattformwechsel oder Transportable Tablespaces. All diese Möglichkeiten können gegebenenfalls noch miteinander kombiniert werden.

Migration? Bitte testen!

Bei einer Migration sollte nicht nur der Ablauf der Migration selbst, sondern auch das Ergebnis getestet werden! Es geht darum, dass die migrierte Datenbank von den Anwendungen genutzt werden kann und dass die Antwortzeiten „unter Kontrolle“ sind beziehungsweise keine Ausreißer festzustellen sind. Daher sollte man nicht als Erstes mit der Migration der produktiven Datenbank loslegen, sondern zum Beispiel zunächst einen Clone der Datenbank auf einem Test-Server migrieren.

Die erforderlichen Schritte zur Migration sind dabei in Skripte zu gießen. Das stellt sicher, dass bei der produktiven Migration exakt die gleichen Schritte verwendet werden wie beim Test. Damit werden Überraschungen bei der produktiven Migration vermieden.

Viele der möglichen Probleme bei einer Migration fallen recht schnell auf, zum Beispiel wenn Passwörter mit alten Kodierungen nicht mehr genutzt werden können. Andere sind eher versteckt im Hintergrund: Wer etwa noch alte DBMS_JOB-Datenbank-Jobs hat, behält diese zwar, sie werden aber nun vom Scheduler ausgeführt und nicht mehr vom „alten Code“. Daher sollten auch diese Jobs genau unter die Lupe genommen werden.

Obwohl 19c also „eigentlich nur ein 12.2er Release“ ist: Tests sind erforderlich!

GOTO <PDB>

Es wurde hier und an vielen anderen Stellen bereits oft erwähnt: Der Wechsel zur CDB-Architektur ist genau mit 19c sehr empfehlenswert. Neben den bekannten Argumenten noch eines, das insbesondere mit der Migration zu tun hat: Das Kommando zum Erstellen einer Pluggable Database als Clone hat mit der Klausel FROM non$cdb@<dblink> die Möglichkeit, als Quelle eine Non-CDB zu ziehen. Die Erfahrung zeigt, dass dies am besten mit Quellversionen ab 12.2 funktioniert.

Die Kopiergeschwindigkeit ist vergleichbar mit dem DUPLICATE ... FROM ACTIVE DATABASE beim RMAN; im Anschluss ist neben dem Upgrade der PDB lediglich ein weiteres Skript (noncdb_to_pdb.sql) auszuführen. Dies ist oft ein guter Start in die CDB-Welt!

Werkzeuge zur Unterstützung

Oracle liefert mit der Datenbank-Software umfangreiche Tools und Hilfsmittel aus, weitere Werkzeuge wie AutoUpgrade sind auf My Oracle Support verfügbar. Zudem gibt es aber auch Werkzeuge von anderen Herstellern, die in vielen Situationen weiterhelfen können.

Insbesondere wenn bereits das HL Monitoring Module und/oder Dbvisit Standby genutzt wird, kann ODBMotion von der Herrmann & Lenz Solutions GmbH genutzt werden, um mit wenigen Klicks Datenbanken zu verschieben oder eine Test-DB zu erstellen. Mit der Test-DB können dann die Migration und weitere Schritte getestet werden.

Bei Struktur- und/oder Plattformwechseln entsteht mit klassischen Mitteln oft eine immens lange Auszeit. Diese kann durch den Einsatz von logischen Replikationswerkzeugen auf wenige Minuten verkürzt werden. Insbesondere mit dem Werkzeug Quest Shareplex haben wir in vielen Projekten sehr gute Erfahrungen gemacht.

Fazit

Bei einer Migration sind viele Aspekte zu beachten! Gute Planung und umfangreiche Tests sind die Basis für eine erfolgreiche Datenbankmigration. Eines ist jedoch sicher: Es wird nicht die letzte sein!

Quellen

[1] Auf dem Papier gibt es den Zustand „ohne Support“ bei Oracle nicht. Es gilt der „Sustaining Support“, was lediglich bedeutet, dass alte Patches und Support-Dokumente verfügbar bleiben.

[2] Für Unicode-Migrationen bietet Oracle den Database Migration Assistant for Unicode an, der gegebenenfalls eine In-Place-Konvertierung der Datenbank in Unicode ermöglicht.

Dierk Lenz
dierk.lenz@hl-services.de

Pro-aktive Performance-Analyse in PostgreSQL

Dirk Krautschick, Trivadis Germany

Eine alltägliche Aufgabe von Datenbank-Administratoren ist die Analyse von Performance-Problemen. Häufig kann dann nicht auf ein bestimmtes SQL Statement fokussiert werden, sondern die Probleme werden unpräzise gemeldet. Wenn das Problem nicht reproduziert werden kann, muss mit präventiven Monitoring- und Analyse-Werkzeugen gearbeitet werden. Oracle bietet für Datenbanken beispielsweise mit dem „Diagnostic Pack" einen umfassenden Werkzeugkasten, um solche Probleme zu analysieren. Beim aktuellen Trend, Oracle-Datenbanken nach PostgreSQL zu migrieren, stellt sich häufig die Frage, ob es bei PostgreSQL ähnliche Hilfsmittel gibt.Dieser Artikel zeigt, wie und mit welchen Tools beziehungsweise Erweiterungen man sich bei PostgreSQL auf Performance-Probleme vorbereiten und in der Praxis vorgehen kann.

Das Performance-Problem

Es gibt kaum einen Datenbank-Administrator, der nicht schon mit Performance-Problemen zu tun hatte. Und hier muss man auch keinen Unterschied zwischen den unterschiedlichen Datenbank-Plattformen machen.

Der erste Schritt besteht darin, das Problem zu identifizieren und einzugrenzen. Dies ist häufig eine schwierige Aufgabe. Die Motivation für detaillierte Beschreibungen sind bei den geneigten und vermutlich genervten Nutzern eher gering und der Datenbank-Administrator muss oft Informationen, wie beispielsweise Beobachtungszeiten oder genauere Symptome, einfordern. Angaben zu den letzten Aktivitäten vor Beginn des Problems traut man sich schon meistens gar nicht zu erfragen. Ist diese Hürde genommen, kann der DBA mit der Analyse beginnen.

Die typischen Ursachen sind nicht optimale Parametrisierungen der Datenbank, eine grundlegende Veränderung des Applikationsverhaltens oder der entsprechenden Daten. Solche Probleme sind auch nach Migrationen häufig ein Thema.

Wie geht man also mit einer Aussage, wie zum Beispiel „aktuell ist alles langsam…", um? Oder was macht man, wenn es heißt: „…gestern um 13.30 Uhr war alles langsam!"

Prävention

Die erste Frage lautet immer, wie man vorbeugend gegen Performance-Probleme vorgehen kann. Die rudimentärste und kostenintensivste Methode wäre hier die scherzhaft genannte Methode „Kill it with Iron" (kurz KIWI, analog auch „Kill it with idle") , die einfach nur aussagt, möglichst viele Hardware-Ressourcen bereitzustellen. Aber auch so kann man nie weitere Probleme ausschließen, obwohl man tief in den Geldbeutel gegriffen hat.

Man sollte dieses Geld besser in eine umfangreichere Testphase – insbesondere während und nach Migrationen – investieren. Diese Tests sollten auch in enger Zusammenarbeit mit Anwendern, Entwicklern und Administratoren durchgeführt werden.

Bei PostgreSQL gibt es keine integrierten Features, um Performance-Problemen automatisch entgegenzuwirken. Hier hat Oracle mit adaptiven Features oder dem Konzept der Autonomous Databases bereits Hilfsmittel integriert.

Einzig ein durchdachtes Sizing der Datenbank-Konfiguration ist eine Möglichkeit, Performance-Problemen entgegenzuwirken. PostgreSQL ist per Default sehr leichtgewichtig konfiguriert und muss daher den reellen Bedürfnissen angepasst werden. Außerdem sollten regelmäßige Reviews der Konfiguration auch den Wandel der Datenbank-Nutzung berücksichtigen. Wenn keine Spezifikationen der Anwendung vorhanden sind, gibt es zu den wichtigen Parametern aus der Praxis empfohlene Initialwerte. Noch einfacher bieten Anbieter im Internet Konfigurations-Tools an, die basierend auf diversen individuellen Settings und Best-Practice-Settings die Konfiguration generieren. Zwei gute Beispiele sind der PGConfigurator von Cybertec *[1] (siehe Abbildung 1)* oder PGTune *[2]*.

Das Handwerk – Statement-basierte Analyse

Die granulare Methode der Untersuchung ist bei allen Datenbanken die Analyse der Ausführungspläne von SQL-Abfragen. Diese Pläne werden in der Regel von einem Teil der Datenbank-Instanz (z.B. bei Oracle der Optimizer oder bei PostgreSQL der Query Planner) basierend auf gesammelten Statistiken, Metadaten und Einstellungen generiert. Die meisten relationalen Datenbanksysteme bieten eine Möglichkeit an, einen Ausführungsplan darzustellen.

Bei PostgreSQL kann man mit dem Befehl EXPLAIN *(siehe Listing 1)* den Ausführungsplan einer SQL-Abfrage erzeugen, ohne dass die Abfrage ausgeführt wird. Es wird lediglich ein erwarteter Plan erzeugt *[3]*.

Wenn man mehr Details benötigt oder die Ergebnisse aus echten Ausführungen analysieren möchte, kann man den EXPLAIN-Befehl bei PostgreSQL durch Beifügen des Parameters ANALYZE erweitern. In der Dokumentation finden sich auch weitere Einstellungen zum Befehl EXPLAIN, mit denen man sowohl die Details als auch die Ausgabeformate variieren kann. Da bei neuen Releases gerne neue Funktionalitäten hinzugefügt werden, hilft auch ein regelmäßiger Blick in diese Dokumentation. Im *Listing 2* findet man eine vollständige Ausgabe des EXPLAIN-Befehls aus dem *Listing 1* mit allen aktuellen Optionen für das PostgreSQL Release 14. Wichtig ist, dass bei der ANALYZE-Angabe die Abfrage auch ausgeführt wird. Dies sollte bei Schreiboperationen oder sehr Last-intensiven Leseabfragen berücksichtigt werden. Im Zweifelsfall gilt: Testumgebung nutzen.

Mit einem vollständigen Ausführungsplan kann man dann die Untersuchung beginnen und Schlüsse darüber ziehen, ob gegebenenfalls Indizes fehlen, das Statement nicht optimal geschrieben ist oder ob es andere Ursachen für eine langsame Ausführung gibt. Häufig erkennt man auch, dass die Erfassung von Statistiken nicht hinreichend ist und man hier Veränderungen vornehmen muss.

Je nach Umfang einer SQL-Abfrage ist das Lesen solcher Pläne sehr mühsam und komplex. Bei großen Plänen kann man nur Block-basiert und sukzessive vorgehen. Ich empfehle im speziellen Fall von PostgreSQL dazu webbasierte grafische Aufbereitungstools, die hilfreich für eine übersichtliche Untersuchung sind. Gute Beispiele dafür sind das Tool von der Firma Dalibo *[4] (siehe Abbildung 2)* oder das von Tatyants *[5]*.

```
scott_db=# explain select e.name, d.name from scott.dept d,scott.emp
e where d.deptno=e.deptno;
                              QUERY PLAN
---------------------------------------------------------------------
 Hash Join    (cost=1.09..2.28 rows=14 width=84)
   Hash Cond: (e.deptno = d.deptno)
   ->  Seq Scan on emp e (cost=0.00..1.14 rows=14 width=42)
   ->  Hash    (cost=1.04..1.04 rows=4 width=50)
         ->  Seq Scan on dept d (cost=0.00..1.04 rows=4 width=50)
(5 rows
```

Listing 1: SQL-Ausführungsplan in PostgreSQL mit EXPLAIN

Abbildung 1: Cybertec PostgreSQL Configurator (Quelle: Cybertec)

```
scott_db=# explain (analyze, verbose, buffers true, wal true, settings
true, format text)
          select e.name, d.name from scott.dept d,scott.emp e where
d.deptno=e.deptno;
                                                        QUERY PLAN
------------------------------------------------------------------------------
-----------------------------------------------
 Hash Join    (cost=1.09..2.28 rows=14 width=84) (actual
time=1.168..1.197 rows=14 loops=1)
   Output: e.name, d.name
   Inner Unique: true
   Hash Cond: (e.deptno = d.deptno)
   Buffers: shared hit=3 read=2
   I/O Timings: read=1.036
   ->  Seq Scan on scott.emp e   (cost=0.00..1.14 rows=14 width=42) (ac-
tual time=0.714..0.720 rows=14 loops=1)
      Output: e.empno, e.name, e.job, e.manager, e.hiredate, e.salary,
e.comm, e.deptno
      Buffers: shared read=1
          I/O Timings: read=0.692
   ->  Hash  (cost=1.04..1.04 rows=4 width=50) (actual time=0.373..0.374
rows=4 loops=1)
          Output: d.name, d.deptno
          Buckets: 1024  Batches: 1    Memory Usage: 9kB
          Buffers: shared read=1
          I/O Timings: read=0.344
          ->  Seq Scan on scott.dept d    (cost=0.00..1.04 rows=4
width=50) (actual time=0.357..0.359 rows=4 loops=1)
                Output: d.name, d.deptno
                Buffers: shared read=1
                I/O Timings: read=0.344
 Planning:
   Buffers: shared hit=91 read=20
   I/O Timings: read=14.577
 Planning Time: 23.734 ms
 Execution Time: 3.090 ms
```

Listing 2: EXPLAIN ANALYZE mit allen Optionen von PostgreSQL 14

Häufig ist es allerdings nicht möglich, Performance-Probleme an einem bestimmten Statement festzumachen oder das betroffene Statement klar zu identifizieren. Es ist auch nicht immer klar, ob es sich um ein einmaliges oder häufigeres Problem handelt. Dazu kommt auch die Schwierigkeit, bestimmte Situationen im Nachgang überhaupt analysieren zu können, da zum Erhalt der Ablaufpläne das genaue Szenario reproduzierbar sein beziehungsweise reproduziert werden muss. Nicht immer hat man die Voraussetzungen, dass exakt gleiche Problem in einem nahezu identischen Testsystem nachvollziehen zu können.

Wie wird man pro-aktiv?

Die Idee hinter pro-aktiver Performance-Analyse besteht darin, auf etwaige Probleme so gut wie möglich vorbereitet zu sein. Das bedeutet, dass ein Administrator in der Lage sein sollte, eine hohe Last und die möglichen Ursachen selbst zu erkennen, bevor es der Anwender meldet. Wenn dann doch solche Beschwerden durch die Nutzer auftreten, sollte es auch unnötig sein, die Fälle reproduzieren zu müssen.

Um dies zu realisieren, muss es möglich sein, dass sowohl aktuelle als auch vergangene Details zur Last und zu den Abfragen in der Datenbank gesammelt und vorgehalten werden. Im Idealfall sollten also zu jeder Abfrage Informationen zu Laufzeiten, Statistiken, Meta-Informationen und im Idealfall auch noch der zugehörige Ausführungsplan gesammelt werden.

Externe Überwachungslösungen können hier Abhilfe schaffen. Das hieße jedoch, dass auch die oben genannten Informationen protokolliert werden müssen. Sonst erhält man am Ende nur Meldungen darüber, dass ein Performance-Problem vorliegt, ohne dass man die Gründe sichten kann. Solche Lösungen müssen dann auch in die administrativen Tabellen und Views der Datenbanksysteme Einsicht haben und abgreifen können.

Ein gutes Open-Source-Beispiel speziell für die Überwachung von PostgreSQL ist PGWatch *[6]*.

Ein weiterer Lösungsansatz bei PostgreSQL ist das sehr umfangreich konfigurierbare Logging-Verhalten. Es ist möglich, sehr viele Informationen zu protokollieren und diese dann auszuwerten. PostgreSQL ermöglicht es, jedes Statement inklusive Ausführungsplan, Laufzeitstatistiken und vieles mehr in die Protokollierung mit aufzunehmen. Außerdem können auch optional Sampling-Raten oder Mindestlaufzeiten definiert werden, um die Menge der Protokollierung einzugrenzen. Insbesondere Letzteres empfiehlt sich, um wirklich nur lang laufende Statements zu erfassen, die in der Regel die Auslöser von Performance-Problemen sind.

Der große Nachteil ist die Verwaltung der Logfiles, die bei vielen aktivierten Details und hoher Last sehr groß werden können. Damit wird nicht nur zusätzliche Last auf dem Storage erzeugt, sondern es muss auch auf den verfügbaren Speicherplatz geachtet werden. Diese Dinge müssen berücksichtigt werden, da PostgreSQL selbst davon abhängig ist und es keine integrierte Überwachung gibt.

Außerdem ist die Auswertung von sehr großen Logfiles unhandlich und oft umständlich. Das Open-Source-Tool PGBADGER *[7]* bietet hier eine gute Möglichkeit, aus PostgreSQL Logfiles einen übersichtlichen und auswertbaren Report zu erzeugen *(siehe Abbildung 3)*.

Grundsätzlich bietet PostgreSQL standardmäßig leider keine nachhaltige Sammlung solcher Informationen ohne Hilfsmittel an. Externes Monitoring oder der Umweg über das Logging-Verfahren sind zwar mögliche Optionen, allerdings nicht immer optimal oder nicht immer verfügbar. Oracle hat hier mit dem Diagnostic und Tuning Pack oder mit dem immer verfügbaren Statspack gute integrierte Lösungen, mit denen man Performance-Analysen durchführen kann.

PostgreSQL hingegen kann zwar sämtliche vergleichbare Informationen bereitstellen, aber wie findet man diese

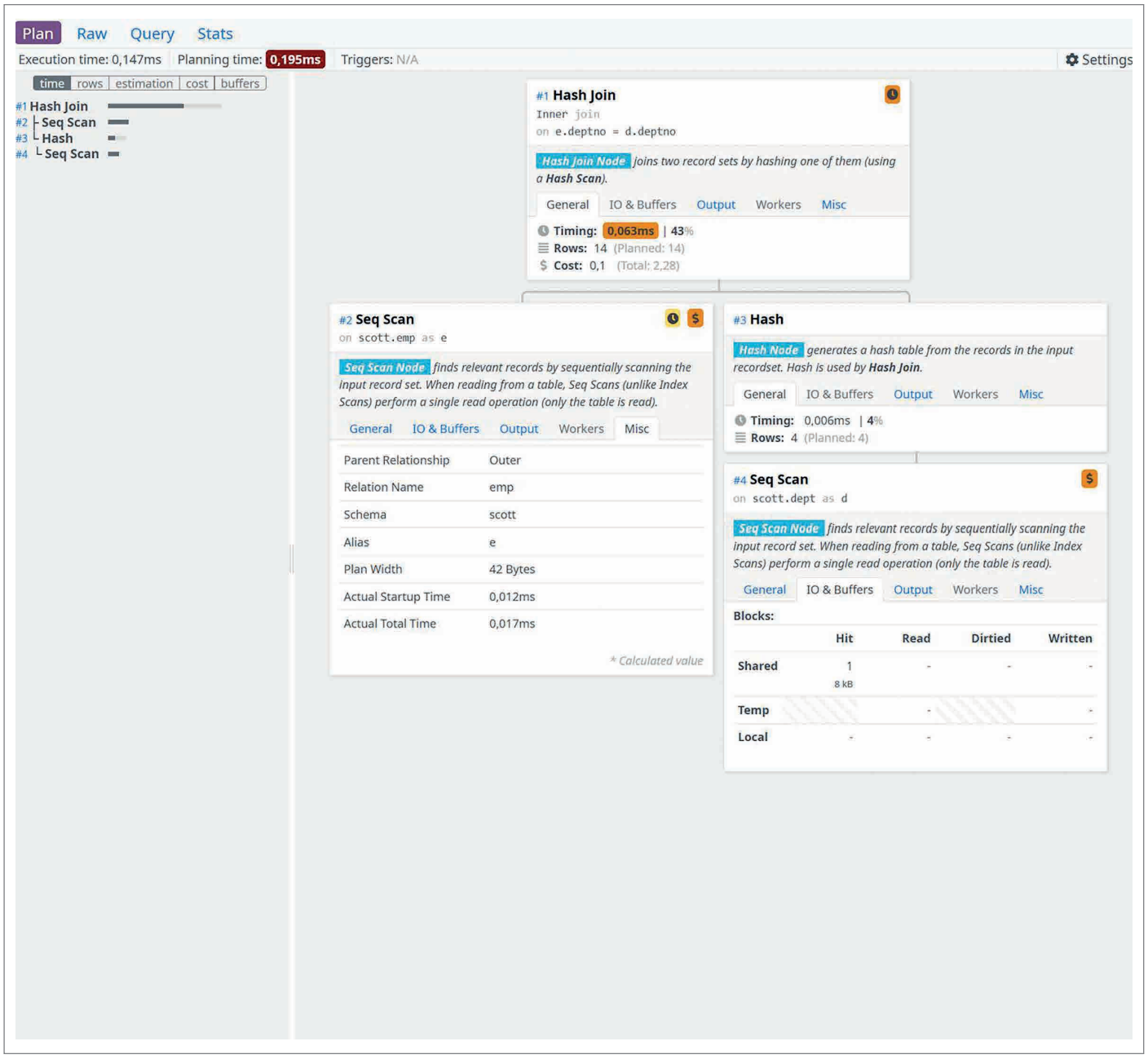


Abbildung 2: Beispiel eines Explain-Plans visualisiert mit explain.dalibo.com (Quelle: Dalibo)

Informationen und wie kann man sie am besten nachhaltig aufbewahren?

PostgreSQL ist ein leichtgewichtiges Datenbanksystem. Viel Funktionalität kann mit Erweiterungen leicht hinzugefügt werden. Dies sorgt dafür, dass die Datenbank-Software nie überfrachtet und unhandlich ist, jedoch trotzdem mit der Erweiterbarkeit ein mächtiges Feature enthalten ist, um die Datenbank funktional immer den eigenen Bedürfnissen anzupassen.

Auch bei der Analyse von Performance-Problemen existieren einige Erweiterungen, die PostgreSQL-Werkzeuge, Funktionen und Views bereitstellen, die stellenweise an die Konzepte des Oracle Diagnostic und Tuning Pack erinnern.

PG_STAT_STATEMENTS

Eine der wichtigsten Erweiterungen ist PG_STAT_STATEMENTS *[8]*. Diese ist im offiziell von der Community verwalteten CONTRIB-Paket enthalten und kann damit leicht installiert werden. Nach der Aktivierung werden – aggregiert pro SQL-Abfrage – detaillierte Statistiken gesammelt, die in einer View abgefragt werden können. *(siehe Listing 3)*

Die Anzahl der gesammelten Abfragen ist standardmäßig auf 5000 begrenzt, damit der Einfluss auf den Datenbankbetrieb gering bleibt. Diese Anzahl lässt sich mit dem Parameter pg_stat_statement.max in der PostgreSQL-Konfiguration verändern. Weitere mögliche Parameter findet man in der Dokumentation. So lässt sich zum Beispiel auch mit dem Parameter pg_stat_statements.track einstellen, ob verschachtelte Abfragen einzeln berücksichtigt werden. Dies kann besonders bei Abfragen innerhalb von Stored Procedures wichtig sein.

Mit dieser Erweiterung kann man auch im Nachgang eine Last-Analyse be-

```
# show shared_preload_libraries;
  shared_preload_libraries | pg_stat_statements
# create extension pg_stat_statements;
# \d pg_stat_statements
                    View "public.pg_stat_statements"
         Column     |          Type      | ...
--------------------+--------------------+-...
 userid             | oid                | ...
 dbid               | oid                | ...
 queryid            | bigint             | ...
 query              | text               | ...
 total_plan_time    | double precision   | ...
…
 calls              | bigint             | ...
 total_exec_time    | double precision   | ...
 min_exec_time      | double precision   | ...
 max_exec_time      | double precision   | ...
 mean_exec_time     | double precision   | ...
 stddev_exec_time   | double precision   | ...
 rows               | bigint             | ...
…                                         ...
 blk_read_time      | double precision   | ...
 blk_write_time     | double precision   | ...
…
```

Listing 3: Aktivierung und Übersicht der Extension/View PG_STAT_STATEMENTS

```
# select
      substring(query, 1, 50) as short_query,
      round(total_exec_time) as total_exec_time, calls,
      round(mean_exec_time) as mean_exec_time,
      round(100 * total_exec_time / (select sum(total_exec_time) from pg_stat_statements)) as percentage
  from
      pg_stat_statements
  order by percentage desc;
  short_query                                        |  total_exec_time |   calls   |mean_exec_time|percentage
  ---------------------------------------------------+------------------+-----------+--------------+-----------
 UPDATE pgbench_branches SET bbalance = bbalance +   |             7114 |   1500    |        5     |         63
 UPDATE pgbench_tellers SET tbalance = tbalance + $  |             2506 |   1500    |        2     |         22
 copy pgbench_accounts from stdin                    |              664 |   1       |      664     |          6
 UPDATE pgbench_accounts SET abalance = abalance +   |              194 |   1500    |        0     |          2
 alter table pgbench_accounts add primary key (aid)  |              193 |   1       |      193     |          2
 vacuum analyze pgbench_accounts                     |              138 |   1       |      138     |          1
…
…
```

Listing 4: Beispiel-Abfrage für PG_STAT_STATEMENTS

treiben und jederzeit schnell die Auswahl der Abfragen generieren, die für die weitere Untersuchung relevant sind. Ein beliebtes Beispiel für eine solche Abfrage zeigt das *Listing 4*, in dem eine Auflistung von Abfragen anhand der prozentual verbrauchten Gesamtlaufzeit deutlich macht, wo die meiste Zeit verbraucht wurde.

