Agiles Datenmanagement in Big-Data-Umgebungen

Pi-Architektur

Unternehmen sehen sich heute mit enormen Volumina an potenziell relevanten Daten konfrontiert, die aus einer laufend steigenden Zahl interner und externer Quellen geliefert werden. Für viele Anwendungsdomänen ist das Sammeln und Analysieren ausschließlich hochstrukturierter Daten aus klassischen relational orientierten Transaktionssystemen nicht mehr zufriedenstellend. Ein Großteil der neuen Daten etwa aus sozialen Medien oder dem Internet der Dinge liegt in einer komplexen semi- oder unstrukturierten Form vor. Es stellt sich die Frage, wie diese neuen „polystrukturierten" Datenquellen bei Bedarf agil in die dispositiven Systeme zur Entscheidungsunterstützung (Business Intelligence, BI) eingebunden werden können, ohne übergreifende Anforderungen an Konsistenz und Integrität zu verletzen.

Vor allem auf der Seite der Fachbereiche besteht ein Bedarf nach Konzepten, die es den Anwendern ermöglichen, eigenständig und weitgehend unabhängig von der IT-Abteilung auf für sie wichtige Informationen zuzugreifen und daraus selbstständig Berichte sowie Analysen zu erstellen (Self-Service-BI). Eine agile Self-Service-BI, die neuartige Datenquellen berücksichtigt, verspricht innovative Analyseszenarien und eine verbesserte Informationsbedarfsdeckung. Um diese Anforderungen umzusetzen, skizziert der Beitrag einen Ansatz, der Ideen aus dem Data-Vault-Ansatz aufgreift, aber strukturierend-integrative Teile des Datenmodells separat von den polystrukturierten Analyseinhalten ablegt und so die Stärken klassisch relationaler Strukturen mit denen von Big-Data-Ansätzen kombiniert. Die entsprechende „Pi-Architektur" – so genannt nach der Form, die in der Architektur sichtbar wird, in Anlehnung an die Bezeichnungen Lambda- und Kappa-Architektur – erlaubt eine benutzer- bzw. fachbereichsgetriebene Ad-hoc-Einbindung polystrukturierter Datenquellen ohne Gefährdung der Kern-BI-Inhalte. Verbunden mit einer Governance-Struktur, wird so der Spagat zwischen Agilität und Effizienz angegangen.

Ausgangssituation

Mit dem Fortschreiten der weltweiten Globalisierung und der steigenden Bedeutung von Themen wie Internet der Dinge oder Industrie 4.0 sehen sich Unternehmen in den unterschiedlichsten Wirtschaftszweigen mit einer Verschärfung der Marktbedingungen konfrontiert. Mit diesen Entwicklungen gehen ein sich verändernder Wettbewerb, neue Märkte und Konkurrenten sowie ein wachsender Time-to-Market-Druck einher. Um die eigene Wettbewerbsfähigkeit zu erhalten und schnell auf Kundenbedürfnisse und die sich ändernden strategischen sowie betrieblichen Rahmenbedingungen reagieren zu können, wird eine kontinuierliche Anpassung der Geschäftsprozesse benötigt. Hierbei fällt häufig der Begriff „Agilität", der die Fähigkeit eines Unternehmens beschreibt, in einem kontinuierlich dynamischen, globalen Marktumfeld profitabel zu operieren. Ziel ist die schnelle und proaktive Anpassung von Unternehmenselementen an unerwartete und unvorhergesehene Ereignisse.

