Pragmatischer Einstieg in Daten-Analyse und Konsolidierung: Hadoop meistert den Spagat zwischen Big Data und Kosteneffizienz

Oliver Herzberg und Carsten Herbe, metafinanz GmbH

Die Finanz- und Versicherungsbranche steht vor großen Herausforderungen. Einerseits müssen kostspielige historische Systeme konsolidiert werden, andererseits steigt mit neuen digitalen Geschäftsmodellen und Regulatorik der Bedarf an Big-Data-Lösungen. Hadoop eröffnet hier neue Chancen, um das Datenmanagement strategisch neu zu überdenken und den Spagat zwischen Innovation und Kosten zu meistern.

Es ist eine alte Grundsatzfrage: Ist die IT in den Unternehmen nur ein Kostenfaktor oder bildet sie die Basis für den Betrieb und stellt als Innovator die Wettbewerbsfähigkeit her? Zumindest in der Finanz- und Versicherungswirtschaft ist diese Frage schon lange entschieden. Ohne IT würde heute der Betrieb zum Erliegen kommen, die Entwicklung von Produkten, die allesamt auf ausgeklügelter Software basieren, wäre nicht mehr vorstellbar. Dennoch sind auch in dieser Branche die Zeiten vorbei, in denen für Technologie nahezu grenzenlose Budgets zur Verfügung stehen und Individualfertigungen als Status quo gelten. Die Zeichen stehen auf Industrialisierung, was für die IT bedeutet, dass sie Service-orientierter wird und im Zusammenspiel mit den Geschäftsbereichen neu organisiert.

Als große Herausforderung für die IT-Verantwortlichen gilt dabei weiterhin die Verteilung der großen Kostenblöcke. Laut der Gartner-Studie „The 2014 IT Agenda" fließen 41 Prozent der IT-Kosten in die Rechenzentren, 26 Prozent gehen in die Netzwerk-Infrastruktur und 20 Prozent in den Betrieb der Clients. Für innovative Themen bleibt dabei nicht mehr viel übrig. Genau an diesem Punkt setzen viele CIOs mit Restrukturierungsmaßnahmen an, wie das aktuelle Beispiel bei einem großen Versicherungskonzern zeigt. Das IT-Management initiierte hier ein Konsolidierungsprojekt, das sich über alle weltweiten Rechenzentren erstreckt, gleichzeitig wer-

