Hadoop-Einstiegshürden meistern – mit Business Case, Skill-Aufbau und der richtigen Technologie zu Big Data

Oliver Herzberg und Slavomir Nagy, metafinanz Informationssysteme GmbH

Für einen kosteneffizienten Einstieg in Big Data verspricht Hadoop mehr Analyse-Potenzial als klassisches DWH. Doch die florierende Open-Source-Technik erfordert ein strategisches und konzeptionelles Umdenken. Der Artikel zeigt, worauf es beim Finden der Business Cases, dem Kompetenz-Aufbau und den Test-Umgebungen in der Cloud ankommt.

Auch wenn Big Data lange als ein etwas nebulöser Begriff galt, so sind zumindest die Herausforderungen für die Geschäftswelt offensichtlich. Unternehmen erhalten immer mehr Daten von Menschen und in letzter Zeit auch vermehrt von Maschinen; für die Analyse dieser Daten sind neue, mächtigere Technologien erforderlich. In diesem neu zu erschließenden Gebiet wurden inzwischen deutliche Schneisen geschlagen. Als wichtigste technische Plattform zur Bewältigung einer zukunftsorientierten Datenbewirtschaftung gilt in IT-Fachkreisen das Framework Hadoop.

Doch dieser Einigkeit zum Trotz wirft die Technologie nach wie vor Fragezeichen auf. Zum einen sind es die ständig neu entstehenden technischen Standards (siehe Kasten „Was ist eigentlich Hadoop?"), die einen klaren Überblick erschweren, zum anderen müssen Einsatzszenarien wohlüberlegt sein, um überzeugende geschäftliche Mehrwerte in BI-gesättigten Umgebungen zu entwickeln. Immerhin kann inzwischen schon eine Reihe von Unternehmen aus unterschiedlichen Branchen mit überzeugenden Anwendungsbeispielen aufwarten, die sich auf Gebieten wie der Betrugserkennung, der Kundenabwanderungs-Analyse oder der Fahrzeugdaten-Auswertung abspielen.

BARC-Studie beleuchtet Status quo

Die Aussicht auf völlig neue Möglichkeiten der Datenauswertung reicht allein nicht aus, um in Unternehmen den Einsatz einer neuen Technologie zu rechtfertigen. Das belegt auch die aktuelle Studie des Marktforschungs- und Beratungshauses BARC mit dem Titel „Hadoop als Wegbereiter für Analytics". Bei vier Prozent der 261 Teilnehmer aus der DACH-Region ist Hadoop bereits Bestandteil der Unternehmensprozesse, weitere 14 Prozent betreiben ein Pilotprojekt und elf Prozent hegen Pläne dafür. 49 Prozent können sich ein Engagement immerhin vorstellen. Der Rest, mit 22 Prozent rund ein Fünftel der Befragten, sieht noch überhaupt keinen Handlungsbedarf.

Bemerkenswert sind auch die Erkenntnisse darüber, wer in den Unternehmen die Treiber von Hadoop sind. Mit einem Anteil von 54 Prozent sticht hier die IT-Abteilung hervor, BI-Verantwortliche wurden bei 42 Prozent der Befragten als Motivator genannt, aus den Fachbereichen kommt in 26 Prozent der Fälle die Initiative. Insgesamt müsse nach den Erkenntnissen der Würzburger Analysten noch viel Überzeugungsarbeit geleistet werden, was den

metafinanz Informationssysteme: Navigator für die Big-Data-Welt

Wie können Unternehmen ihre Daten gewinnbringend nutzen? Welche Technologien eignen sich, welche Kompetenzen sind vonnöten, wo ist der Start-, wo der Endpunkt? Bei all dem Hype, der sich im Moment um das Thema aufgebaut hat, ist es schwierig, einen Überblick zu bekommen. Die Strategie und die Technologie gibt es nicht aus einem Guss, sondern beides muss individuell aus verschiedenen Teilen zusammengesetzt werden. Big Data fordert das Business und die IT gleichermaßen heraus. Beratungsunternehmen wie die metafinanz bringen die Kunden gerade bei kritischen Themen wie Skill-Entwicklung und effizienzorientierten Einstiegsszenarien sicher ans Ziel. Wir greifen dabei auf einen umfangreichen Erfahrungsschatz zurück, denn wir halten uns kontinuierlich auf dem Laufenden, was die neuen Technologien können, was sie bringen und wo sie sich einsetzen lassen. Gleichzeitig beraten wir unsere Kunden auch im Hinblick auf ihre Geschäftsprozesse, denn diese sind unzertrennlich mit der technologischen Entscheidung verknüpft. Wir unterstützen unsere Kunden dabei, lohnende Analyseprojekte zu ermitteln, Big-Data-Szenarien anhand von Prototypen zu testen und das Alignment zwischen Business und IT zu organisieren.