Eine Möglichkeit, die Agilität innerhalb des Unternehmens zu erhöhen, ist der verstärkte Einsatz innovativer Informationstechnologien innerhalb der einzelnen Geschäftsprozesse. Speziell zur Verbesserung der Informationsbedarfsdeckung in den Fachbereichen bieten sich Systeme zur Entscheidungsunterstützung an. Sie sind idealerweise Teil eines unternehmensweiten Ansatzes zur Informationsversorgung, der die Erkennung relevanter Geschäftsvorfälle unterstützt und über das Konzept der Business Intelligence (BI) realisiert wird. BI wird dabei verstanden als integrierter, unternehmensspezifischer, IT-basierter Gesamtansatz zur betrieblichen Entscheidungsunterstützung, der den Benutzern bei der Synthese von Informationen aus einer Vielzahl an Daten mit Hilfe von Analysen, Reporting und Abfrage-Tools assistiert.

Die rasante Dynamik treibt dabei auch den Bedarf der BI selbst (Agile Business Intelligence), was eine agile Einbindung neuer Datenquellen, deren Integration in übergreifende Datenarchitekturen, die Erstellung von Datentransformationsstrecken sowie die Generierung und Verteilung von Analysen und Reports beinhaltet. Treiber sind vor allem neue Anforderungen der Benutzer, wechselnde technologische Potenziale, wechselnde Prozessrahmenbedingungen sowie unerwartete Ereignisse in operativen Prozessen.

Eine agile BI kann unter anderem durch agile BI-Architekturen, BI-Organisationsstrukturen und BI-Prozesse gefördert werden. Während klassische Core-DWHs, die ohne Beteiligung der Fachbereiche verwaltet werden und ihren Fokus auf ein Standardberichtswesen legen, üblicherweise nur ein geringes Maß an Agilität bieten, gehen Ansätze, die etwa Fachbereichs-Sandboxes mit aus der agilen Softwareentwicklung stammenden Vorgehensmodellen kombinieren, bereits deutlich weiter [ZBK12].

Ein potenzieller Baustein für eine agile BI ist daneben insbesondere auch eine Self Service Business Intelligence (SSBI), die es den Anwendern aus den Fachbereichen erlaubt, eigenständig und weitgehend unabhängig von der IT-Abteilung auf relevante Informationen zuzugreifen und diese aufzubereiten. SSBI erlaubt so ad hoc fachspezifische Analysen aus einer Vielzahl unterschiedlicher Daten ohne die Einbindung einer zentralen IT- oder BI-Einheit. Gerade Power-User können mit SSBI ihre Aufgaben einfacher und schneller umsetzen [AlS16]. Als Folge verschieben sich immer mehr Datenaufbereitungsprozesse aus der IT in den Fachbereich. Idealerweise ist neben der Analyse auch die