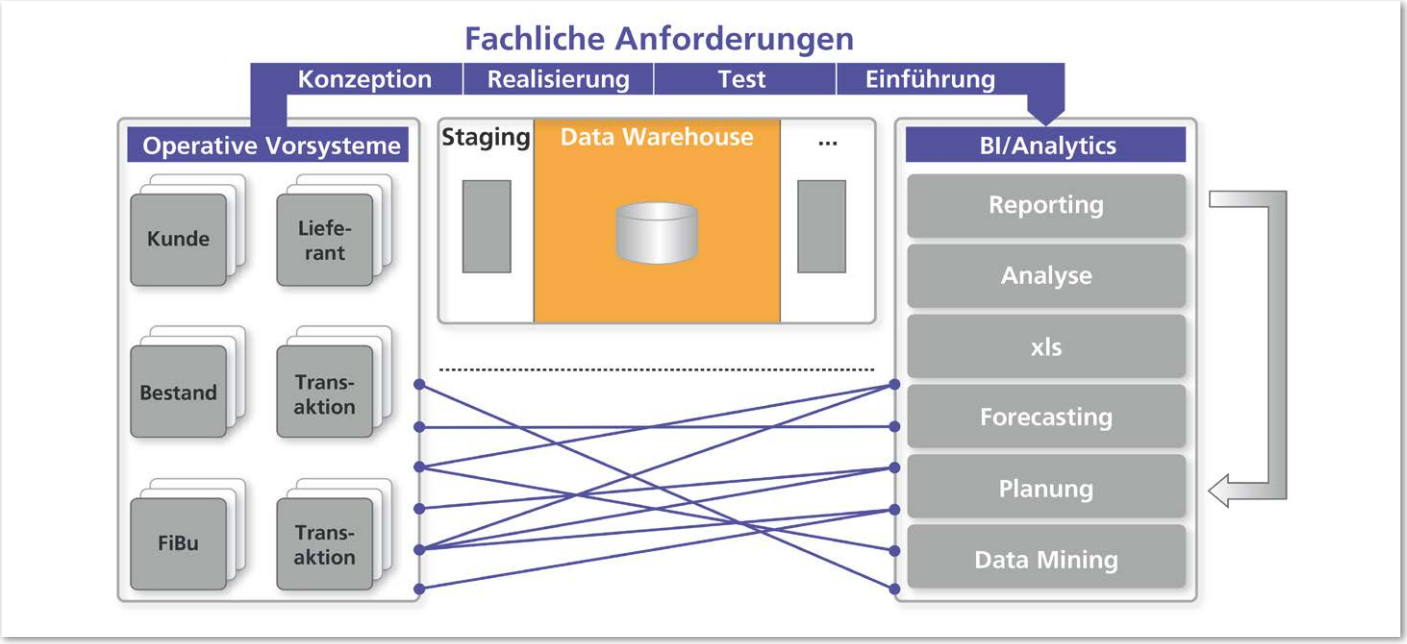

Abbildung 1: Leistungsfähige, flexible und elastische DWH- und BI-Infrastrukturen wirken vor allem bei künftigen Change-Kosten

den auch die Netzwerk-Infrastruktur erneuert und die Arbeitsplätze auf zentral bereitgestellte virtuelle Clients umgestellt.

Datenbeschaffung sorgt weiterhin für hohe Kosten

Auch die Modernisierung historisch gewachsener, teilweise veralteter IT-Anwendungen steht in der Finanzwirtschaft seit einiger Zeit ganz oben auf der Agenda. Viele Unternehmen konsolidieren ihre Kernsysteme und betreiben mit großem Aufwand die Erneuerung ihrer Data Warehouses. Hier rechnen Experten auch weiterhin mit steigenden Kosten; als dominierende Bereiche gelten leistungsfähige Data Warehouses und BI-Infrastrukturen *(siehe Abbildung 1)*. Den Löwenanteil der Kosten verschlingen mit 60 Prozent die Datenbeschaffung, -validierung und -transformation, zudem schlagen aufwändige Testphasen überdurchschnittlich zu Buche.

Man ist sich in der Branche einig, dass die Datenmengen weiterhin stark wachsen und damit die Kosten weiter nach oben treiben. Als Hauptgrund gelten die umfangreichen regulatorischen Anforderungen wie Solvency II, Basel III, IFRS, EBA Reporting oder aktuell BCBS239. Wie das Marktforschungsunternehmen Lünendonk herausfand, wird allein die Umsetzung der Regulierungsauflagen von Solvency II die Versicherungswirtschaft bis zum Jahr 2020 belasten. Neben der Beherrschung der Datenvolumina wird es aber auch verstärkt um den Ausbau der Analyse-Anwendungen gehen, mit denen Versicherer beispielsweise Kunden- oder Risikogruppen besser segmentieren können.

Solvency II & Co. verursachen enorme Datenmengen

Zu den derzeit größten Herausforderungen zählt die Umsetzung der wachsenden Zahl an Compliance-Vorgaben. Diese Aktivitäten binden Ressourcen in der IT und den Fachbereichen; die daraus entstehenden Kosten belasten die Gesamtbudgets und schränken den Spielraum für Innovationen weiter ein. Da meist auch ein hoher Umsetzungsdruck besteht, kommen oft Zwischenlösungen zum Einsatz, die im Betrieb und in der späteren finalen Lösung weitere Ausgaben nach sich ziehen.

Wie sich die Anforderungen an eine zukünftige IT im Versicherungsgewerbe ändern, hat die Lünendonk-Studie ebenfalls untersucht. Demnach wird es bis zum Jahr 2020 weitere Firmenfusionen geben und die Kooperationen mit Banken werden ausgeweitet, um das Geschäft mit dem Asset- und Vermögensmanagement auszubauen. Zunehmend werden mobile Technologien zum Einsatz kommen, um etwa die Kunden-Kommunikation über Apps anzubieten und den direkten Vertrieb von Versicherungsprodukten auszubauen.