Mehrwert von Big Data und Analytics betreffe. Denn drei Viertel der befragten Unternehmen stuften Hadoop noch als weniger oder gar nicht wichtiges Thema ein.

Hadoop adressiert Versäumnisse beim Daten-Management

Womit ein zentraler Punkt genannt wäre: Was sind die überzeugenden Gründe für den Einsatz von Hadoop? Entscheidend dürfte sein, dass die traditionellen Technologien für Daten-Management in den vergangenen drei Jahrzehnten nicht die Erwartungen der Anwender erfüllt haben. Kernsegmente wie BI und Data Warehouse kratzten nur an der Oberfläche und förderten aus den Daten nicht im erwarteten Maße nützliche Informationen und handlungsmotivierende Erkenntnisse zutage. Der Einsatz von Hadoop kann ein Ansatz sein, diese Herausforderungen im klassischen BI-Umfeld zu meistern.

Um es klarzustellen: Es ist nicht gemeint, nun alle historisch gewachsenen Daten-Management-Strukturen über Bord zu werfen. Der parallele Betrieb einer schnellen und einer langsamen Infrastruktur ist eine sinnvolle Vorgehensweise für den Einstieg in Big Data. Alle strukturierten, unstrukturierten und teilstrukturierten Daten im Unternehmen, die beispielsweise nicht für das Enterprise Data Warehouse geeignet sind, lassen sich komplett in einen Hadoop-basierten Daten-Hub übernehmen. Auf diese Weise werden die Daten für Berichtswesen und Analysen verfügbar – und zwar wahlweise mit bestehenden BI- oder mit neuen Hadoop-Tools.

Nicht übersehen sollten Unternehmen dabei allerdings den Umstand, dass Hadoop völlig andere Techniken erfordert als klassische Business Intelligence, die auf strukturierten, integrierten und bereinigten relationalen Datenbank-Management-Systemen basiert. Vor allem aber braucht der Einstieg in die neue Welt auch ein Business-Szenario.

Die Technik muss das Business unterstützen

Vor der Einführung neuer Technologien steht der Nachweis der Rentabilität innerhalb eines bestimmten Zeitraums oder ein darauf aufbauendes Geschäftsmodell. Unternehmen steigen ein, indem sie lohnende Analyse-Projekte ermitteln und Big-Data-Szenarien anhand von Prototypen testen. Und, indem sie einen Transfer zwischen Business und IT organisieren. Denn es geht nicht um die Technik, sondern um die Frage, wie die Technik das Business unterstützen kann.

Zur Ermittlung des Business Case bieten sich diverse Grundüberlegungen zu dem gewünschten Ziel an – selbstverständlich abhängig vom jeweiligen Geschäft; dennoch gibt es einige allgemeine Fragen, die sich jedes Unternehmen stellen sollte: Sollen Kosten gesenkt und die Profitabilität des Geschäfts gesteigert werden? Etwa durch die Beschleunigung bestimmter Datenauswertungen. Sollen neue Formen der Kundenkommunikation gefunden werden? Wie durch die Auswertung von Kundenfeedbacks. Oder sollen Produkte um Services ergänzt werden? Durch die Nutzung von Vorhersagen beispielsweise lassen sich Wartungen proaktiv gestalten