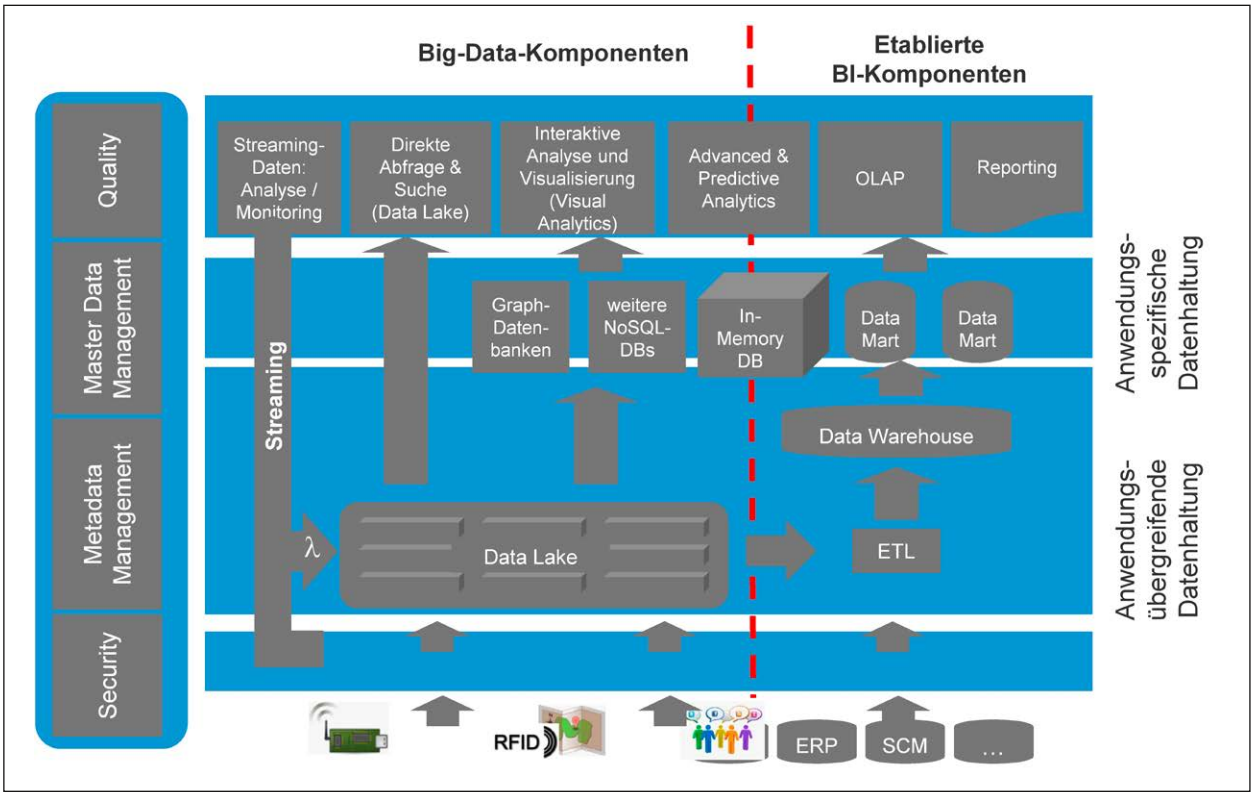

Abb. 1: Erweiterter BI-Architekturrahmen [BaK15]

Datenaufbereitung ein innerhalb des Fachbereichs aufgehängter Prozessschritt (Self Service Data Preparation).

Eine besondere Herausforderung besteht darin, eine solche Agilität auch angesichts der neuen Datenquellen mit ihren teils erheblichen Datenvolumen und einer rapiden Datenanfallsfrequenz umzusetzen. Für viele Anwendungsdomänen ist das Sammeln und Analysieren von reinen strukturierten und transaktionsorientierten Daten nicht mehr zufriedenstellend, da ein Großteil gerade der neu generierten Daten komplex, semi- oder unstrukturiert ist. Als übergreifende Bezeichnung für die entstehende Vielfalt an Datenquellen wird der Begriff „polystrukturiert“ genutzt, während Speicherung und Verarbeitung großvolumiger, polystrukturierter Datenbestände in Literatur und Praxis unter den Begriffen „Big Data“ und „Big Data Analytics“ subsumiert werden.

Ein offenes Thema besteht hierbei insbesondere in der Frage nach Lösungskonzepten für die agile Einbindung polystrukturierter Big-Data-Quellen in eine agile BI. Typischerweise streben BI- und IT-Abteilungen dazu, ihre Daten für die Entscheidungsunterstützung in einem zentralen Repository zu konsolidieren (Data Warehouse, DWH). Die daraus resultierenden Architekturen sind in ihrer Umsetzung aber häufig sehr starr und Reaktionszeiten auf neue oder geänderte Anforderungen können Wochen und Monate in Anspruch nehmen. Gleichzeitig sind die verwendeten relationalen Datenmodelle gängiger DWH-Lösungen nur bedingt für eine direkte Integration semi- und unstrukturierter Daten ausgelegt. In aller Regel werden diese stattdessen in sogenannten Content und Document Management Systems vorgehalten und sind normalerweise nur unzureichend mit Systemen zur Aufbereitung und Analyse strukturierter Daten aus dem Data Warehouse verzahnt.