Explodierende Datenmengen kommen aber auch von der generellen Digitalisierung immer weiterer Lebensbereiche. Unternehmen sind auch hier gefordert, solche Quellen gewinnbringend zu analysieren und zu veredeln. Daten mutieren damit zum neuen Rohstoff, den sich Unternehmen im Sinne neuer Geschäftsmodelle zunutze machen müssen. Das strategische Management interner und externer Daten gewinnt weiter an Bedeutung, Datenmanagement wird zu einem elementaren Bestandteil der IT-Strategie von Unternehmen. Zusammenfassend ergeben sich aus den geänderten Rahmenbedingungen des Marktes und der IT folgende drei Herausforderungen für CIOs:

- Sie müssen für die Umsetzung regulatorischer Vorgaben einen großen Aufwand betreiben, was mit hohen Kosten ohne Ertrag verbunden ist
- Sie müssen die Digitalisierung aller gesellschaftlichen und wirtschaftlichen Bereiche erkennen und sich zunutze machen, um neue Märkte zu erschließen und Kunden zu binden
- Sie müssen Initiativen ergreifen, um die gewaltigen, weiter wachsenden Datenschätze zu heben und für geschäftliche Szenarien zu nutzen

Das Big-Data-Missverständnis

Die Auflistung macht deutlich, dass praktisch alle Konsolidierungs- und Innovationsthemen unmittelbar oder indirekt mit Daten-Management zu tun haben. Seit einigen Jahren kursiert dafür das Schlagwort „Big Data" – als Schublade für alles, was mit umfangreichen Datenvolumina und Analysen zu tun hat. Allerdings hat sich nach der ersten Big-Data-Welle auch etwas Ernüchterung breitgemacht. Zu oft

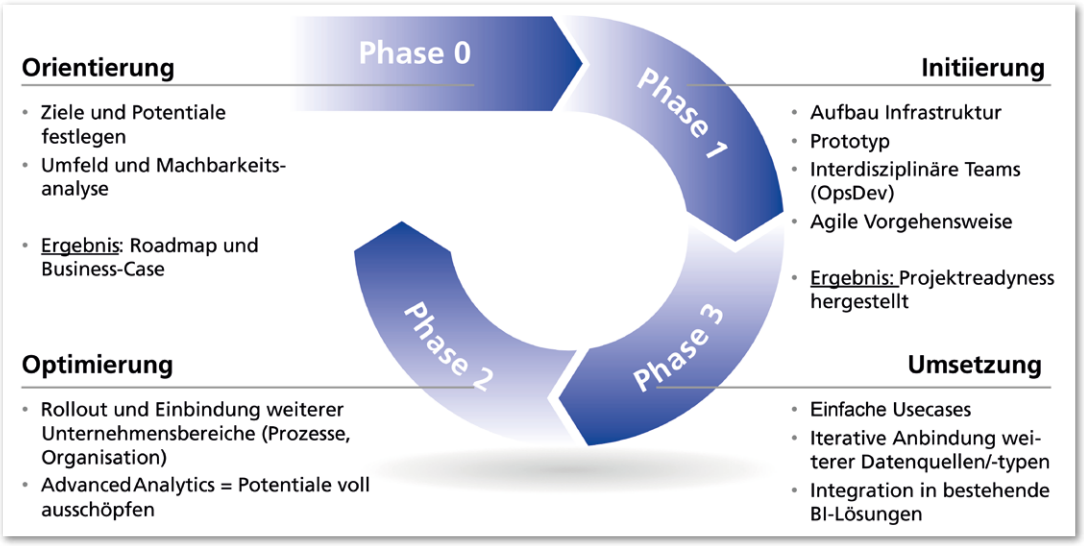

Abbildung 2: Ein iteratives Vorgehensmodell zur Einführung von Hadoop umfasst vier Kernphasen

wurden prestigeträchtige Megaprojekte vorgeführt, mit dem Effekt, dass viele Entscheider heute Big Data mit kostspieligen Initiativen assoziieren, die erst nach längerer Zeit geschäftliche Mehrwerte und einen Return on Investment einfahren.

Aus Sicht der CIOs jedenfalls sind diesbezüglich die Prioritäten klar definiert, wie eine andere Gartner-Studie herausfand. Gefragt nach den größten Herausforderungen bei Big Data und Analytics nannten die Entscheider mit großer Mehrheit die Antwort: „Wie man am schnellsten Mehrwerte aus großen Datenbeständen generieren kann."