Um sich bei einem reproduzierbaren Szenario ein klareres Bild zu machen oder um einfach einen sauberen Zustand wiederherzustellen, kann man mit der Funktion PG_STAT_STATEMENTS_RESET() auch die View zurücksetzen. Optional kann auch durch die Parameterübergabe von Datenbank, User ID oder Query ID dafür gesorgt werden, dass bestimmte Inhalte entfernt werden.

Ein bleibender Nachteil ist der fehlende zeitliche Bezug. Man sieht innerhalb der gesammelten Daten die Situation, jedoch kann man keine Korrelation zum jeweiligen Zeitpunkt herleiten.

Dennoch zeigt die Praxis, dass mit dieser View sehr effektiv lang laufende und problematische Abfragen identifiziert und genauer betrachtet werden können.

Die Erweiterung PG_STAT_STATEMENTS ist für viele PostgreSQL-Administratoren ein essenzieller Bestandteil und wird gerne als Standard bei jeder Installation aktiviert. Tatsächlich bauen auch viele Monitoring Tools und andere Erweiterungen auf den Informationen aus den PG_STAT_STATEMENTS-Views auf oder setzen diese voraus. Hier wird auch durch frequentiertes Abgreifen der Inhalte der zeitliche Bezug, der vorab als

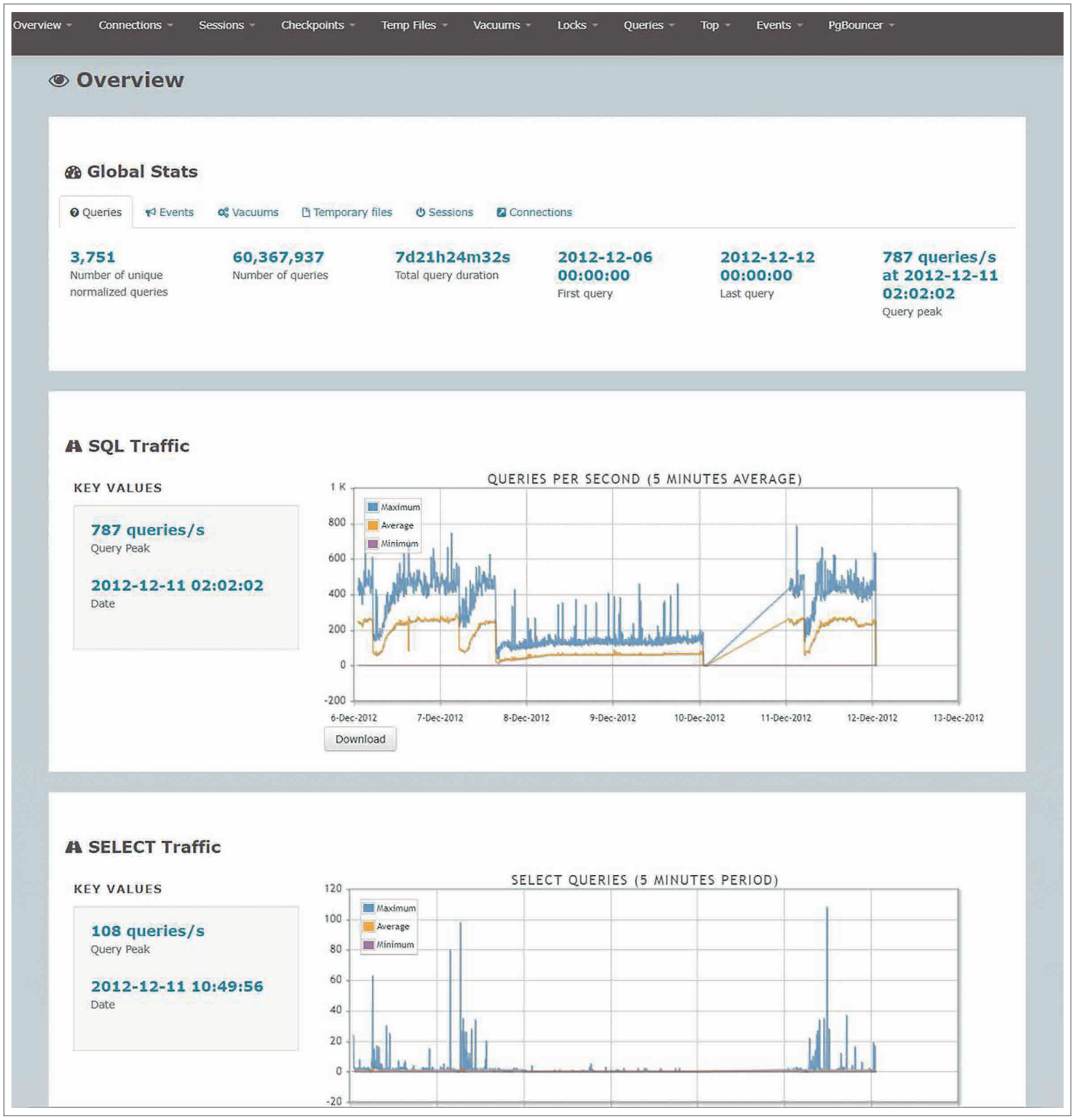


Abbildung 3: Beispiel einer Logging-Visualisierung mit PGBadger (Quelle: https://pgbadger.darold.net/samplev7.html)

Nachteil bezeichnet wurde, wiederhergestellt.

PG_WAIT_SAMPLING

Bisher haben wir gezielt nach Abfragen als grundsätzliche Ursache für Performance-Probleme gesucht. Ein anderer Ansatz ist die Analyse der Details über die verbrauchten Zeiten. Wie bei Oracle-Datenbanken gibt es bei PostgreSQL die sogenannten WAIT_EVENTS, die genau diese Details definieren. Eine Beschreibung aller WAIT_EVENTS findet man in der PostgreSQL-Dokumentation übersichtlich aufgelistet *[9]*.

Betrachtet man in PostgreSQL die View PG_STAT_ACTIVITIES (das PostgreSQL-Äquivalent zur View V$SESSION in Oracle-Datenbanken), findet man genau diese Angabe. Nur ist diese Angabe stark dynamisch und eine Momentaufnahme. Um diese Information sinnvoll nutzen zu können, muss sie auch hier wieder für die weitere Aufbereitung nachhaltig aufbewahrt werden.

Genau dies übernimmt die Erweiterung PG_WAIT_SAMPLING und stellt dazu mehrere Views bereit, um Auswertungen vorzunehmen *[10]*.

- PG_WAIT_SAMPLING_CURRENT
- PG_WAIT_SAMPLING_HISTORY
- PG_WAIT_SAMPLING_PROFILE

In der View PG_WAIT_SAMPLING_HISTORY sind alle gesammelten Informationen enthalten. Da es sich hier um ein Sampling handelt, wird mit einer definierbaren Frequenz das aktuelle WAIT_EVENT inklusive WAIT_EVENT_TYPE pro Prozess und Abfrage gelistet. Der Standardwert der Frequenz beträgt 10 ms und ist konfigurierbar. Vergleichbar mit der Erweiterung PG_STAT_STATEMENTS nutzt auch PG_WAIT_SAMPLING einen Ring-Buffer und hält standardmäßig 5000 Einträge vor. Auch dies lässt sich mit Parametern konfigurieren.

In der Praxis ist aber meist die View PG_WAIT_SAMPLING_PROFILE relevant. Hier findet man die Summe der WAIT_EVENTS in Bezug zum jeweiligen Prozess und zur Abfrage. So lässt sich explizit für eine Verbindung – wenn man diese per Prozess-ID identifiziert hat – eine vollständige Wartezeitensituation darstellen. Bei Bedarf kann hier auch noch eine Einschränkung auf eine bestimmte Abfrage hinzugefügt werden.

```
         View "public.pg_wait_sampling_profile"
  Column |   Type    |
-------------+----------+
pid          |   integer|
event_type   |   text    |
event        |   text    |
queryid      |   bigint  |
count        |   bigint  |

     View "public.pg_wait_sampling_history"
  Column     |          Type              |
-------------+----------------------------+
pid          | integer                    |
ts           | timestamp with time zone   |
event_type   | text                       |
event        | text                       |
queryid      | bigint                     |

     View "public.pg_wait_sampling_current"
  Column     |          Type              |
-------------+----------------------------+
pid          | integer                    |
event_type   | text                       |
event        | text                       |
queryid      | bigint                     |…
```

Listing 5: Übersicht über die Views von PG_WAIT_SAMPLING

Weitere Erweiterungen

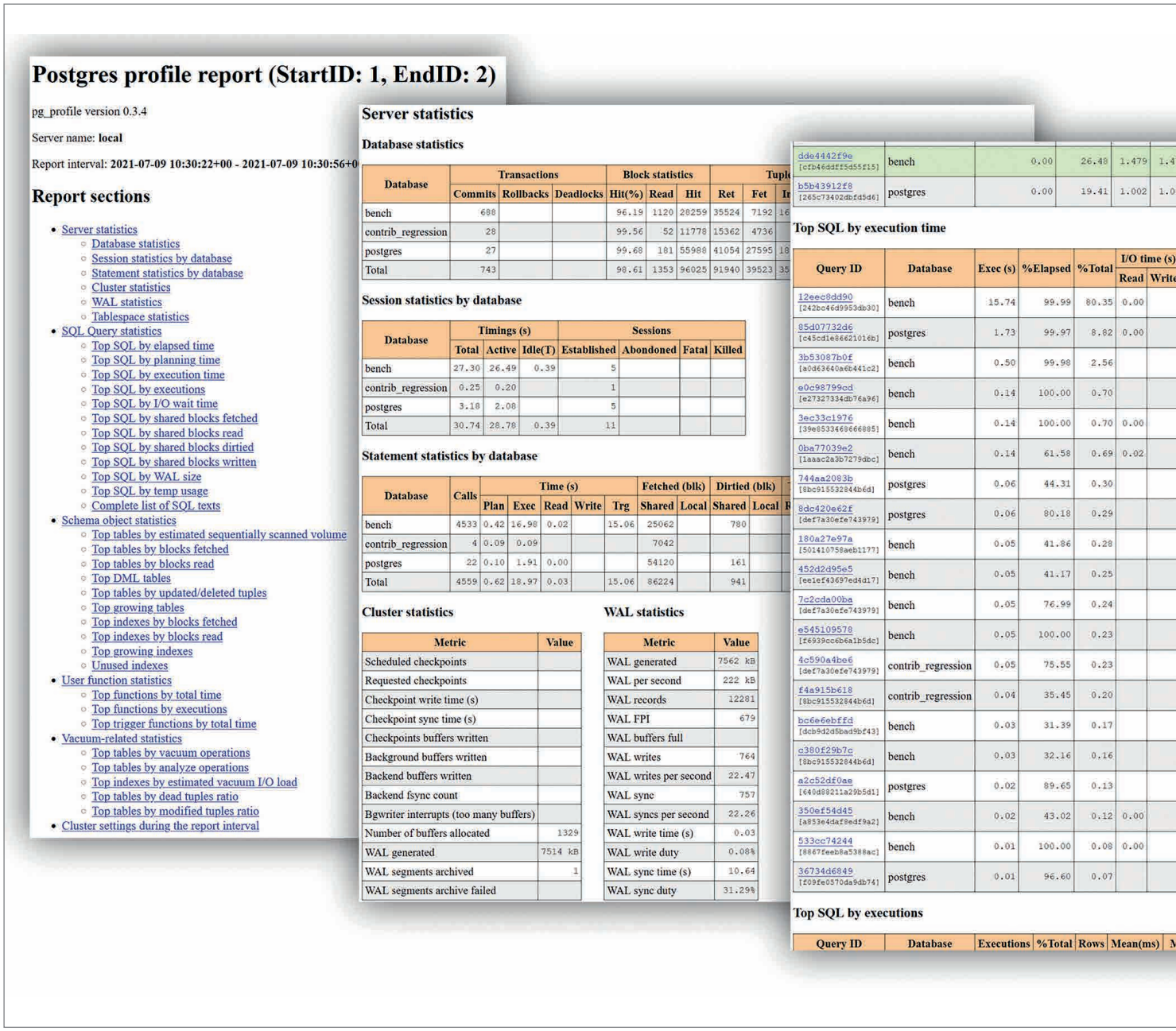
In der Praxis ist die Aktivierung von PG_STAT_STATEMENTS Standard und PG_WAIT_SAMPLING sehr empfehlenswert. Da Letzteres nicht Teil des CONTRIB- Pakets ist, sind hier häufig vor der Installation ein paar Hürden zu überwinden. Insbesondere beim Einsatz von DBaaS-Lösungen sind solche Werkzeuge nicht immer verfügbar. Da die Untersuchung der WAIT_EVENTS jedoch sehr effektiv ist und häufig zu schnellen Lösungen führt, kann man auch mit simplen Skripten zumindest zeitweise ein derartiges Tracking nachbauen und bei Bedarf einsetzen. Dies widerspricht dann zwar dem proaktiven Ansatz, ist aber dennoch hilfreich.

Beide Erweiterungen haben einen vernachlässigbaren Einfluss auf den Betrieb, wenn die parametrisierbaren Einschränkungen nicht völlig ausgehebelt werden.

Zusätzliche Erweiterungen können die Analyse-Möglichkeiten noch ergänzen. Mit PG_STAT_KCACHE werden die Inhalte der hier vorausgesetzten PG_STATS_STATEMENTS um mehr Details

```
postgres=# select
              pid,
              event_type,
              event,
              count
           from
              pg_wait_sampling_profile
           where
              pid = 3300169;
pid       |   event_type |      event           |   count
----------+--------------+----------------------+----------
3300169   |   IO         |   DataFileRead       |   63
3300169   |   IO         |   BufFileWrite       |   3620
3300169   |   IO         |   BufFileRead        |   1931
3300169   |   IO         |    DataFileExtend    |   21
3300169   |   LWLock     |   WALWrite           |   3
3300169   |   IO         |   DataFileWrite      |   1
3300169   |   IO         |   WALWrite           |   1
3300169   |   IO         |   WALSync            |   1
3300169   |   LWLock     |   WALBufMapping      |   1
3300169   |   Client     |   ClientRead         |   1
…
```

Listing 6: Beispiel-Abfrage für PG_WAIT_SAMPLING



Postgres profile report (StartID: 1, EndID: 2)

pg_profile version 0.3.4

Server name: **local**

Report interval: **2021-07-09 10:30:22+00 - 2021-07-09 10:30:56+0**

Report sections

- Server statistics
  - Database statistics
  - Session statistics by database
  - Statement statistics by database
  - Cluster statistics
  - WAL statistics
  - Tablespace statistics
- SQL Query statistics
  - Top SQL by elapsed time
  - Top SQL by planning time
  - Top SQL by execution time
  - Top SQL by executions
  - Top SQL by I/O wait time
  - Top SQL by shared blocks fetched
  - Top SQL by shared blocks read
  - Top SQL by shared blocks dirtied
  - Top SQL by shared blocks written
  - Top SQL by WAL size
  - Top SQL by temp usage
  - Complete list of SQL texts
- Schema object statistics
  - Top tables by estimated sequentially scanned volume
  - Top tables by blocks fetched
  - Top tables by blocks read
  - Top DML tables
  - Top tables by updated/deleted tuples
  - Top growing tables
  - Top indexes by blocks fetched
  - Top indexes by blocks read
  - Top growing indexes
  - Unused indexes
- User function statistics
  - Top functions by total time
  - Top functions by executions
  - Top trigger functions by total time
- Vacuum-related statistics
  - Top tables by vacuum operations
  - Top tables by analyze operations
  - Top indexes by estimated vacuum I/O load
  - Top tables by dead tuples ratio
  - Top tables by modified tuples ratio
- Cluster settings during the report interval

Server statistics

Database statistics

Database	Transactions			Block statistics			Tuple		
	Commits	Rollbacks	Deadlocks	Hit(%)	Read	Hit	Ret	Fet	In
bench	688			96.19	1120	28259	35524	7192	16
contrib_regression	28			99.56	52	11778	15362	4736	
postgres	27			99.68	181	55988	41054	27595	18
Total	743			98.61	1353	96025	91940	39523	35

Session statistics by database

Database	Timings (s)			Sessions			
	Total	Active	Idle(T)	Established	Abondoned	Fatal	Killed
bench	27.30	26.49	0.39	5			
contrib_regression	0.25	0.20		1			
postgres	3.18	2.08		5			
Total	30.74	28.78	0.39	11			

Statement statistics by database

Database	Calls	Time (s)					Fetched (blk)		Dirtied (blk)		W
		Plan	Exec	Read	Write	Trg	Shared	Local	Shared	Local	
bench	4533	0.42	16.98	0.02		15.06	25062		780		
contrib_regression	4	0.09	0.09				7042				
postgres	22	0.10	1.91	0.00			54120		161		
Total	4559	0.62	18.97	0.03		15.06	86224		941		

Cluster statistics

Metric	Value
Scheduled checkpoints	
Requested checkpoints	
Checkpoint write time (s)	
Checkpoint sync time (s)	
Checkpoints buffers written	
Background buffers written	
Backend buffers written	
Backend fsync count	
Bgwriter interrupts (too many buffers)	
Number of buffers allocated	1329
WAL generated	7514 kB
WAL segments archived	1
WAL segments archive failed	

WAL statistics

Metric	Value
WAL generated	7562 kB
WAL per second	222 kB
WAL records	12281
WAL FPI	679
WAL buffers full	
WAL writes	764
WAL writes per second	22.47
WAL sync	757
WAL syncs per second	22.26
WAL write time (s)	0.03
WAL write duty	0.08%
WAL sync time (s)	10.64
WAL sync duty	31.29%

dde4442f9e [cfb46ddff5d55f15]	bench	0.00	26.48	1.479	1.4
b5b43912f8 [265c73402dbfd5d6]	postgres	0.00	19.41	1.002	1.0

Top SQL by execution time

Query ID	Database	Exec (s)	%Elapsed	%Total	I/O time (s)	
					Read	Write
12eec8dd90 [242bc46d9953db30]	bench	15.74	99.99	80.35	0.00	
85d07732d6 [c45cd1e86621016b]	postgres	1.73	99.97	8.82	0.00	
3b53087b0f [a0d63640a6b441c2]	bench	0.50	99.98	2.56		
e0c98799cd [e27327334db76a96]	bench	0.14	100.00	0.70		
3ec33c1976 [39e85334686668b5]	bench	0.14	100.00	0.70	0.00	
0ba77039e2 [1aaac2a3b7279dbc]	bench	0.14	61.58	0.69	0.02	
744aa2083b [8bc915532844b6d]	postgres	0.06	44.31	0.30		
8dc420e62f [def7a30efe743979]	postgres	0.06	80.18	0.29		
180a27e97a [501410758aeb1177]	bench	0.05	41.86	0.28		
452d2d95e5 [ee1ef43697ed4d17]	bench	0.05	41.17	0.25		
7c2cda00ba [def7a30efe743979]	bench	0.05	76.99	0.24		
e545109578 [f6939cc6b6a1b5dc]	bench	0.05	100.00	0.23		
4c590a4be6 [def7a30efe743979]	contrib_regression	0.05	75.55	0.23		
f4a915b618 [8bc915532844b6d]	contrib_regression	0.04	35.45	0.20		
bc6e6ebffd [dcb9d2d5bad9bf43]	bench	0.03	31.39	0.17		
c380f29b7c [8bc915532844b6d]	bench	0.03	32.16	0.16		
a2c52df0ae [640d88211a29b5d1]	postgres	0.02	89.65	0.13		
350ef54d45 [a853e4daf8edf9a2]	bench	0.02	43.02	0.12	0.00	
533cc74244 [8867feeb8a5388ac]	bench	0.01	100.00	0.08	0.00	
36734d6849 [f09fe0570da9db74]	postgres	0.01	96.60	0.07		

Top SQL by executions

Query ID	Database	Executions	%Total	Rows	Mean(ms)	M

Abbildung 4: Auszug aus Beispiel-Report mit PG_PROFILE (Quelle: https://github.com/zubkov-andrei/pg_profile/tree/master/report_examples)

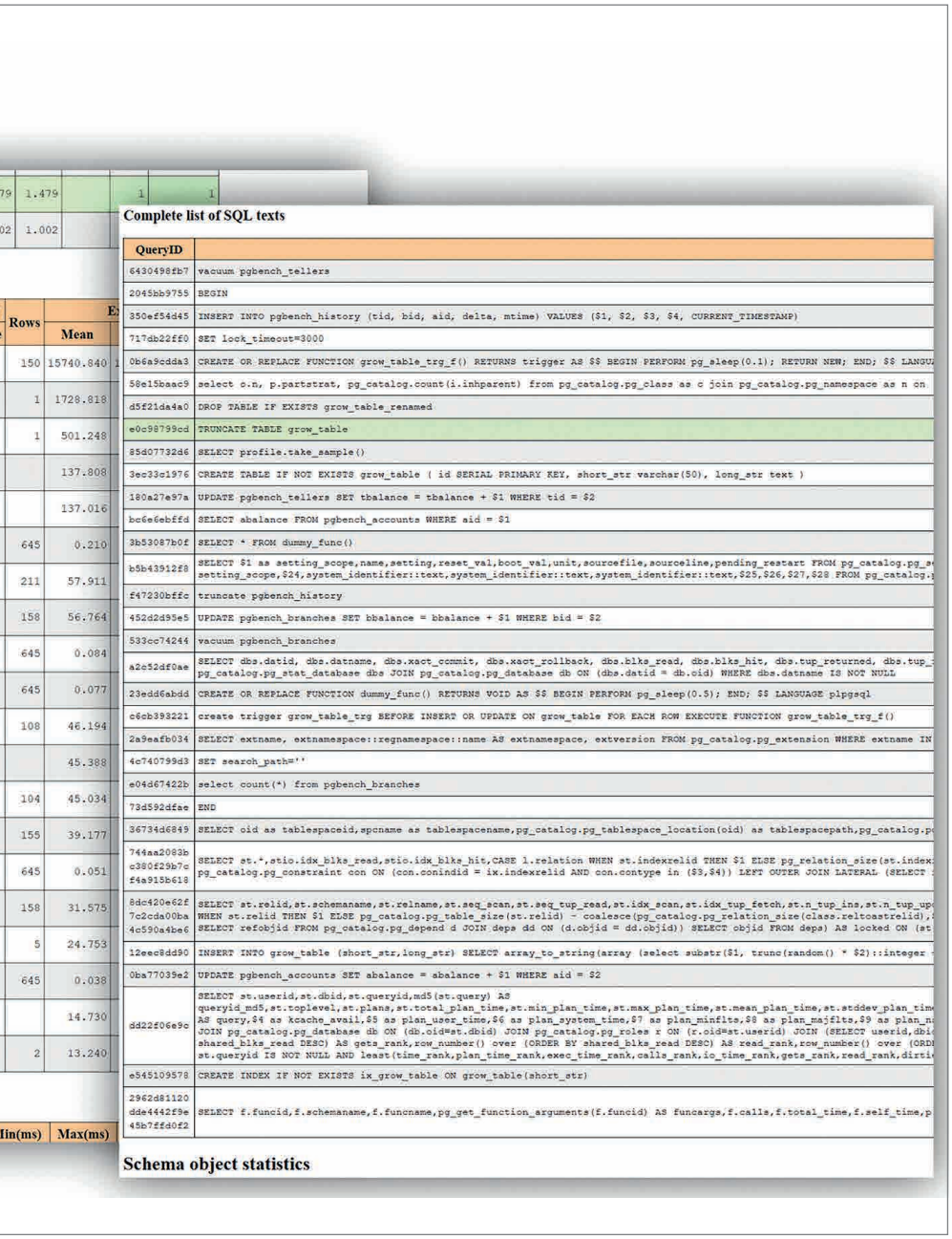
und Statistiken zu den Lese- und Schreiboperationen auf dem Dateisystem angereichert.

Um die Inhalte zu konsolidieren und die fehlende zeitliche Korrelation wieder herzustellen, können mit der etwas umfangreicheren Erweiterung PG_PROFILE *[11]* ein Snapshot-basiertes Repository angelegt und daraus Reports erzeugt werden, in denen die Inhalte von PG_STAT_STATEMENTS und PG_STAT_KCACHE übersichtlich in einer HTML-Ansicht dargestellt werden können. Das Grundprinzip erinnert sehr an das Oracle Statspack oder Oracle Automatic Workload Repository (AWR). Leider ist derzeit die Integration der Inhalte von PG_WAIT_SAMPLING in solchen Reports noch nicht möglich, aber die Idee besteht schon als mögliches Feature für die Zukunft. So könnte gegebenenfalls ein Äquivalent zu den Top 10 Wait Events aus den Reports von Oracle auch bei PostgreSQL sehr hilfreich sein. *Abbildung 4* zeigt Auszüge aus einem Beispiel-Report.

Vorgehensweisen

Die vier genannten Beispiele sind nur ein kleiner Auszug aus den Erweiterungsmöglichkeiten für Performance-Analysen, die es für PostgreSQL gibt. Zusammen mit dem Einsatz einer geeigneten Monitoring-Lösung haben sich zumindest für mich in der Praxis diese Erweiterungen als sinnvoll und bis heute ausreichend erwiesen.