Der hier vorgestellte Ansatz zielt darauf ab, Big-Data-Komponenten in eine agile Business Intelligence einzubetten und hierfür Self-Service-Konzepte zu unterstützen. Der Mehrwehrt einer solchen Kombination ergibt sich aus der Ermöglichung neuer Analyseszenarien und der verbesserten Informationsbedarfsdeckung durch BI-Systeme in den Fachbereichen. Beispielsweise wird der Einfluss sozialer Medien auf das Kaufverhalten von Kunden erst aus der Kombination von Verkaufstransaktionsdaten mit Daten und Einträgen aus sozialen Netzwerken ersichtlich.

Integration von Big-Data-Komponenten in die BI

Polystrukturierte Datentypen aus unterschiedlichen Quellen im Unternehmen werden zunehmend unter Verzicht auf aufwendige Datenbereinigungs- und harmonisierungsprozesse in ihrer Rohform gespeichert und unmittelbar für Analysen verfügbar gemacht. So entsteht eine neue Art von Datenhaltungskomponente abseits der klassischen DWH-Komponenten als eigenständiger „Big Data Store“ (vgl. Abbildung 1). Im hier vorgestellten Ansatz wird ein solcher mit bestehenden DHWs verlinkt, sodass Nutzer übergreifende Analysen über die Daten beider Systeme hinweg durchführen können.

Für bestimmte Fragen kann eine direkte, grobe Exploration des Datenbestands im Big Data Store erfolgen, zum Beispiel zeitkritische Abfragen oder ein Abruf bestimmter Einzelinformationen. Big Data Stores sind kein Ersatz für ein DWH, da bislang maßgebliche Funktionen für die Integration, die Historienbildung, die Administration und die Qualitätssicherung der Daten fehlen. Die Kombination aus unstrukturierten Inhalten aus einem Big Data Store mit hochstrukturierten