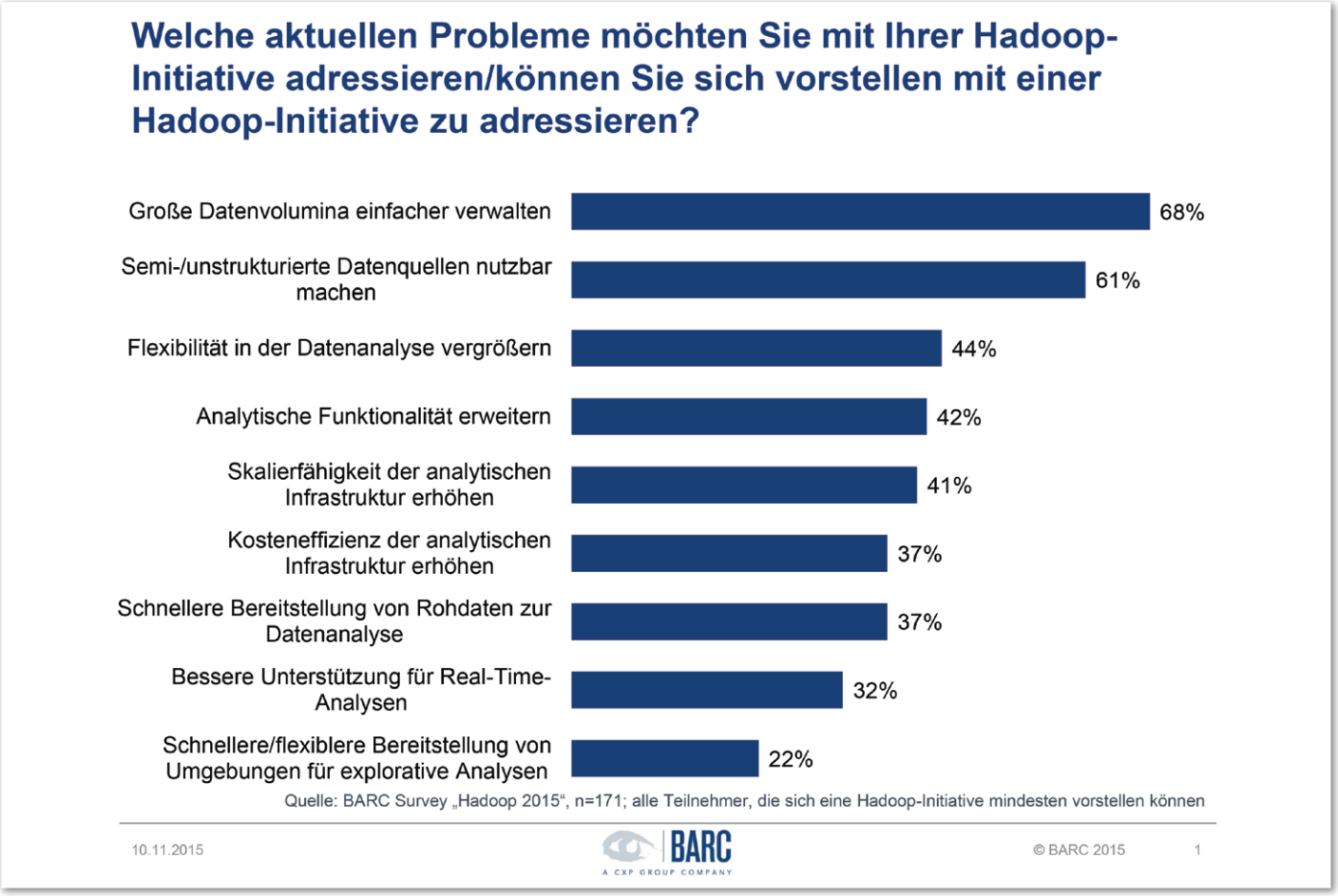

Abbildung 1: Nehmen Sie sich die Zeit, die Leute und die Sorgfalt, um zu Beginn die richtigen Fragen zu stellen.

– das Ersatzteil ist dann schon beim Kunden, bevor die Maschine den Geist aufgibt. Oder möchte das Unternehmen vielleicht ein völlig neues Geschäft starten? Indem es beispielsweise seine Produkte oder Services stärker individualisiert.

Auch innerhalb der IT-Organisation lassen sich solche Gedanken durchspielen. Wenn die Datenmanagement-Infrastruktur historisch bedingt zu aufwändig in der Pflege und Weiterentwicklung geworden ist, hilft eine agile Technologie sicherlich, das Dilemma zwischen Kostendisziplin und dem Wunsch nach Innovationen aufzulösen. Denn im Gegensatz zu den hohen Aufwänden und langen Zeiträumen, die im Bereich Datenmanagement selbst bei Einstiegsszenarien benötigt werden, lassen sich mit Technologien wie Hadoop in der Regel binnen weniger Wochen oder höchstens Monaten Testszenarien durchspielen.