Der Vorteil der Modularität mit den Erweiterungen liegt darin, dass die Wahl besteht, welche Funktionalität geeignet ist. Für kleinere und unkritische Datenbanken kann man gegebenenfalls vollständig auf solche Mittel verzichten. Intensiver genutzte Datenbanken sollten gegebenenfalls unter einer derartigen Beobachtung mit möglichst vielen gesammelten Informationen stehen. So kann man jederzeit auf Probleme oder auch pro-aktiv



Complete list of SQL texts

QueryID	
6430498fb7	vacuum pgbench_tellers
2045bb9755	BEGIN
350ef54d45	INSERT INTO pgbench_history (tid, bid, aid, delta, mtime) VALUES ($1, $2, $3, $4, CURRENT_TIMESTAMP)
717db22ff0	SET lock_timeout=3000
0b6a9cdda3	CREATE OR REPLACE FUNCTION grow_table_trg_f() RETURNS trigger AS $$ BEGIN PERFORM pg_sleep(0.1); RETURN NEW; END; $$ LANGU
58e15baac9	select c.n, p.partstrat, pg_catalog.count(i.inhparent) from pg_catalog.pg_class as c join pg_catalog.pg_namespace as n on
d5f21da4a0	DROP TABLE IF EXISTS grow_table_renamed
e0c98799cd	TRUNCATE TABLE grow_table
85d07732d6	SELECT profile.take_sample()
3ec33c1976	CREATE TABLE IF NOT EXISTS grow_table (id SERIAL PRIMARY KEY, short_str varchar(50), long_str text)
180a27e97a	UPDATE pgbench_tellers SET tbalance = tbalance + $1 WHERE tid = $2
bc6e6ebffd	SELECT abalance FROM pgbench_accounts WHERE aid = $1
3b53087b0f	SELECT * FROM dummy_func()
b5b43912f8	SELECT $1 as setting_scope,name,setting,reset_val,boot_val,unit,sourcefile,sourceline,pending_restart FROM pg_catalog.pg_s setting_scope,$24,system_identifier::text,system_identifier::text,system_identifier::text,$25,$26,$27,$28 FROM pg_catalog.
f47230bffc	truncate pgbench_history
452d2d95e5	UPDATE pgbench_branches SET bbalance = bbalance + $1 WHERE bid = $2
533cc74244	vacuum pgbench_branches
a2e52df0ae	SELECT dbs.datid, dbs.datname, dbs.xact_commit, dbs.xact_rollback, dbs.blks_read, dbs.blks_hit, dbs.tup_returned, dbs.tup_ pg_catalog.pg_stat_database dbs JOIN pg_catalog.pg_database db ON (dbs.datid = db.oid) WHERE dbs.datname IS NOT NULL
23edd6abdd	CREATE OR REPLACE FUNCTION dummy_func() RETURNS VOID AS $$ BEGIN PERFORM pg_sleep(0.5); END; $$ LANGUAGE plpgsql
c6cb393221	create trigger grow_table_trg BEFORE INSERT OR UPDATE ON grow_table FOR EACH ROW EXECUTE FUNCTION grow_table_trg_f()
2a9eafb034	SELECT extname, extnamespace::regnamespace::name AS extnamespace, extversion FROM pg_catalog.pg_extension WHERE extname IN
4c740799d3	SET search_path=''
e04d67422b	select count(*) from pgbench_branches
73d592dfae	END
36734d6849	SELECT oid as tablespaceid,spcname as tablespacename,pg_catalog.pg_tablespace_location(oid) as tablespacepath,pg_catalog.p
744aa2083b c380f29b7c f4a915b618	SELECT st.*,stio.idx_blks_read,stio.idx_blks_hit,CASE l.relation WHEN st.indexrelid THEN $1 ELSE pg_relation_size(st.index pg_catalog.pg_constraint con ON (con.conindid = ix.indexrelid AND con.contype in ($3,$4)) LEFT OUTER JOIN LATERAL ($SELECT
8dc420e62f 7c2cda00ba 4c590a4be6	SELECT st.relid,st.schemaname,st.relname,st.seq_scan,st.seq_tup_read,st.idx_scan,st.idx_tup_fetch,st.n_tup_ins,st.n_tup_up WHEN st.relid THEN $1 ELSE pg_catalog.pg_table_size(st.relid) - coalesce(pg_catalog.pg_relation_size(class.reltoastrelid), SELECT refobjid FROM pg_catalog.pg_depend d JOIN deps dd ON (d.objid = dd.objid)) SELECT objid FROM deps) AS locked ON (st
12eec8dd90	INSERT INTO grow_table (short_str,long_str) SELECT array_to_string(array (select substr($1, trunc(random() * $2)::integer +
0ba77039e2	UPDATE pgbench_accounts SET abalance = abalance + $1 WHERE aid = $2
dd22f06e9c	SELECT st.userid,st.dbid,st.queryid,md5(st.query) AS queryid_md5,st.toplevel,st.plans,st.total_plan_time,st.min_plan_time,st.max_plan_time,st.mean_plan_time,st.stddev_plan_tim AS query,$4 as kcache_avail,$5 as plan_user_time,$6 as plan_system_time,$7 as plan_minflts,$8 as plan_majflts,$9 as plan_n JOIN pg_catalog.pg_database db ON (db.oid=st.dbid) JOIN pg_catalog.pg_roles r ON (r.oid=st.userid) JOIN (SELECT userid,dbi shared_blks_read DESC) AS gets_rank,row_number() over (ORDER BY shared_blks_read DESC) AS read_rank,row_number() over (ORD st.queryid IS NOT NULL AND least(time_rank,plan_time_rank,exec_time_rank,calls_rank,io_time_rank,gets_rank,read_rank,dirti
e545109578	CREATE INDEX IF NOT EXISTS ix_grow_table ON grow_table(short_str)
2962d81120 dde4442f9e 45b7ffd0f2	SELECT f.funcid,f.schemaname,f.funcname,pg_get_function_arguments(f.funcid) AS funcargs,f.calls,f.total_time,f.self_time,p

Schema object statistics

Die hier vorgestellten Beispiele bieten aus eigener Erfahrung bereits genug, um den größten Teil aller Performance-Probleme pro-aktiv analysieren und lösen zu können.

Quellen

[1] *http://pgconfigurator.cybertec.at*
[2] *https://pgtune.leopard.in.ua/#/*
[3] *https://www.postgresql.org/docs/14/sql-explain.html*
[4] *https://explain.dalibo.com/*
[5] *https://tatiyants.com/pev*
[6] *https://pgwatch.com*
[7] *https://pgbadger.darold.net*
[8] *https://www.postgresql.org/docs/14/pgstatstatements.html*
[9] *https://www.postgresql.org/docs/14/monitoring-stats.html#WAIT-EVENT-TABLE*
[10] *https://github.com/postgrespro/pg_wait_sampling*
[11] *https://github.com/zubkov-andrei/pg_profile*

Über den Autor

Dirk Krautschick arbeitet als Berater für die Trivadis GmbH. Als Diplom-Informatiker war er für neun Jahre verantwortlicher Datenbankadministrator und Quality Support Engineer bei einem Hersteller für Optimization-Software im Luftfahrtbereich, wo er sein Know-how für den Betrieb von Applikationsservern und Oracle- beziehungsweise PostgreSQL-Datenbanken aufgebaut hat. 2017 hat er sich dann dazu entschieden, die Begeisterung für Datenbanktechnologien und seine Fähigkeiten als Consultant unter anderem im Finanz- und Energiesektor einzusetzen und stetig weiter zu vertiefen.

auf Belastungen eingehen und mit Applikations-Nutzern oder -Entwicklern in die Diskussion gehen.

Grundsätzlich bieten dann PG_STAT_STATEMENTS und PG_WAIT_SAMPLING (wenn erlaubt oder als Beispiel für die Auswertung der WAIT_EVENTS) die beiden für mich wichtigsten Säulen, mit denen aus zwei verschiedenen Richtungen untersucht werden kann.

Sollte dies aus unterschiedlichen Gründen nicht möglich sein, bietet sich als eingeschränkte, aber mögliche Option das Logging-Verhalten von PostgreSQL an. Hier sollte man jedoch zwischen einer dauerhaften und einer gegebenenfalls temporär zuschaltbaren Konfiguration unterscheiden, um den Einfluss auf den Betrieb so gering wie möglich zu halten.

Fazit

Häufig hört man viel Kritik über einen weitaus geringeren Umfang an Möglichkeiten für Performance-Analysen von PostgreSQL im direkten Vergleich zu kommerziellen Datenbanken, wie zum Beispiel dem umfangreichen Diagnostic und Tuning Pack von Oracle. Die Vielzahl an optionalen Erweiterungen wird dabei zu Unrecht ignoriert oder übersehen.

Dirk_Krautschick
dirk.krautschick@trivadis.com

Dbvisit StandbyMP – Von der Evolution zur Revolution

Rainier Kaczmarczyk, Opitz Consulting

Seit mehr als 15 Jahren ist Dbvisit Standby auf dem Oracle-Datenbank-Markt etabliert. Da Oracle Data Guard in der Oracle Standard Edition 2 (SE2) nicht verfügbar ist, bietet sich Dbvisit Standby als Lösung für SE2-Installationen an. An dieser Stelle stellt der Autor die neue Software vor und erklärt, warum das eine R(evolution) ist.

Was ist Dbvisit Standby?

Seit mehr als 15 Jahren ist Dbvisit Standby auf dem Oracle-Datenbank-Markt etabliert. Da Oracle Data Guard in der Oracle Standard Edition 2 (SE2) nicht verfügbar ist, bietet sich Dbvisit Standby als Lösung für SE2-Installationen an.

Wie funktioniert Dbvisit Standby?

Das Hauptsystem, die sogenannte Primary-Datenbank, wird 1:1 auf ein Remote-System – genannt Standby – repliziert. Sinnvollerweise befindet sich das Standby-System an einem anderen Standort, um die Verfügbarkeit im Fehlerfall (Feuer, Wasser etc.) zu gewährleisten. Auch eine hybride Lösung ist möglich, bei der sich das Standby-System in der Cloud befindet.

Das Ganze ist eine physikalische Replikation. Das bedeutet, dass Betriebssystem und Datenbankversion identisch sein müssen. Migrationen auf neue Versionen sind mit dieser Lösung nicht möglich. Wohl aber zum Beispiel das Einspielen eines Betriebssystem-Patches auf dem Standby-System. Nach einem Rollentausch, einem sogenannten Switchover, kann dann auf dem zweiten System der Patch installiert werden. Dann kann, nach einem weiteren Switchover, auf das alte Primary-System zurückgeschaltet werden. Vorteil: Schlägt der Patch-Prozess fehl, ist immer noch das Primary-System verfügbar.

Warum Evolution?

Seit der ersten Produktivsetzung wurde Dbvisit Standby kontinuierlich weiterentwickelt. Damit existierte ein sehr ausgereiftes und zuverlässiges Produkt für Oracle-Datenbanken. Das ist die Evolution.

Warum Revolution?

Dbvisit StandbyMP (MultiPlatform) unterstützt seit Anfang des Jahres neben Oracle auch Microsoft-SQL-Server! Diese neue Lösung ist in einem zentralen Webinterface für beide Datenbanksysteme integriert. Hier können alle Funktionen von Dbvisit StandbyMP gesteuert werden. Also ein Switchover sowohl einer Oracle-Datenbank als auch einer Microsoft-SQL-Server-Datenbank. Und das in parallelen Prozessen für jede einzelne der definierten Replikationen. Es kann dazu alternativ auch ein CLI (Command Line Interface) benutzt werden.

Bemerkenswertes?

Bei einer Oracle-Replikation dauert der Switchover (Rollentausch) auf einem Standard-PC ca. vier Minuten.

Mehr als beeindruckend ist, dass dieser Rollentausch bei Microsoft SQL Server keine fünf Sekunden dauert. „Hut ab“ an Microsoft!

Extras?

Für Oracle-Datenbanken unter Linux gibt es eine sehr interessante, kostenfreie Funktionalität: die Snapshot-Option. Das Problem des Standby-Systems ist, dass es im Normalbetrieb für nichts nutzbar ist. Kurz: Es kostet Strom und Lizenzgebühren.

Mit der Snapshot-Option kann das Standby-System ähnlich wie ein Oracle Active Data Guard für Abfragen als Reporting-System genutzt werden. Das Ganze basiert auf dem Linux-Filesystem, das Snapshots ermöglicht. Es gibt allerdings Einschränkungen:

- Die Datenbank muss auf EINEM logischen Filesystem installiert sein.
- Da eine zweite oder auch mehrere Instanzen mit der Snapshot-Option betrieben werden, ist der Bedarf an Speicher (Platte und Memory) deutlich höher. Dafür verfügt der Nutzer aber über eine Reporting-Datenbank, die die produktive Datenbank (Primary) nicht belastet.
- Die Daten sind nicht „Real Time“. Je nach Konfiguration gibt es einen Unterschied in der Aktualität der Daten auf der Reporting DB von wenigen Minuten bis zu einer Stunde oder auch mehr.

Fazit

Dbvisit StandbyMP ist eine kostengünstige Software, die einfach zu bedienen ist und nun zwei Datenbank-Systeme unterstützt.

MP? Also Multi? Ja, Multi bedeutet mehr als Eins. Das können jedoch auch drei oder noch mehr unterstützte Datenbankprodukte sein. Als Insider kann man sich leicht vorstellen, welche Plattformen in naher Zukunft weiter hinzukommen.

Anwender, die diverse Applikationen mit unterschiedlichen Datenbank-Systemen betreiben, haben in Zukunft eine einheitliche Oberfläche, um für alle unterstützten Systeme eine einfach zu bedienende Hochverfügbarkeit zu gewährleisten.

Über den Autor

Rainier Kaczmarczyk arbeitet seit 35 Jahren im Umfeld der Oracle-Datenbank. Darunter fünf Jahre beim Hersteller selbst. Seine Schwerpunkte sind Performance Tuning, Migrationen, Upgrades, Administration und – natürlich – Hochverfügbarkeit. Seit fast 15 Jahren ist er in diesem Bereich für Opitz Consulting tätig

Rainier Kaczmarczyk
Rainier.Kaczmarczyk@opitz-consulting.com

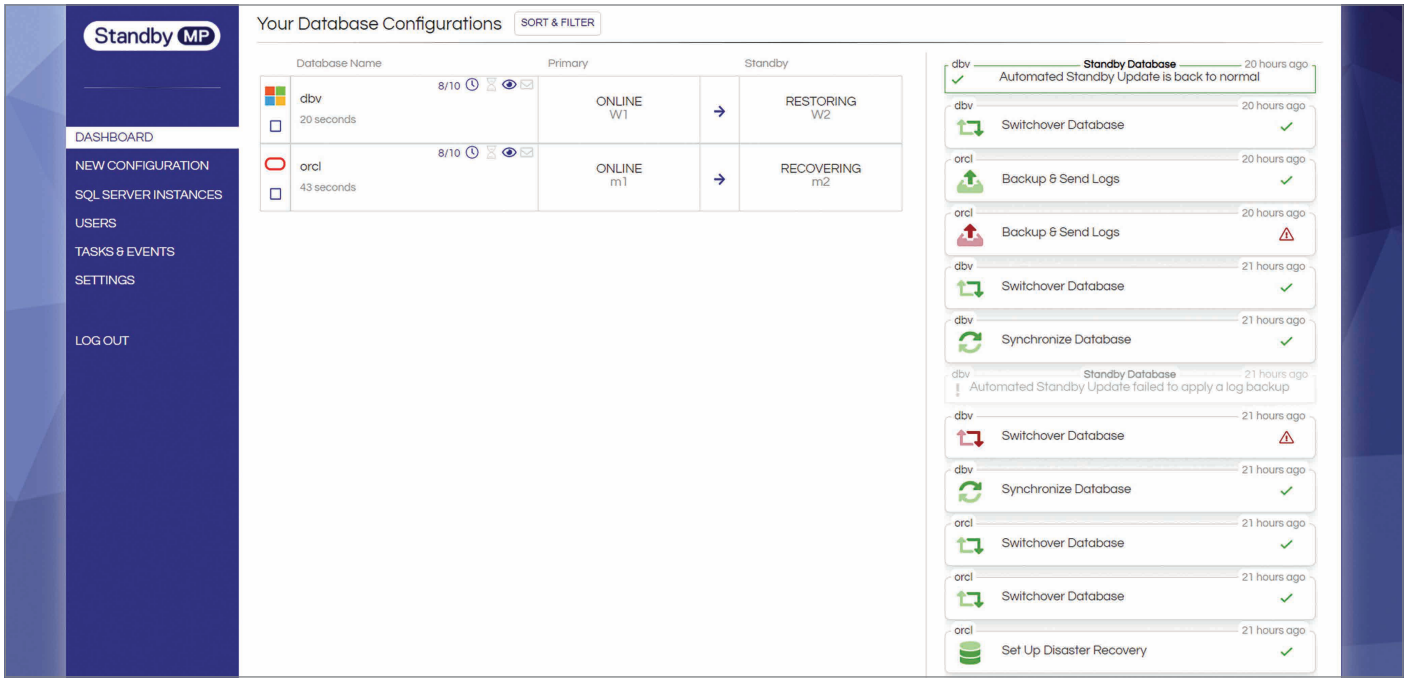
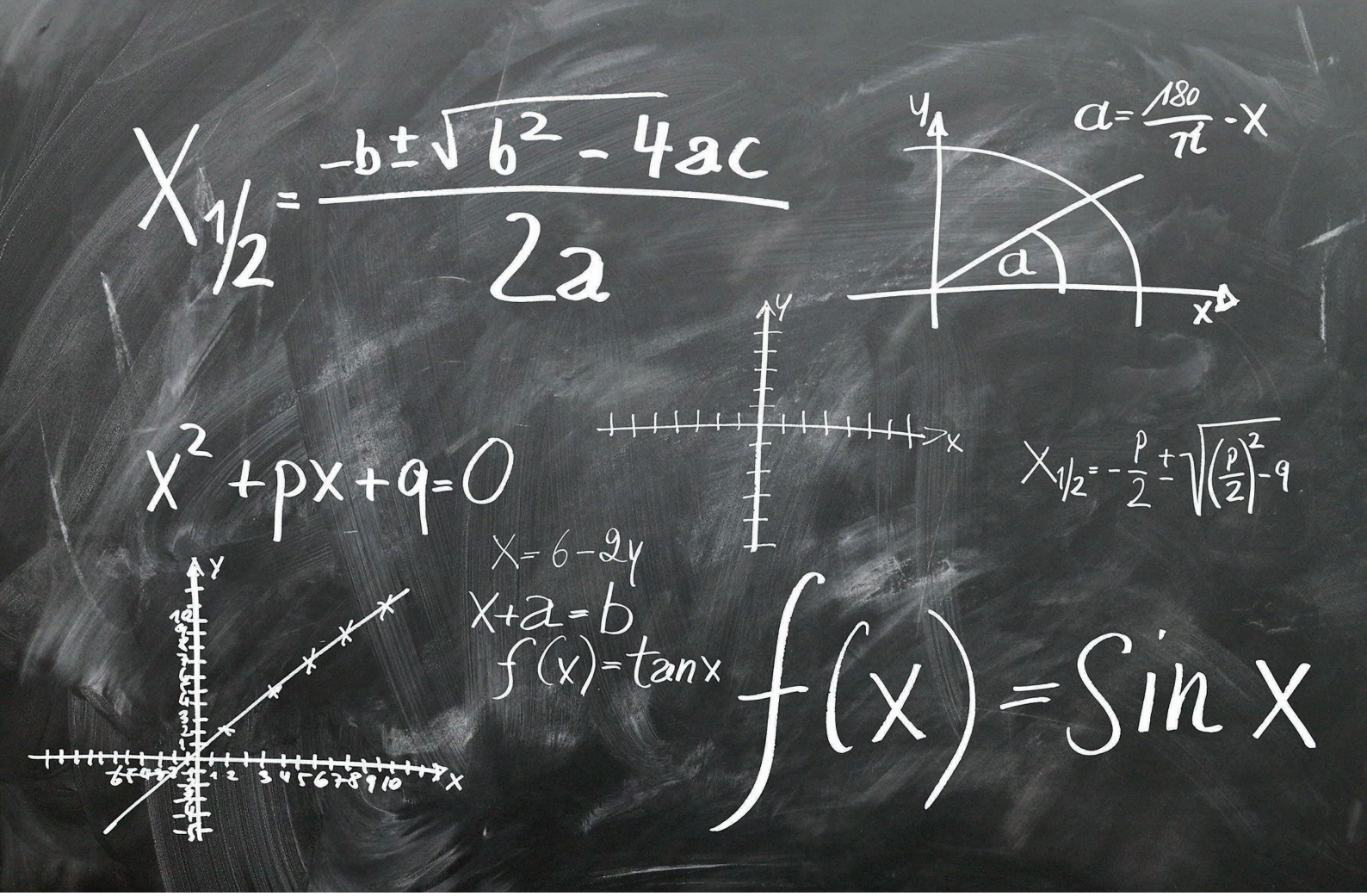


Abbildung 1: Database Configurator (Quelle: Dbvisit Software)



Deterministische Funktionen

Jürgen Sieben, ConDeS

Dieser Artikel ist ein weiterer aus einer kleinen Reihe von Fundstellen, die mich bei der Lektüre von Fachliteratur zu Widerspruch gereizt haben. Mir geht es natürlich nicht darum, mit dem Finger auf andere Autoren zu zeigen, sondern sie stellen den Anker dar, an denen ich ein Thema erläutern möchte, von dem ich glaube, dass es für viele Leser interessant sein könnte. Hier nun einige Überlegungen zur Deterministic-Klausel.

Die Fundstelle

Naja, diesmal war es keine Veröffentlichung, sondern in einer Liste von Themen enthalten, die ich von einem Kunden bekam, um eine Inhouse-Schulung vorzubereiten, die ich halten sollte. Es ist nur ein Spiegelstrich in einer längeren Liste, aber so ist das halt manchmal: Daran entzündete sich die Idee für diesen Artikel. Der Spiegelstrich lautet:

- deterministic function
- soweit ich weiß, ergibt dies keinen Sinn bei Nutzung in PL/SQL, nur beim Aufruf von SQL

Das stimmt fast, oder sagen wir besser, es hat einmal gestimmt. Doch das Thema ist vielschichtig: Was denn nun: Deterministic, Result Cache, Relies-on-Klausel … Sehen wir uns einmal den aktuellen Stand und die Optionen an.

Das Problem

Dem alten Performance-Tuner-Spruch *„How can I make things go faster? – Don't!"* folgend, bemühen wir uns bei der Programmierung natürlich, unnötige Berechnungen auch nicht durchführen zu lassen. Das ist auch in PL/SQL nicht anders. Wenn eine Funktion aufgerufen wird, die für einen gegebenen Parameter stets das gleiche Ergebnis liefern wird, wie das etwa

für die Funktionen *lower, upper* etc. gilt, sollten wir doch nicht gezwungen sein, die immer gleichen Berechnungsschritte auszuführen. Insbesondere gilt dies im Umfeld von SQL, weil hier die Funktion millionenfach ausgeführt werden kann, je nachdem, wie die Funktion aufgerufen wird und wie viele Zeilen die Tabelle enthält. Doch auch in reinem PL/SQL, in einer Schleife etwa, stellt sich das Problem. Hier vielleicht nicht durch die reine Anzahl der Aufrufe, aber zum Beispiel aufgrund der komplexen, zeitaufwendigen Berechnungen, die in der Funktion ausgeführt werden müssen.

Allerdings sollten wir uns vorab klarmachen, wie groß das Problem eigentlich ist. Durch alle Optionen, die wir im Folgenden besprechen, beschleunigen wir die PL/SQL-Funktion, indem wir unnötige Berechnungen unterbinden. Damit wir jedoch auch einen Performanz-Effekt spüren, muss in PL/SQL auch signifikant viel Arbeit erbracht werden. Die Funktion *lower()* ist insofern kein wirklich gutes Beispiel, es sei denn, Sie rufen sie viele, viele, viele (!) Male und dann auch noch mit den gleichen Parametern auf.

Lösungsansätze

Sie haben also eine Funktion, für die gilt: Sie macht erheblich Arbeit, liefert für eine gegebene Kombination von Parametern aber immer ein identisches Ergebnis. Zudem wird diese Funktion potenziell sehr häufig benötigt, und zwar nicht aus SQL heraus (dann müssen wir das Problem der Umgebungswechsel lösen), sondern aus PL/SQL heraus. Viele Wenn und Aber, doch so sind nun einmal die Regeln. Erfüllen Sie diese Regeln nicht, ist das Folgende vielleicht eher von allgemeinem Interesse, wirklich schneller wird Ihre Anwendung durch diese Möglichkeiten wohl eher nicht.

Erste Variante: Die Klausel DETERMINISTIC

Zunächst können Sie die Funktion (und nur eine Funktion, für Prozeduren gilt dies naturgemäß nicht) als deterministisch kennzeichnen, indem Sie die Klausel deterministic hinter der Parameterdeklaration einfügen *(siehe Listing 1)*.

Sie teilen damit der Datenbank mit, dass diese Funktion für die gleichen Parameter stets die gleichen Resultate liefern wird. Das ist insofern wichtig, als die Datenbank dies nicht weiter prüft, sondern sich auf Ihr Wort verlässt. Leider, das soll bereits hier gesagt werden, sind gar nicht so viele Funktionen deterministisch, wie Sie sich vielleicht wünschen: Aufrufe der Funktionen *user, dbms_random* oder *sysdate* innerhalb der Funktion verbieten sich ebenso wie Resultate, die auf der Abfrage von Tabellendaten beruhen, selbst wenn dies Stammdaten sind, die sich nicht häufig ändern. Auch die – in APEX-Kreisen allgegenwärtige – Funktion *v*, die Daten aus dem Session State liest, ist nicht deterministisch.

```
create or replace function deterministic_test(
  p_in number)
  return number
  deterministic
as
  ...
end deterministic_test;
```

Listing 1: Kennzeichnung der Funktion als deterministisch

```
SQL> create or replace package utl_timer
  2  is
  3    procedure start_timer;
  4    function get_duration
  5      return varchar2;
  6  end utl_timer;
  7  /
Package wurde erstellt.

SQL> create or replace package body utl_timer
  2  is
  3
  4    l_point_in_time number := null;
  5
  6    procedure start_timer
  7    is
  8    begin
  9      l_point_in_time := dbms_utility.get_time;
 10    end;
 11
 12    function get_duration
 13      return varchar2
 14    is
 15      l_duration number;
 16    begin
 17      l_duration := (dbms_utility.get_time - l_point_in_time) / 100;
 18      return l_duration || ' Sek.';
 19    end;
 20
 21  end utl_timer;
 22  /
Package Body wurde erstellt.
```

Listing 2: Hilfspackage, um die Zeit zu stoppen

Nebenbei bemerkt muss eine Funktion deterministisch sein, wenn Sie auf dieser Funktion einen Index oder eine virtuelle Spalte aufsetzen möchten. Achten Sie in diesem Zusammenhang auch darauf, dass sich die Implementierung der Funktion über die Zeit ändern und damit die Ergebnisse ebenfalls andere sein könnten. Das bekommt weder die virtuelle Spalte noch der funktionsbasierte Index mit!