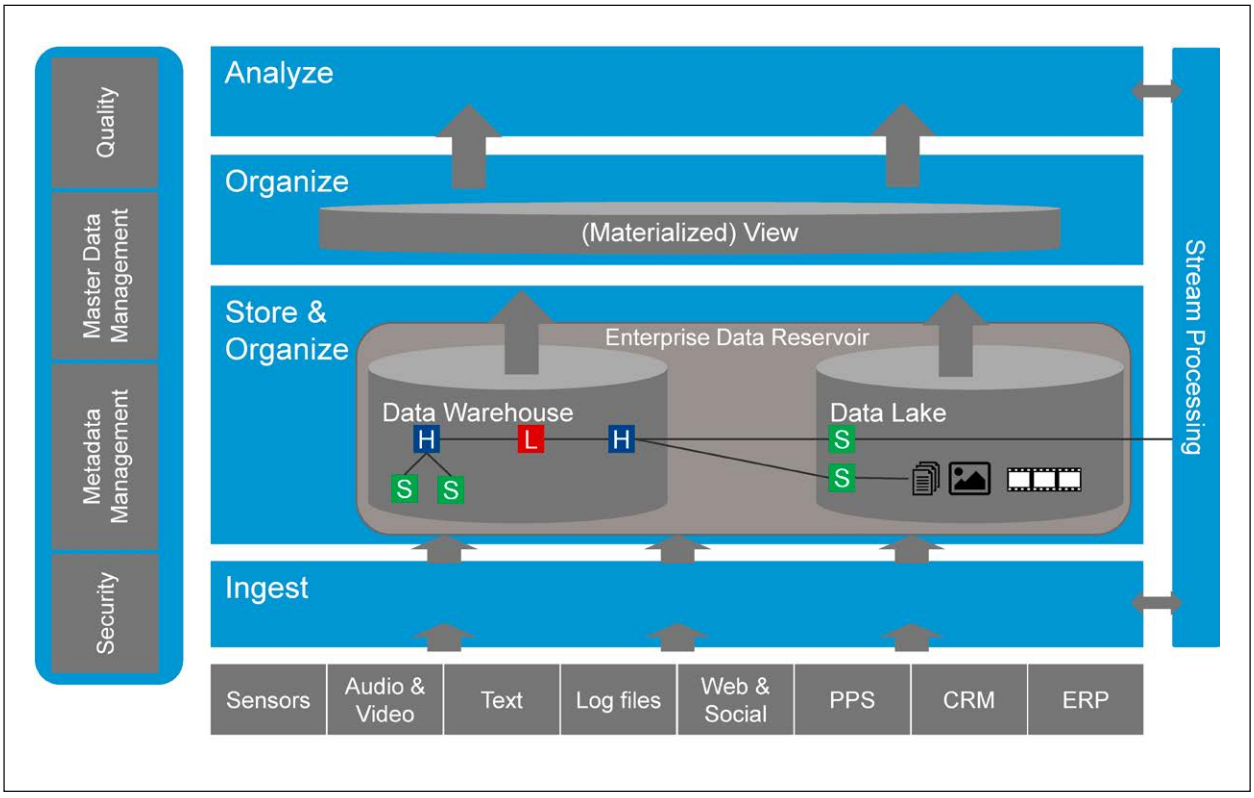

Abb. 2: Pi-Architektur

und wohldefinierten Kennzahlen aus dem DWH stellt daher eine besondere Herausforderung dar. Gleichzeitig erlaubt ein Big Data Store den Unternehmen, Daten vorzuhalten, deren Nutzungskontext noch nicht abschließend feststeht und deren Vorhaltung im DWH deshalb nicht sinnvoll ist. So können ad hoc neue Analysen auf Datenbeständen ergänzt werden, deren Relevanz im Vorfeld nicht absehbar ist [BaK15].

Die führenden BI-Werkzeuganbieter warten bereits mit Umgebungen auf, die sowohl Big-Data- als auch BI-Komponenten umfassen. Diese werden üblicherweise jedoch im Verantwortungsbereich der IT verortet und lassen Anwendern nur einen eingeschränkten Zugriff für eine selbstständige Datenintegration und -aufbereitung. Eine Self Service Data Preparation für integrierte DWH- und Big-Data-Anwendungen ist so nur in einem eingeschränkten Rahmen möglich. Wir schlagen als Erweiterung bestehender Ansätze eine „Pi-Architektur" vor. Diese Vorgehensweise fußt auf Ideen aus dem Data-Vault-Modell, erweitert diese jedoch dadurch, dass strukturierend-integrative Modellbestandteile separat von polystrukturierten Inhalten abgelegt und so die Stärken klassisch relationaler Strukturen mit denen von Big-Data-Ansätzen kombiniert werden. Diese beiden Säulen des Pi-Modells werden um eine integrative Abfrageschicht ergänzt, die gewissermaßen das Dach auf dem Pi darstellt (vgl. Abbildung 2). Die Modifizierbarkeit des Data-Vault-Modells stellt dabei sicher, dass Kern-BI-Inhalte nicht gefährdet werden.

Trennung von Verantwortungsbereichen in der Pi-Architektur

Unser Konzept sieht des Weiteren eine Trennung verschiedener Datenhaltungs-Layer mit jeweils unterschiedlichen Verantwortungsbereichen vor, um eine iterative Ad-hoc-Einbindung neuer polystrukturierter Datenquellen durch Benutzer und Fachbereiche zu unterstützen.

Die Grundlage des Lösungsansatzes bildet eine von der IT angebotene, zentrale Informationsplattform als „Enterprise Data Reservoir", die eine Kombination klassischer DWH-Technologien mit einem Data Lake darstellt, die über eine übergreifende Abfrageschicht ergänzt werden. Durch die Trennung von Verantwortungsbereichen erlaubt das Konzept den Anwendern, selbstständig neue Datenquellen zu definieren. Diese können mit Hilfe von Logiken und Extraktions- und Transformationsregeln von feingranularen Daten in zunehmend konsolidiertere Sichten überführt werden. Anstatt Integrationsverfahren für die Daten vorzugeben, bietet die IT den Fachbereichen Tools und Prozesse an, um ihre bestehenden Datenmodelle selbstständig mit weiteren Analyseobjekten und -attributen zu erweitern und diese auf Basis ihrer Anforderungen zu ergänzen. Ausgangspunkt ist hierbei weiterhin ein zentral von der IT gemanagtes Datenmodell mit qualitätsgesicherten Daten.