Den Spagat zwischen Innovation und Kosten meistern

Mehrwerte aus großen Datenbeständen – aber ohne Großprojekte und Riesenbudgets? Dass dieser vermeintliche Widerspruch auflösbar ist, beweist die noch relativ junge Technologie „Hadoop". Deren großer Vorteil als Daten-Plattform liegt darin, dass Einführungen mit kleinen Projekten starten können, die zunächst einmal auf aktuelle und pragmatische Ziele wie Konsolidierung und Kostensenkung fokussiert sind. Bei Hadoop handelt es sich um ein Open-Source-Framework, das sich vor allem durch zwei Bereiche auszeichnet:

- Es bietet neue, bisher nicht umsetzbare Analytics-Möglichkeiten
- Es ermöglicht konkrete Einsatz-Szenarien, um die Infrastruktur- und Speicherkosten zu senken

Die Stärken von Hadoop liegen vor allem im dezentralen Speichern und parallelen Verarbeiten sehr großer Datenmengen, wobei hier das horizontal verteilte Dateisystem HDFS zum Einsatz kommt. Es besteht aus einem Cluster mit Standardservern und kann daher beliebige Datenformate in beliebigen Größenordnungen speichern. Ein weiteres Charakteristikum ist das dazugehörige MapReduce-Framework, das eine parallele Verarbeitung der Daten ermöglicht. Somit realisiert Hadoop ein kostengünstiges Speichern von beliebigen strukturierten und nicht-strukturierten Daten sowie eine parallele Verarbeitung riesiger Datenmengen.

Hadoop-Cluster skalieren übrigens linear, was bedeutet, dass zehn Prozent mehr Knoten die Speicherkapazität um zehn Prozent erweitern und damit die Leistung um zehn Prozent steigt. Ein weiterer Vorteil von Hadoop ist das mittlerweile große Ökosystem an Tools, das vielfältige Erweiterungsmöglichkeiten wie beispielsweise Datenauswertungen mit SQL bietet.

Einsatzszenarien von Hadoop

Die möglichen sinnvollen Einsatzszenarien von Hadoop im Banken- und Versicherungsumfeld lassen sich in folgende drei Bereiche einteilen:

- *RDBMS Offload*
  Hadoop-Cluster ermöglichen die Optimierung von Speicherplatz, um etwa bestehende, komplexe ETL-Prozesse oder klassische Datenbank-Systeme zu ersetzen
- *DWH-Extension*
  Bestehende Data Warehouses lassen sich durch Hadoop funktional erweitern, indem vorgeschaltete ODS-Datensammler eingesetzt werden
- *Big Data Exploration*
  Einsatz von Big-Data-Anwendungen, die Daten unterschiedlichster Quellen sammeln sowie Data Mining und Machine Learning ermöglichen

Auch wenn Hadoop von der Zielsetzung her für viele klassische BI-Szenarien infrage kommt, erfordert es teilweise völlig andere Methodiken und Herangehensweisen. Typisch für die klassische BI-Welt sind beispielsweise klar definierte Projektziele – ob es sich dabei um den Aufbau eines Near-Time-Reportings, eines Management-Cockpits oder eines Dashboards für das GuV-Stress-Testing handelt. Hadoop-Projekte hingegen gestalten sich zumindest am Anfang oftmals eher als Forschungs- und Entwicklungsprojekte, beispielsweise wenn es um Themen wie „Data Exploration" geht. Die IT stellt hierbei eine Plattform zur Verfügung, auf der das Business herausfinden kann, welche Fragen sinnvoll sind.

Beispiel-Szenarien für erfolgreiche Hadoop-Einführungen

Um Hadoop erfolgreich als Big-Data-Lösung einzuführen, empfiehlt sich ein iteratives Vorgehensmodell, das aus folgenden vier Phasen besteht *(siehe Abbildung 2)*:

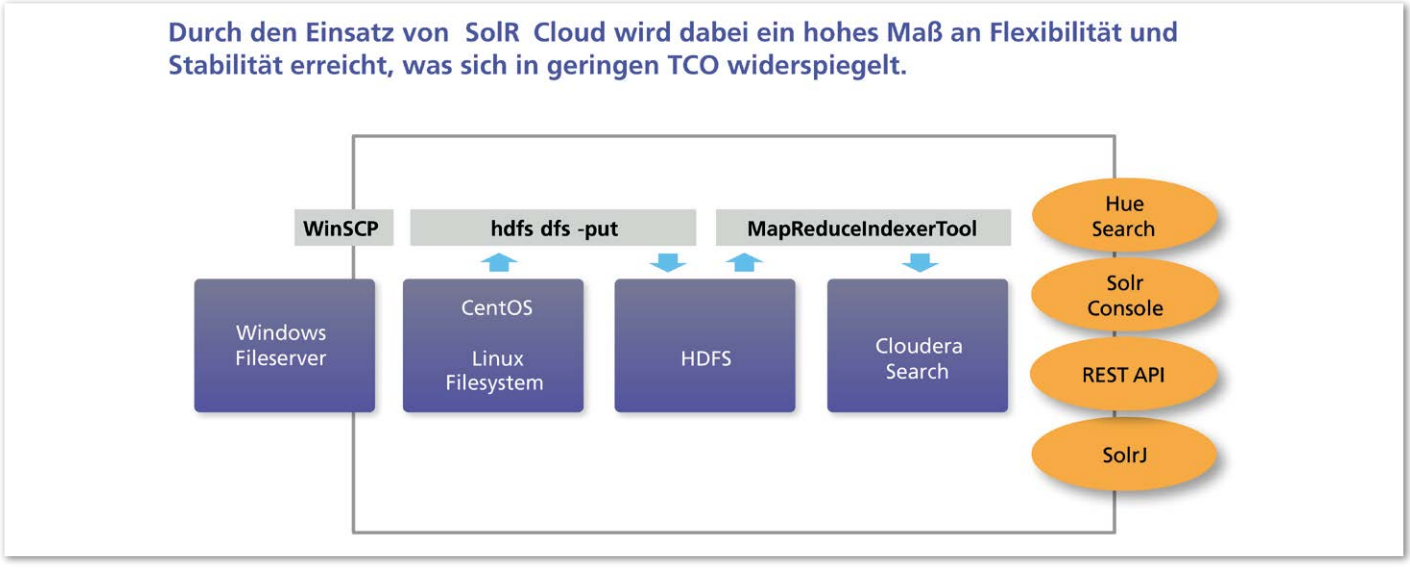

Abbildung 3: Das Hadoop-Search-API analysiert performant den Personalstamm

- Orientierungsphase: Strategie-Entwicklung, Ermittlung des Business-Value und Beispiele für Use-Cases
- Aufbau der Infrastruktur, Prototyp, Durchstich „front to back“
- Umsetzung mit einfachen Use-Cases für einen Quick-Win, stufenweiser Ausbau (Daten) und Integration (Prozesse und Organisation)
- Optimierung: unternehmensweite Integration von Big Data und Advanced Analytics

Anhand von Beispielszenarien lassen sich die Möglichkeiten von Hadoop sowie deren konkreter geschäftlicher Nutzwert anschaulich machen. Die folgenden drei Szenarien bieten einen Einblick in den Charakter typischer Hadoop-Lösungen.

Szenario 1: Proof of Concept „HR – strategische Personalentwicklung“

Bei der Personalsuche für IT-Projekte muss eine HR-Abteilung schnell auf Kundenanfragen reagieren und die vorhandenen Skills abfragen können. Wenn es um die Entwicklung des eigenen Personals geht, müssen die Skills der Mitarbeiter ständig mit den Marktbedürfnissen verglichen und daran ausgerichtet werden.

Informationen darüber liegen sowohl strukturiert als auch unstrukturiert in Projekt-Historien, Schulungsplänen, Gehaltsspiegeln der Branche und internen wie externen Stellenausschreibungen vor. Zu Projektbeginn ist unklar, ob die Daten sich in Beziehung setzen lassen.

Im ersten Schritt wird ein Prototyp auf Hadoop-Basis in einer definierten Projektumgebung entwickelt. Dabei wird auf den Einsatz von Open-Source-Technologien geachtet, um ein Vendor-Lock-in zu vermeiden. Mitarbeiterprofile werden in das Hadoop Distributed File System (HDFS) geladen und mit MapReduce indiziert. Auswertungen sind über Velocity, Solr Console, REST API und SorJ auf den Originaldaten möglich. Durch den Einsatz von SolR Cloud wird ein hohes Maß an Flexibilität und Systemstabilität erreicht mit der Folge geringer TCO. Bereits in dieser Phase kann das Unternehmen bei Kundenanfragen feststellen, ob das Know-how vorhanden ist.

Im zweiten Schritt werden die Daten aus dem ERP-System integriert, um Mitarbeiter mit Projekt und Laufzeit zu ermitteln. Im dritten Schritt integriert man Daten aus öffentlichen Quellen wie Stellenbörsen. Das Unternehmen kann dann schnell auf angefragte Ressourcen reagieren und Skills strategisch und an den Kundenbedürfnissen orientiert weiterentwickeln *(siehe Abbildung 3)*.