Was aber tut die Datenbank, wenn die Funktion als deterministisch deklariert wurde?

Zunächst einmal recht wenig. Bis Version 9 sogar offensichtlich gar nichts,

dann jedoch wurde ein Cache eingeführt, der beim Aufruf einer PL/SQL-Funktion aus SQL Ergebnisse zwischenspeichern und somit wiederverwenden konnte. Erst ab Version 11.2 wurde ein solcher Cache auch für die Arbeit in PL/SQL selbst eingeführt, sodass diese Klausel sich nun zunehmend als sinnvoll herausstellt.

Wird also innerhalb von PL/SQL dieser Cache eingerichtet, hat dies zur Folge, dass die Datenbank Ergebnisse bereits berechneter Prozeduraufrufe im Rahmen des jeweiligen Server-Calls speichert (das wird später noch bedeutsam). Weitere Aufrufe der Funktion mit gleichen Parametern werden also nicht mehr neu berechnet, sondern aus dem Cache geliefert. In SQL gilt der Fokus analog: Die Ergebnisse werden innerhalb einer Abfrage im Cache vorgehalten, nicht darüber hinaus.

Vielleicht ein Beispiel. Ich verwende hier die einfachste Implementierung, die mir einfiel, um den Effekt zu zeigen. *Listing 2* zeigt zunächst ein kleines Hilfspackage, um die Zeit zu stoppen.

Nun folgt der eigentliche Test. Die Funktion nimmt einen Parameter entgegen, wartet die angegebene Zeit und meldet sich zurück. Die Funktion wird einmal mit, einmal ohne den Hint *deterministic* implementiert, wie in *Listing 3* gezeigt.

Der Test zeigt, dass die deterministische Funktion nur einmal, die nicht deterministische Funktion hingegen mehrfach aufgerufen wird *(siehe Listing 4)*.

Diese Effekte gelten nur im Umfeld eines Server-Calls, wie Sie erkennen können, wenn wir in der gleichen Session die Funktion erneut aufrufen *(siehe Listing 5)*.

<Wäre ein Session-Cache erstellt worden, hätte die Ausführung der deterministischen Funktionsvariante nicht 0,5 Sekunden gedauert.

Diese Optimierung ist, wie gesagt, mit Version 11.2 neu in die Datenbank gekommen. Vorher wurde diese Optimierung nur dann durchgeführt, wenn die PL/SQL-Funktion aus SQL aufgerufen wurde, und dies auch erst ab Version 10.

```
SQL> create or replace function test_deterministic(
  2    p_n number)
  3    return number
  4    deterministic
  5  is
  6  begin
  7    dbms_lock.sleep(p_n);
  8    return p_n;
  9  end test_deterministic;
 10  /
Funktion wurde erstellt.

SQL> create or replace function test_non_deterministic(
  2    p_n number)
  3    return number
  4  is
  5  begin
  6    dbms_lock.sleep(p_n);
  7    return p_n;
  8  end test_non_deterministic;
  9  /
Funktion wurde erstellt.
```

Listing 3: Test der deterministischen Funktion einmal mit, einmal ohne Implementierung des Hint deterministic

Zweite Variante: Function Result Cache

Der Function Result Cache bezeichnet eine Funktionalität, die es der Datenbank ermöglicht, Ergebnisse von PL/SQL-Prozeduren in einen Cache zu speichern, der nicht nur Session-bezogen, sondern Instanz-bezogen in der SGA vorgehalten wird. Eine aufwendige PL/SQL-Funktion kann so einmal gerechnet werden, die Daten werden zentral zur Verfügung gestellt und können direkt aus anderen Sessions genutzt werden. Analog zum *deterministic*-Hint wird einfach die Klausel *result_cache* an gleicher Stelle der Funktionsdeklaration eingesetzt, um der Datenbank mitzuteilen, dass diese Funktion für ein Caching infrage kommt (und um Oracle mitzuteilen, dass Sie gern die Enterprise-Edition-Lizenzkosten zahlen möchten). Diese Funktionalität bezieht sich nicht nur auf deterministische, sondern auch auf nicht-deterministische Funktionen. Oracle prüft selbstständig die Abhängigkeiten, denen die Funktion unterliegt (insbesondere Tabellendaten, auf die die Funktion zugreift). Ändern sich diese, wird das Ergebnis der Berechnung invalidiert, die Funktion rechnet erneut.

In diesem Zusammenhang war bis Version 11.2 die Klausel *relies on* wichtig, denn mit dieser Klausel konnte definiert werden, von welchen Tabellen das Ergebnis der Funktion abhängig ist. Ändern sich diese Tabellen, werden also auch die Funktionsergebnisse im Cache invalide. Wie bereits erwähnt, ist diese Klausel mit Version 11.2 nicht mehr erforderlich und tut einfach gar nichts mehr.

Für diese Funktionalität gelten folgende Warnhinweise:

- Sie ist derzeit an die Enterprise Edition gebunden.
- Sie vermeidet keine Umgebungswechsel zwischen SQL und PL/SQL, im Zweifel ist also eine skalare Unterabfrage billiger und schneller.
- Da Umgebungswechsel dennoch erforderlich sind, bietet sich die Kombination mit einer skalaren Unterabfrage an. Hier werden der Cache auf der PL/SQL-Seite genutzt und die Umgebungswechsel durch die skalare Unterabfrage genutzt.
- Es scheint, als habe der Result Cache einigen Overhead bezüglich der Cacheverwaltung, was auffällt, wenn Sie eine Funktion sehr oft aufrufen, die Rechenzeit in der Funktion aber klein ist. Daher lohnt diese Technologie vor allem dann, wenn die Rechenzeit der Funktion groß gegen diesen Overhead ist.

Man kann diese Funktionalität empfehlen, wenn komplexe, lange laufende Berechnungen in PL/SQL dadurch eingespart werden können, nicht jedoch als Standardmittel.

Anderen wird durch die Beschränkung auf die Enterprise Edition dieses Feature *verlieden*. Aber vielleicht wird diese Technik ja, wie ehedem die funktionsbasierten Indizes, aus der Enterprise Edition in die Standard Edition aufgenommen.

Zusammenfassung

Was lernen wir daraus? Die eigentliche Nachricht ist, dass wir – wieder einmal – der Datenbank so genau wie möglich sagen sollten, was wir über den Code wissen. Es ist nicht realistisch, dass Ihre Beschäftigung mit einer neuen Version der Datenbank stets so intensiv ist, dass Sie Änderungen wie das geänderte Verhalten des *deterministic-Hint* mitbekommen und Sie danach sofort all Ihren Code refaktorisieren. Doch diesen Hint gibt es schon lange. Und wenn man auch in vielen Blogs lesen kann, er solle nicht genutzt werden, da er nichts bringe und daher nur verwirre: Ich bin anderer Meinung. Wenn eine Funktion deterministisch ist, sagen Sie das. Ob die Datenbank eine Optimierung daraus ableiten kann oder nicht, entscheidet die Datenbank. Aber selbst, wenn die aktuelle Version keine Optimierung anbieten kann, eine zukünftige wird es können, ansonsten wäre die Einführung eines entsprechenden Hint nicht plausibel. Ähnlich können Sie es übrigens mit dem – seit Version 12c verfügbaren – Pragma *UDF* (User Defined Function) halten, mit dem Sie kennzeichnen, dass eine Funktion hauptsächlich aus SQL heraus aufgerufen wird. Auch dieses Pragma bringt im Moment noch nicht viel, aber das muss nicht so bleiben.

Im Gegensatz hierzu ist die Klausel *result_cache* ein echtes Tuning, das einerseits die Enterprise-Edition voraussetzt und andererseits genau daraufhin durchgesehen werden muss, ob der Einsatz wirklich etwas bringt. Ob tatsächlich eine Beschleunigung eintritt oder diese durch zusätzlichen Verwaltungsaufwand der Datenbank wieder einkassiert wird, müssen Sie testen. Im Gegensatz zu den anderen Lösungen übersteht diese Option sogar ein Ab- und Anmelden, da der Cache in der SGA gelagert und daher allen angemeldeten Benutzer zur Verfügung gestellt wird. Auch das kann eine interessante Beschleunigung Ihrer Anwendung sein.

Ein interessantes Feature also, aber sicher keine Killerfunktionalität. Es bleibt dabei: Diese Optimierungen können im ein oder anderen Fall wirklich drastische Wirkungen haben, im Regelfall stellen sie allerdings wohl eher eine Evolution und keine Revolution dar. Cool ist und bleibt der Trick mit der skalaren Unterabfrage, er kostet nichts und bringt viel – aber eben auch nur im Zusammenspiel zwischen SQL und PL/SQL, nicht in reinem PL/SQL.

```
SQL> set serveroutput on
SQL> declare
  2    l_n number;
  3  begin
  4    utl_timer.start_timer;
  5    for indx in 1 .. 10
  6    loop
  7      l_n := test_deterministic (0.5);
  8    end loop;
  9    dbms_output.put_line('Dauer deterministisch: '
 10      || utl_timer.get_duration);
 11
 12    utl_timer.start_timer;
 13    for indx in 1 .. 10
 14    loop
 15      l_n := test_non_deterministic (0.5);
 16    end loop;
 17    dbms_output.put_line('Dauer nicht deterministisch: '
 18      || utl_timer.get_duration);
 19  end;
 20  /
Dauer deterministisch: ,5 Sek.
Dauer nicht deterministisch: 5,02 Sek.

PL/SQL-Prozedur erfolgreich abgeschlossen.
```

Listing 4: Test, bei dem die deterministische Funktion nur einmal, die nicht deterministische Funktion hingegen mehrfach aufgerufen wird

```
SQL> r
  ...
Dauer deterministisch: ,5 Sek.
Dauer nicht deterministisch: 5,02 Sek.
```

Listing 5: Erneutes Aufrufen der Funktion in der gleichen Session im Umfeld eines Server-Calls

Jürgen Sieben
j.sieben@condes.de

BUSINESS NEWS

MAI
03

AUSBILDUNGSBERUFE IN DER IT

FELIX HUCHZERMEYER, CGS MBH

IT-Berufe – Ausbildung und Studium im Praxisverbund

FELIX HUCHZERMEYER
fhuchzermeyer@cgs-online.de

Schon lange ist das klassische Universitätsstudium nicht mehr der einzige Bildungsweg in der Informatik. Allerdings schrecken gerade kleinere Firmen noch häufig von der Ausbildung im eigenen Betrieb zurück. Der Betreuungsaufwand, die notwendigen Investitionen und die Kundenakzeptanz sind häufig schwer einzuschätzen. In folgendem Artikel beschreiben wir unsere Erfahrung aus Unternehmenssicht und aus der Perspektive der Auszubildenden.

Als mittelständisches Softwareunternehmen hängt der Geschäftserfolg der CGS mbH im Wesentlichen von unserem Personal ab. Wir können immer nur so gut sein wie die Summe unserer Mitarbeiterinnen und Mitarbeiter. Deshalb stehen bei uns die Themen Recruiting, Weiterbildung und langfristige Mitarbeiterbindung an oberster Stelle.

Die CGS ist mit ihren ca. 50 Mitarbeiterinnen und Mitarbeitern in drei Geschäftsfeldern, die jeweils unterschiedliche Rahmenbedingungen aufweisen, aktiv. Unser Hauptgeschäft ist die Individualentwicklung von komplexen Web-Applikationen, vorwiegend im Java EE- und Python-Umfeld. Dabei spielen auch Themen wie Cloud, Datenintegration und mobile Anwendungen eine Rolle. Daneben haben wir mit RIAS* eine eigene Standardsoftware, die ebenfalls auf Java EE basiert und auf eine Oracle-Datenbank aufbaut.

* RIAS – *Managementsystem für die Interne Revision*
Als webbasierte Revisionsmanagementlösung unterstützt RIAS die Interne Revision in ihrer täglichen Arbeit [siehe auch www.rias-revision.de]. Dabei reichen die Funktionen vom Audit Universe und der Risikobewertung, über die risikoorientierte Mehrjahresplanung, die Prüfungsdurchführung mit Arbeitsprogramm und Berichtserstattung bis hin zur Verwaltung der Maßnahmen und deren Nachverfolgung (Follow-up-Prozess).

RIAS entwickeln wir als Produktgeschäft ständig weiter und verkaufen entsprechende Lizenzen. Daneben haben wir in den letzten fünf Jahren unser immenses Know-how im Bereich Datenbanken und Datenmanagement strategisch weiterentwickelt und den Bereich Data Solutions geformt. Hier bieten wir Beratung in den Bereichen Datenbanken, Daten-Management, Datenintegration, Data Quality bis hin zu Data Analytics und KI-Themen.

Für diese Geschäftsfelder brauchen wir Beschäftigte, die fachlich interessiert, technisch gut ausgebildet und hoch motiviert sind. Solche Mitarbeiter sind schon seit vielen Jahren nicht nur in Deutschland sehr knapp. Deshalb bemühen sich alle Unternehmen um eine hohe Mitarbeiterzufriedenheit, denn erfahrene Informatikerinnen und Informatiker sind nur selten wechselwillig. Daraus folgt, dass der Aufbau neuer Fachkräfte für das Unternehmenswachstum dementsprechend schwierig ist.

FACHKRÄFTEMANGEL – NICHT JAMMERN, SONDERN AUSBILDEN

Es bringt leider nichts, über den Fachkräftemangel zu jammern – auch nicht darüber, dass große Marketingkampagnen und Imagewerbung für kleinere Unternehmen kaum zu bezahlen sind.

Deshalb haben wir uns bereits 2012 dazu entschieden, junge Mitarbeiterinnen und Mitarbeiter selbst auszubilden und zu fördern. Überwiegend machen wir dies durch das duale Studium mit der Ostfalia Hochschule für angewandte Wissenschaften in Wolfenbüttel. In Zusammenarbeit mit der Universität realisieren wir die integrierte Ausbildung Fachinformatik Anwendungsentwicklung.

Darüber hinaus haben wir auch eine Auszubildende mit dem Schwerpunkt Fachinformatikerin für Systemintegration, die sich überwiegend mit der IT-Infrastruktur und Administration der Systeme beschäftigt.

FACHINFORMATIK SYSTEMINTEGRATION

Seit August 2020 bilden wir erstmalig eine Fachinformatikerin für Systemintegration aus. Die Ausbildung ist dual, das heißt, sie erfolgt parallel in unserem Unternehmen und in der Berufsschule. In dieser Ausbildung liegt der Fokus auf der IT-Infrastruktur und deren Administration, wobei das Aufgabenspektrum weit gestreut und abwechslungsreich ist. Die Lösung vielschichtiger System- und Anwendungsprobleme gehört zum Tagesgeschäft, ebenso die Administration von Servern, Clients und Telefonanlagen. Ein weiteres Aufgabenfeld ist die Vernetzung von Hard- und Softwarekomponenten mit komplexen Systemen sowie die Installation und Pflege der IT-Infrastruktur. Hierbei spielt die IT-Sicherheit eine ganz wichtige Rolle, und regelmäßig werden neue Technologien bewertet, um eine kontinuierliche Weiterentwicklung der IT-Infrastruktur sicherstellen zu können

FACHINFORMATIK ANWENDUNGSENTWICKLUNG

Im Rahmen des dualen Studiums nehmen die Studierenden auch an der IHK-Prüfung teil und erwerben dadurch auch die Ausbildung zur Fachinformatikerin beziehungsweise zum Fachinformatiker für Anwendungsentwicklung. Für das Jahr 2023 planen wir, auch eine klassische Ausbildung in diesem Bereich anzubieten, das heißt ohne duales Studium. Bei der Ausbildung Fachinformatik für Anwendungsentwicklung steht die Softwareentwicklung im Vordergrund. Von der Planung, Konzeption, Programmierung und der Testung bis hin zur Dokumentation werden alle Aspekte gelehrt. Im Gegensatz zum dualen Studium findet hier die Ausbildung überwiegend im Betrieb statt, unterstützt durch die Berufsschule.

Alice F. – seit 2020 Auszubildende zur Fachinformatikerin Systemintegration bei CGS mbH

„Für die Ausbildung zur Fachinformatikerin für Systementwicklung hatte ich mich entschieden, da ich schon während meiner Schulzeit viel Spaß an IT-Administration und technischen Systemen hatte.

Der betriebliche Teil der Ausbildung macht mir persönlich am meisten Spaß, da ich viel lerne und auch sehr früh eigene Verantwortung und Projekte übernehmen konnte. In der Berufsschule gibt es einige Lernfelder, dich ich nicht ganz so interessant finde, die meisten sind aber spannend – ich konnte bereits einiges lernen. Trotz pandemiebedingter Probleme mit der Berufsschule kann ich die Ausbildung zur Fachinformatikerin für Systemintegration nur empfehlen. Ich habe bisher sehr gute Erfahrungen gemacht.

Mein Projektleiter hat sich Zeit für meine Fragen genommen und mir fehlendes Wissen schnell beigebracht. Viele Aufgaben durfte ich daraufhin selbstständig umsetzen, obwohl ich damals erst im 1. Ausbildungsjahr war. Ich wurde von Anfang an als vollwertige Mitarbeiterin behandelt und meine Projektleiter und Ausbilder waren jederzeit für mich da. Wir haben gemeinsam nach Lösungen gesucht, wenn es Probleme gab.“

STUDIUM IM PRAXISVERBUND – DAS DUALE STUDIUM

Seit 2012 bieten wir gemeinsam mit der Ostfalia und der IHK Braunschweig ein Studium im Praxisverbund im Bereich Informatik mit der Vertiefungsrichtung Software Engineering an. Das duale Studium dauert sieben Semester, also dreieinhalb Jahre, und beginnt jedes Jahr zum Wintersemester im

Stefan – Studium Informatik im Praxisverbund

„Der firmeninterne Anteil meines dualen Studiums hat mich sehr gut auf meinen jetzigen Berufsalltag vorbereitet. Ich habe in den Praxisphasen an echten Projekten mitarbeiten können und bereits früh an Kundengesprächen teilgenommen. Über die gesamte Studienzeit habe ich Einblick in fast alle Bereiche der Softwareentwicklung erhalten und hatte die Möglichkeit, Aufgaben eigenverantwortlich zu bearbeiten.

Das Arbeitspensum war dadurch in einigen Semestern sehr straff, sodass eine hohe Eigenmotivation und die Fähigkeit, sich Inhalte und Arbeitsweisen eigenständig zu erarbeiten, eine wesentliche Voraussetzung waren. Es gibt allerdings eine Vielzahl verschiedener Studienmodelle, die sich je nach Studienort unterscheiden.

Als zweiter dualer Student im Unternehmen habe ich die Organisation zu Beginn noch als etwas holprig wahrgenommen. Gerade in kleineren oder mittelständischen Unternehmen, die keinen dedizierten Prozess oder personelle Ressourcen für die Ausbildung der Studierenden eingerichtet haben, sollten der Betreuungsaufwand in den Projekten nicht unterschätzt und ausreichend zeitliche Ressourcen zur Verfügung gestellt werden. Mittlerweile wurde hierfür eine Koordinationsstelle geschaffen, die sich in den Praxisphasen teilzeitlich um alle Belange der Studierenden kümmert.

Wer sich für ein Studium in einem kleineren oder mittelständischen Unternehmen interessiert, sollte sich zudem darauf einstellen, dass gewisse Vorkenntnisse (beispielsweise Beherrschung von Programmiersprachen, Arbeitserfahrung durch private Projekte) gewünscht oder gar gefordert werden.“

Stefan Gaertner hat zwischen 2014 und 2018 sein Informatikstudium im Praxisverbund an der Ostfalia in Kooperation mit der CGS absolviert und arbeite seitdem in der agilen Produktentwicklung der RIAS Software. Da er schon während seinem Studium mehrfach im RIAS Team gearbeitet hat, ist er ein wichtiger Know-how-Träger mit viel Erfahrung. Neben der Programmierarbeit ist er ein wichtiger Treiber in der Optimierung unserer Entwicklungsprozesse und Testautomatisierung. Zusätzlich agiert er als Betreuer der dualen Studenten der CGS.


Ostfalia

August oder September. Die Studierenden starten zunächst mit einem sechsmonatigen Einsatz bei uns. Dabei erhalten sie Einblicke in die Projekte bei der CGS.

Danach beginnt das erste Semester an der Hochschule. Im ersten Teil des Studiums absolvieren die Studierenden insgesamt vier Theoriesemester, bevor sie die Prüfung zur Fachinformatikerin beziehungsweise zum Fachinformatiker vor der IHK ablegen. Damit haben sie dann bereits einen vollwertigen Berufsabschluss. In der zweiten Hälfte des dualen Studiums folgen zwei weitere Semester mit Vorlesungen. Im siebten und letzten Semester schreiben unsere Studierenden ihre Bachelor-Arbeit bei uns in der CGS. Wir finden für jeden ein interessantes und praxisnahes Thema.

Über das gesamte Studium arbeiten die Studierenden in den vorlesungsfreien Zeiten bei uns an echten Projekten mit. Sie werden von einer Ausbilderin beziehungsweise einem Ausbilder und der jeweiligen Projektleiterin beziehungsweise dem jeweiligen Projektleiter persönlich betreut. Anspruchsvolle Aufgaben entsprechend dem Kenntnisstand der Studierenden sind für uns selbstverständlich.

Die Anzahl der auszubildenden Studierenden haben wir auf ein bis zwei pro Jahr begrenzt, da wir als kleines bis mittelständisches Unternehmen eine entsprechend gute Betreuung bei einer größeren Anzahl nicht sicherstellen könnten. Mittlerweile haben fünf Beschäftigte das duale Studium mit uns absolviert. Wir sind sehr stolz, dass alle unsere Studierenden überdurchschnittlich gute Abschlüsse erzielen konnten. Die fünf Absolventinnen und Absolventen sind heute noch bei uns in Festanstellung und leisten hervorragende Arbeit in unseren Projektteams. Gerade durch die Abwechslung von Studienphasen mit Praxisphasen ist die Lernkurve sehr hoch. Gelerntes kann häufig direkt in den Projekten angewandt und damit gefestigt werden. Für uns ist dieser Weg des Mitarbeiteraufbaus eine absolute Erfolgsgeschichte.

Hierbei ist die Win-Win-Situation für das Unternehmen und die Studierenden wichtig, denn nur wenn beide Parteien gleichermaßen davon profitieren, kann eine nachhaltige Beziehung aufgebaut werden. Die Vorteile für die Studierenden sind nicht nur, dass die Studiengebühren und andere Kosten durch die CGS übernommen werden. Sie schließen auch einen Vertrag mit der CGS und erhalten ein monatliches Festgehalt, sowohl in den Praxis- als auch in den Studienphasen. Dadurch muss man keine anderen Nebenjobs annehmen, um sich das Studium zu finanzieren. Die Studierenden haben also mehr Zeit, sich auf die Informatik zu konzentrieren, sowohl in der Theorie als auch in der Praxis. Diese Kombination macht es für viele Studierende auch interessant, denn die praktischen Anteile bringen eine passende Abwechslung zum Studienalltag.

Sebastian – Studium Informatik im Praxisverbund

„Das duale Studium bot mir die Möglichkeit, die akademische Bildung mit einer klassischen Ausbildung zu kombinieren. Ich konnte damit sowohl einen begehrten Abschluss als Bachelor of Science der Informatik erwerben als auch die praktische Erfahrung der Ausbildung erlangen.

Der Studiengang an der Ostfalia sah dabei vor, dass der akademische Teil im normalen Semester abgehandelt wird und der praktische Teil in den Semesterferien. Dadurch konnte in dieser Zeit in realen Projekten mitgewirkt werden. Wichtig ist hier, dass das betreuende Unternehmen geeignete Projekte für die Praxisphase findet. Das Projekt sollte innerhalb des Zeitraumes abgeschlossen werden können und dem Wissensgrad der Studierenden entsprechen. Es ist zudem vorteilhaft, wenn die Projekte einen überschaubaren Technologie-Stack haben, sodass eine selbstständige Einarbeitung möglich ist.

Der akademische Anteil des Studiums erfolgte ähnlich wie in einem normalen Studium – mit der Ausnahme, dass die Prüfungen am Ende des Semesters geschrieben wurden. Durch das duale Studium hatte ich hier den Vorteil, dass ich sämtliche praktische Anteile in der Firma durchführen konnte. Dies schließt auch das Schreiben der Bachelorarbeit mit ein. Hier ergaben sich für mich große Vorteile, da ich die Bachelorarbeit in meinem Unternehmen durchführen konnte, welches ein erhebliches Interesse an meinem Erfolg hat und mich somit optimal unterstützen konnte. Alles in allem kann ich das duale Studium allen empfehlen, die praktische Erfahrung neben dem Studium sammeln möchten.“

Sebastian Schubert hat zwischen 2012 und 2016 sein Informatikstudium im Praxisverbund an der Ostfalia in Kooperation mit der CGS absolviert. Für die CGS war er der erste Student im Praxisverbund und damit der Auftakt zu einer erfolgreichen Zusammenarbeit. Er ist einer der zentralen Know-how-Träger nicht nur innerhalb der Produktentwicklung der RIAS Software, sein Steckenpferd ist die Software-Architektur. Seine Erfahrungen gibt er über technische „Deep Dives“ oder interne Vorträge im Rahmen unserer Entwicklerrunde an die Kollegen weiter.