Der Prozess basiert dabei auf folgenden vier Schritten, die einen Ansatz von Topchyan aufgreifen (vgl. Abbildung 3) [Top16]:

1. Identifikation und Beschreibung potenzieller neuer und gegebenenfalls polystrukturierter Datenquellen durch die Fachbereiche und deren Speicherung „as is" im **Raw Data Layer.** Aufgrund wachsender regulatorischer Anforderungen an Datenqualität, Nachvollziehbarkeit und Verantwortlichkeiten sind bereits hier eindeutige Verantwortlichkeiten („Stewardship") und (fachliche und technische) Metadaten zu definieren, sofern diese später auch von anderen Anwendern genutzt werden sollen.

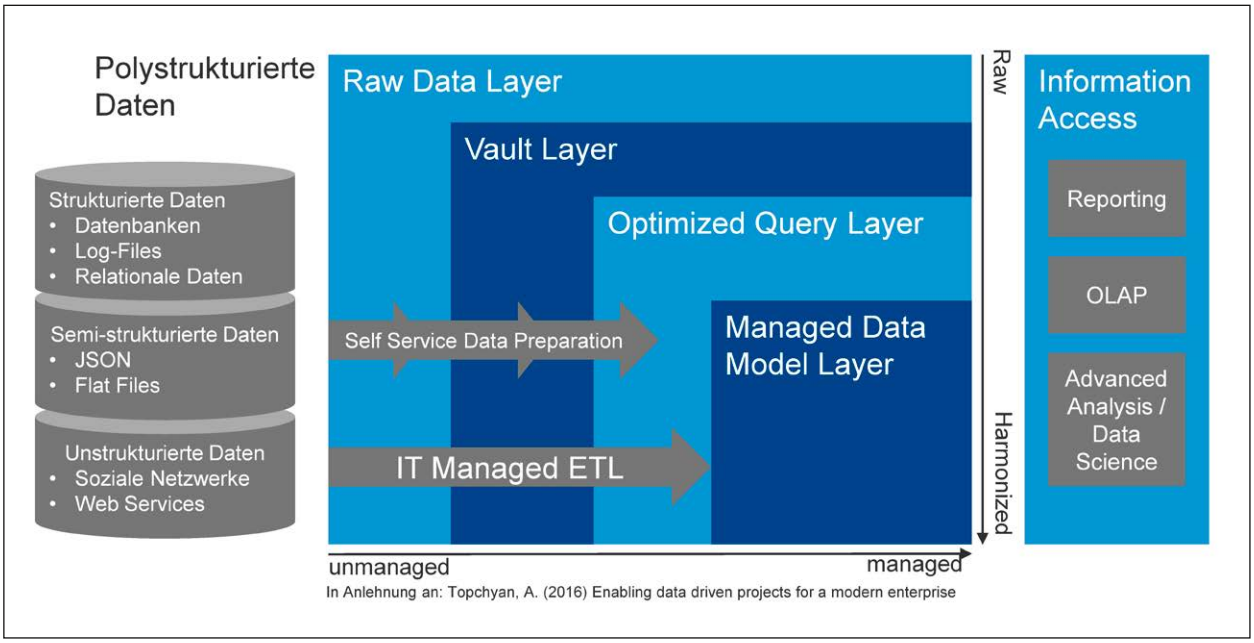

Abb. 3: Klare Trennung von Verantwortungsbereichen [vgl. Top16]