Traditionelles BI stößt an seine Grenzen

Der Forrester-Analyst Boris Evelson stellt in seinem Blog fest, dass die traditionellen Technologien für Datenmanagement in den letzten drei Jahrzehnten nicht die Erwartungen der Anwender erfüllt haben. Kernsegmente wie BI und Data Warehouse kratzen nur an der Oberfläche und fördern nicht im erwarteten Maße nützliche Informationen und handlungsmotivierende Erkenntnisse aus den Daten zutage. Er nennt dafür folgende Gründe:

- *Unternehmen werten zu wenige Daten aus*
  Unternehmen ziehen nur 40 Prozent ihrer strukturierten Daten für strategische Entscheidungen heran.
- *Unstrukturierte Daten werden lediglich zu einem Drittel in Analysen berücksichtigt*
- *BI ist zu komplex*
  Frühere sowie einige aktuelle Plattformen für Business Intelligence sind zu kompliziert. Sie erfordern die Integration von Dutzenden Komponenten unterschiedlicher Hersteller. Deshalb dauert es so lange und ist so kostspielig, aus der Gesamtheit der Daten eine „single version of truth" herauszudestillieren.
- *BI-Architekturen sind unflexibel*
  Viele Unternehmen brauchen zu lange, um das Idealziel einer zentralisierten BI-Umgebung zu erreichen. Wenn sie glauben, angekommen zu sein, liegen schon wieder neue Datenquellen vor, neue Regulatoriken oder neue Kundenwünsche.

Größte Herausforderung ist der Know-how-Aufbau

Ist die Entscheidung für den Schritt in die neue Welt gefallen, wartet schon die nächste Herausforderung: der Aufbau der richtigen Mannschaft. Das richtige Know-how zu entwickeln, ist derzeit eine der größten Hürden, wie auch aus der BARC-Studie hervorgeht. Gefragt nach den Haupthindernissen bei der Einführung von Big Data nannten 68 Prozent der Befragten das fehlende technische und 65 Prozent das fehlende fachliche Know-how.

Derartige Schwierigkeiten sind auch uns aus eigener Erfahrung bekannt. Die Autoren befassen sich schon seit Längerem mit Hadoop und mussten einiges lernen. Nachdem am Markt kaum Experten zu finden waren, stellten sie sich die Frage, welche IT-Expertengruppe die besten Voraussetzungen für die Skill-Entwicklung zum Hadoop-Spezialisten mitbringt. Zur Wahl standen Java- oder Datenbank-Entwickler sowie Linux-Administratoren. Idealerweise sollten die Kandidaten alle drei Bereiche abdecken, aber diese findet man nur selten und wenn, dann sind sie teuer. Die Entscheidung fiel damals auf Datenbank-Entwickler, schließlich brachten sie bereits Wissen im Umgang mit großen Datenmengen beziehungsweise mit dem Management von Daten mit.

Doch schnell machten sich kulturelle Probleme bemerkbar. Richtig ist, dass es um die Verarbeitung von Datenmassen geht. Das kennen und können Datenbank-Entwickler. Dennoch, neu ist für sie die bisher ungewohnte Art zu arbeiten. Hadoop basiert auf einer in Java entwickelten Umgebung. Das Arbeiten ist schneller und nicht so klar abgegrenzt wie mit klassischen DWH- oder BI-Systemen.

Auch ist die Technologie jünger und der Markt sehr dynamisch. Die Open-Source-Community entwickelt sie stetig weiter –

Fünf Empfehlungen für den risikofreien, schnellen und günstigen Einstieg in Hadoop

1. Gewinnbringendes Szenario auswählen
2. Die richtigen Menschen finden und ausbilden
3. Großes Augenmerk auf Security legen
4. Die Cloud als Chance für schnelle Resultate nutzen
5. Datenqualität im Auge behalten

und das mit einer hohen Geschwindigkeit. Lehrgänge oder aktuelle Literatur gibt es kaum, die Experten diskutieren ihre Lösungsszenarien vor allem auf Veranstaltungen und in Internetforen.

Hier die richtigen Leute zu finden und Skills aufzubauen, ist daher weniger eine Frage des vorhandenen technischen Knowhows als einer gewissen Mentalität. Ohne Begeisterung geht es nicht. Ohne Beharrlichkeit auch nicht. Denn während klassische Datenbank-Experten gewohnt sind, mit kurzen, klar identifizierbaren Fehlercodes zu arbeiten, müssen Hadoop-Fachleute mit Java-Fehlermeldungen klarkommen, die schon einmal mehrere Bildschirme füllen. Es sind Menschen gefragt, die sich aus eigener Kraft in das Thema einarbeiten, durch Probleme beißen und auch einmal Rückschläge hinnehmen können. Unsere Erfahrung hat außerdem gezeigt, dass die Kollegen ständig am Ball bleiben müssen. In der Welt von Big Data und Hadoop kann der Anschluss schon nach wenigen Monaten verloren gehen.