AUSBILDUNG IM MITTELSTAND

Sowohl die Ausbildung als auch das duale Studium läuft in einem mittelständischen Unternehmen wie der CGS etwas anders ab als in einem Großunternehmen. Bei großen Betrieben gibt es oft mehrere Studierende und Auszubildende, die gemeinsam als Gruppe an Ausbildungsprojekten arbeiten. Bei uns sind die Auszubildenden und Studierenden in die echten Kundenprojekte einbezogen und arbeiten im jeweiligen Projektteam mit. Zu Beginn werden einfachere Aufgaben erledigt und häufig im Pair Programming mit erfahrenen Kollegen umgesetzt. Nach ersten Erfahrungen gibt es aber auch komplexere Entwicklungen und es wird schrittweise eigenständiger gearbeitet.

Dieses Vorgehen hat zwei wesentliche Vorteile, denn die Auszubildenden lernen von Beginn an die Arbeitsweise in echten

Nach seinem Studium der Informatik und der anschließenden Promotion arbeitete Bernd Müller für die IBM und die HDI Informationssysteme. Er ist Professor für Software-Engineering an der Ostfalia, Autor mehrerer Bücher zu den Themen JSF und JPA sowie Speaker auf nationalen und internationalen Konferenzen. Er veröffentlicht regelmäßig Artikel in der DOAG-Zeitschrift Java aktuell.

bernd.mueller@ostfalia.de

Ostfalia Hochschule für angewandte Wissenschaften – Studium Informatik im Praxisverbund

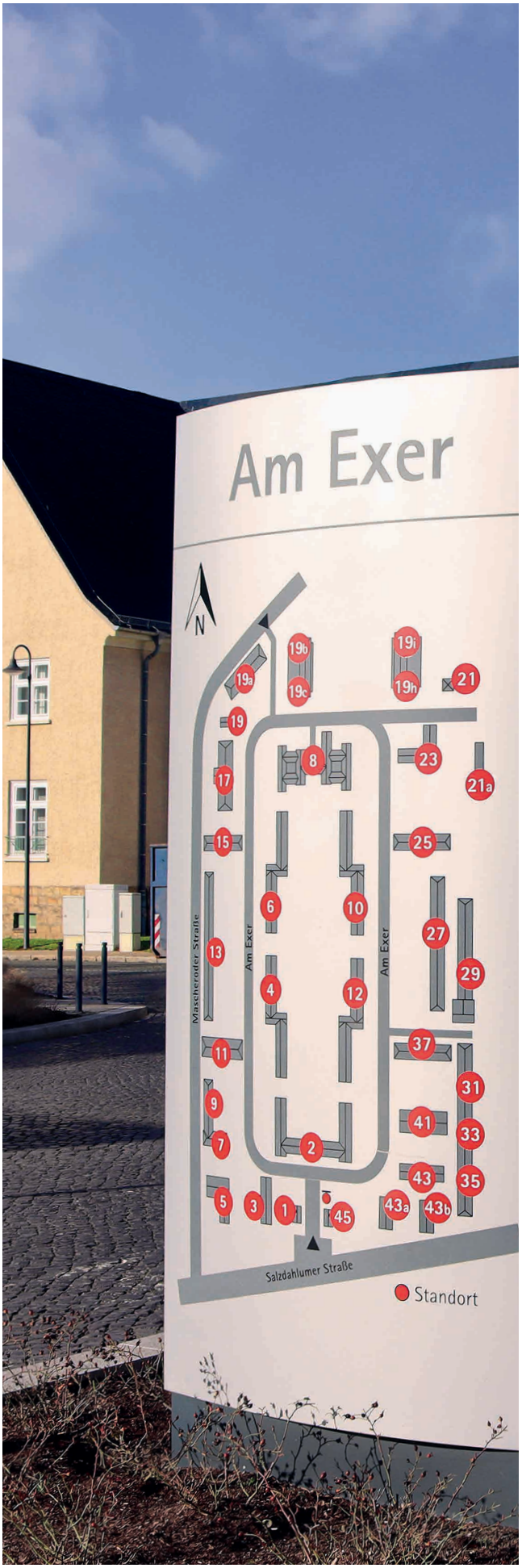
„Die Fakultät Informatik, damals noch Fachbereich Informatik, der Hochschule Braunschweig/Wolfenbüttel entstand 1992 als Ausgliederung aus dem Fachbereich Elektrotechnik. 2009 gab sich die Hochschule den Namenszusatz Ostfalia Hochschule für angewandte Wissenschaften, kurz Ostfalia, mit dem sie nun in der Region bekannt, um nicht zu sagen renommiert ist.

Die Zielsetzung von Fachhochschulen, anwendungsorientierte und praxisbezogene Lehre und Forschung auf wissenschaftlicher Grundlage zu betreiben, zeigt sich insbesondere in den sogenannten dualen Studiengängen. Diese firmieren mittlerweile unter der konkreteren Bezeichnung Studium im Praxisverbund. 2008 entschloss sich die Fakultät Informatik, den Studiengang Informatik im Praxisverbund einzuführen, und startete mit acht Kooperationspartnern. Wenige Jahre später folgte der Studiengang Wirtschaftsinformatik im Praxisverbund. Heute sind es über 60 Kooperationspartner, die CGS ist einer davon.

Die Studiengänge der Informatik unterscheiden nicht zwischen Praxisverbund oder nicht Praxisverbund. Es gibt nur eine Studien- und eine Prüfungsordnung, wobei nur aufgrund der um das Unternehmenssemester längeren Studiendauer bezüglich der Studiengänge unterschieden wird. In den Lehrveranstaltungen befinden sich Studentinnen und Studenten mit und ohne Praxisverbund Seite an Seite. Sie nehmen an denselben Übungen und Tutorien teil und schreiben dieselben Prüfungen.

Aus meiner persönlichen Dozentensicht ist allerdings bei der Noteneingabe, die nach Studiengängen getrennt erfolgt, ein Unterschied festzustellen. Die durchschnittlich erreichten Benotungen sind in der Regel im Praxisverbund besser. Dies wird von vielen Kolleginnen und Kollegen bestätigt. Der Autor kann sich zu den tatsächlichen Gründen nicht äußern, da wir keine entsprechenden, verlässlichen Informationen haben, vermutet dahinter aber die von den Kooperationsfirmen geleistete Unterstützungs- und Motivationsarbeit. Ebenfalls nur eine subjektive Wahrnehmung des Autors ist die folgende: Studentinnen und Studenten im Praxisverbund zeigen generell eine höhere Motivation, aber auch eine höhere Erwartungshaltung bezüglich der Qualität und des Niveaus von Lehrveranstaltungen. Durch die betrieblichen Erfahrungen, die zeitlich häufig vor einer entsprechenden Lehrveranstaltung stattgefunden haben, sind die Studentinnen und Studenten zudem auch durchaus kritischer gegenüber Dozenten, Lehrinhalten und Lehrmethoden eingestellt. Sie haben zum Teil bereits erlebt, wie man etwas konkret realisieren kann, und hinterfragen daher häufiger den vom Dozenten vorgeschlagenen Weg. Aus Sicht des Autors ist dies eher eine Bereicherung als ein Malus.

Für die Fakultät Informatik sind die Praxisverbundstudiengänge ein Gewinn. Motivierte Studentinnen und Studenten mit, in der Regel, besseren Studienleistungen bereichern und intensivieren den Hochschulalltag. Die Fakultät wird auch zukünftig versuchen, einen hohen Anteil der Studentenschaft durch ihre Praxisverbundstudiengänge zu erlangen, vorausgesetzt die Bewerberlage bleibt so, wie sie seit einigen Jahren ist.“



Projekten kennen und können „real-life"-Erfahrungen sammeln. Oft ist es auch spannender, an echten Kundenanforderungen zu arbeiten als an rein theoretischen Übungsprojekten. Bei uns findet die Ausbildung beziehungsweise die Praxisphase sehr stark in den Projekten statt, also „learning on the job", oft auch mit erstem Kundenkontakt.

Unsere Kunden informieren wir selbstverständlich über die Mitarbeit der Studierenden und Auszubildenden und bisher haben alle Kunden sehr positiv darauf reagiert. Eine Fakturierung der Projektstunden von Studierenden und Auszubildenden ist in der Regel nicht möglich, sondern man muss diese Aufwände als Ausbildungskosten verstehen.

Natürlich bringt dieses Vorgehen nicht nur Vorteile. Grundlegendes theoretisches Wissen kann im Projektalltag nur schwer vermittelt werden, da die Zeit hierfür fehlt. Die Auszubildenden werden also auch mal ins „kalte Wasser geschmissen" und müssen schon gewisse Grundkenntnisse mitbringen. Bei uns ist es also vielleicht etwas spannender, manchmal aber auch etwas herausfordernder.

Dementsprechend legen wir im Bewerbungsprozess auf drei Aspekte besonderen Wert:

- Leidenschaft für Software, Daten und/oder IT
- eamfähigkeit
- technisches Talent und Grundwissen

Die Reihenfolge entspricht auch unserer Priorität. Denn wenn jemand Leidenschaft und Interesse für das Thema hat, dann ist er auch lernbereit und hat den entsprechenden Ehrgeiz. Da wir in der Regel nicht allein in einem Projekt arbeiten, ist die Teamfähigkeit absolut wichtig. Die Ausbildung wird bei uns eben nicht durch eine Ausbilderin oder einen Ausbilder durchgeführt, sondern durch das gesamte Team, in dem die Praxisphase absolviert wird.

Wenn jemand technisches Talent und gewisse Grundkenntnisse der Informatik mitbringt, teamfähig und mit Leidenschaft bei dem Thema ist, kann er bei uns in den Projektteams alles lernen. Die CGS investiert dabei nicht unerheblich in die Ausbildung; inklusive des Gehalts mit Lohnnebenkosten, der Semesterbeiträge, der Prüfungsgebühren etc. summiert sich dies im Laufe des gesamten Studiums auf ca. dreißigtausend Euro. Beachtet man die hohen Vermittlungsgebühren der Personalvermittler und die Einarbeitungskosten für neue Beschäftigte, ist dieser Weg nicht teurer als die Einstellung einer Junior Developerin beziehungsweise eines Junior Developers über eine Personalvermittlung. Dennoch ist eine Bindung an das Unternehmen Bestandteil des Vertrags zwischen dem Studierenden

und der CGS. Bisher haben wir sehr positive Erfahrungen mit diesem Konzept sammeln können und dabei überdurchschnittlich gute Mitarbeiterinnen und Mitarbeiter gewinnen und halten können. Natürlich stehen wir auch hier im Wettbewerb zu anderen, teils deutlich größeren Unternehmen.

Es gibt viele Schülerinnen und Schüler, die nach ihrem Studium lieber bei einem großen, namhaften Konzern ihr duales Studium beginnen. Denn diese Großunternehmen bieten häufig Prestige und Sicherheit. Die Bewerbenden sind oft noch sehr jung und die Eltern haben nicht selten Einfluss auf die Entscheidung. Wir punkten hier häufig mit unserer familiären Unternehmenskultur, kollegialem Umgang und flachen Hierarchien. Bewerberinnen und Bewerber mit stärkeren Vorkenntnissen finden unsere Vorgehensweise mit der Beteiligung an echten Kundenprojekten oft sehr attraktiv, sodass wir immer wieder sehr gute Studierende für uns gewinnen können. Das Niveau der Vorkenntnisse ist auch oft sehr unterschiedlich. Es gibt die Schülerinnen und Schüler, die sich bisher kaum mit Informatik beschäftigt haben und nach einem sicheren Job mit guter Bezahlung suchen. Dies ist in der Informatik aktuell durchaus gegeben. Andere entwickeln schon seit ihrer Kindheit und haben dabei schon mehr Java-Know-how als manche Absolventin oder mancher Absolvent nach seinem Studium – beide verbindet der Wunsch, in erster Linie coole Software zu entwickeln.

Um als CGS bei den Schülerinnen und Schülern als potenzieller Arbeitgeber wahrgenommen zu werden, veranstalten wir jedes Jahr einen Zukunftstag, an dem wir die Berufe und die Ausbildungsmöglichkeiten vorstellen. Hier gibt es keine trockenen Power-Point-Präsentationen, sondern gemeinsame Übungsaufgaben und interaktiven Austausch. Das ist unser Weg. Neben den Vorteilen für das Unternehmen macht es uns auch Spaß, jungen Menschen eine berufliche Perspektive zu geben und einen gemeinsamen Ausbildungsweg zu beschreiten. Dieser gemeinsame Ausbildungsweg verbindet; die Absolventinnen und Absolventen des dualen Studiums beziehungsweise der Ausbildung sind sehr loyale und motivierte Teammitglieder.

Wir können es nur empfehlen, sich mit den Themen Ausbildung und Studium im Praxisverbund zu beschäftigen – auch bei kleinen IT-Firmen. Aus meiner Sicht überwiegen die Vorteile eindeutig.

Felix Huchzermeyer ist seit 2017 Geschäftsführer der CGS mbH und seit über 15 Jahren in der Technologie- und IT-Branche tätig. Durch seine unterschiedlichen Managementpositionen bei der Siemens AG im In- und Ausland kennt er sowohl die Ausbildungsmöglichkeiten im Mittelstand als auch die in großen Konzernen.

ELINA HATTENDORF, GYMNASIUM GROSSE SCHULE WOLFENBÜTTEL

Von der Motivation, nach der Schulbank einen Ausbildungsberuf in der IT zu ergreifen

Wie kommt man auf die Idee, ein duales Studium zu machen? Die Antwort darauf kann so unterschiedlich sein wie die Menschen, die sie geben. Hier können Sie meine ganz persönliche Sicht auf die Dinge kennenlernen. Ich werde berichten, wo für mich alles angefangen hat und welche Möglichkeiten ich in der Schule hatte, mich darauf vorzubereiten. Mein Weg begann schon weit vor meinem Praktikum und führte schlussendlich zu meiner erfolgreichen Bewerbung.

ELINA HATTENDORF
LunaeEclipsis@gmx.de

Wenn man aus einer Familie kommt, in der es Common Sense ist, gemeinsam Computerspiele zu spielen, und einer der erfreulichen Punkte am Älterwerden ist, dass man den Eltern endlich beim Spielen von Diablo III über die Schulter sehen darf, dann kommt das Interesse an dem, was hinter allem steckt, fast schon automatisch. Zumindest bei mir war es so und ich empfinde es als Alltag, dass wir Teile eines Computers „mal eben" selbst austauschen, wenn sie kaputt gehen oder zu alt werden. Ziemlich genau da kam ich mit diesem Randbereich der Informatik das erste Mal in Kontakt, denn welches Kind ist nicht von der Arbeit der Eltern fasziniert?

Auch die Ostfalia *(Anm. d. Red.: siehe Leitartikel ab S. 66)* war schon immer mehr oder weniger ein Teil meines Lebens, denn sowohl mein Weg zum Kindergarten als auch der zur Grundschule führten mich beinahe acht Jahre zwei Mal pro Tag am Hauptgebäude der Hochschule vorbei. Die Male, die ich meinen Vater zur Außenstelle der Ostfalia Am Exer begleitete, habe ich mich jedes Mal fasziniert mit großen Augen umgesehen und ganz fest daran geglaubt, dort eines Tages, wie mein Vater, Informatik zu studieren.

GRAFISCHES PROGRAMMIEREN IN DER 10. KLASSE

Die Idee zum dualen Studium habe ich schon deutlich länger und ich habe bereits vor meinem Eintritt in die Oberstufe auf dieses Ziel hingearbeitet. Ob mit der Wahl von zusätzlichen Kursen oder den Zukunftstagen, an der Großen Schule Wolfenbüttel hatte ich, was das anging, durchaus verschiedene Optionen. Ich hatte zwar nicht die Möglichkeit, ein naturwissenschaftliches Profil vor der Oberstufe zu belegen, dafür wurde der „MINT-Kurs" eingeführt, den ich dann zum Beginn des 7. Schuljahres wählen konnte. So konnte ich jedes Halbjahr entscheiden, ob mich der angebotene Kurs aus den Bereichen Mathematik, Informatik, Naturwissenschaften, Technik und


YOU SHALL NOT PASS
apex.world
ā'pěks
ASUS
ā'pěks
(#orclapex)
utPLSQL
TESTING FRAMEWORK FOR PL/SQL

Die Male, die ich meinen Vater zur Außenstelle der Ostfalia Am Exer begleitete, habe ich mich jedes Mal fasziniert mit großen Augen umgesehen und ganz fest daran geglaubt, dort eines Tages, wie mein Vater, Informatik zu studieren.

Erdkunde interessiert und ich ihn besuchen möchte. Es war viel Verschiedenes dabei, vom Kurs in 7.1, bei dem wir uns mit unterschiedlichen Verschlüsselungen beschäftigt und ein eigenes Morsegerät gebastelt haben, über ein Jahr, in dem wir uns mit Bodenarten beschäftigt haben, bis hin zu einem der Kurse in der 10. Klasse, in dem wir grafisches Programmieren mit Java gelernt und es außerdem geschafft haben, einen Traktor zum Fahren zu bringen.

Auch die bundesweiten Zukunftstage, die in allen Klassenstufen zwischen 5 und 9 stattfinden, boten jedes Jahr wieder eine Möglichkeit, sich umzusehen und interessante Erfahrungen mitzunehmen. 2016 lernte ich in der IT-Abteilung der Hoyer Group in Hamburg anhand eines Beispiels, wie weit die Vorstellung des Auftraggebers manchmal von dem abweicht, was von den Entwicklern verstanden wird und was am Ende dabei herauskommen könnte, wenn man nicht aufpasst. 2017 lernte ich die Ostfalia von innen kennen und 2018 verbrachte ich schließlich den Zukunftstag in der CGS *(Anm. d. Red.: siehe vorangegangenen Artikel)*, also bei meinem zukünftigen Arbeitgeber.

EIN PRAKTIKUM ALS BESTÄRKUNG

In der 10. Klasse konnte man sich schließlich entscheiden, im darauffolgenden Schuljahr seine zweite Fremdsprache zu ersetzen. Dieses Angebot kam mir aus zweierlei Gründen sehr gelegen. Zum einen war ich mehr als froh, nach meinem kleinen Latinum – und somit sechs Jahren des Lateinlernens – endlich von dieser Sprache „erlöst“ zu sein. Sie war interessant, aber lag nie wirklich innerhalb meiner Begabung. Zum anderen war es die Möglichkeit, Latein durch Informatik und Erdkunde zu ersetzten. Das war eine mehr als nur angenehme Entscheidung für mich, von der ich bis heute überzeugt bin. In der 11. Klasse war es dann schließlich auch so weit, dass man sich um einen Platz für das dreiwöchige Schülerpraktikum im Frühling kümmern musste, und mir wurde zu meiner großen Freude das Praktikum in der CGS ermöglicht. Obwohl ein Schülerpraktikum in der Firma bis dahin nicht angeboten wurde, hat man für mich eine Ausnahme gemacht, da ich ja bereits zum Zukunftstag zwei Jahre zuvor da gewesen war.

Das dreiwöchige Praktikum, das ich dann in der CGS absolvieren durfte, kann ich durchweg als positiv beschreiben. Ich habe in der Zeit über das, was ich später machen wollte, mehr gelernt als in meiner gesamten Zeit in der Schule. Ich wurde so gut, wie es eben ging, in den Alltag eingebaut und durfte sogar mit Code arbeiten, der – nach einem Review durch einen anderen Softwareentwickler – auch genutzt wurde. Jedes der Teammitglieder vermittelte mir einen anderen Teil der Arbeit, vom Scrummaster über Entwicklung bis hin zum Testen. Ich habe auch festgestellt, dass der Arbeitsalltag eine völlig andere Anstrengung ist, als wenn man in der Schule sitzt. Viel Zeit habe ich damit verbracht, an tatsächlichen Aufgaben zu arbeiten.

Etwas, dass ich ebenfalls gerne erwähnen möchte, ist, dass ich während meines Praktikums die Möglichkeit hatte, mit einem dualen Studenten gemeinsam zu arbeiten und mir einige seiner Erfahrungen anzuhören. Das war auf jeden Fall etwas, dass mich in meinen Plänen für die Zukunft weiter bestärkt hat.

PERFEKTE SITUATION, ABER ZWEIFEL SIND WICHTIG

Die Frage, warum ich mich zuerst bei der CGS beworben habe, ist damit auch schon in weiten Teilen beantwortet. Neben der Erfahrung im Praktikum, dass sehr viel Wert auf Zusammenarbeit im Team gelegt wird, steht dabei noch sehr weit oben auf der Liste der Vorteile, dass die Firma kein großer Konzern ist und man eine realistische Chance hat, die meisten der Kollegen kennenzulernen. Die Ostfalia war für mich aus bereits genannten Gründen und aufgrund der guten Erfahrungsberichte ebenfalls die naheliegende Entscheidung.
So perfekt diese Situation klingt, auch Zweifel sind wichtig. Besonders in der Zeit, als ich beginnen wollte, meine ersten Bewerbungen zu schreiben, war ich mir auf einmal nicht mehr sicher, ob dieses Studium tatsächlich das ist, was ich gerne möchte. Ratsuchend habe ich mich mit der Unsicherheit an meinen Vater gewendet und mir auf seinen Vorschlag hin eine (wahnsinnig lange) Liste aller Ausbildungs- und Studien-Optionen heruntergeladen. Ich war drei Tage damit beschäftigt, die Liste durchzugehen und alles zu streichen, was mir auf den ersten Blick nicht so zugesagt hat. Am Ende sind dann etwa 15 Optionen übriggeblieben, von denen fünf Teilgebiete der Informatik waren. Nachdem ich mir aus jeder Kategorie, unter der die Vorschläge aufgeführt waren, einen ausgesucht hatte, waren für mich noch fünf Studiengänge aus den Fachrichtungen Informatik,

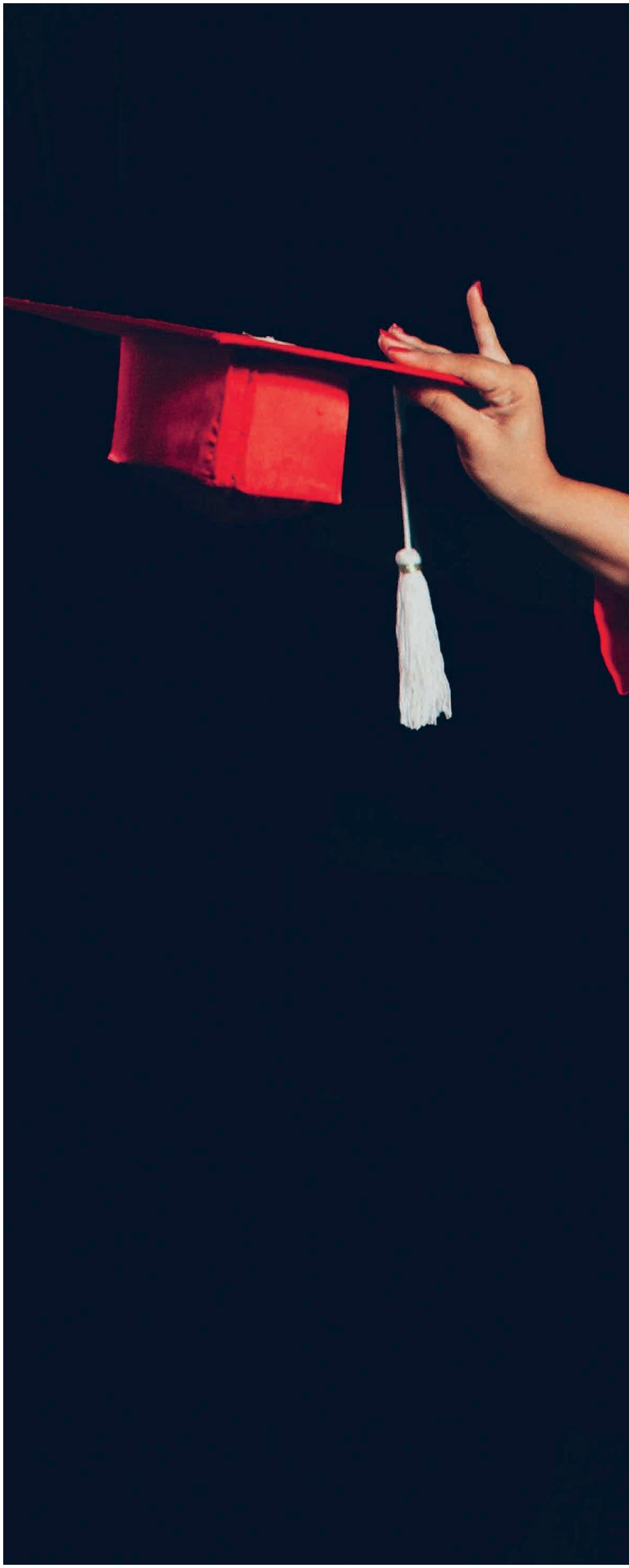
Architektur und Physik übrig. Man kann schon erkennen, worauf das hinausläuft, denn relativ bald habe ich gesehen, dass ich immer eher auf die dualen Studiengänge aus dem Bereich Informatik geschaut habe, bis ich die anderen Optionen schließlich ausgeschlossen hatte.