2. Überführung und Integration der feingranularen Daten in eine integrierte und stärker managementorientierte Sicht auf der Ebene eines **Vault Layer**; Basis bildet dabei eine Data-Vault-Modellierung. Ausgehend von einem qualitätsgesicherten Datenmodell als Informationsnukleus können Anwender hier ihre neuen Rohdaten selbstständig zu neuen Analyseobjekten und -attributen erklären und mit diesen das Kerndatenmodell so für ihre eigenen Zwecke erweitern. Insbesondere können neue Attribute als „Satelliten" (in der Terminologie des Data-Vault-Modells) angehängt werden, das heißt als Tabellen mit ergänzenden beschreibenden Daten. Neue Analyseobjekte werden als sogenannte „Hub-Tabellen" ergänzt und über „Link-Tabellen" mit anderen Hubs verzahnt. Durch diese Art der Modellierung sind Änderungen flexibel möglich, sodass in der Regel keine bestehenden Tabellen angepasst werden müssen, sondern lediglich neue Tabellen hinzugefügt werden. Mit diesem umfassenden Ansatz gewährleistet das Data-Vault-Konzept effektiv die von der Fachseite geforderte Agilität. Idealerweise werden die Anwender bei der Integration durch Assistenzsysteme unterstützt, sodass sie neue Daten ohne tiefere Data-Vault-Kenntnisse mit dem bestehenden Datenmodell zusammenführen können. Hierbei sind möglichst automatisch ergänzende Metadaten zu den Verknüpfungen bereitzustellen. In unserem Ansatz werden polystrukturierte Daten dabei als Satelliten in einer Big-Data-Umgebung hinterlegt, während die Hub- und Link-Tabellen, in denen die Zusammenhänge zwischen den Objekten dokumentiert sind, in einer relationalen Umgebung gespeichert werden (vgl. Abbildung 2). Dies erlaubt es insbesondere, dass die über die Links und Hubs erreichte Zusammenführung der Strukturdaten größtenteils im relationalen Kontext erfolgen kann, der deutlich besser auf Join-intensive Abfragen ausgelegt ist. Der Aufruf großer polystrukturierter Daten verbleibt jedoch in den Satelliten der Big-Data-Systeme – die gegebenenfalls auch physisch lokal im Fachbereich lokalisiert sein können.
3. Optimierte Abfragen auf die Data Vault-Tabellen. Hierfür können ergänzend in einem **Optimized Query Layer** zusätzliche Views definiert werden. Sollen die verschiedenen Entitäten zusammengeführt werden, um sie zum Beispiel in denormalisierte Dimensionstabellen zu laden, ist dies über Views möglich. Ergänzend kann bereits ein Teildatenbestand für einen bestimmten Organisationsbereich oder eine bestimmte Anwendung oder Analyse in entsprechenden Data Marts erstellt werden. Solche Schritte bieten sich vor allem bei Abfragen auf den neuen Daten durch mehrere Benutzer an und legen bereits deutlich stärker eine Unterstützung durch die IT nahe. Spätestens hier müssen auch zusätzliche Daten zur Informationsherkunft, ihrer Qualität und ihrer Verwendbarkeit bereitgestellt werden.
4. Im letzten Schritt erfolgt die vollständige Übernahme der Daten durch die IT in einem **Managed Data Model Layer**. Wenn der Analysefokus vollständig definiert ist und häufig wiederkehrende, standardisierte Abfragen erfolgen, ist ein von der IT gepflegtes, konsolidiertes Datenmodell nach wie vor das Mittel der Wahl. Ebenso ist es das bevorzugte Konzept, wenn eine konsolidierte Gesamtsicht im Vordergrund steht, die von unterschiedlichen Anwendern konsistent abgefragt wird.