Auch im Management sind neue Skills gefragt. Der Einstieg in Big Data ist mit Risiken behaftet, die in Kauf genommen werden müssen. Und es gibt keine Standards, auf die man sich verlassen kann. Auch aus Managementsicht ist also Ausdauer gefragt, ebenso wie die Bereitschaft und Fähigkeit, mit Risiken und Unsicherheit umzugehen.

Für Projektmitarbeiter wie für Manager kommt es also weniger auf eine bestimmte fachliche Expertise an, sondern auf die Bereitschaft, ausgetretene Pfade zu verlassen, sich mit Neuem auseinanderzusetzen. Da sich Business und IT in Einklang setzen müssen, empfehlen wir nicht zuletzt außerdem den Aufbau einer interdisziplinären Mannschaft. Optimal sind Teams mit Menschen, die kommunizieren können, sich mit Datenauswertungen auskennen, das Geschäft verstehen und selbstverständlich auch technisches Verständnis besitzen *(siehe Abbildung 1)*.

Cloud-Umgebungen senken die Einstiegshürden

Treiber der Technologie und damit letztlich des Marktes ist die weltweit verteilte Community. Sie ist es, die diverse Verbesserung und Neuerungen entwickelt. Prominente Unterstützung gibt es unter anderem von Playern wie Yahoo, Facebook oder Google. Es kann jedoch weder von einem Standard die Rede sein, noch ist erkennbar, welche Entwicklungen von Dauer sein werden. Das bedeutet, dass manche der neuen Ansätze weiterverfolgt werden wie zum Beispiel „Spark“, andere hingegen wieder verschwinden.

Ein schneller Ein- und ein schneller Ausstieg müssen also möglich sein. Daher bieten sich Cloud-Umgebungen an, um beweglich zu bleiben, kostengünstig in Hadoop einzusteigen und mögliche Szenarien auszuprobieren. Amazon etwa stellt hier mit AWS eine inzwischen sehr ausgereifte Virtualisierungsumgebung bereit, die sich für das Erstellen von Machbarkeitsstudien (Proof of Concept) nutzen lässt. Als weiterer Massenanbieter mischt inzwischen auch Microsoft mit Azure HDInsight in der Big-Data-Szene mit. Ein weiterer Player auf dem globalen Markt ist Google. Die Company stellt die für eigene Services genutzten Infrastrukturen wie YouTube Shared auch an-

Was genau ist Hadoop?

Hadoop gilt derzeit als eine der Kerntechnologien im Big-Data-Umfeld. Doch anders als der Begriff suggeriert, steckt dahinter eine schnell wachsende Vielfalt von Technologien, Projekten und Anbietern, die eine klare Definition nahezu unmöglich macht. Einige Experten wie der Gartner-Analyst Merv Adrian und der Cask-Manager Andreas Neumann haben sich in jüngster Zeit damit befasst, eine grobe Übersicht zu vermitteln.

Als erster Ausgangspunkt dient in solchen Fragen Wikipedia, aber der dortige Eintrag erweist sich als sehr vage: „Apache Hadoop ist ein freies, in Java geschriebenes Framework für skalierbare, verteilt arbeitende Software. Es basiert auf dem MapReduce-Algorithmus von Google sowie auf Vorschlägen des Google-Dateisystems und ermöglicht es, intensive Rechenprozesse mit großen Datenmengen auf Computerclustern durchzuführen“. Laut Neumanns Recherche eigenen sich jedoch weder „Open Source“ noch „Java“ zu einer eindeutigen Charakterisierung, weil es hiervon einige Abweichler im Markt gibt. Auch ein Muster an Standardkomponenten für eine Hadoop-Distribution ist nicht zu erkennen. Am Ende vergleicht er Hadoop mit einem Puzzle aus vielen Technologien: Um es zu lösen, fehlen immer wieder einzelne Teile, andere Teile scheinen zu einem anderen Puzzle zu gehören und die Anbieter im Markt spielen die Rolle, das Puzzle für die Kunden zu lösen.

Adrian verfolgt einen Definitionsansatz, bei dem er untersucht hat, welcher Hadoop-Anbieter welches der vielen Hadoop-bezogenen Projekte unterstützt. Ausgangspunkt sind die drei auf der Apache-Hadoop-Hauptsite genannten Projekte HDFS, YARN und MapReduce. In seiner Matrix mit den relevanten Herstellern ergänzt er die obigen drei mit HBase, Hive, Pig, Spark und Zookeeper, die von allen gelisteten Herstellern unterstützt werden. Es folgen diverse Projekte, die jeweils für Teile der Anbieter relevant sind, und so summiert sich am Ende die Zahl der Hadoop-bezogenen Projekte auf 48. Ein Ende der Fahnenstange ist damit allerdings keinesfalls erreicht.