Als nächster Schritt begann schließlich die Suche nach dem zukünftigen Arbeitsplatz für mich. Um einem ähnlichen Problem wie bei der Auswahl des Studiengangs entgegenzuwirken, bin ich dabei von Anfang an – soweit es ging – ergebnisoffen herangegangen. Ich habe mir die vielen verschiedenen Angebote angesehen, wobei schon ein sehr großer Teil einfach dadurch weniger attraktiv war, dass ich beschloss, in meiner Heimatstadt zu bleiben. Aus der langen Liste an Firmen, die ein duales Studium anbieten, entschied ich mich am Ende für drei, um mich dort zu bewerben. Diese Firmen hatte ich bereits 2019 auf der Informationsveranstaltung der Ostfalia zum dualen Studiengang Informatik kennengelernt, auf dem sich eine ganze Reihe von Firmen vorgestellt haben, und sie hatten einen besonders guten Eindruck auf mich gemacht. Falls es bei keinem meiner drei Wunscharbeitgeber klappen würde, wollte ich zunächst ein Freiwilliges Soziales Jahr machen.

Später ging es dann recht schnell. Ich hatte kaum eine Antwort der anderen Firmen erhalten und deren Termine für Bewerbungsgespräche lagen noch Monate in der Zukunft, da hatte ich das Gespräch bei der CGS schon hinter mir und knapp zwei Wochen später die Zusage. Mitten im Stress für die Vorbereitungen der ersten Klausuren unter Abiturbedingungen im späten Herbst – in einigen Bundesländern werden die Klausuren im Winter vor dem Abitur schon unter ähnlichen Bedingungen geschrieben wie später die schriftlichen Abiturprüfungen – fiel mir ein Stein vom Herzen.

FREUDE AUF DAS DUALE IM STUDIUM

Abschließend kann ich dazu sagen, dass mich der Gedanke an ein duales Studium insbesondere bei der CGS schon einen großen Teil meines Lebens begleitet, und ich habe mit den Zukunftstagen, dem Praktikum und der Auswahl meiner Leistungskurse darauf hingearbeitet. Ich sehe mit einem positiven Bauchgefühl in die Zukunft und in Richtung meines Studiums und freue mich auch schon auf die erste vorlesungsfreie Zeit, in der ich dann beginne zu arbeiten. Genauso freue ich mich auf meine zukünftigen Kollegen in der Firma und auf das Team.



Elina Hattendorf ist Abiturientin am Gymnasium Große Schule in Wolfenbüttel im Jahrgang 13, leidenschaftliche Computerspielerin und zukünftige duale Studentin an der Ostfalia Hochschule für angewandte Wissenschaften und in der Consulting Gesellschaft für Systementwicklung, CGS mbH.

NADINE WAGNER UND MICHAEL PERGANDE, PROMATIS GRUPPE, ETTLINGEN (TECHNOLOGIEREGION KARLSRUHE)

Individuelle Weiterentwicklung von Talenten – Die Schlüsselstrategie für erfolgreiches Unternehmenswachstum

NADINE WAGNER
nadine.wagner@promatis.de

MICHAEL PERGANDE
michael.pergande@promatis.de

Sie sind vielversprechend, profitabel und angesagt: Immer mehr Jobs drehen sich um die IT. Zudem hat Corona einen Digitalisierungsschub ausgelöst – der Bedarf an IT-Fachkräften wird also zukünftig weiter steigen, da digitale Geschäftsmodelle, Technologien und Prozesse entsprechende Expertise zugrunde legen. Wie meistern jedoch IT-Beratungsunternehmen diesen Spagat zwischen geringem Angebot und erhöhter Nachfrage?

Gerade die Digitalisierung von Wirtschaft und Gesellschaft hält uns den Spiegel des Fachkräftemangels vor Augen, denn das Defizit bei Berufen, die spezielles Wissen in Mathematik, Informatik, Naturwissenschaft und Technik (MINT) erfordern, ist besonders groß. Seit mehr als zwei Jahrzehnten ist PROMATIS als global tätiger Implementierungspartner eng mit der nationalen und internationalen Oracle-Organisation verbunden. Eingesetzt werden Applikations- und Technologieprodukte des Weltmarktführers, die vom zertifizierten und mehrfach ausgezeichneten Oracle-Partner und Cloud Excellence Implementer nicht nur in Kundenprojekten eingeführt werden, sondern die man auch im eigenen Unternehmen erfolgreich nutzt. Folglich ist der zunehmende Fachkräfteengpass besonders spürbar, wenn es darum geht, geeignetes IT-Beratungspersonal am Arbeitsmarkt zu finden. Durch die duale Berufsausbildung kann jedoch Rekrutierungsproblemen vorgebeugt sowie der Fachkräftebedarf gedeckt werden, ohne sich vom Arbeitsmarkt abhängig zu machen. So profitieren Unternehmen davon, dass die Auszubildenden den Betrieb, die Produkte und Dienstleistungen sowie relevante Techniken von der Pike auf kennenlernen, rasch selbstständige Arbeitsaufgaben übernehmen und erfolgreich in die Prozesse des Unternehmens eingebunden sind.

DUALE AUSBILDUNG NAH AN DER PRAXIS

Der Spezialist für intelligente Geschäftsprozesse und Oracle-Digitalisierungslösungen – im Herzen der TechnologieRegion Karlsruhe – setzt auf die duale Berufsausbildung als einen der bedeutendsten Erfolgsfaktoren. Gerade die Kombination aus Theorie und Praxis bereitet Auszubildende besonders gut auf die Herausforderungen vor, die im täglichen Berufsleben auf sie warten: nicht nur Fachwissen, sondern auch praktische Erfahrungen darin, dieses Wissen anzuwenden. So bietet die duale Ausbildung – um die uns vermutlich viele europäische Nachbarn beneiden – zahlreiche Chancen für einen erfolgreichen Berufseinstieg und eine solide berufliche Karriere. Im Fokus der betrieblichen Berufsausbildung steht das Lernen in qualifizierenden und wertschöpfenden Arbeitsprozessen. Besonders bei der Ausbildung als Fachinformatiker (Fachrichtung Systemintegration, aber auch in der Anwendungsentwicklung) wird ein hohes Maß an Problemlösungskompetenz voraussetzt. Denn auf der einen Seite

wird der Auszubildende im Rahmen von Helpdesk-Tätigkeiten – je nach Know-how der Kolleginnen und Kollegen – mit unterschiedlichen Supportanfragen und somit ganz vielschichtigen Aufgaben und Fragestellungen konfrontiert. Auf der anderen Seite spielt im Oracle-Beratungshaus das enorm große Produktportfolio des amerikanischen Software-Riesen eine bedeutende Rolle, mit dem sich der Azubi technisch auseinandersetzen muss, wie es beispielsweise im Zuge von Wartungstätigkeiten – oder sogar beim Aufsetzen – interner Test- und Demo-Systeme der Fall ist. In diese Prozesse ist der Auszubildende bereits zu Beginn seiner Ausbildung eingebunden und erlangt schon früh ein breites Wissensspektrum, resultierend aus den Kundenfragestellungen und eingesetzten Technologien. Zudem ist der Anteil an simplen Standardtätigkeiten eher gering – es gibt keine fingierten Übungen von der Stange, die losgelöst vom operativen Tagesgeschäft gestellt werden. So tragen die Auszubildenden bei PROMATIS schon während der Ausbildung zur Produktivität bei. Ferner hat der Auszubildende in der Anwendungsentwicklung die Möglichkeit, an der „echten" Entwicklung von Applikationen beteiligt zu sein. Diese sinnhafte Tätigkeit, gerade im Backend von Kundenprojekten, fördert nicht nur die persönliche Entfaltung, sondern liefert dem Mitarbeitenden tatsächliche Erfolgserlebnisse. Das sind wesentliche Highlights, die diesen Berufsausbildungszweig so interessant machen, besonders wenn das Lösen von Softwareproblemen und damit einhergehend das Lob des Kunden im Vordergrund stehen.

Der Beweis für die Attraktivität der dualen Berufsausbildung konnten wir vor allem in den letzten Jahren feststellen, da sich tendenziell mehr Abiturientinnen und Abiturienten für eine Ausbildung als Fachinformatiker entscheiden. Für diesen Ausbildungsberuf sind vorzugsweise analytisches Denken und ein Verständnis für Mathematik elementare Voraussetzungen. Ein nicht zu unterschätzender Aspekt, gerade wenn es um die Bindung

Abbildung 1: Learning Center im PROMATIS Intranet "inside" (© PROMATIS)

und Weiterentwicklung des Human Capital geht, liegt allerdings ebenso in der nachhaltigen Zufriedenheit der Belegschaft. Hierbei steht das nach innen gerichtete Employer Branding auf der Agenda, das sich in der intrinsischen Motivation niederschlägt.

PERSONALENTWICKLUNG – MOTIVATION LEICHT GEMACHT

Personalentwicklung erfüllt das Bedürfnis eines jeden Mitarbeitenden nach stetiger beruflicher Entfaltung. Um das Potenzial, Talent, aber auch die Leistung jedes Einzelnen einzuordnen und weiterzuentwickeln, werden Zielvereinbarungen ausgearbeitet, die vorrangig die persönlichen Fähigkeiten berücksichtigen. Transparent mit dem Vorgesetzten kommuniziert und dokumentiert werden klar messbare und erreichbare Ziele gesetzt.

Durch die regelmäßigen Feedbackgespräche werden zudem Aktivitäten und Fortschritte formuliert, die beim Jahresgespräch noch einmal angesprochen und untermauert werden. Mittlerweile finden auch Zwischengespräche halbjährlich statt, um den jeweiligen Entwicklungsprozess eines Mitarbeitenden besser beobachten und beurteilen zu können. Denn in einer agilen Umgebung der VUCA-Welt (VUCA = volatility, uncertainty, complexity und ambiguity) wird es für uns immer interessanter, die Mitarbeitenden zur persönlichen Förderung für zeitlich begrenzte Einsätze und Aufgaben heranzuziehen und zusammenzubringen. Hierbei müssen die Mitarbeitenden für Projekte, Wissensaustausch oder auch für Mentoring und Coaching als richtiges „Match" gefunden werden. Es braucht also den Überblick, über welche Fähigkeiten, Interessen und Motivation jeder einzelne Mitarbeitende im Unternehmen verfügt.

WISSENSTRANSFER UND INDIVIDUELLE WEITERBILDUNG ALS SCHLÜSSEL ZUM ERFOLG

Die Weiterbildung der Mitarbeitenden ist im Unternehmen ein zentrales Thema, denn die Investition in Wissen und Qualifikation ist sowohl für jeden einzelnen Mitarbeitenden als auch für das gesamte Unternehmen ein Gewinn. Zentraler Ort im PROMATIS Intranet „inside" ist hierzu das Learning Center, an dem sich Angebot und Nachfrage rund um die Schlüsselressource Bildung treffen. Aus dieser Bildungsquelle schöpfen die Mitarbeitenden wertvolle Potenziale für ihre ganz individuelle Weiterentwicklung. So finden beispielsweise regelmäßig Inhouse-Schulungen statt, unsere sogenannte „Lernfabrik", in der Spezialisten ihr Know-how teilen. Weiterhin werden interne und externe Trainings- und Zertifizierungsmöglichkeiten sowie Zusammenstellungen von unterschiedlichen Wissensartefakten zum Selbststudium angeboten.

Die betriebliche Weiterbildung umfasst neben klassischen Kursen auch Besuche von externen Seminaren, Fachmessen und Tagungen. Fachliche Weiterbildungen und Qualifizierungen sind wichtige Voraussetzungen für eine erfolgreiche Personal- und dadurch Unternehmensentwicklung. Dies geschieht durch Oracle-Zertifizierungen und interne Zertifizierungen bezüglich des PROMATIS-Vorgehensmodells. Darüber hinaus können jedoch auch Weiterbildungsmöglichkeiten im Zusammenspiel mit verschiedenen Hochschulen und Universitäten mit den Mitarbeitenden individuell abgestimmt werden. Dies sind beispielsweise Modelle für ein Masterstudium oder gar eine Promotion in einer sinnvoll abgestimmten Kombination mit dem Arbeitsleben.

Diese Form der Weiterentwicklung der eigenen Mitarbeitenden kommt natürlich auch den Kunden und Geschäftspartnern zugute. Eine entsprechend umgesetzte individuelle Personalentwicklung führt dadurch auch automatisch zu einer Weiterentwicklung der gesamten Unternehmensorganisation und zu einer Verbesserung der Qualität der angebotenen Produkte und Services. Somit können höchste Qualitätsstandards gewährleistet, das Produkt- und Leistungsportfolio sukzessiv erweitert und die Lösungskompetenz sowie Project Excellence permanent verbessert werden.

Die Auszeichnungen als TOP-Arbeitgeber für IT-Jobs (2021) und „Leading Employer 2022" spiegeln diese vertrauensvolle und nachhaltige Erfolgsgeschichte des Unternehmens und seiner Mitarbeitenden wider, denn qualifizierte Fachkräfte sind der elementare Schlüssel einer aufstrebenden Unternehmensentwicklung.

Nadine Wagner ist Director Human Capital Management der PROMATIS Gruppe mit Schwerpunkt Talent Acquisition. Nach ihrem Studium der Wirtschaftspsychologie in Heidelberg war sie bereits in verschiedenen Positionen im Personalwesen tätig und hat die Themen der nachhaltigen Personalentwicklung und Weiterbildung vertieft.

Michael Pergande ist Executive Vice President und Mitglied des Management Board der PROMATIS Gruppe. Als Leiter der internen Systemadministration ist er ebenso verantwortlich für die Auszubildenden der technischen Fachrichtungen. Seit seinem Studium der Informatik an der Universität Karlsruhe (TH) mit Abschluss Diplom-Informatiker hat er über 20 Jahre Erfahrung in den Bereichen Geschäftsprozesse, IT-Infrastruktur und Informationssysteme sowie ausgeprägte Kenntnisse im Einsatz neuer Technologien beim Aufbau von Web-Portalen, bei Integrations- und Collaboration-Lösungen, Workflow-Systemen und hoch transaktionsorientierten Systemen im Internet gesammelt.


DOAG 2022
Konferenz + Ausstellung
In Nürnberg
20.-23. SEPT.
Die Oracle-ANWENDERKONFERENZ
anwenderkonferenz.doag.org
Eventpartner:
AOUG AUSTRIAN ORACLE USER GROUP
SOUG swiss oracle user group
iJUG Verbund

LORENA HOFFMANN, GEA GROUP AG

Von der Motivation, neben dem Beruf ein Masterstudium zu ergreifen

Die Überlegung, ein Masterstudium zu machen, beschäftigt viele in meinem Umfeld. Ist es doch, einmal im Berufsleben angekommen, gar nicht so selbstverständlich, wieder „die Schulbank zu drücken". Mit einem Rückblick über mein Studium im Bereich Business Consulting & Digital Management und der Reflexion über das Studium hinsichtlich meiner täglichen Aufgaben im Beruf sowie der uns alle betreffenden digitalisierten Welt kann ich sagen: Es lohnt sich.

Mache ich noch meinen Master? Was erwarte ich von einem Masterstudium? Und was mache ich dann damit? Drei Fragen, die ich mir immer wieder stellte und deren Antworten ich mittlerweile kenne. Ich bin 29 Jahre alt und befinde mich gerade in der finalen Phase meines Masterstudiums, genau genommen in den letzten vier Wochen meiner Masterarbeit. Insofern kann die erste Frage hier direkt mit einem Ja beantwortet werden. Das Studium habe ich im Herbst 2019 neben meinem Beruf als Projektmanagerin im Bereich Sales Excellence und Product Ownerin einer globalen Analytics-Plattform für Sales & Marketing mit weltweit über 1200 Usern eines internationalen Unternehmens im Maschinen- und Anlagenbau begonnen.

Ich habe das Studium an der FOM Hochschule mit dem Titel „Business Consulting & Digital Management" gewählt. Die FOM ist „Deutschlands Hochschule für Berufstätige". Ich habe mir davon erwartet, dass es mir in meinem Beruf hilft, aktuelle Themen und Projekte im Bereich der Digitalisierung und Innovation sowie der Beratung unserer Landesgesellschaften effizient anzugehen und voranzutreiben. Ebenso habe ich mir versprochen, neue Ideen für Geschäftsmodelle zu generieren und zu bearbeiten. Ich wollte nicht „einfach nur den Master machen, damit er gemacht ist". Das war mir wichtig.

Die Inhalte des Studiums erschienen mir hierfür ein perfektes Match zu sein: Es geht um Grundlagen der Digitalisierung und damit einhergehende Veränderungen für Organisationen und Unternehmen, daraus entstehende Notwendigkeiten der digitalen Transformation, Methodiken und Ansätze für Geschäftsmodellinnovationen sowie Themen wie New Work, Design Thinking, dabei geltendes IT-Recht und Compliance. Fünf Semester an der FOM in Mannheim am Donnerstag und Freitagabend sowie samstags sollten dies ermöglichen.

LORENA HOFFMANN
lorena-hoffmann@web.de

DIGITALE DATEN UND DIGITALISIERTE PROZESSE VERSTEHEN

Starten möchte ich an dieser Stelle mit Modulen aus den ersten beiden Semestern, nämlich Grundlagen im Bereich Digital Business und der Digitalisierung, wobei notwendige Definitionen (jeder redet von Digitalisierung, doch was ist das nun genau?) und resultierende Auswirkungen auf die Wirtschaft sowie Unternehmen behandelt wurden. Dazu zählen mögliche Strategien und die Entwicklung von neuen,

innovativen Geschäftsmodellen, um erfolgreich am Markt bestehen zu bleiben und Herausforderungen meistern zu können. Konkret ging es dann um Daten, also den Einsatz von Big Data und diverse Anwendungen der künstlichen Intelligenz, mögliche Einsatzfelder und Ansätze zur Nutzung und Verwertung der Daten, aber auch bestehende Richtlinien und Herausforderungen im Umgang mit den Daten. Weitere Inhalte bildeten Aktivitäten und Ansätze im Bereich des eBusiness als Teil von umfassenden Omni-Channel-Strategien sowie IT-Recht und Compliance, die solche Themen begleiten und die auf den ersten Blick trocken erscheinen mögen. Doch man wird sich wundern, wie oft wir im privaten Alltag genau diesen Themen gegenüberstehen, zum Beispiel beim simplen Einkauf auf Websites von Bekleidungsgeschäften oder ehemaligen Buchhändlern, die mittlerweile gefühlt alles anbieten und innerhalb eines Tages liefern. Retourenfälle und DSGVO-Verstöße aus genau diesen bekannten Bereichen halfen, die Thematik zu begreifen. Und natürlich finden diese Themen auch im geschäftlichen Alltag der Unternehmen statt.

Um den Erfolg solcher Unternehmen zu verstehen, war ein weiterer Schwerpunkt im Studium die Thematik von bekannten, traditionellen Geschäftsmodellen im Vergleich zu neuen, innovativen Geschäftsmodellen. Sie etablieren sich mehr und mehr durch diesen Wandel und verändern die Welt auf „disruptive“ Art und Weise – man hat diesen Ausdruck in Verbindung mit Facebook, Uber, Airbnb und Co. sicher schon öfters gehört. Also standen Themen wie digitale Plattformen als erfolgreiche Form von Geschäftsmodellen sowie Modelle wie X as a Service, Pay-Per-Demand und Outcome etc. auf dem Plan. Nun konnte man Fragen wie „Wieso sind die denn so erfolgreich?“ und „Wieso sind die so viel Wert?“ fundiert beantworten. Auch hier verstand man diverse TV-Streaming-Dienste plötzlich neu und die Quintessenz lautete kurz und knapp: Digitale Daten und digitalisierte Prozesse sind überall und ein gutes Verständnis darüber zur Ausschöpfung unzähliger Möglichkeiten ist mindestens so wichtig wie die Kenntnis und das Auseinandersetzen damit einhergehender Herausforderungen.

WIESO DAS VERSTÄNDNIS ÜBER DIGITALE DATEN UND PROZESSE FÜR MICH UND MEINEN BERUF WICHTIG IST

Als Projektmanagerin im Bereich Sales Excellence und Business Intelligence konnte ich neben den privaten Gesichtspunkten mit solchen Inhalten vor allem viele Bezugspunkte zu meiner täglichen Arbeit herstellen. Darüber hinaus konnte ich auch Ideen generieren sowie eigene Ansätze für künftige Projekte weiterentwickeln. Denn auch innerhalb meiner Projekte geht es um die Digitalisierung der Customer Journey unserer Kunden. Unsere globale Analytics-Plattform SHARE, für die ich Product Ownerin bin, nutzt verschiedenste Daten aus den Bereichen Sales, Marketing sowie Performance-Daten unserer Maschinen. Die Erkenntnisse dieser Daten verwende ich für diverse BI- und Analytics-Anfragen, weltweite Trainings sowie im Zuge von Strategiediskussionen.

Genannte innovative Geschäftsmodelle spielen sowohl für mein Unternehmen als auch für meine tägliche Arbeit eine Rolle, da mein Unternehmen selbst diese Trends identifiziert hat und innerhalb der aktuellen Strategie verfolgt. Neuartige Geschäftsmodelle wie beispielsweise Equipment as a Service oder auch Subscription genannt, bilden aktuelle Ausrichtungen unserer Produktstrategie. Dabei bildet ein Leistungsversprechen, das wir als Lösungsanbieter an unsere Kunden geben und wofür die Auswertung von Daten als Beratungsdienstleistung genutzt wird, einen entscheidenden Faktor.

Die Module des Studiums passten insofern nicht nur sehr gut zu meinem beruflichen Alltag, sondern greifen auch ineinander beziehungsweise bauen aufeinander auf. Ich konnte mir dementsprechend ein umfassendes Verständnis von Geschäftsmodellen verschaffen, vor allem aber neue Formen davon im Kontext der Digitalisierung einordnen und auf unsere Strategie übertragen. Durch das Bearbeiten und Diskutieren von diversen Beispielen digitaler Geschäftsmodelle im Kurs konnte ich die eigenen Erfahrungen außerdem erweitern und im Vergleich zu anderen Industrien einschätzen.

Offensichtlich war nach den ersten Semestern also bereits ein Transfer zwischen Theorie und Praxis möglich – so konnte es weitergehen.

PRAXISORIENTIERTE PROJEKTE IN VUCA-ZEITEN

Unter dem Einfluss der Pandemie fanden die nächsten Semester fortan remote statt. Die Inhalte hier waren die organisationale Transformation, die mit den beschriebenen Veränderungen der Welt und der Wirtschaft für Unternehmen notwendig wird. Hinzu kamen existierende Entscheidungstheorien in unterschiedlich komplexen Umgebungen und Managementmethoden im Umgang solcher Veränderungen. Ironischerweise waren diese Veränderungen durch die bestehende Pandemie spürbarer, als man das erwartet hätte.

In den Modulen des Studiums lernten wir solche Umweltveränderungen als VUCA kennen. VUCA bedeutet in diesem Kontext instabil, unsicher, komplex und mehrdeutig. Genau diesen Herausforderungen sahen sich nicht nur wir als Studenten konfrontiert, sondern auch mein Unternehmen. Dieses bewegt sich seit Jahren in einem sich stark verändernden, das Thema der Konsolidierung steht dabei ganz oben. Damit einher gehen veränderte, meist gestiegene Anforderungen unserer Kunden. Diese resultieren wiederum in notwendigen Anpassungen unsererseits. Daneben trug die Corona-Pandemie dazu bei, dass stabil erscheinende Kunden innerhalb kürzester Zeit Probleme bei Auftragseingängen hatten. Und auch wir befanden uns monatelang in Kurzarbeit – dies hat natürlich Einfluss auf das Tagesgeschäft, das trotzdem funktionieren muss. Entsprechend waren viele Vergleiche zwischen erlernten VUCA-Inhalten möglich. Ein weiterer Aspekt, den ich mit mei-

nem Tagesgeschäft optimal verknüpfen konnte, waren die Vorstellung des New-Work-Konzepts sowie entsprechende Methoden und Initiativen, da wir genau dies bei der zwei Jahre zuvor gegründeten Digital Unit leben. Ich bin selbst Teil unseres interdisziplinären Teams der Digital Unit, wo wir demokratisch über unser Arbeitsumfeld, die Einrichtung der Büroräume sowie soziale und methodische Aspekte entscheiden. Die Kenntnis über New Work und die damit verbundenen Initiativen und Arbeitsweisen helfen mir, diese in unserem New-Work-Konzept einzubringen, neue Ideen zu liefern und Verbesserungspotenziale aufzuzeigen.

Neben genannten Inhalten, die wir in den Vorlesungen erlernen, bilden Praxisprojekte und praxisnahe Aufgabenstellungen einen weiteren großen Teil des Studiums, vor allem zum Ende hin. Diese Inhalte sollen innerhalb eines Projektteams gewählt und erarbeitet werden, was aus mehreren Gründen eine effiziente Vorgehensweise darstellt. Zunächst setzt man sich mit der Erarbeitung der Themen an einem realen Beispiel nochmal mehr mit dem Thema auseinander, als beispielsweise Inhalte „nur" durch Lesen zu lernen. Außerdem gewinnt man relevante Einblicke aus der Praxis, je nach Projekt entweder von bereits bekannter Praxis oder im Falle von Inhalten von Studienkollegen bisher völlig fremde. Man spiegelt solche Erkenntnisse mit den bereits erlernten Inhalten aus vergangenen Semestern wider und diskutiert diese außerdem mit anderen Studierenden. Insofern waren anstehende Projekte fordernd, aber auch fördernd. Und trotz weiterhin anhaltender Remote-Organisation des Studiums und dahingehend auch lediglich digitalen Projektmeetings war dennoch eine gewinnbringende Erarbeitung der projektbezogenen Inhalte, beispielsweise die Digitalisierung von Geschäftsprozessen oder die Erarbeitung von Geschäftsmodellen, möglich. Dieser Weg erforderte zudem eine höhere Anpassungsfähigkeit und Ehrgeiz – Voraussetzungen, die keiner Karriere schaden.