Fazit

Die Einführung eines Data Lake und die damit zu erzielende schnelle Integration neuer Daten ist bloß ein Teil einer agilen Lösung. Erst die End-to-End-Betrachtung der gesamten Daten-Wertschöpfungskette und die übergreifenden Data-Governance-Prozesse stellen sicher, dass die beschriebene Datenversorgung zur Effektivitätssteigerung für die Anwender führt. Fragen nach der Quelle, der Vertrauenswürdigkeit,

dem Schutz und dem Lifecycle-Management der Daten sind wichtiger denn je. Welche Daten sind im Repository vorhanden, wie ist deren Definition und in welchem Kontext stehen sie zueinander? All dies sind Informationen, die einen weiteren Erkenntnisgewinn erst ermögliche und deren Bereitstellung im Rahmen der beschriebenen vier Schritte berücksichtigt werden muss.

Die durch den Einsatz eines umfassenden „Enterprise Data Reservoir" auf Basis einer Pi-Architektur gewonnene Agilität bleibt eine große Herausforderung aus Governance-Sicht. Diese beschränkt sich dabei nicht nur auf die Sicherheit, sondern umfasst auch Aspekte wie Nachvollziehbarkeit der Prozesse, Dokumentation der Dateninhalte und Interpretation oder aber auch Maskierung von Daten für bestimmte Nutzergruppen. Eine effektive Governance erfordert einen ganzheitlichen Ansatz über den gesamten Prozess und Technologiegrenzen hinweg, um ein komplettes Bild des „Puzzles" zu erhalten.

Die vorgestellte Pi-Architektur kann einem breiten Personenkreis im Unternehmen Zugang zu Daten und Analysen ermöglichen. Mit Hilfe des vorgestellten Self-Service-Data-Preparation-Ansatzes der Pi-Architektur fällt es Benutzern aus den unterschiedlichsten Bereichen leicht, neue Datenquellen zu integrieren und aufzubereiten. Die selbstständige, schnelle und einfache Anpassung von KPIs, Reports und Analysen zur Beantwortung relevanter Business-Fragen steht dabei im Mittelpunkt. Der Datenaufbereitungsprozess verschiebt sich hierfür zunehmend von der Seite der IT in Richtung des Fachbereichs. Gleichzeitig erlaubt sie Erkenntnisse jenseits der standardisierten BI-Reports und lädt zum kreativen Arbeiten mit den Daten ein.

[Literatur]

[AlS16] Alpar, P. / Schulz, M.: Self-service business intelligence. In: Business & Information Systems Engineering, 58. Jg., Nr. 2, 2016, S. 151–155

[BaK15] Baars, H. / Kemper, H.-G.: Integration von Big Data-Komponenten in die Business Intelligence. In: Controlling, 27. Jg., Nr. 4–5, 2015, S. 222–228

[Top16] Topchyan, A.: Enabling data driven projects for a modern enterprise. In: Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS), 28. Jg., Nr. 3, 2016, S. 209–230

[ZBK12] Zimmer, M. / Baars, H. / Kemper, H.-G.: The impact of agility requirements on business intelligence architectures. In: 45th Hawaii International Conference on System Science (HICSS), 2012, S. 4189–4198

Timm Marschall ist wissenschaftlicher Mitarbeiter am Lehrstuhl für ABWL und Wirtschaftsinformatik 1 der Universität Stuttgart. Im Rahmen seiner Masterarbeit beschäftigt er sich mit Anwendungen, Potenzialen und Lösungskonzepten für eine agile analyseorientierte Datenhaltung in Big-Data-Umgebungen. **E-Mail: Marschall@wi.uni-stuttgart.de**

Dr. rer. pol. Henning Baars ist Akademischer Oberrat am Lehrstuhl für ABWL und Wirtschaftsinformatik 1 der Universität Stuttgart und Sprecher der Fachgruppe „Business Intelligence" der Gesellschaft für Informatik. Seit 2003 ist er an der Universität Stuttgart tätig. Aktuelle Forschungsthemen sind „Agile Business Intelligence", „BI und Big Data", „BI in the Cloud" sowie „BI und Analytics im Internet of Things". **E-Mail: baars@wi.uni-stuttgart.de**