Von einer eindeutigen Definition ist man mit dieser Erkenntnis weiter weg denn je. Aber gleichzeitig zeigt sie: Kaum ein IT-Segment scheint aktuell dynamischer und innovativer als der Hadoop-Markt zu sein.

deren Unternehmen zur Verfügung. Wie bei allen Cloud-Szenarien kommt aber dem Thema „Datensicherheit und Datenschutz“ eine geschäftskritische Bedeutung zu. Deswegen wird der Einsatz gerade in Deutschland kritisch und zum Teil kontrovers diskutiert

Die Plattformen stellen Infrastruktur as a Service sowie das Hadoop-Framework zur Verfügung, um Big-Data-Anwendungsfälle durchzuspielen. Der Vorteil liegt klar darin, dass sich die Nutzer voll auf den fachlichen Use Case konzentrieren können, statt sich um den Aufbau und den Betrieb der Infrastruktur zu kümmern. „Pay per use“, also nur für die wirkliche Zeit der Nutzung zu bezahlen, ist ein weiterer attraktiver Aspekt, der für den Einsatz spricht.

Da die Datenverarbeitung in der Cloud für viele Unternehmen aus datenschutzrechtlichen Gründen ein kritisches Unterfangen ist, müssen bei dieser Lösung natürlich besondere Vorsichtsmaßnahmen getroffen werden. Um hier jegliche Risiken zu vermeiden, gilt es im ersten Schritt, frühzeitig die genutzten Daten zu klassifizieren und individuelle Schutzmechanismen (organisatorisch und technisch) aufzusetzen. Eine frühzeitige Einbindung der Sicherheits- und Datenschutzbeauftragten im Unternehmen stellt sicher, dass das Vorhaben nutzenstiftend eingesetzt werden kann. Denn nicht alle Daten sind personenbezogen oder besonders schutzbedürftig.

Für den Umgang mit sensiblen Daten empfiehlt sich, sie in einer dedizierten Umgebung im Unternehmen zu halten oder mit mittlerweile am Markt verfügbaren Verschlüsselungslösungen zu schützen – und die Cloud für die Verarbeitung nicht sensibler Daten zu nutzen, zumindest für den Einstieg.

Oliver Herzberg
oliver.herzberg@metafinanz.de

Slavomir Nagy
slavomir.nagy@metafinanz.de

Aufbau eines Semantic-Layers zwischen Datenbank und Hadoop

Matthias Fuchs, capgemini Nürnberg

Kein Business-Intelligence-Hersteller kommt momentan am Thema „Hadoop“ vorbei. Eine Integration ist Pflicht. Wie können Daten aus Hadoop eingebunden werden? Wie wirkt sich die Einbindung der Daten auf die Abfragegeschwindigkeit aus? Welche Art von Daten können aus BI-Sicht verwendet werden? Welche Möglichkeiten ergeben sich im Oracle-Umfeld?

Hadoop ist ein Framework, das es ermöglicht, verteilt arbeitende Rechenprozesse mit großen Datenmengen auf einem Cluster aus mehreren Servern durchzuführen. Zu den Basis-Komponenten gehören ein Dateisystem HDFS („Hadoop Distributed File System“) und die Möglichkeit, nach dem MapReduce-Algorithmus Berechnungen über mehrere Knoten durchzuführen. Erst durch die Entstehung von meist Open-Source-Applikationen auf Basis von Hadoop wie Hive oder Spark wurde Hadoop zu einer Plattform, um vielfältige Daten zu speichern, zu analysieren und zu verarbeiten.

Der Semantic-Layer

Im Rahmen von Business Intelligence versteht man unter einem Semantic-Layer eine Repräsentationsebene von Daten, die es Endbenutzern ermöglicht, ohne technisches Wissen Daten einfach abzufragen. Der Begriff geht auf Business Objects zurück. Der Layer soll es ermöglichen, komplizierte Bezeichnungen beziehungsweise Beziehungen auf Objekte wie Produkte, Kunden oder Umsatz zurückzuführen. Der Semantic-Layer ist zentraler Bestandteil in den meisten BI-Tools von SAP, IBM, MicroStrategy oder Oracle.