ERARBEITUNG DIGITALER GESCHÄFTSMODELLE DURCH BEWÄHRTE METHODEN MEINER TÄGLICHEN ARBEIT

Besonders lehrreich empfand ich hier das Modul „Consultingprojekt: Digital Business Design" aus dem vierten und damit vorletzten Semester unter der Leitung von Herrn Dr. Reinhard Ematinger [Anm. d. Red.: Lesen Sie auch den Beitrag von Dr. Reinhard Ematinger in der Red Stack inkl. Business News Nr. 01/2022: „Brücken bauen: Mit LEGO SERIOUS PLAY® auf die Zukunft vorbereitet sein", ab S. 88]. Inhalt und Aufgabe der Vorlesung war es, ein digitales Geschäftsmodell zu einem gewählten Thema im Verlauf des Design-Thinking-Prozesses zu erarbeiten und dabei zumindest den Verlauf und die Entstehung eines Geschäftsmodells durch die Phasen der Ideation, Generation sowie Evaluation zu untersuchen. Damit wurden sowohl Erkenntnisse im Bereich digitale Geschäftsmodelle als auch Methodiken des Design Thinking erarbeitet und neu gewonnen. Unser gewähltes Thema war dabei

die „Förderung des informellen Austauschs in Zeiten der Pandemie“. Es ging also um die Erarbeitung von Ideen zu Geschäftsmodellen, die den informellen Austausch trotz der andauernden Home-Office-Situation in Unternehmen unter Kollegen und Teams fördern sollen, um dadurch die Motivation bei den Mitarbeitern, den Zusammenhalt und damit einhergehend auch die Produktivität und Effektivität der Arbeit zu steigern. Ziel war es, Überlegungen zu einem Prototyp einer solchen Lösung zu erstellen und diese mit unterschiedlichen Werkzeugen wie beispielsweise dem Business Model Canvas sowie die Customer Journey und die Persona Canvas zu bearbeiten. Ebenso sollten aufgestellte Hypothesen und Ideen bestimmten Tests unterzogen und somit geprüft werden. Das Modul half neben der Generierung von Ideen in einer bisher neuen Thematik auch dabei, verschiedene Werkzeuge auszuprobieren und zu vergleichen und anschließend zu beurteilen, wann welches Werkzeug am ehesten geeignet ist. Ein Verständnis und eine Vergleichbarkeit verschiedener Werkzeuge wurde geschaffen. Weiterhin wurden im Zusammenhang agiler Arbeitsmethoden weitere Ansätze gelehrt, allen voran Scrum.

Auch für meine tägliche Arbeit sehe ich mit dem Erlernen und Festigen dieser Tools konkrete Einsatzmöglichkeiten, beispielsweise in der aktuellen Migration der Customer-Relationship-Management-Plattform. Dafür ist das Verständnis über relevante Prozesse aus der Sicht mehrerer Stakeholdergruppen erforderlich, wofür wir bereits die Persona-Methodik nutzen. Auch die Tatsache, dass das Projekt in enger Zusammenarbeit mit der IT anhand der Scrum-Methodik durchgeführt wird, baut in meinem Falle auf erlernte Inhalte des Studiums auf und ermöglicht für mich einen fundierten Einsatz der Methodik und damit sicheren Umgang. Die Verantwortung als Product Ownerin der globalen Analytics-Plattform meines Unternehmens und dafür regelmäßige Scrum-Methodiken zur Weiterentwicklung der Plattform bedingen ebenfalls den professionellen Umgang mit solchen Methoden.

DER ABSCHLUSS DES STUDIUMS UND EINE MASTERTHESIS ZUM THEMA „WACHSTUMSSTRATEGIEN DIGITALER GESCHÄFTSMODELLE“

In jedem Studium steht nach bestandenen Projekten und Klausuren eine Abschlussarbeit an. In meinem Falle geht es um Wachstumsstrategien digitaler Geschäftsmodelle im B2B-Bereich, konkret um die Frage, wie Unternehmen die kritische Masse hierfür erreichen können. Die kritische Masse an Kunden, die wiederum Daten liefern, damit bestimmte digitale Geschäftsmodelle „fliegen“. Eine theoretische Grundlage schaffe ich hier durch das im Studium Erlernte, ein Zusammenspiel relevanter Inhalte wie die digitale Transformation, innovative Geschäftsmodelle und den Einsatz von Daten und künstlicher Intelligenz. Dazu sollen qualitative Interviews mit Vertretern verschiedener Unternehmen aus dem Industrie- und Anlagenbau weitere Erkenntnisse liefern. Mit knapp 15 Interviews bin ich gerade mitten in der Transkription und Auswertung. Ich bin immer wieder erstaunt, wie viel ich doch durch die vergangenen Semester zu diesem Thema schon lernen durfte und wie gut sich das auf die Masterarbeit auswirkt.

Zum Zeitpunkt des Verfassens dieses Artikels sind es noch vier Wochen bis zur Abgabe meiner Masterarbeit. Das ist ein guter Zeitpunkt, um meine drei Fragen im Hinblick auf ein Masterstudium zu beantworten. Mache ich den Master? Offensichtlich ja. Und mein Ziel ist es auch, ihn fertig zu machen. Was erwarte ich mir davon? Meine genannten Erwartungen wurden zweifellos erfüllt, vielleicht sogar übertroffen. Ein Transfer war immer möglich und gewinnbringende Erkenntnisse daraus ebenso. Ein wichtiger Faktor war hier sicherlich, dass die Inhalte so gut zu meinem Beruf gepasst haben. Daher empfehle ich, den Master nicht direkt nach dem Bachelor zu machen, sondern dann, wenn man bereits ein paar Jahre in dem Bereich arbeiten konnte, der einem Spaß macht, um genau jene Themen vertiefen zu können, von denen man tagtäglich umgeben ist und mit denen man etwas erreichen möchte.

Mit den behandelten Themen und Schwerpunkten bleiben für mich vor allem die Bedeutung und das Potenzial des vielfältigen Einsatzes von Daten hängen. Wir befinden uns in einer Welt, die man sich ohne digitale Daten und deren elektronische Verarbeitung nicht mehr vorstellen kann. Und das wird wohl so weitergehen – im privaten und auch professionellen Umfeld, wenn man beispielsweise die Potenziale der künstlichen Intelligenz und der digitalisierten Prozesse betrachtet, auf die viele der genannten Geschäftsmodelle aufbauen. Die Chancen, aber auch Herausforderungen dahingehend zu kennen, ist absolut wichtig. Ein Studium, das einen Schwerpunkt auf IT-relevante Themen legt, ist meiner Meinung nach für meine und die kommenden Generationen ein Erfolgsfaktor und je mehr Gesichtspunkte und Themen dabei übermittelt werden, desto gewinnbringender können Absolventen dieses Wissen einsetzen.

Bleibt noch die Antwort auf die Frage, was ich nun damit mache. Als Product Ownerin eines digitalen Geschäftsmodells durfte ich mit dem Wechsel zur GEA Group AG kürzlich eine neue, spannende Herausforderung annehmen. Dieser neue Job wäre ohne das Studium wohl so nicht möglich geworden oder zumindest deutlich „weiter weg“. Und genau das soll so ein Masterstudium doch tun – uns helfen, voranzukommen und Ziele zu erreichen.

Lorena Hoffmann stieg nach ihrem Bachelorabschluss mit der Fachrichtung International Business im Bereich Sales Excellence bei der Heidelberger Druckmaschinen AG ein, um gemeinsam mit den Landesgesellschaften im Zuge einer effizienten Vertriebssteuerung zusammenzuarbeiten. Projekte zur Digitalisierung der Customer Journey und sowie der Geschäftsmodellentwicklung weckten dabei ihr Interesse an den Themen Digitalisierung und Innovation. Mit dem Abschluss des Master of Science zum Thema Digital Management erweiterte sie ihr Wissen in diesem Gebiet und widmet sich mit dem Wechsel zur GEA Group AG als Product Ownerin seit Mai 2022 der Entwicklung digitaler Geschäftsmodelle im Bereich des Energiemanagements

KAI-UWE FISCHER, LOGICALIS GMBH

Oracle Cloud Infrastructure meets Microsoft Azure – wie Unternehmen von den Vorteilen zweier Cloud-Dienste profitieren

KAI-UWE FISCHER
kai-uwe.fischer@logicalis.de

Skalierbarkeit auf Abruf, flexible Abrechnungsmodelle, bessere Performance und keine Investitionskosten – das sind nur einige der Vorteile für Unternehmen, die sich für eine Public Cloud entscheiden. Doch damit nicht genug: Verknüpfen sie zwei Public Clouds zu einer Multi Cloud, ergeben sich weitere interessante Synergien – beispielweise bei der Kopplung der Oracle Cloud Infrastructure (OCI) mit der Microsoft Azure Cloud; hier können Unternehmen von den Vorteilen des Oracle-Datenbankbetriebs im Database-as-a-Service-Modell (DBaaS) in der OCI profitieren. Dank der Multi Cloud ist es möglich, Oracle- und Nicht-Oracle-Applikationen sowie das Backend und Frontend in genau der Cloud zu betreiben, die dafür am besten geeignet ist. Doch welche Vorteile bietet die Multi Cloud in der Praxis genau, und was müssen Unternehmen bei der Einführung beachten?

IT-VERANTWORTLICHE IN DER ZWICKMÜHLE – PUBLIC CLOUD SCHAFFT ABHILFE

CIOs und IT-Verantwortliche in Unternehmen sitzen auch weiterhin zwischen den Stühlen: Einerseits müssen sie durch die laufende Aktualisierung der IT-Infrastruktur die nötige Performance der Datenbanken und Applikationen gewährleisten, niedrige Latenz- und Ausfallzeiten sowie die erforderliche Bandbreite zur Übertragung großer Datenmengen zur Verfügung stellen.

Auf der anderen Seite belasten laut einer aktuellen Capgemini-Studie *[1]* Ausgaben für den Erhalt und die Modernisierung das IT-Budget. Dabei schwinden Ressourcen und IT-Fachkräfte gleichermaßen, Datenmengen und Technologie-Stacks hingegen steigen. Um die Verantwortlichen zu entlasten und sich zukunftssicher aufzustellen, bieten Public-Cloud-Angebote verschiedene Benefits:

- Kontinuierliche Investitionen: Anbieter von Public-Cloud-Lösungen investieren stetig in State-of-the-Art-Technologien und bieten ein umfassendes Portfolio an Dienstleistungen und Services.
- **Flexibel und skalierbar:** Sowohl die Speicherkapazität als auch die Rechenleistung lassen sich in einer Public Cloud flexibel und nahezu unbegrenzt skalieren. Damit stellen sich die Nutzer agil für den digitalen Wandel auf.
- Hohe Verfügbarkeit: Service Level Agreements (SLAs) garantieren eine hohe Verfügbarkeit sowie feste Reaktionszeiten bei Fehlern und Problemen.
- **Faires Preismodell:** Dank eines Pay-as-you-go-Modells zahlen Unternehmen nur die Ressourcen und Services, die sie auch tatsächlich nutzen.

Führen Unternehmen mehrere unterschiedliche Technologie-Stacks aus, ist es ratsam, eine integrierte Multi-Cloud-Strategie zu nutzen – damit lässt sich für jeden Anwendungsfall der optimale Nutzen aus der Public Cloud ziehen. Beispiel: Die Verbindung der OCI und der Microsoft Azure Cloud zu einer Multi Cloud. Diese bietet Kunden die Möglichkeit, Azure-Dienste mit Oracle-Cloud-Diensten zu verbinden. In dieser Kombination profitieren die Unternehmen von einem performanten, hochverfügbaren und ausfallsicheren IT-Betrieb.

OCI UND AZURE – DIE VERBINDUNG MACHT'S

Per Private Network Interconnect (PNI) sind die OCI und Azure über eine dedizierte, virtuelle private Verbindung gekoppelt; dies gewährleistet eine hohe Bandbreite und Performance sowie geringe Latenzen. Verknüpft sind die Clouds über den OCI / Azure Interconnect, der dank FastConnect auf Oracle-Seite und ExpressRoute auf Azure-Seite eine schnelle, latenzarme Netzwerkverbindung zwischen den beiden Cloud-Anbietern ermöglicht.

IT-Verantwortliche, die einen OCI / Azure Interconnect aufbauen möchten, müssen wie folgt vorgehen. Der Aufbau des Interconnect wird am Beispiel der Region Amsterdam dargestellt. Die Vorgehensweise in einer anderen Region wie beispielsweise Frankfurt erfolgt analog.

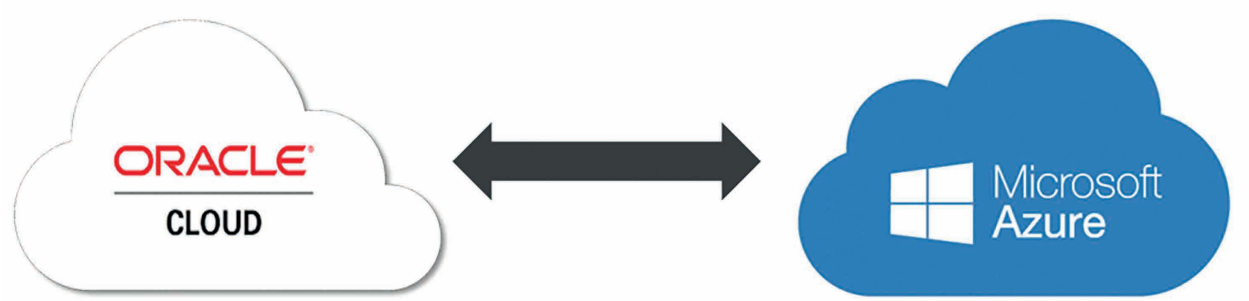
Auf der Microsoft-Seite werden ein beliebiges Azure Virtual Network (VNET) mit mindestens einem CLIENT Subnetz, ein Subnetz für das Virtual Network Gateway und ein VPN Gateway benötigt *(siehe Abbildung 1)*.
Auf der Oracle Seite sieht es ähnlich aus; hier werden ein Virtual Cloud Network (VCN) mit mindestens einem privaten Datenbank-Subnetz sowie ein Dynamic Routing Gateway benötigt *(siehe Abbildung 2)*.

Eine wichtige Sache gilt es dabei zu beachten: Die IP-Adressen der beiden virtuellen Netzwerke dürfen sich nicht überlappen. Beispielsweise kann für das Azure Netzwerk der CIDR-Block 10.10.0.0/16 und für Oracle der CIDR-Block 10.100.0.0/16 verwendet werden.

Sind diese Voraussetzungen gegeben, sind lediglich die folgenden Schritte für den Aufbau des Interconnect notwendig:

- Konfiguration einer ExpressRoute in Azure, dabei wird Oracle Cloud FastConnect als Provider ausgewählt. Zusätzlich müssen die Peering Location, in diesem Fall Amsterdam2, und die Bandweite festgelegt werden *(siehe Abbildung 3)* Nachdem die ExpressRoute angelegt wurde, muss sich der Nutzer aus der Übersicht den Service Key kopieren *(siehe Abbildung 4)*.

- Im nächsten Schritt wechselt der Nutzer auf die OCI-Console, erzeugt dort eine FastConnect-Verbindung und wählt als Partner Microsoft Azure ExpressRou-



Oracle Cloud Infrastructure meets Microsoft Azure

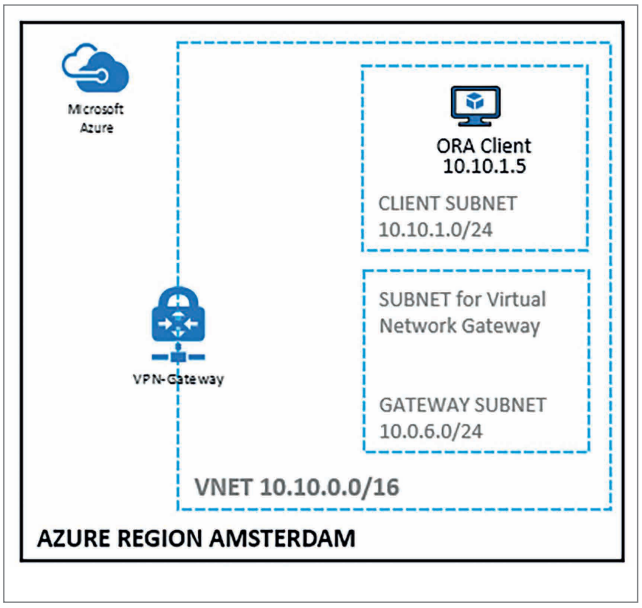


Abbildung 1: Azure VNET (© Kai-Uwe Fischer, Logicalis)

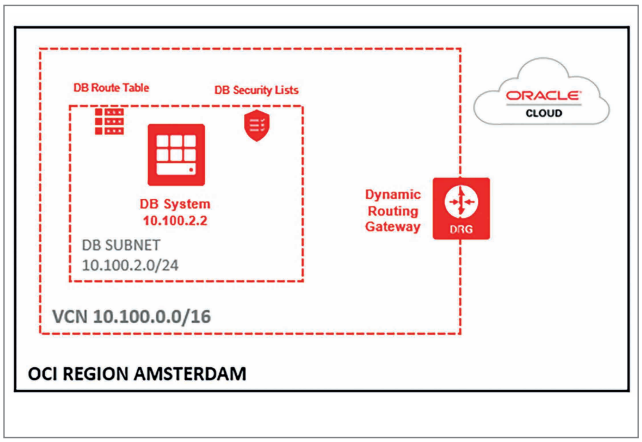


Abbildung 2: OCI VCN (© Kai-Uwe Fischer, Logicalis)

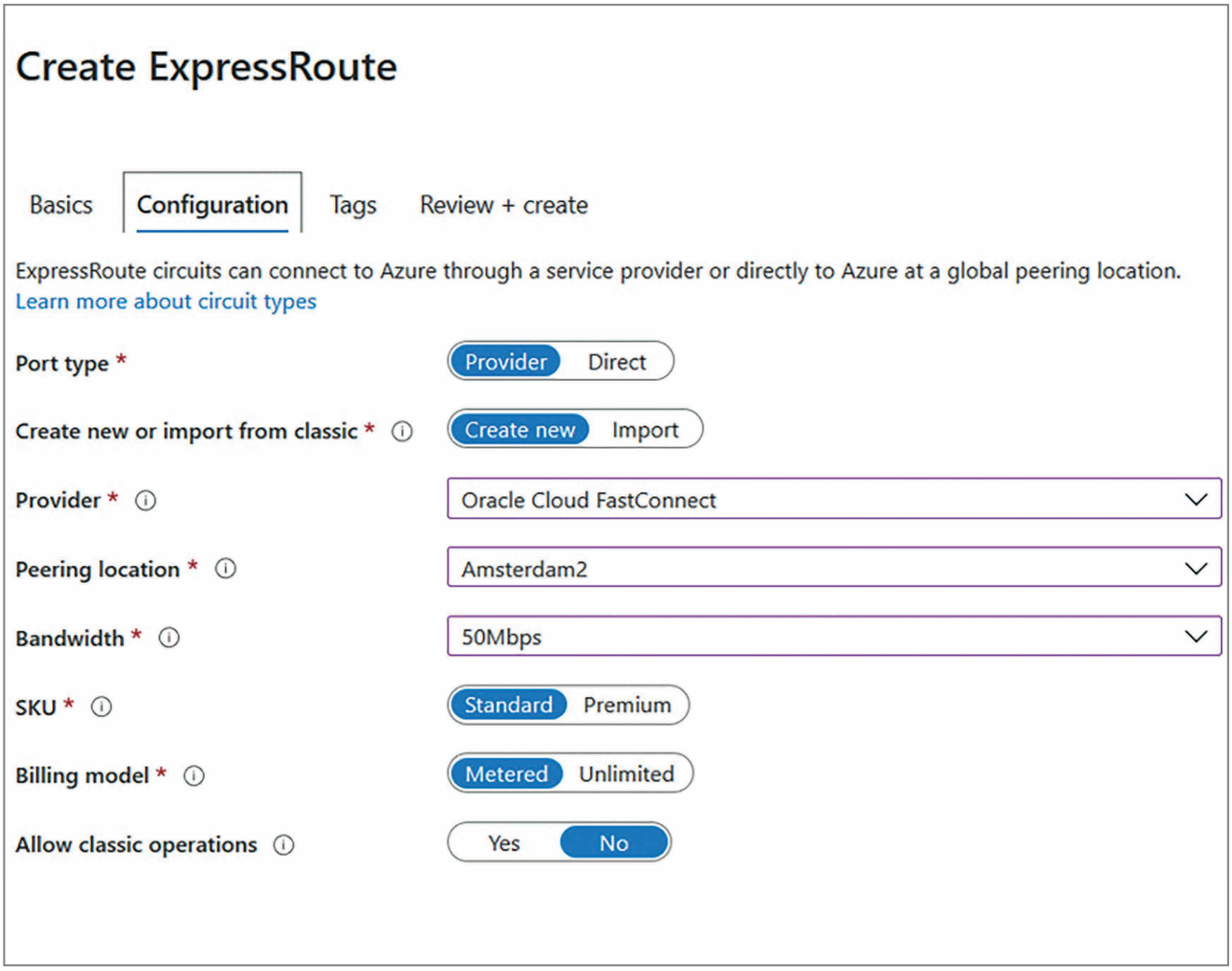


Abbildung 3: Create ExpressRoute (© Kai-Uwe Fischer, Logicalis)

te aus. Zudem kopiert er den Azure Service Key in das Feld PARTNER SERVICE KEY *(siehe Abbildung 5).*

- **Im Gegensatz zu Azure beträgt die kleinste zu konfigurierende Bandbreite 1 Gbps.**
- Nach dem Anlegen der FastConnect-Verbindung steht der Lifecycle Status in OCI auf PROVISONED *(siehe Abbildung 6).* Bei erfolgreichem Aufbau des Interconnect ändert sich der ExpressRoute Provider Status ebenfalls auf PROVISONED *(siehe Abbildung 7).*
- Im letzten Schritt muss der Nutzer einen Link zwischen dem Azure Virtual Network und dem ExpressRoute Circuit herstellen *(siehe Abbildung 8).*

Zusätzlich müssen auf beiden Cloud-Seiten die Route Tables und Security Groups beziehungsweise die Security Lists angepasst werden.

Da der Datenverkehr nicht internetbasiert, sondern ausschließlich über vertrauenswürdige Endpunkte erfolgt, ist er absolut sicher. Darüber hinaus ist kein zusätzlicher Netzwerk-Provider zum Einrichten der Verbindung zwischengeschaltet, was wiederum eine Kosteneinsparung für Unternehmen bedeutet.

Weitere Vorteile der Verbindung sind die zentrale und einheitliche Identitäts- und Zugriffsverwaltung per Single Sign-on (SSO) im Azure Active Directory (Azure AD) sowie ein automatisiertes User Provisioning zur Cloud-übergreifenden Verwaltung von Ressourcen mit Terraform-Skripten.

VON DER THEORIE ZUR PRAXIS: OCI UND AZURE IN DER ANWENDUNG

Unternehmen können mit der Multi Cloud die Vorzüge des Oracle-Datenbankbetriebs in der OCI im PaaS-Modell genießen. Das ist für viele Firmen interessant – nicht zuletzt, weil das Datenbankmanagementsystem von Oracle laut DB-Engines-Ranking *[2]* für Oktober 2021 weiterhin das beliebteste ist. Tauchen Unternehmen in die beiden Cloud-Welten ein, erwarten sie in der alltäglichen Arbeit wertvolle Benefits – konkrete Kundenanwendungsfälle zeigen, warum:

1. MIGRATION UND AUSFÜHRUNG VON FULL-STACK-ANWENDUNGEN ÜBER ORACLE UND AZURE HINWEG (SIEHE ABBILDUNG 9).

In diesem Fall laufen die Full-Stack-Anwendungen von Oracle in der OCI genauso wie die Full-Stack-Anwendungen in Azure autonom.

Die gemeinsam genutzten Daten werden untereinander latenzarm und sicher über einen privaten Interconnect ausgetauscht. Damit lassen sich von Oracle unterstützte SaaS-Applikationen mit einer Oracle-Datenbank in der OCI betreiben. Dazu laufen eigenentwickelte Anwendungen mit einem Fokus auf Microsoft-Schnittstellen oder Azure-Full-Stack-Anwendungen in der Azure Cloud.

Vorteile auf einen Blick: Es lassen sich Lizenzkosten optimieren und die Best Practices des jeweiligen Herstellers berücksichtigen. Ohne Neuarchitektur kann die hochleistungsfähige Konnektivität zwischen abhängigen Anwendungen in der Cloud aufrechterhalten werden.

2. MIGRATION UND AUSFÜHRUNG VON SPLIT-STACK-ANWENDUNGEN ÜBER ORACLE UND AZURE HINWEG

Der Anwendungs-Stack wird so aufgeteilt, dass die Applikationen und Oracle-Datenbank jeweils in der am besten geeigneten Cloud-Umgebung ausgeführt werden – das sorgt für eine optimale Performance. In der Azure Cloud laufen eine benutzerdefinierte .NET- und native Cloud-Lösungen. Diese greifen auf eine Oracle-Datenbank zu, die in der OCI betrieben wird.