Szenario 2: Zentralisierung der Archivierungsmedien

Zur Erfüllung der Grundsätze ordnungsgemäßer Buchführungssysteme (GOBS) oder der Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GdPdU) müssen geschäftsrelevante Daten zehn Jahre archiviert werden. Die Archivierung von Daten erfolgt oft im Data Warehouse

oder in speziellen Archivierungs-Marts und verursacht so hohe Kosten. Durch Weiterentwicklung von Software und Dateiformaten steigt der Aufwand, gespeicherte Daten erneut zu laden und zu verarbeiten.

Die in den entsprechenden Data-Mart gehaltenen Daten werden in den Hadoop-Cluster repliziert. Der Quickwin besteht darin, dass sich die Speicherkosten reduzieren, da alle älteren Daten aus dem Data-Mart gelöscht werden. Im nächsten Schritt lädt man die Archivdaten, die in unterschiedlichsten Formaten vorliegen, in das Hadoop-Cluster, um sie zu analysieren und zu verarbeiten. In der Regel erfolgt dies mit den gleichen BI-Tools und Reports, mit denen auch schon die Daten in der relationalen Datenbank ausgewertet werden. Möglich machen dies Technologien wie Hive, für das viele Tool-Hersteller schon eine Schnittstelle anbieten *(siehe Abbildung 4)*.

Szenario 3: Hadoop als Staging-Area im ETL-Prozess

Im Staging- und ETL-Prozess eines Finanzdienstleisters werden im Rahmen der Tagesend-Verarbeitung große Datenmengen zwischengespeichert. In der Praxis erweisen sich die vorhandenen Daten-Infrastrukturen als nicht ausreichend, um den Bedarf an schneller und genauer Information abdecken zu können. Daher setzen die Fachbereiche oft individuelle Lösungen ein – die

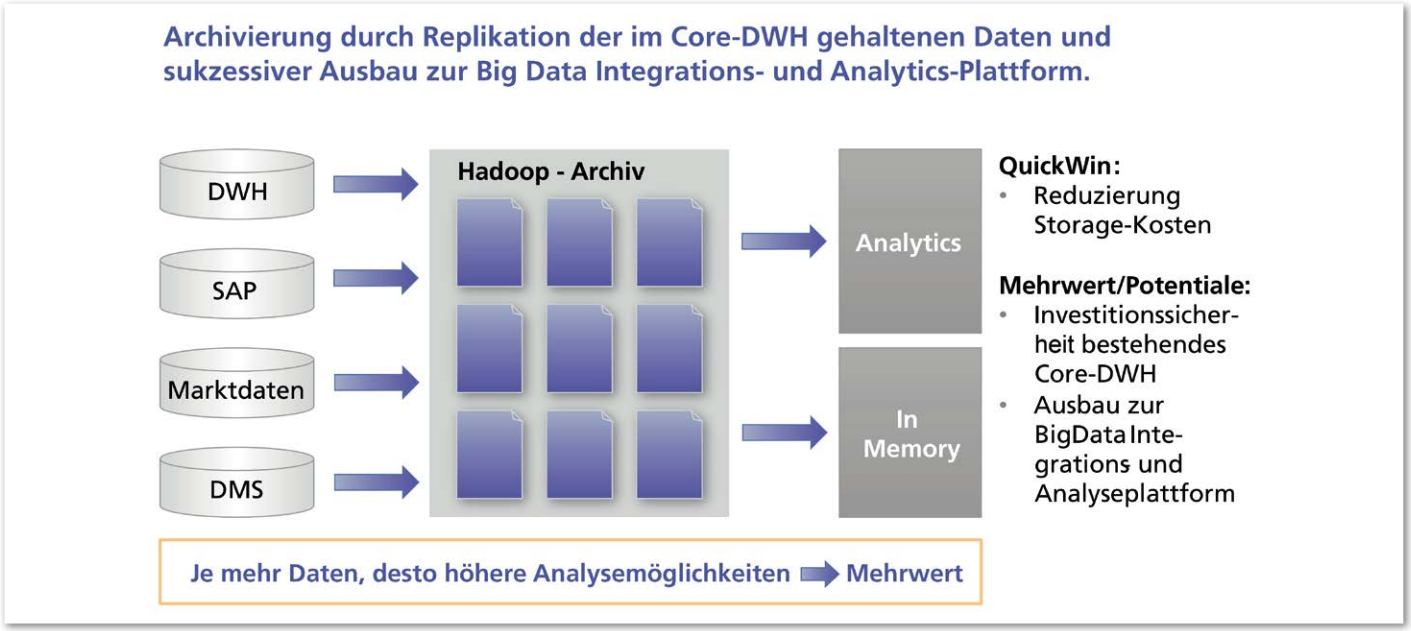

Abbildung 4: Zentrale Archivierung mit Hadoop

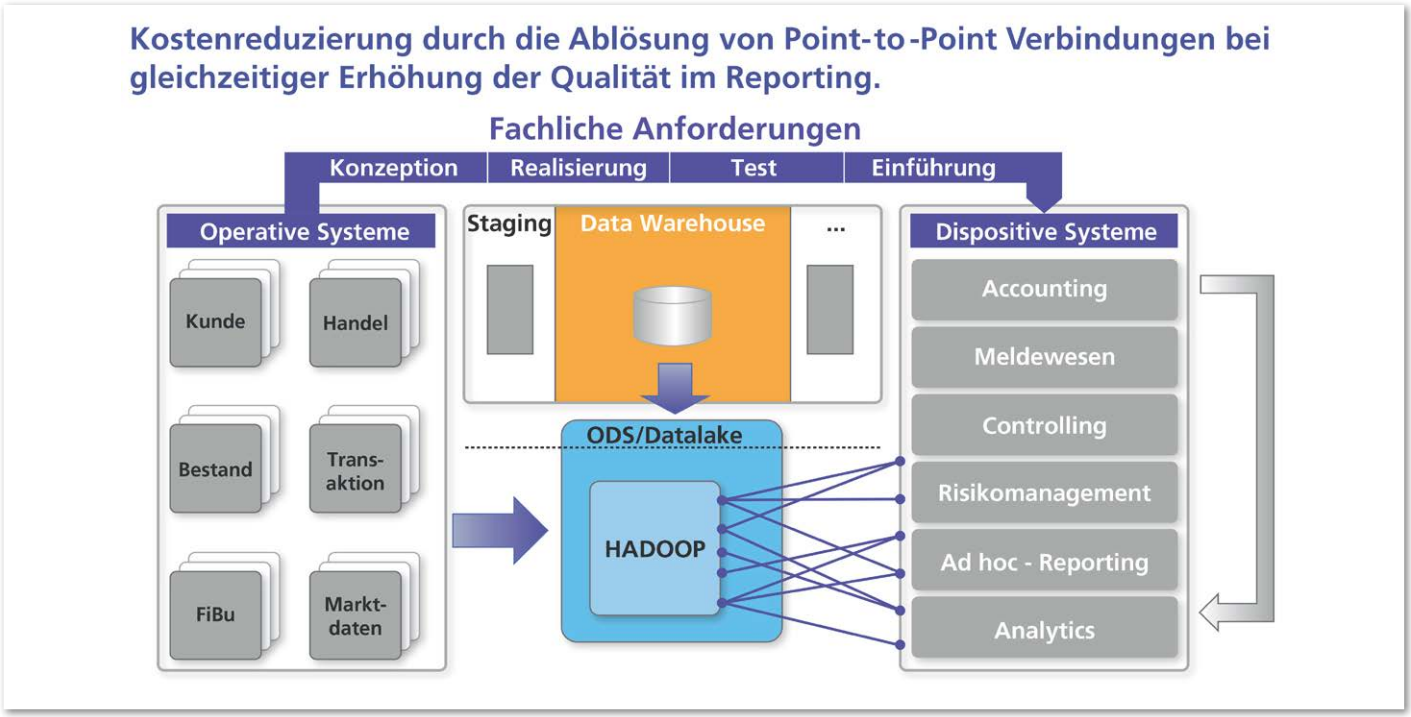

Abbildung 5: Ergänzung des DWH durch den Aufbau eines Hadoop ODS

allerdings den regulatorischen Anforderungen nur bedingt entsprechen. Strukturelle Änderungen des Datenmodells an den operativen Vorsystemen, zum Beispiel zur Anpassung an regulatorische Anforderungen, sind im Change sehr teuer, da die komplette ETL-Strecke konzipiert, geändert und durch alle Nutzer getestet werden muss.

Ziel ist ein dem Core-DWH vorgelagerter Data-Lake – ein Speicher-Repository, das zunächst große Mengen an Rohdaten speichert. Dezentrale Anwendungen in den Fachbereichen können über klare Schnittstellen sowohl auf Rohdaten aus den Vorsystemen als auch auf die Daten des DWH und seines Data Mart zugreifen. Die fachlich erforderliche Flexibilität bleibt *(siehe Abbildung 5)*.