Vorteile auf einen Blick: Die für den Betrieb einer Oracle-Datenbank notwendigen Lizenzkosten können reduziert und Oracle-Datenbanken als PaaS genutzt werden. Durch die Ausführung der Anwendungskomponenten in der optimalen Umgebung werden eine hohe Leistung und Gewinne erzielt.

3. INTEROPERABILITÄT DER SERVICES IN DEN CLOUDS, ENTWICKLUNG NATIVER CLOUD-ANWENDUNGEN

Die Verbindung von OCI und Azure ermöglicht es, native und performante Cloud-Anwendungen zu entwickeln. Konkret bedeutet das: Eine First-Party-Applikation wird in der OCI erstellt, beispielsweise in der Abstraktionsschicht der autonomen Datenbank (Database Abstraction Layer), während das Frontend auf dem Webhosting-Dienst Azure App Service als PaaS-Modell betrieben wird. Die Interoperabilität der Cloud-Services von Oracle und Microsoft sorgt dafür, dass sich Informationen aus einer Oracle-Datenbank und der OCI per Selfservice mit dem Analysetool Power BI in Azure abfragen, auswerten und visualisieren lassen.

Vorteile auf einen Blick: Cloud-übergreifendes Mischen und Zuordnen der Services ermöglichen eine optimale Leistung und Gewinne. Die konkreten Anwendungsfälle zeigen: Durch die Multi Cloud kann die Oracle Database in der OCI als genuine DBaaS-Lösung genutzt werden – das wiederum bietet zahlreiche Vorteile im Vergleich zu einem Betrieb im Infrastructure-as-a-Service-Modell (IaaS) in Azure.

EINSPARUNGEN DANK MULTI CLOUD

Mit der Multi-Cloud-Lösung können Unternehmen Lizenz- und Betriebskosten sparen. Denn: Die Kosten für eine Oracle-Datenbank in der OCI sind deutlich geringer als beim Betrieb in Azure. Bei Verwendung des flexiblen DBaaS- Abonnements lassen sich die Datenbankkosten deutlich senken, wenn bei Nichtnutzung der Services die Datenbank gestoppt wird. Dies ist vor allem für Test- sowie Development-Datenbanken interessant. Zudem können durch die Skalierung von Datenbank-OPCU-Lizenzen die Anwendungsanforderungen erfüllt und gleichzeitig die Kosten minimiert werden. Für eine FastConnect-Leitung zahlen User lediglich einen Flatrate-Preis für die genutzten Port-Stunden. Somit fallen weder Gebühren für die Einrichtung eines Ports noch für ein- und ausgehenden Datentransfer an.

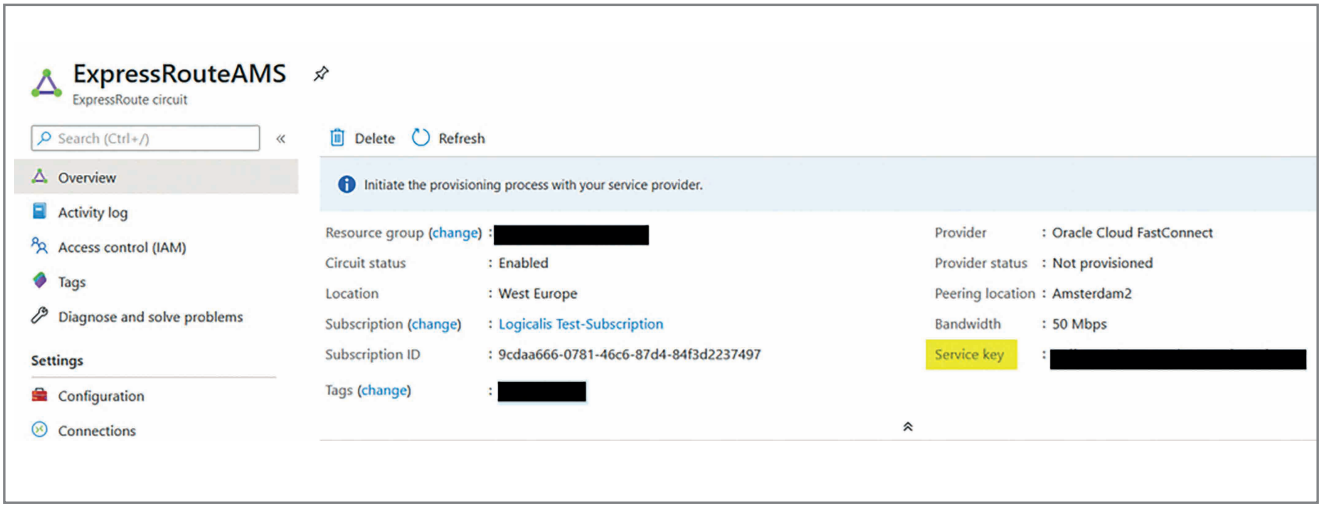


Abbildung 4: ExpressRoute Service Key (© Kai-Uwe Fischer / Logicalis)

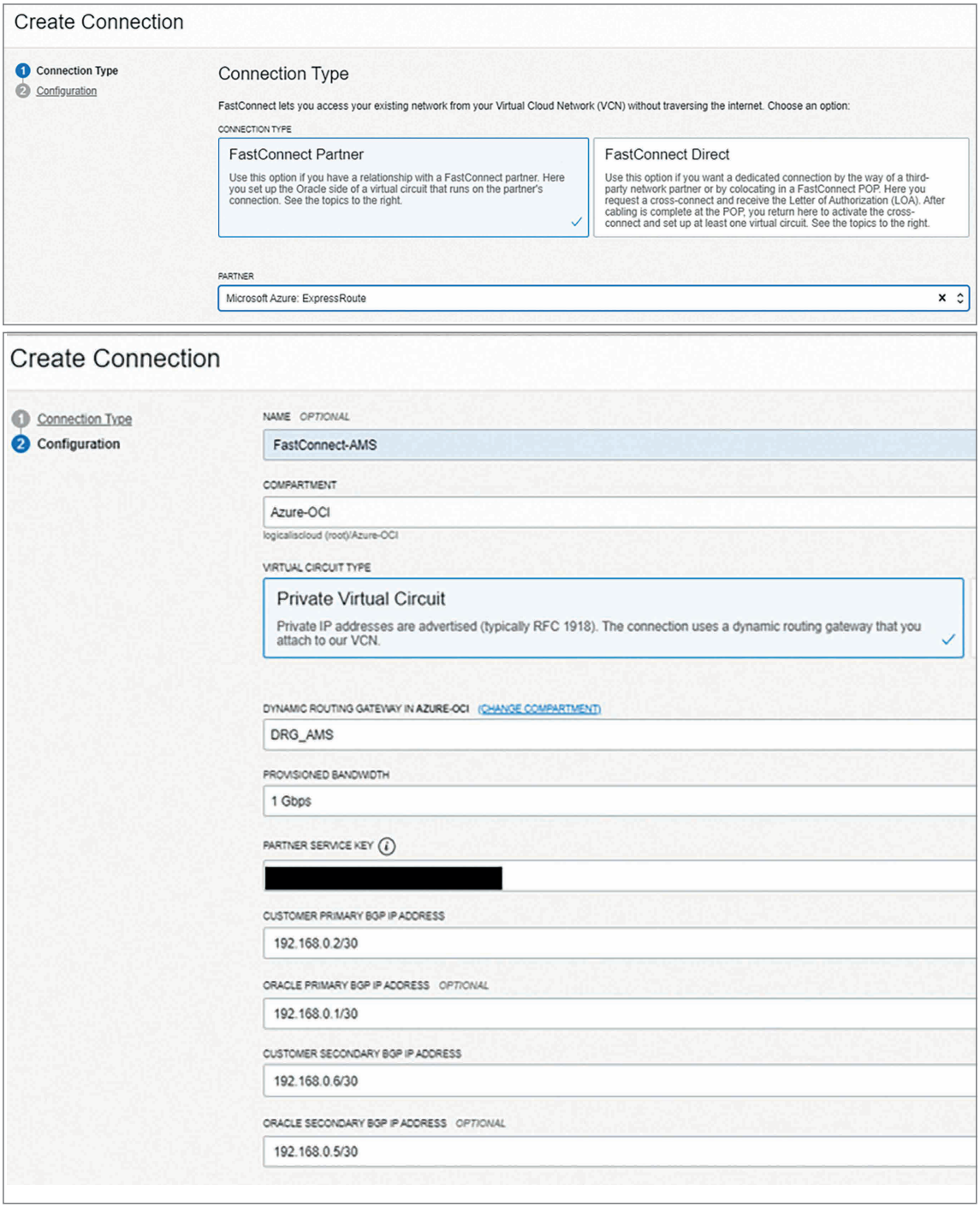


Abbildung 5: Create FastConnect (© Kai-Uwe Fischer, Logicalis)

Abbildung 6: Lifecyle Status in OCI (© Kai-Uwe Fischer, Logicalis)

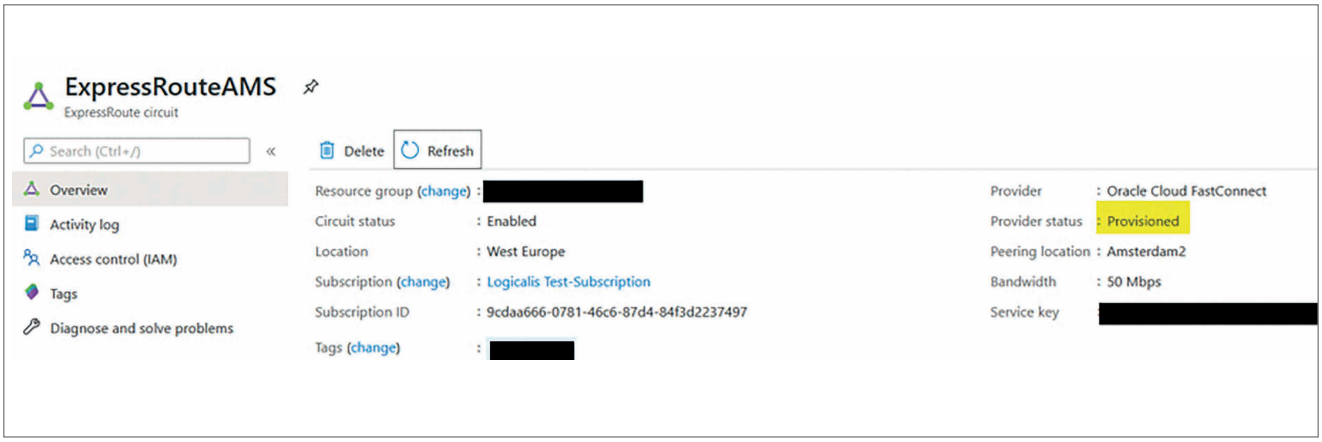


Abbildung 7: ExpressRoute Provider Status (© Kai-Uwe Fischer, Logicalis)

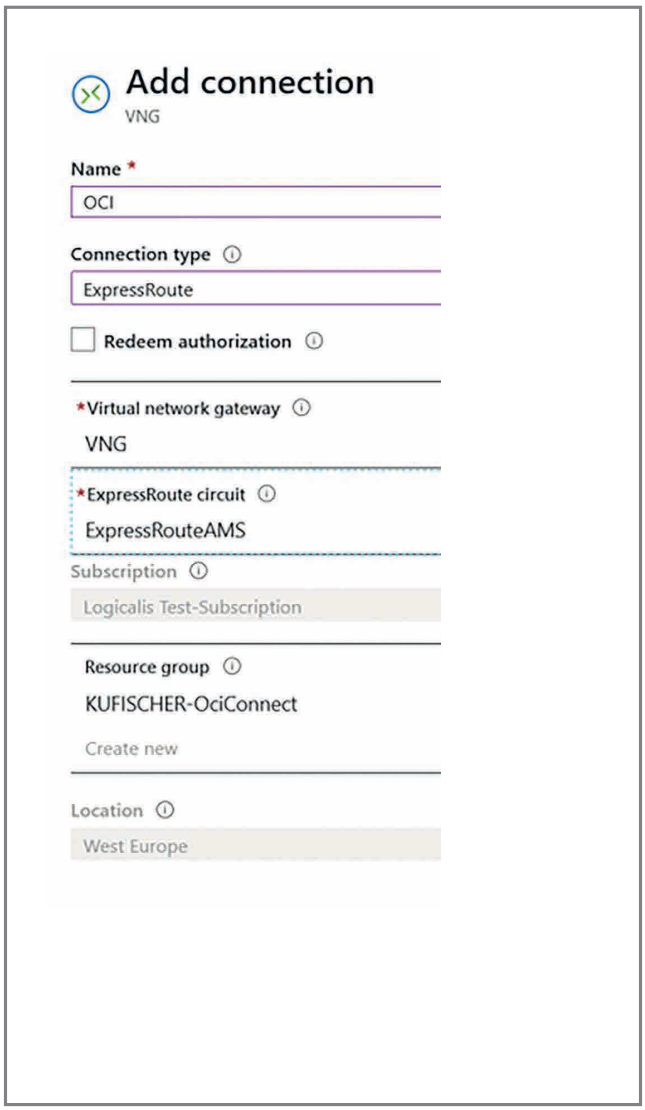


Abbildung 8: Link zwischen Azure VNET und ExpressRoute (© Kai-Uwe Fischer, Logicalis)

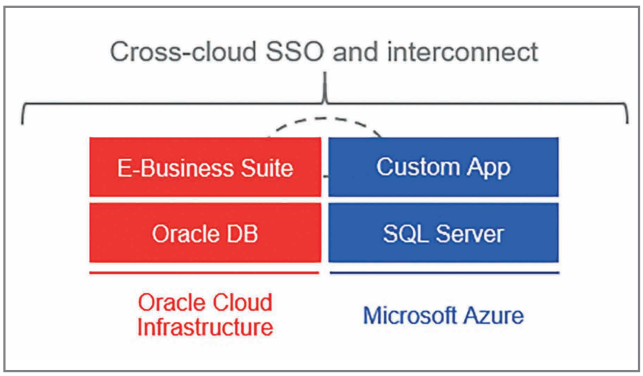


Abbildung 9: Full-Stack (© Kai-Uwe Fischer, Logicalis)

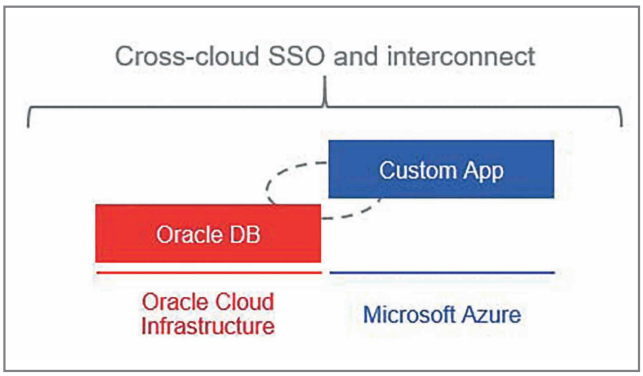


Abbildung 10: Split-Stack (© Kai-Uwe Fischer / Logicalis)

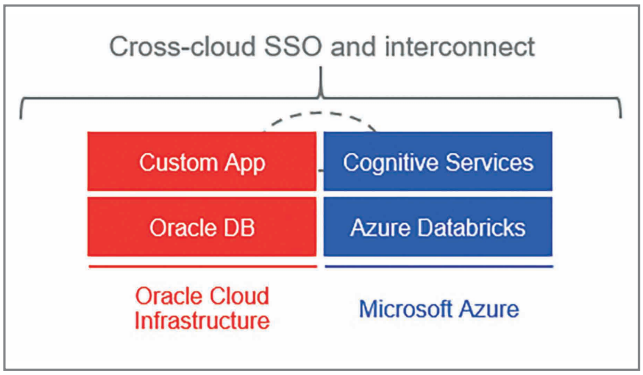
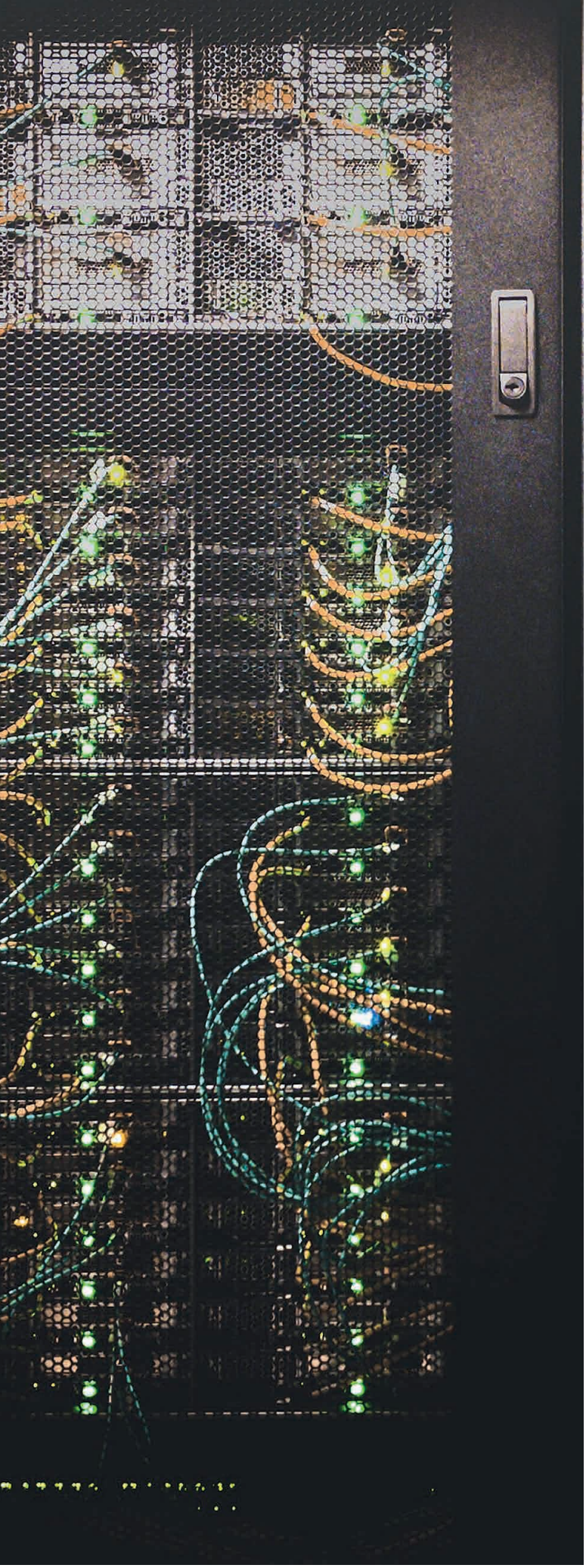


Abbildung 11: Native Cloud-Anwendung (© Kai-Uwe Fischer, Logicalis)



HOHE SICHERHEIT DURCH TRANSPARENT DATA ENCRYPTION (TDE)

Zudem erfüllt die OCI höchste Sicherheitsansprüche. Denn die Oracle-Datenbank – dies gilt sowohl für die Standard Edition als auch für die Enterprise Edition – wird „out of the box" per Transparent Data Encryption (TDE) verschlüsselt. Um TDE auch in Azure nutzen zu können, muss dort, anders als in der OCI, die Advanced-Security-Option für die Oracle-Datenbank erworben werden. Eine TDE- Verschlüsselung der Standard Edition ist hier nicht möglich.

ADMINISTRATIONSARBEITEN LEICHT GEMACHT

Einspielen von Patches und Upgrades, Erstellen von Hochverfügbarkeitslösungen und Backups: Über das bei PaaS nutzbare Cloud Tooling in der OCI sind solche Arbeiten einfach zu händeln. So werden zum Beispiel aktuelle Datenbank-Patches in der Konsole automatisch zur Verfügung gestellt und können bei Bedarf auf Knopfdruck eingespielt werden. Bei der Nutzung der Autonomous Datenbank erfolgt dies dagegen sogar vollautomatisiert. Fehler in der Administration werden so eliminiert und die Verfügbarkeit der Datenbanken erhöht. Die positive Folge: Der Datenbankadministrator ist von derlei Aufgaben befreit und hat mehr Raum für seine Kernaufgaben. Darüber hinaus lassen sich in der OCI – anders als in der Azure Cloud – auch hochverfügbare Oracle-Real-Application-Cluster-Umgebungen (Oracle RAC) problemlos erstellen und ohne Einschränkungen beim Support ausführen. In der Azure Cloud kann RAC zwar aktiviert werden, Oracle bietet dafür aber keinen Support an.

FAZIT UND EMPFEHLUNG

Um vor der endgültigen Entscheidung sicher zu sein, ist es für Unternehmen ratsam, die Multi-Cloud-Infrastruktur in einem Proof of Concept (POC) ausführlich zu testen. Werden die angestrebten Ziele erreicht, kann die POC-Umgebung in den Regelbetrieb überführt werden.

QUELLEN

1. Capgemini-Studie: *https://www.capgemini.com/de-de/news/it-trends-studie-2021-it-budgets-steigen-trotz-pandemie/*
2. DB Engines-Ranking: *https://db-engines.com/de/ranking*

Kai-Uwe Fischer ist seit mehr als 20 Jahren im Oracle-Datenbank-Umfeld als Consultant unterwegs. Seit sechs Jahren ist er als Senior PreSales Consultant bei Logicalis tätig. Neben der PreSales-Tätigkeit beschäftigt er sich intensiv mit dem Oracle-Cloud-Thema und ist verantwortlich für Kunden-Proof-of-Concepts sowie Migrationen in die Oracle Cloud Infrastructure.

Wir begrüßen unsere neuen Mitglieder

Korporative Mitgliedschaften:

- Miele & Cie. KG, Repräsentant: Heinrich Lotz
- menten GmbH, Repräsentant: Ralph Menten
- databee UG (haftungsbeschränkt), Repräsentant: Harald Sellmann

Natürliche Mitglieder:

- Nils Werkmeister
- Frank Prechtel
- Andreas Müller
- Marco Kurmann
- Xaver Schiener
- Steven Grzbielok
- Arvid Regenberg
- Patricia Hoikins
- Alexander Berkmann
- Tim Bell
- Christopher Berg
- Victoria Meier
- Ridvan Bulduk
- André Monson

Termine

Mai 05

17.05.2022
DOAG Dev Talk zum Thema: APEX Upgrade
Online

19.05.2022
IMC-WebSession: DB-Systems auf ODAs
Online

30.05.2022
DOAG 2022 Datenbank
Düsseldorf

Juni 06

01.06.2022
Regioday 2022 (KickOff der Regionalgruppen)
Zeitgleich an über 10. Standorten

02.06.2022
Migration von APEX Apps
Online

10.06.2022
Nach der Härtung - Database Auditing
Online

14.06.2022
Praxisworkshop Oracle Database Appliance
Online

21.06.2022
Oracle PL/SQL Performance Tuning
Berlin

28.06.2022
Integration mittels Boomi
Online

29.06.2022
CloudLand 2022
Brühl

Juli 07

08.07.2022
New Features in Multitenant mit 21c
Online

Impressum

Red Stack Magazin inkl. Business News wird gemeinsam herausgegeben von den Oracle-Anwendergruppen DOAG Deutsche ORACLE-Anwendergruppe e.V. (Deutschland, Tempelhofer Weg 64, 12347 Berlin, *www.doag.org*), AOUG Austrian Oracle User Group (Österreich, Lassallestraße 7a, 1020 Wien, *www.aoug.at*) und SOUG Swiss Oracle User Group (Schweiz, Dornacherstraße 192, 4053 Basel, *www.soug.ch*).

Red Stack Magazin inkl. Business News ist das User-Magazin rund um die 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.

Red Stack Magazin inkl. 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: Björn Bröhl. 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: Fried Saacke
Redaktionsleitung Red Stack Magazin:
Martin Meyer
Redaktionsleitung Business News:
Marcos López
Kontakt: redaktion@doag.org
Weitere Redakteure (in alphabetischer Reihenfolge): Dirk Andres, Andreas Buckenhofer, Kai-Uwe Fischer, Lothar Flatz, Oliver Fuhrmann, Elina Hattendorf, Germans Hirsch, Lorena Hoffmann, Felix Huchzermeyer, Rainier Kaczmarczyk, Dirk Krautschick, Dierk Lenz, Marcos López, Martin Meyer, Jan Ott, Michael Pergande, Alfred Schlaucher, Detlef E. Schröder, Jürgen Sieben, Wolfgang Taschner, Nadine Wagner

Titel, Gestaltung und Satz:
Diana Tkach
DOAG Dienstleistungen GmbH
(Anschrift s.o.)

Fotonachweis:
Titel: © Storyset | *www.freepik.com*
S. 12: © AaronJOlson | *www.pixabay.com*
S. 16: © Ian Arlett | *www.flickr.com*
S. 22: © Bessi | *www.pixabay.com*
S. 28: © Xresch | *www.pixabay.com*
S. 36: © Geralt | *www.freepik.com*
S. 42: © Daniele Franch | *https://unsplash.com*
S. 48: © Umihir | *www.pixabay.com*
S. 50: © PublicDomainPictures | *www.pixabay.com*
S. 60: © John Loannidis | *www.pixabay.com*
S. 62: © Geralt | *www.pixabay.com*
Titel S. 66: © Umberto | *www.unsplash .com*
S. 69: © Arne Hattendorf
S. 72: © Arne Hattendorf
S. 73: © Arne Hattendorf
S. 75 © Arne Hattendorf
S. 77: © Honey-Yanibel-Minaya-Cruz | *www.unsplash .com*
S. 84: © Freepik | *www.freepik.com*
S. 92: © Taylor Vick | *www.unsplash .com*
S. 93: © Pch Vector | *www.freepik.com*

Anzeigen:
sponsoring@doag.org

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

Druck:
WIRmachenDRUCK GmbH,
www.wir-machen-druck.de

Inserentenverzeichnis