Durch den Einsatz von Hadoop und HDFS lassen sich die Kosten der auf günstigen File-Servern und SANs abgelegten Daten deutlich reduzieren. Gleichzeitig stehen sie für performante Auswertungen zur Verfügung. Der Vorteil für Analytics besteht darin, dass die Daten durch Hadoop in der Granularität vorliegen, wie sie aus Sicht der fachlichen Nutzer benötigt werden. Somit lassen sich dezentrale Anwendungen reduzieren.

Fazit

Wie am Vorgehensmodell und an den Beispielszenarien zu erkennen ist, erfordert der Umgang mit Hadoop ein Umdenken und die Bereitschaft, unbekanntes Terrain zu beschreiten. Best Practices müssen sich erst etablieren oder an die jeweiligen Unternehmensziele angepasst werden. Generell lautet die Empfehlung, klein anzufangen, ein Gefühl für die Technologien und ihre Möglichkeiten zu entwickeln und darauf aufbauend eine dem Unternehmen adäquate Strategie zu erarbeiten.

Keinesfalls werden klassische Data Warehouses oder relationale Datenbanken überflüssig, stattdessen werden sie ergänzt und erweitert. Damit sind die Investitionen in bestehende DWH-, BI- und Reporting-Infrastrukturen gesichert. Gleichzeitig eröffnet Hadoop neue Potenziale für Analytics auf der einen und Kosteneinsparungen auf der anderen Seite. Vor dem Hintergrund steigender Ansprüche im Bereich Datenmanagement empfiehlt es sich damit als ideale Plattform für Konsolidierungen, innovative Lösungen und Geschäftsmodelle der Zukunft.

Oliver Herzberg
oliver.herzberg@metafinanz.de

Carsten Herbe
carsten.herbe@metafinanz.de

Die (Oracle-)Welt wächst zusammen: Meeting mit der Taiwan Java User Group in Taipei

Gunther Pippèrr, freiberuflicher Berater bei GPI Consult, verbrachte im Oktober 2014 zwei Wochen in Taiwan und initiierte dort ein Treffen mit der Taiwan Java User Group. Neben einem regen Erfahrungsaustausch mit der taiwanesischen Community konnte er auch viele kulturelle Eindrücke sammeln.

Dass er ausgerechnet am anderen Ende der Welt auf Gleichgesinnte treffen würde, die genau wie er große Anhänger des NoSQL-Ansatzes sind, hätte sich Pippèrr nun wirklich nicht träumen lassen. Doch bei genauerer Betrachtung sei das in einem Land wie Taiwan gar nicht so weit hergeholt: „Das schnelle Marktwachstum sowie der hohe Marktdruck in den asiatischen Ländern erfordern, alles möglichst schnell umzusetzen und möglichst schnell fertig zu werden – ganz anders als in Deutschland, wo Wartbarkeit, Sorgfalt und niedrige Betriebskosten entscheidend sind“, berichtet er von seinen Beobachtungen. „Gerade das könnte ein Grund dafür sein, weshalb der NoSQL-Ansatz in den asiatischen Ländern so beliebt ist“, vermutet er. Denn auch bei NoSQL gehe es darum, schnell Lösungen für Applikationen zu finden, die in der Regel nur kurz zum Einsatz kommen.

Pippèrrs Reise begann eigentlich mit der Einladung eines befreundeten Software-Entwicklers, der schon seit einigen Jahren in Taiwan lebt. Pippèrr hatte bald eine Idee: Zusammen mit seinem Reisegefährten, Slava Schmidt, ebenfalls Software-Entwickler, wollte er ein Treffen mit der Oracle Anwendergruppe vor Ort organisieren. Auf der Suche nach einem Ansprechpartner wandte sich Pippèrr an Dr. Dietmar Neugebauer, Vorstandsvorsitzender der DOAG. Dieser freute sich über das Engagement: „Der von Pippèrr initiierte Austausch ist ein gutes Beispiel für eine gelungene Kommunikation über Ländergrenzen hinweg und zeigt, wie die Usergruppen auf internationaler und persönlicher Ebene zusammenarbeiten können. Die DOAG unterstützt bei solchen Vorhaben gerne bei der weltweiten Kontaktvermittlung.“ So konnte er Pippèrr nach kurzer Zeit einen Kontakt zur taiwanesischen Java User Group vermitteln.

Marina Fischer
marina.fischer@doag.org